An Investigation into Multi-Word Expressions in Machine Translation

Lifeng Han B.Sc., M.Sc.

Joint Supervisors: Prof. Gareth J. F. Jones and

Prof. Alan F. Smeaton

A thesis presented for the degree of Doctor of Philosophy (Ph.D.)

School of Computing

Dublin City University

January 2022

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy, is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge, breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Lifeng Han

Student ID No.: 17210060

4/January/2022

To Sonia and her family for their support, love and friendship I received.

Acknowledgements

I thank Professor Qun Liu for recruiting me onto a PhD position in the ADAPT Centre. I am very grateful to my supervisors Professors Alan Smeaton and Gareth Jones who assumed academic supervision of my PhD when my former supervisor departed from DCU without whom this thesis would not have been finished.

I thank Paolo and Shaohui for helping me out with some experimental setups for the investigations into Chinese radical knowledge and MWEs. Thanks to my colleagues who were involved in collaborations: Yvette, Ash, Alfredo, Ahmed, Erwan, Pan, Qinyuan, Ning, Daniela, Vanessa, Gultekin, Tereska, Dexmont, Yi, Serge, Irina, and Gleb. Thanks to Professors Tomas Ward and Dorothy Kenny for helping to shape the thesis structure as my mid-term PhD Transfer Evaluation committees, and Prof. Cathal Gurrin who chaired the talk. Thanks also go to Prof. Bogdan Babych from Heidelberg University who is my thesis external examiner and again to Prof. Dorothy Kenny who is my thesis internal examiner for their valuable comments on the thesis structure and their hard work reading through each line helping to improve the content. Thanks to Dr. Alessandra Mileo for chairing my viva.

I thank Ricardo, Anne, Owen, and Frank who gave me huge support without whom I would not even have had the chance to write this thesis, and I am in debt to them on this. I thank Gideon for a long-term friendship, and thank Clark, Peter (Brennan), James, Paloma, Terry, Peter (Berry), Dylan, Antonio, Giuseppe, Prashant, Acho; Mark, Rob, Myrthe and her Mom Julia for caring and help. I thank Eoin, Minkun, Lorin, Anwesha, Artem, Manoj, Esma, Zhengwei, Tengqi, Mikael, and ADAPT colleagues for the time spent together.

I thank former colleagues from ILLC, Uni. of Amsterdam, who made a lovely goodbye party for me moving to Ireland to continue my PhD journey, especially Jenny and Karine who organised this, as well as Joost who organised a gift package for me. I also thank Tanja, Khalil and other colleagues who always cared about me. I am grateful to my land-lords/ladies during this journey: Feargus/Eileen, Hongliang/Juan, Ultan/Peipei; housemates/neighbours Cosmina, Olivia, Longhao, GPG, Fan, Han, Xingyang, Vee, Ellie, Tiago, Rosin, Yong, Sheila, and my girlfriend Gabrielluccia. I thank nurse Sínead from the student health centre for caring.

I would like to acknowledge the support from the following funding source: The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. I would like to finish this section with an old Irish/Gaelic blessing from Celtic literature:

May the road rise up to meet you.
May the wind be always at your back.
May the sun shine warm upon your face;
the rains fall soft upon your fields
and until we meet again,
may God hold you in the palm of His hand.

Foreword

序言

< 念 - 七載之殤 >

春花鞦月，故國之道。春寒料峭，秋已入冬，島國之殤。尚記七載之初，獨身入歐，西行之旅。乘俄航，轉阿姆，求學行。心蕩漾，雖抓狂，夢在泱，道路長。不慎入情，欣喜若狂，未熟犢牛，不知何從。巨山突下，心魂崩塌，學業之殤，如若斷腸。

遇奇之女子，先荷後波，終華沙之行，告別死生。文化之旅，陪伴之行，且行且笑，新喜之生。

轉戰島國，不知所從。每日相片，綠意蔥蔥。人心向背，不慎入圍。欺凌之賊，日日盛行。直至甚晚，不知所終。幸與英雄，拔刀助行。身臨深淵，不易不生。兄弟而上，眾人賜贈。妖魔奮戰，終不勝正氣。又上正規，求學且長。路漫漫，道且遠，日日新，長新魂。終通暢，見曙光。今已辯，贏紙張。念七載之行，何去何從。古往今生，來去匆匆。

Contents

Acknowledgements ii

Foreword i

List of Abbreviations vi

List of Figures ix

List of Tables xi

Abstract xiii

1 Introduction and Background 1

1.1 Computer Powered Artificial Intelligence (AI) 1

1.2 Automatic Processing of Natural Languages 3

1.2.1 Natural Language Processing (NLP) in General 3

1.2.2 Machine Translation (MT) 5

1.2.3 NLP & MT Organisations and Events 7

1.3 MT Applications 8

1.4 Research Background and Thesis Hypotheses 9

1.4.1 Research Background 9

1.4.2 Thesis Hypotheses and Research Questions (RQs) 10

1.5 Thesis Structure 12

2 Machine Translation: Development, Issues and Challenges 14

2.1 MT Paradigms: from Rules to Neural . . . 15
2.1.1 Statistical vs. Neural Machine Translation: SMT vs NMT . . 16
2.1.2 Summary . . . 23
2.2 MT Evaluation Methodologies . . . 23
2.2.1 Background . . . 24
2.2.2 Human Evaluation Methods . . . 26
2.2.3 Automatic Evaluations of MT . . . 33
2.2.4 Quality Estimation Methods . . . 46
2.2.5 Meta-evaluation . . . 47
2.2.6 Discussion . . . 52
2.2.7 Summary . . . 54
2.3 MT Broad Issues and Challenges . . . 54
2.3.1 MT Modelling . . . 55
2.3.2 Linguistic Awareness . . . 56
2.3.3 Corpus Issues . . . 57
2.3.4 Language Bias and Low Resources . . . 58
2.3.5 Evaluating MT . . . 59
2.3.6 Summary and Thesis Focus . . . 60
2.4 Multi-word Expressions (MWEs) as a Bottleneck in MT . . . 61
2.4.1 Chinese-to-English Examples of MWEs in MT . . . 61
2.4.2 English-to-Chinese MT Issues with MWEs . . . 64
2.5 Summary . . . 73

3 Literature Review of Addressing Multi-Word Expressions in Machine Translation 74

3.1 Activities and Background Related to MWEs in MT . . . 74
3.2 Literature Review of MWE Investigations in MT . . . 77
3.3 Discussion: Revisiting MT Issues and Thesis Hypotheses . . . 80

4 Experimental Investigations into Multi-Word Expressions in Machine Translation Modelling **82**

4.1 Pilot Study with the State-of-the-Art Models 83

4.1.1 Bilingual MWE (BiMWE) Extraction Models 84

4.1.2 Experimental Examinations of MWEs in NMT 89

4.1.3 Discussion 93

4.1.4 Summary 94

4.2 Chinese Character Decomposition Models Addressing MWEs 96

4.2.1 Motivation for this Work 97

4.2.2 Chinese Character Decomposition for NLP 99

4.2.3 Two-Step Investigations 103

4.2.4 Model I: Chinese Radicals as NMT Features 103

4.2.5 Model I: Experimental Investigations 106

4.2.6 Model II: Chinese Character Pieces for NMT 113

4.2.7 Model II: Experimental Settings 115

4.2.8 Model II: Automatic Evaluation Using BLEU 117

4.2.9 Model II: Crowd-sourced Human Evaluation 119

4.2.10 Model II: Expert Validation - Phase One 120

4.2.11 Model II: Expert Validation - Phase Two 123

4.2.12 Discussion 127

4.3 Summary 129

5 Machine Translation Quality Assessments: Looking into Multi-Word Expressions **132**

5.1 Critique of MT Issues in Assessment Methods 133

5.2 AlphaMWE: A Multilingual Corpus with MWEs 135

5.2.1 Motivations 135

5.2.2 Related Work on Corpus Construction 137

5.2.3 AlphaMWE Corpus Construction 138

5.2.4 MT Issues with MWEs during AlphaMWE Construction . . . 143

5.2.5 Discussion . . . 154
5.2.6 Summary of AlphaMWE . . . 156
5.3 HiLMeMe: *H*uman-*i*n-the-*L*oop *M*T *E*valuation with *M*W*E*s . . . 157
5.3.1 HiLMeMe Model Design . . . 157
5.3.2 HiLMeMe Implementation and Platform . . . 160
5.3.3 Summary of HiLMeMe Evaluation Methodology . . . 161
5.4 Chapter Summary . . . 162

6 Conclusions **164**
6.1 Revisiting Hypotheses and Research Questions (RQs) . . . 164
6.2 Conclusions and Future Work . . . 166

Appendices **170**

A Further Details on NN and NMT Models **171**

B Publications on Work from this Thesis **177**

C Created Open-sourced Resources and Toolkits **181**

References **230**

Abbreviations

AAMT: Asian-Pacific Association for MT

ACL: Association for Computational Linguistics

AE: Automatic Evaluation

AI: Artificial Intelligence

ALPAC: Automatic Language Processing Advisory Committee (US)

AMTA: Association for MT in the Americas

ATS: Automatic Text Summarization

AlphaMWE: Multilingual parallel corpus with MWE annotations

BLEU: Bilingual Evaluation Understudy (metric)

BPE: Byte Pair Encoding

CCD: Chinese Character Decomposition

CJK: Chinese, Japanese, and Korean (script)

CL: Computational Linguistics

CNER: Chinese Named Entity Recognition

CNN: Convolutional Neural Network

CRF: Conditional Random Field

CWS: Chinese Word Segmentation

DA: Direct Assessment

DARPA: Defense Advanced Research Projects Agency (US)

DL: Deep Learning

EAMT: European Association for MT

EN-DE/ZH/PL/ES/IT/RU/FR/CS/LV: English-German /Chinese /Polish /Spanish /Italian /Russian /French /Czech /Latvian

HA: Human Assessment

HiLMeMe: Human-in-the-loop MT evaluation with MWEs

HOPE: **H**uman-Centric Evaluation Framework Based **o**n **P**rofessional **P**ost-**e**diting

HPPR: Harmonic Mean of n-gram Position penalty, Precision, and Recall

HTER: Human-targeted Translation Error Rate

IAMT: International Association for MT

ICCL: International Committee on Computational Linguistics (ICCL)

IE: Information Extraction

IR: Information Retrieval

LEPOR: **Le**ngth Penalty, **P**recision, n-gram **Po**sition difference Penalty and **R**ecall

*h*LEPOR: *h*armonic LEPOR (metric)

*cush*LEPOR: customised hLEPOR (metric)

LSTM: Long-Short Term Memory

METEOR: Metric for Evaluation of Translation with Explicit ORdering (metric)

ML: Machine Learning

MMT: Multi-modal Machine Translation

MT: Machine Translation

MTE: Machine Translation Evaluation

MTurk: Amazon Mechanical Turk

MWE: Multi-word Expression

MultiMWE: Multi-lingual Multi-word Expression corpus

NB: Naive Bayes

NER: Named Entity Recognition

NIST: National Institute of Standards and Technology (US, also metric)

NLP: Natural Language Processing

NMT: Neural Machine Translation

OOV: out-of-vocabulary (words)

PARSEME: Parsing and Multiword Expression EU project

PBSMT: Phrase-based Statistical Machine Translation

PER: Position-independent Word Error Rate

POS: Part-of-Speech

QE: Quality Estimation

RNN: Recurrent Neural Network

SMT: Statistical Machine Translation

SRL: Semantic Role Labelling

STC: Short Text Categorization

SVM: Support Vector Machine

TER: Translation Edit Rate

THUMT: TsingHua University Machine Translation system

TQA: Translation Quality Assessment

WER: Word Error Rate

WMT: Workshop on Statistical Machine Translation

WSR: Web Search Ranking

List of Figures

1.1 Chinese-English MT example from Báihuà (白話) 6

2.1 Typical phrase-based SMT workflow and components 17
2.2 SMT branches from string level to the sophisticated 18
2.3 NMT components 19
2.4 NMT development and branches: MWEs in the semantic sector. . . 20
2.5 Dependency example in a Chinese sentence with English translation from Wu et al. 2017 (IJCAI). 21
2.6 Classification of Human Assessment Methods 27
2.7 Automatic Quality Assessment Methods for MT 34
2.8 Meta-eval of MTE: applying to general NLP tasks 48
2.9 Chinese-English MT example from poem (詩歌 shī gē) 63
2.10 Chinese-English MT example from proverb (文言 Wényán, Classical Chinese) 64
2.11 English-to-Chinese MT: named entity 65
2.12 English-to-Chinese MT: *Engli*-Chinese 67
2.13 English-to-Chinese MT: misusing MWEs 67
2.14 English-to-Chinese MT: document-level vs sentence-level 68
2.15 English-to-Chinese MT: Pinyin on MWEs 70
2.16 English-to-Chinese MT: Mandarin vs Cantonese, Simplified vs Traditional Chinese. 72

4.1 MultiMWE corpora extraction workflow. 84
4.2 Samples of MPAligner aligned bilingual MWEs (no filtering). 88

4.3 Extracted Zh-En MWEs without pruning. 89
4.4 Samples of DE-EN MWEs after pruning with threshold 0.70. 90
4.5 Zh-En MT Examples: Baseline vs Base+MWE 94
4.6 Radical 刂 (Dāo): (leftmost) pictogram to regular script (rightmost) containing only two strokes. 99
4.7 Examples of the decomposition of Chinese characters. 99
4.8 Chinese word to character level MT. 100
4.9 A mapping between English and Chinese in composition 102
4.10 Example of a Chinese radical as an independent character 104
4.11 Example of a Chinese radical as a non-independent character 104
4.12 Architecture of NMT with multi-embedding. 105
4.13 Chinese character examples from the IDS dictionary. 115
4.14 Learning curves from different models with BLEU metric 119
4.15 Samples of the ZH-EN MT output at 100K learning steps. 122
4.16 Samples of the ZH-EN MT output at 180K learning steps. 124
4.17 Character decomposition examples of qiáo and liáng 126

5.1 Workflow to prepare the AlphaMWE corpus. 140
5.2 Sample comparison of outputs from four MT models. 142
5.3 AlphaMWE corpora samples with two sentences 144
5.4 MT issues with MWEs: common sense. Pinyin offered by GoogleMT with post-editing. 145
5.5 MT issues with MWEs: super sense. 146
5.6 MT issues with MWEs: abstract phrases 147
5.7 MT issues with MWEs: idioms 148
5.8 MT issues with MWEs: metaphor 149
5.9 MT issues with MWEs: context-unaware ambiguity 150
5.10 MT issues with MWEs: social/literature-unaware ambiguity 150
5.11 MT issues with MWEs: coherence-unaware ambiguity 151
5.12 Additional vMWEs or MWEs 155

5.13 HiLMeMe Algorithms 161
5.14 HiLMeMe Platform Workflow 162

A.1 Recurrent Continuous Translation Models (RCTM) and RLM model from [131]. 172
A.2 Convolutional sentence embedding example from [131]. 173
A.3 Fully RNNs encoder-decoder NMT from [263]. 173
A.4 GRU (left) [45] vs LSTM (right) [122]. 173
A.5 BiRNN for encoder and RNN for decoding with attention (soft-alignment) from [9]. 174

List of Tables

4.1 Added Chinese Patterns for MWEs from the LCMC Tags 86

4.2 DE-2-EN NMT BLEU Scores with 20k Transformer Learning Steps . 91

4.3 Zh-2-En NMT BLEU scores with 20k Transformer learning steps. . . 93

4.4 Model Settings for Radical as MT Features 107

4.5 BLEU Scores on the NIST06 Development Data 108

4.6 NIST Scores on the NIST06 Development Data 109

4.7 Broader Metric Scores on the NIST06 Development Data 110

4.8 BLEU Scores on NIST08 Test Data 110

4.9 NIST Scores on NIST08 Test Data 111

4.10 Broader Metrics Scores on NIST08 Test Data 112

4.11 Chinese → English NMT BLEU scores for increasing learning steps. 117

4.12 Human evaluation results for systems using Direct Assessment 120

Abstract

An Investigation into Multi-Word Expressions in MT

Lifeng Han

Multi-word Expressions (MWEs) present challenges in natural language processing and computational linguistics due to their popular usage, richness in variety, idiomaticity, and non-decompositionality, which are present in the text content in which they are used. This is a typical level of expectation in the machine translation (MT) field where we require algorithms to perform a translation from one human language to another automatically while requiring high-quality output including features such as adequacy, fluency, and keeping the same or making creative and correct style decisions in that output.

In this thesis, we carry out an extensive investigation into MWEs in Neural MT. Firstly, we carry out a review of relevant literature which includes experimental work on re-examining state-of-the-art models that combine knowledge of MWEs into MT systems, but with new language pairs setting to see what gaps might exist in the published literature. Secondly, we propose our new models on how to address MWE translations. This includes a design where we treat MWEs as low-frequency words and phrases translation issues, by integrating language-specific features such as strokes and radicals representation of Chinese characters into the learning model, expecting that this will facilitate improved accuracy. Thirdly, to properly examine different MT models' performances in the context of MWEs, we need to carry out a new evaluation methodology, and in light of this, we create a multilingual parallel corpus with MWE annotations (AlphaMWE). During the creation of this corpus, we classify the MT issues on MWE-related content into several categories with the expectation that this will help future MT researchers to focus on one or some of these in order to achieve a new state of the art in MT performance, ultimately moving towards human parity. Finally, we propose a new methodology for human in the loop MT evaluation with MWE considerations (HiLMeMe).

Chapter 1

Introduction and Background

"路漫漫其修遠兮，吾將上下而求索。"
lù màn màn qí xiū yuǎn xī, wú jiāng shàng xià ér qiú suǒ.
–屈原 *(qū yuán, 340BC - 278BC)*

This chapter covers an introduction to the thesis and some of the background knowledge behind the research topic. It contains the following sub-sections: Computer Powered AI (artificial intelligence), Automatic Processing of Natural Languages, MT (machine translation) Applications, Summary and Thesis Structure.

1.1 Computer Powered Artificial Intelligence (AI)

Artificial Intelligence (AI) is a very broad topic that relies on many research fields such as reasoning, computational modelling, automation, cognitive modelling, mathematical modelling, engineering, psychology, philosophy, linguistics, and ethics. AI adopts a strategy of mimicking human intelligence (HI) such as cognitive brain and learning processes and tries to achieve certain levels of Machine Intelligence (MI) in corresponding tasks, which are often specified as domain and application specific AI, or in general which are known as general AI and can perform at levels even beyond human experts.

AI became a research discipline back in the 1950s when scientists were still considering other alternative names for it, such as Thinking Machines, Cybernetics, Automata Theory, Complex Information Processing, etc., and the term "AI" was commonly acknowledged to have been coined by the Brain Storm workshop "Dartmouth Summer Research Project on Artificial Intelligence" organised by mathe-

matician and computer scientist John McCarthy in 1956 who was then an assistant professor in Dartmouth College, New Hampshire, USA. These alternative names at that time reflect how researchers tried to tackle the problems and create a new research topic, as well as the different backgrounds and disciplines that they were from. For instance, *Thinking Machines* indicates the ideology of "thinking", "mind" and "cognitive" ability mimicking the human brain, and this inspires the methodology from cognitive science and neural-biology related disciplines; *Cybernetics* is from communications and automatic control systems; *Automata Theory* involves theoretical computer sciences, such as Finite State Machines (FSM, or infinite cases), with "automata" as a borrowed word from Greek meaning "self-making"; and *Complex Information Processing* indicates the perspectives from Mathematics and Information Theory.

AI started to draw scientists' attention in the middle of the last century due to the newly developed technologies of computing machinery, especially the programmable ones at that time, their demonstrated capabilities, and their potential on what human experts could expect them to achieve. These technologies and the later developments, are what people called computer technologies or computer science.

From the invention of rotor-based cipher machine for Enigma cipher code in 1915, to the starting concept of AI in the 1950s, to IBM's Deep Blue computer which won a chess series against world champion Garry Kasparov under regular time controls in 1997, to IBM's Watson computer which won the quiz show Jeopardy against champions Brad Rutter and Ken Jennings in 2011, to Deep Mind's AlphaGo which won the Go game against world champion Ke Jie in 2017, there have been many exciting points in AI development.

Computer powered AI has very broad research fields and applications which include reasoning, knowledge representation, planning, learning, natural language processing, perception, motion and manipulation, social intelligence, general intelligence, AI for gaming, speech recognition, image processing, self driving cars, intelligent robots, human computer interaction (HCI), medical diagnosis, AI generated

art (e.g. paintings), AI for maps (e.g. street view), search engines, information retrieval, multi-modal applications, spam filtering, etc.

Different AI tasks also impact and interact with each other. For instance, the street view function from AI maps will depend on image processing and cloud computing technologies; intelligent robotic control needs speech recognition and pattern recognition; multimodal IR includes IR, video and image processing, etc. [3, 250]. All theses AI developments have been closely connected, attached to, or rely upon computer hardware, software, and computational power evaluations, in addition to the learning and reasoning methodologies themselves.

IBM's Watson is a question-answering (QA) computer system which belongs to the category of automatic processing of human languages, also known as natural language processing (NLP). In the next section, we introduce the topic and research fields of NLP and from there we elicit our thesis topic on *machine translation* (MT) and the research focus of this thesis.

1.2 Automatic Processing of Natural Languages

Natural language processing (NLP) is one of the initial goals of AI, as well as one of the most important applications of it. NLP includes natural language understanding (NLU), inference (NLI), translation and generation (NLG).

Since the second World War, computers have been used for de-cypherment, communications, and information processing, which all involved pioneering NLP tasks. Up until the present date known as the big-data era with the popularity of information on the internet and the support of computer servers, there have been an increasing number of research topics in NLP and computational linguistics.

1.2.1 Natural Language Processing (NLP) in General

From a broader viewpoint, natural language understanding (NLU) and inference (NLI) deals with machine reading comprehension and this involves many sub-tasks including syntactic sentence parsing (e.g. constituency phrase-structure parsing),

semantic sentence parsing (e.g. dependency parsing), entity recognition (e.g. person, location, organisation), entity linking, phonology, morphology and word segmentation, paraphrasing, text entailment, lemmatisation, semantic role labelling, grammatical error correction, sentiment analysis, emotion detection, and text classification. NLU and NLI impact natural language generation (NLG) and NLP applications, such as the tasks of text summarisation, question answering, machine translation and multilinguality, poem and story generation, information retrieval and text mining, information extraction and knowledge acquisition, discourse and pragmatics, dialogue and interactive systems, and computational social science.

In recent years, the topic of multi-modal NLP research connects image processing and text processing models, such as image captioning where a summary of text needs to be generated from a image, multi-modal machine translation where both images and texts are used as input for learning models, and speech translation where speech recognition and text translation are included in model components.

Ethics and social issues in NLP is another very recent research topic which is also linked to trustworthy AI[1]. This includes the privacy protection and de-identification of private information on NLP data, such as clinical data and medical records.

Among all mainstream NLP tasks, machine translation (MT) is one of the earliest AI tasks, which formed into a research topic from the 1950s [283] and was reviewed in the ALPAC progress report in "Languages and Machines" from the US National Academy of Sciences describing it as a very difficult and challenging AI task in 1966 [4]. During its development, MT has been a very sophisticated research topic that involves both NLU and NLG, and has connected many other NLP tasks such as sentence parsing, word segmentation, entity recognition, paraphrasing and text summarisation. In this thesis, we will tackle the challenges and issues that exist in present day MT research.

[1]Ethics in NLP including workshop series at https://aclweb.org/aclwiki/Ethics_in_NLP and ISO trust-worthy AI guide https://www.iso.org/news/ref2530.html

1.2.2 Machine Translation (MT)

Machine Translation (MT) is a task to translate a text in one human language into another [138]. MT researchers try to use computational modelling to address the conventional linguistic text translation problems. MT is an artificial work with much human and computer intelligence involved. It is a very challenging task for MT models to achieve accuracy at the level of a human's expert performance or what linguists expect as output. From early 1950s to nowadays, MT has experienced several very different research methods, from hand crafted rules, to example and translation memories, to statistical and hybrid methods, to machine learning (ML) and neural network models. This evolution of techniques used in MT reflects the evolution of programming languages from imperative to declarative, as well as that of computer hardware, memories, and computational support. Thanks to the massive available multilingual corpora on the Internet and the continuously increasing computational power of computer servers, MT outputs have improved from nonsense to making sense, to getting correct. However, it is still a long way to reach real "human parity" and in this thesis we will discuss the current problems in MT and focus on one of the bottlenecks namely Multiword Expressions (MWEs) which will be explained at length in future chapters (Section 2.4).

As one simple example, let us examine the Chinese-to-English sentence translation from "白話" (bái huà, daily Chinese) in Figure 1.1 [109]. Chinese characters commonly have two writing systems, i.e. simplified and traditional, with the simplified Chinese characters being deployed in mainland China only in recent history since the founding of P. R. China in 1949 and the traditional characters still being used in Hong Kong, Macau, Taiwan, and other areas, in addition to historical documents. The traditional Chinese characters in this example sentence are annotated with pronunciation in Pinyin format, i.e., "小明去學校上課了 Xiǎo míng qù xué xiào shàng kè le". Pinyin is a phoneticism system created with alphabets to represent the pronunciations of Chinese characters, including five different tone patterns, flat (e.g. ā), up-rising (e.g. á), down-up (e.g. ǎ), down (e.g. à), and light (e.g. a).

ZH source:	小明去學校上課了
ZH pinyin:	Xiǎo míng qù xué xiào shàng kè le
EN reference:	Xiao Ming went to school to attend classes
EN MT output:	Xiao Ming went to school

Figure 1.1: Chinese-English MT example from Báihuà (白話)

In this example, we give one reference translation in English, and one MT output in English where the MT model produced a non-perfect translation, i.e., some meaning loss. This example sentence is in modern Chinese called as "白話" (bái huà), also as "普通話" (pǔ tōng huà) or "現代漢語" (xiàn dài hàn yǔ), which is different from ancient Chinese, or ancient scholarly Chinese, named as "文言" (wén yán, Wenyan) that was used by Chinese scholars both in spoken and writing systems in the historical dynasties of the Chinese empire. In this thesis, we will mostly study MT examples in Chinese-English, and slowly increase the difficulty level with Chinese poems and proverbs from Classical Chinese ("文言" wén yán), since a main part of the experimental examinations in the thesis will be carried out on Chinese-English bilingual corpora.

In the example translation in Figure 1.1, we show how important it is to understand Chinese pattern expressions in order to translate the correct tense and overall information in a sentence. The MT output has lost the *aim* of Xiao Ming's action to go to school, i.e., what is his purpose to go there (*to attend classes*). This reflects an overall loss of *adequacy* in MT. In Chinese, there is no direct past tense in a verb, so an MT model needs to acquire the knowledge of language expression patterns to be able to translate the tense information and the purpose of the action here. The Chinese pattern "去 (qù) ... 了 (le)" is a simple dis-continuous Chinese multi-word expression indicating to express a past tense action (went to do something, went to somewhere, or went to somewhere for something).

1.2.3 NLP & MT Organisations and Events

There have been a remarkable number of large projects and organisations focusing on NLP and MT research. The "Cracking the Language Barrier" is a federation of European projects and organisations working on technologies for a multilingual Europe, aiming at development of cross-lingual and multi-lingual language technologies via shared scientific tasks, evaluation campaigns, strategy papers, data management, resource and technology repositories, etc. [2] It is joined by many organisations including the Translation Automation User Society (TAUS, a forum formed in 2004), World Wide Web Consortium (W3C, founded in 1994), Multilingual Europe Technology Alliance (META-NET, a Network of Excellence), European language resource association (ELRA), Globalisation and Localisation Association (GALA), and Conversational Interaction Technology Innovation Alliance (CITIA), etc.

Other large organisations include CLARIN (Common Language Resources and Technology Infrastructure) which is a European research infrastructure for humanities and social sciences aiming at improving Europe's multilingual competence, EC-SPM (European civil society platform for multilingualism), which advocates practices of and research into multilingualism, EFNIL (European Federation of National Institutions for Language), which encourages the study of the official European languages and a coordinated approach towards mother-tongue and foreign-language learning, and ELEN (European Language Equality Network), which is dedicated to the protection and promotion of regional, minority, and endangered languages.

In addition to these organisations, there are specifically Machine Translation (MT) focused groups, such as the International Association for MT (IAMT) and its branches including the European Association for MT (EAMT), the Association for MT in the Americas (AMTA), and the Asian-Pacific Association for MT (AAMT). Together with the Association for Computational Linguistics (ACL) and International Committee on Computational Linguistics (ICCL), these groups and

[2] Home page: http://www.cracking-the-language-barrier.eu

professional societies have organised MT-focused workshops and conferences, including the International Workshop on MT (WMT, now at conference level), NIST MT challenges, International Workshop on Spoken Language Translation (IWSLT), and affiliated shared tasks focusing on MT models, evaluations, and quality estimations, which usually attract system participation from MT groups all around the world.

1.3 MT Applications

MT has broad applications, especially as its quality is getting much improved and more reliable. We list some of these from four different aspects, MT for a). social good, b). for healthcare, for c). broad humanities, and d). MT impacts on other AI research. These four examples sometimes have interactions with each other, e.g. a) and c).

For social good, MT is one of the widely used technologies for helping people. People from different language backgrounds have always sought communication with each other, from the Babel Tower story to the modern era. MT helps the field of secondary language learning for minorities, immigrants, dis-advantaged people and in general [162, 126]. It also helps multilingual social network communications, for instance, the automatic translation option in many social network applications.

MT is also being used in healthcare systems. For instance, MT research has been applied to medical and clinical data processing to help the doctors understand more about their patients; MT technology is one important part in chat-bot research among other NLP tasks, which can help doctors in psychological treatments in a multilingual setting.

For broader humanities, MT helps international business, economic growth, tourism, cross-language communication with customers. It supports professional translators and language service providers, e.g. via computer aided translation tools, as well as helping international family members who do not speak common languages.

Finally, MT impacts many other AI topics, such as cross-lingual natural language

processing (NLP), information retrieval (IR) and information extraction (IE), speech recognition and translation, image captioning and translation, etc. [255, 21]. State of the art MT models have also been applied to traditional AI tasks such as optical character recognition (OCR) to help the digitisation process [67].

Notable MT and NLP applications include Grammarly (paraphrasing, text entailment, anti-plagiarism detection, grammar checking), Google, Baidu, Microsoft Bing, and Yandex (their search engines and MT services).

So far, in this chapter, starting from the concept of Computer-Powered AI, we have introduced NLP as one of the earliest AI tasks, a major branch being MT, and introduced many notable NLP-MT organisations and international conferences and events. In Section 1.3, we listed some examples of the broad application of MT technology in our society. In the next section, we will briefly lay out the research background, hypotheses and corresponding research questions of this thesis.

1.4 Research Background and Thesis Hypotheses

1.4.1 Research Background

From rule based MT to statistical (SMT) and neural MT (NMT) models, some challenges remain in front of MT researchers, such as how to best integrate linguistic knowledge, how to tackle low-frequency words and phrases translation, and how to address idiomatic multi-word expressions (MWEs). One of the bottlenecks in the development of MT models is the popularity and variety of multi-word expressions (MWEs) being used in our language, both verbally and in written text [235, 10, 125].

In this thesis we focus on the challenge of MWEs in MT, since it has a broader connection and influence. For instance, it is connected to linguistic awareness, related to low-frequency phrases translation from statistical point of view, and as well as related to idiomaticity in translation.

Various definitions of MWEs have included both syntactic structure and semantic viewpoints from different researchers covering syntactic anomalies, non-

compositionality, non-substitutability and ambiguity [53]. For instance, [11] **define MWEs** as "lexical items that: (i) can be decomposed into multiple lexemes; and (ii) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". MWEs have a broad coverage of linguistic phenomena, both syntax and semantics, such as compound nouns, named entities, verb particle constructions, discourse markers, collocations, lexical bundles, idioms and metaphors, and more. MWEs can appear in unexpected syntax and can lead to ambiguity. The syntactic richness and idiomaticity of MWEs also inspires MT and NLP researchers to design new methodologies, in addition to the challenges it presents [53].

As examples of MWEs in English, there are verb-noun patterns "jump the gun" and "kick the bucket", noun-noun combinations "hit man" and "word salad", adjective-noun combinations like "cold shoulder" and "hot dog", binomial expressions such as "by and large" and "nuts and bolts", just to list a few[3].

MWEs can be *continuous* where all component words form an n-gram together without gap words. However, they can also be *dis-continuous* with some other words inserted in-between. For instance, in our MT example sentence from Figure 1.1, the Chinese MWE (underlined) "去 (qù) ... 了 (le)" meaning "went to ...(do something, or somewhere)" has contextual words inserted into the MWE which carry the objects of the MWE. In this case, "去 (qù)" is a Chinese word indicating "go to something, or go to do something", and "了 (le)" is a marking character to be affiliated to other words indicating a past or perfect tense. The combination of these two Chinese characters form a discontinuous MWE pattern indicating "(someone) went to somewhere or to do something".

1.4.2 Thesis Hypotheses and Research Questions (RQs)

Our hypotheses and research questions are set out below:

[3]Some MWE examples are taken from the work in [252].

Our research hypothesis can be stated simply as follows *"MWEs set challenges for MT and investigation into MWEs can help to improve MT from both translation modelling and translation evaluation perspectives."* Expanding on this we propose that:
Hypothesis.a - *Current methodologies addressing MWEs in MT can be improved to include language-specific characteristics and features.*
Hypothesis.b - *MWEs have a high presence in general usage; however, due to their unexpected syntactic patterns and idiomaticity, exact MWEs do not have high statistical frequency in repeated appearance in the same text segments. Thus they can be addressed using technologies inspired by tackling rare word and OOV words issues.*
Hypothesis.c - *Current MT quality assessment models and test suites have disadvantages and we can tackle these by introducing better test suites from the data sanity point of view and the design of new evaluation methodologies that deploy MWEs as important features.*

Based on our three research hypotheses, we propose the following research questions (RQs) to investigate:

RQ-I:
We assume that the proper integration of MWE knowledge into MT models can help to improve translation quality of MWEs in the source text and further improve the overall text translation. To investigate this we propose a to do list:
- Validate and re-examine the state-of-the-art of existing models of integrating MWEs into MT but with different language pairs for experimental work.
- Design new methods for integrating MWEs into MT by addressing weak points of existing models, focusing on exact language pairs.
- Verify the proposed models with quantitative and qualitative evaluations, having experts' validation wherever possible.

RQ-II:
We assume a new test suite with MWE annotations can be created for better assessment of state of the art MT models, and an improved MT evaluation methodology can be achieved with MWEs in consideration and human-in-the-loop workflow. To investigate this we propose a to do list:
- Revisit the problems and challenges in MT modellings and their assessment, perform a critique review.
- Create a corresponding new test suite with MWE annotations in bilingual and multilingual settings for better MT quality assessments towards real human parity.
- Address the issues in conventional MT assessment methods, both human judgement and automatic metrics, and design new methodologies that incorporate MWEs into the evaluation system.

Regarding RQ-I on translation modelling for the pilot study, we will re-examine the integration of MWEs into MT as augmented data (addition of training corpus) in the form of bilingual aligned MWE terms (*BiMWEs*). Then, we will propose

our own language specific model focusing on Chinese-English language pair using *character decomposition* technology for Chinese towards better MWEs acquisition and translation accuracy. In the next section, we present the rest of the thesis structure.

1.5 Thesis Structure

The rest of this thesis consists of the following chapters:

(i) **MT Problem Analysis** (Chapter 2): We discuss MT and its problems, provide an MT and MT evaluation (MTE) literature review, and finish by highlighting MWEs as one of the bottlenecks in MT with examples.

(ii) **Literature Review** (Chapter 3): this chapter presents the literature review on the state of the art models investigating MWEs in MT, with discussions, and leading to introduction of our hypothesis and research questions for the thesis.

(iii) **Experimental work on MT** (Chapter 4): this chapter starts with a pilot study re-examining the very recent models on integrating bilingual MWEs (BiMWEs) into MT, but with new language pairs and high-performing scenarios, as well as new NMT structure. Then, we propose a new methodology that addresses MWE translation in a language-specific condition, i.e. Chinese-to-English, by two-step investigations: looking into Chinese linguistic features and in-depth decomposition models to tackle MWEs from a low-frequency phrases point of view, as well as the interaction of decomposition models and BiMWEs.

(iv) **Further Developments on Evaluating MT** (Chapter 5): this chapter further investigates the research questions by exploring a human in the loop methodology in evaluating MT outputs looking at MWEs, with the AlphaMWE multilingual parallel corpora we created as a first step.

(v) **Conclusions** (Chapter 6): this final chapter concludes the thesis by summarising the research work carried out, to what degree we addressed the hypotheses and research questions, and presenting possible future work.

Chapter 2

Machine Translation: Development, Issues and Challenges

In this chapter, we first present a brief literature review of machine translation (MT) technology, its evolution from rule-based methods to example-based, statistical methods, hybrid models, and the state-of-the-art neural network models which our experimental work will rely upon. Then, we present a survey of work on MT evaluation methodologies from conventional human judgements and their development, to automatic evaluation metrics, and quantity estimation (QE) models that do not rely on reference translations, as well as meta-evaluation, i.e. evaluating the evaluation methods. Thirdly, we discuss some of the outstanding issues in MT and MT evaluation paradigms, and from here, we elicit the topic of handling multi-word expressions (MWEs) as one of the challenges in MT modelling, as well as how MWEs can play an important role in MT evaluation, and this thesis' focus on MWE investigations in MT including the aspects of both translation model itself and translation evaluation. Our next chapter will focus on MT modelling aspect while the chapter after will focus on MT evaluation part both looking into MWEs.

2.1 MT Paradigms: from Rules to Neural

Machine Translation (MT) is a apparently straightforward, but in practice very complex application of computing. We will utilise state-of-the-art MT models in the work reported here, but we first give a brief review of the development of MT technologies, and then we outline various assessment methodologies for MT, in order to set the context for our work.

Machine Translation (MT) has a long history dating from the 1950s [283] as one of the earliest topics in artificial intelligence (AI) and intelligent machines. It began with rule-based MT (RBMT) systems that apply human-defined syntactic and semantic rules of source and target languages to the machine, then moved to example-based MT (EBMT), statistical MT (SMT), Hybrid MT (e.g. the combination of RBMT and SMT) and then in recent years to Neural MT (NMT) models [207, 37, 142, 131, 45, 9, 273].

MT gained much more attention from researchers following the publishing of IBM mathematical models I-to-V proposed in the 1990s [27]. Representative SMT work includes the word alignment models [210], introducing of Minimum Error Rate Training (MERT) [208], phrase-based SMT model [147], hierarchical structure models [44], and large parallel standard data development, e.g. [141], etc.

The increasing complexity of MT technologies encouraged many research groups to develop open source tools to promote the advance MT technologies, such as the Moses system featuring statistical phrase-based MT [144], Joshua featuring parsing-based translations [168], Phrasal incorporating arbitrary model features [41], CDEC favouring finite-state and context-free translation Models [68], and NiuTrans featuring syntax-based models [290]. Some advanced information technology companies also built their own MT systems, such as the MT engines by Google[1], Baidu[2], Yandex[3], and Microsoft Bing[4].

[1] translate.google.com
[2] translate.baidu.com
[3] translate.yandex.com
[4] www.bing.com/translator

A key development in recent years is the development of Neural MT, primarily thanks to the work on word-to-vector embedding from [195], NMT was introduced in [131, 45, 9] which deploys both deep learning (DL) and word representation (WR) approaches. In addition to the above mentioned IT companies who have moved their MT systems into neural models, there are also other NMT platforms emerging from other institutes, such as Nematus from the Edinbrugh NLP group [244], OpenNMT from Harvard NLP [140], THUMT from the TsingHua University MT group [294] which is an implementation of Google's all-attention based Transformer model [273].

In earlier work in the mid-1990s, the NMT structure [203] did not work in terms of the quality of output, which may be due to the limitations of computational power of machines and the amount of available corpora for training at that time, though neural networks were also explored later as sub-components in SMT pipelines, e.g. to smooth or re-rank system output candidates as language models [242, 17]. One of the driving forces of NMT research was the launch of the NMT Workshop by Google [5], in addition to the traditional WMT workshops [6] that Incorporated NMT sections or separate NMT translation tasks. MT systems submitted to WMT shared tasks now mostly use neural models.

There are some other advanced MT topics and issues which remain as challenges to MT such as multi-modal [71, 124, 30], multi-lingual [31, 129] and syntactic [15, 1, 165] NMT, but these are outside the scope of this thesis.

2.1.1 Statistical vs. Neural Machine Translation: SMT vs NMT

In traditional SMT, especially phrase-based SMT (PBSMT), there are very clear components in the model pipelines, such as parallel corpora cleaning and preparation, word alignment in the bilingual corpora, phrase table extraction from the aligned corpora to generate translation pairs, language model generation from mono-

[5] sites.google.com/site/acl17nmt/home
[6] www.statmt.org/wmt17/

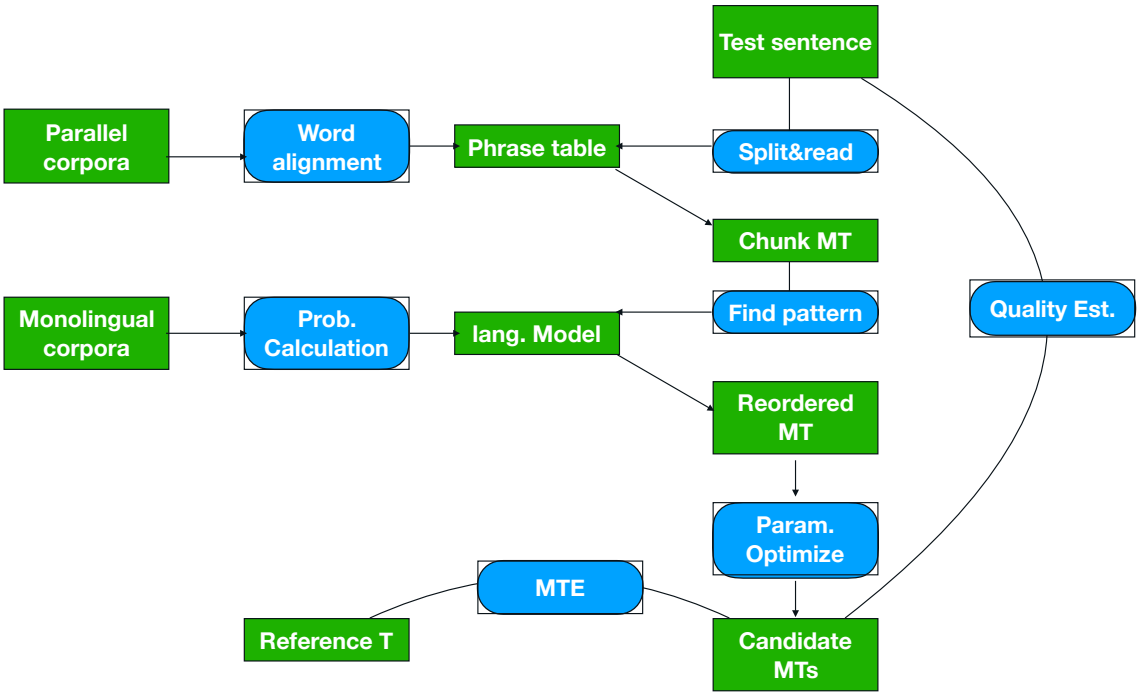

Figure 2.1: Typical phrase-based SMT workflow and components

lingual corpora, word reordering with the trained language model, parameter tuning e.g. minimum-error-rate-training, and translation quality estimation from the system generated candidate translations (see Figure 2.1 "Typical PBSMT workflow and components"). Extended and more sophisticated work beyond these core elements included hierarchical SMT structure with linguistic patterns, Tree-to-String, String-to-Tree, and Tree-to-Tree syntactic MT, and dependency parsing based MT, etc (see Figure 2.2 "SMT branches").

NMT models treat the task of MT as an encoder-decoder workflow, which is very different from the conventional SMT (especially PBSMT) structure [46, 142]. The encoder is applied in the source language side, turning the sentences into vector representations, while the decoder is applied in the target language side generating the words from the target side vectors. Recurrent Neural Network (RNN) models are usually used for both encoder and decoder, though there is some research employing convolutional neural networks (CNN), e.g. [45, 131]. The hidden layers in the neural nets are designed to learn and transfer information [204]. Later development of NMT has seen application of Long-Short-Term-Memory (LSTM), Gated Recurrent Units

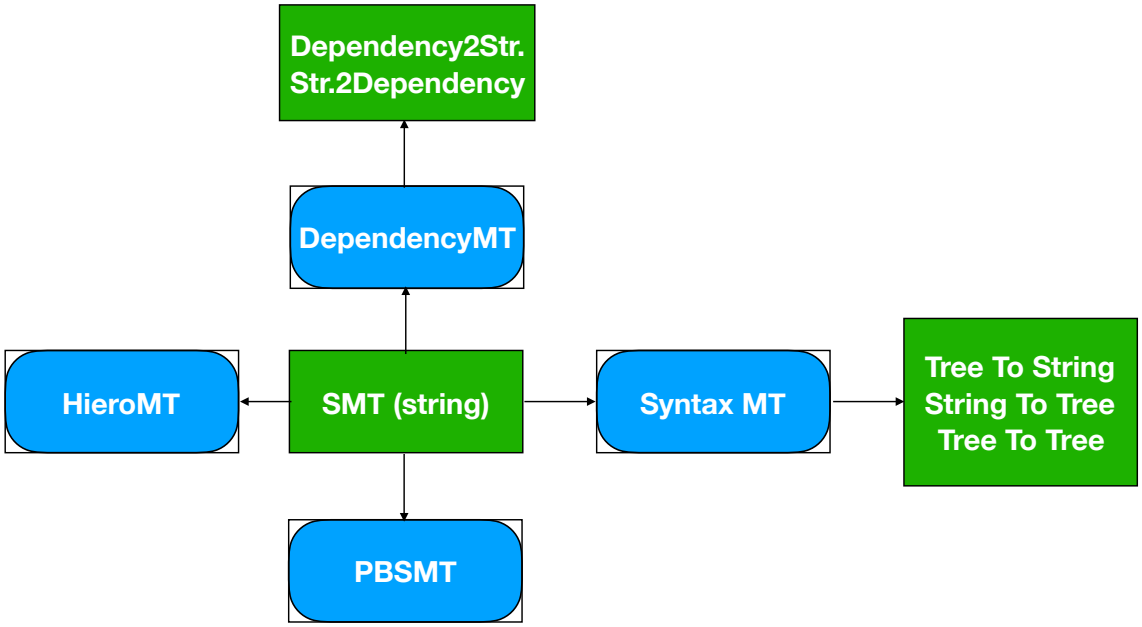

Figure 2.2: SMT branches from string level to the sophisticated

(GRU) [49, 48].

One attraction of these forms of neural modelling is that they can produce higher fluency target language output compared to traditional methods such as SMT. They also get high scores in automatic evaluation challenge tasks for metrics such as BLEU, METEOR, and LEPOR [114, 102], and even with human judgement from the annual WMT shared task.

There are some drawbacks to NMT models including lack of alignment information between source and target side, and less transparency. To address these, an attention mechanism was introduced into the decoder, first by [9] to pay attention to parts of the source sentence selectively, instead of the whole sentence always, when the model is doing translation. This idea is similar to alignment functions in SMT and what human translators usually do when they undertake a translation task, and follows the fact that earlier, attention mechanisms were applied to neural nets for image processing tasks [157, 58]. Recently, attention-based models have appeared in most NMT projects, such as the investigation of global attention-based architectures [176] and target information [218] for pure text NMT, and the exploration of Multi-modal NMT [124]. To generalise the attention mechanism in the

source language side, a coverage model is introduced to balance the weights of different parts of sentences into NMT by [268, 193]. Lately development of NMT saw the introduction of the all/only attention based model Transformer [273]. In addition, researchers are discussing how to use a pre-trained model called BERT [59], the Bidirectional Encoder Representations from Transformers [153]. This is shown in Figure 2.4 "NMT development and branches" where we list only some sample sub-topics, with many still left to explore.

Some of the components and strategies from SMT are also very useful in NMT methodologies. For instance bilingual phrase table utilisation in NMT alignment which is referred to as attention in NMT (mentioned above), and the syntax tree structure translation. Some **traditional MT research issues** are the same or similar in NMT also, such as the **disambiguation** task and accurate translation of certain ambiguous words, phrases and multi-word expressions (including metaphor, proverb and poetic expression), and the addressing of **unseen or rare words** translation.

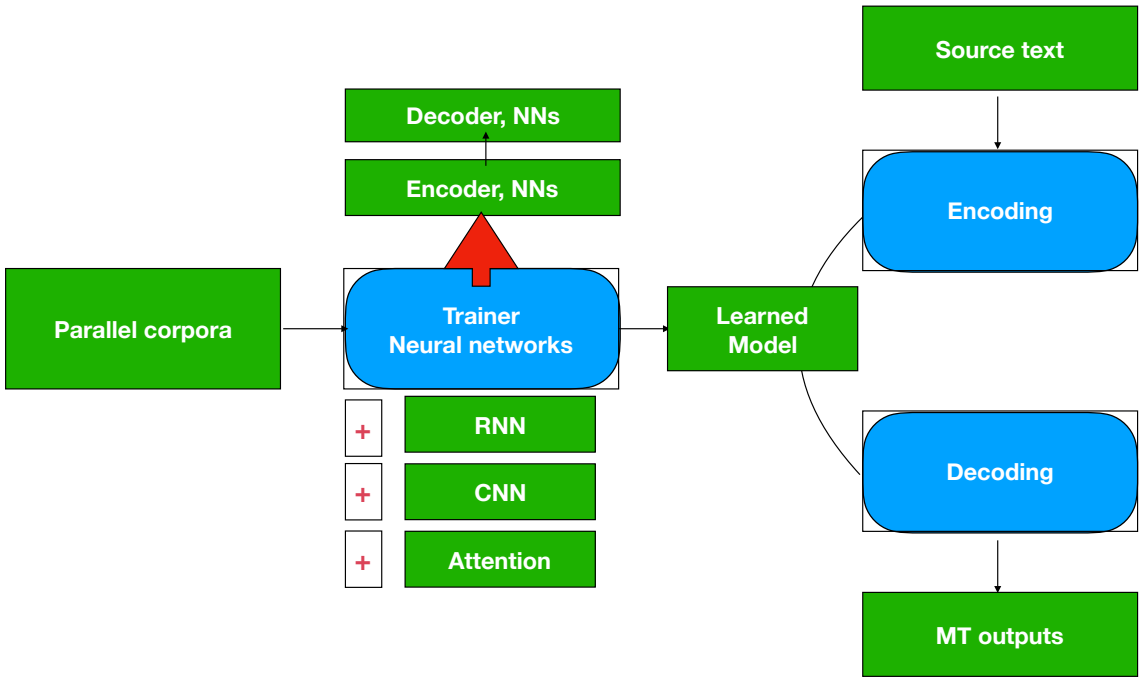

Figure 2.3: NMT components

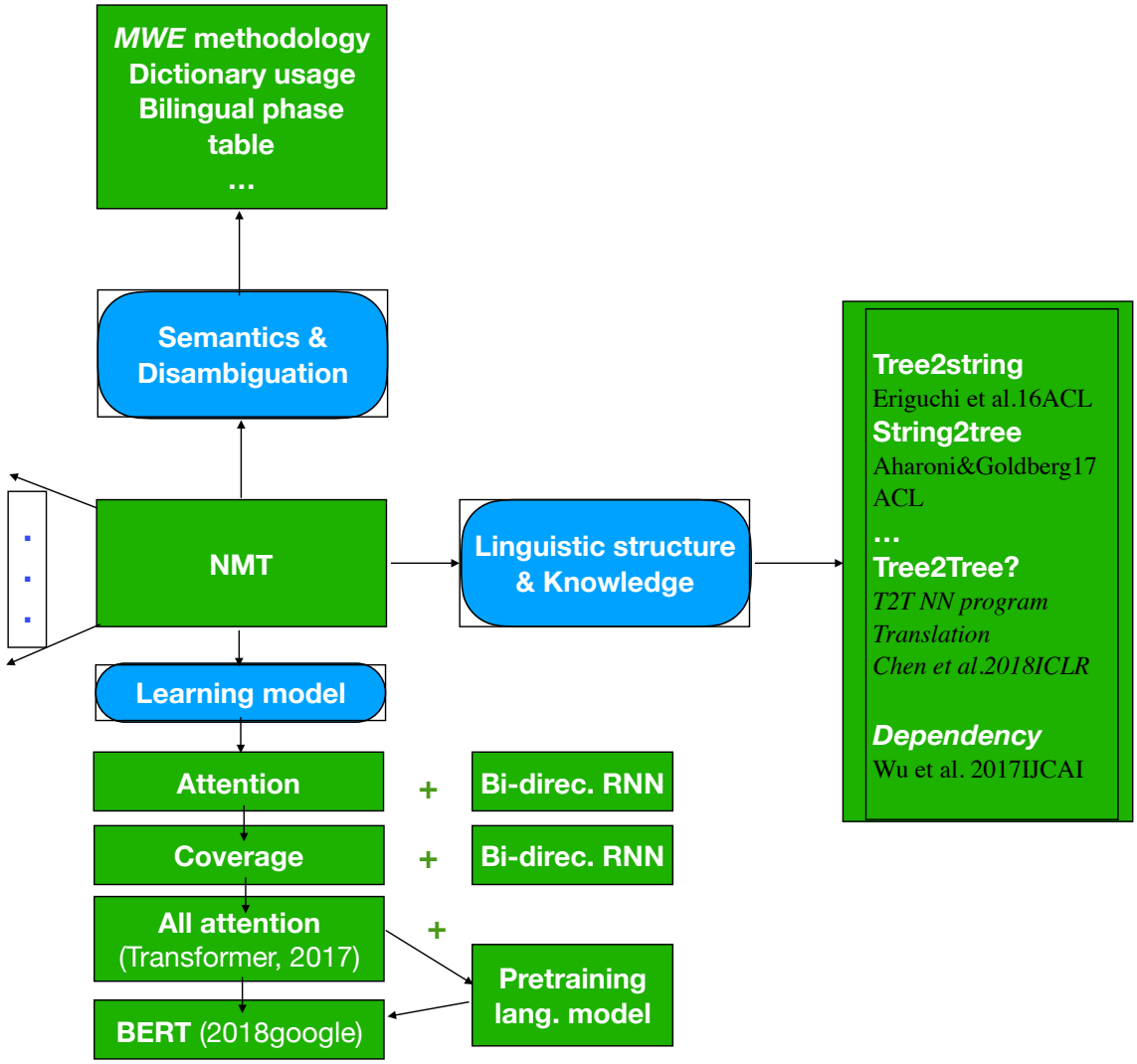

Figure 2.4: NMT development and branches: MWEs in the semantic sector.

Linguistic Structure within NMT

Similar to SMT, researchers have previously tried to integrate **syntax**, linguistic sentence structure and knowledge into NMT systems. There have been explorations of Tree-to-string [72, 165], String-to-Tree [1, 282], and dependency structure NMT [287] which applied dependency (Figure 2.5) to the source side. One remaining open research topic for this is the exploration of the **Tree-to-Tree structure NMT** learning model (see Figure 2.4 on NMT development). One closely related study is [43] that applied a T2T structure Neural attentional translation model for program (coding language) translation tasks.

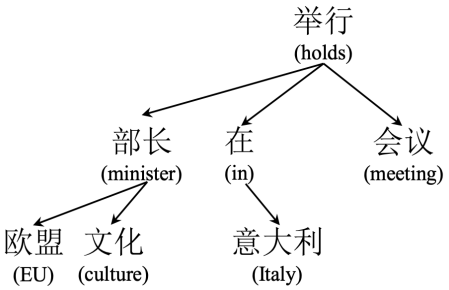

Figure 2.5: Dependency example in a Chinese sentence with English translation from Wu et al. 2017 (IJCAI).

Comparison among MT paradigms

Some research work was carried out on the comparisons between different MT paradigms. For instance, [18] argued that on a case study of translation quality on English-to-German language pair using the data from IWSLP2015, LSTM based NMT with attention model produces translation output that improves word order in placement of verbs with a large winning margin in comparison to traditional PB-SMT model. NMT also produced less morphology and lexical errors than PBSMT in certain degrees. However, as they discussed, NMT still struggles in the aspects of handling long sentences, as well as the correct reordering of "particular linguistic constituents" that needs a deep semantic understanding.

MT and language service provider (LSP) Tilde also made comparative evaluation of their in-house built SMT vs NMT [7]. Using professional translators, their evaluation results carried out on Baltic languages such as EN↔LV and EN↔ET show that NMT system produces more fluent outputs, and handles better word ordering and "morphology, syntax and agreements" in comparison to SMT.

Similar findings on the strength of NMT was reported by [39] in an educational domain (on EL, DE, PT, RU) that NMT produced better results regarding fluency, error-type markup, and post-editing effort. However, the authors mentioned that the NMT systems they used did not show much improvement on adequacy level as well as document-level MT performance, in comparison to PBSMT.

Contrarily to the findings from Tilde, the experimental evaluation carried out on

[7] https://www.tilde.com/about/news/316 last visit 2021.Nov.30th

German-to-French by [202], showed that NMT produced outputs with less agreement on differences to be corrected, in comparison to SMT, using customised MT systems for Swiss Post's language service. They indicate that NMT have the ability of producing correct paraphrases, thus BLEU is not adequate for NMT evaluation due to its strict n-gram matching mechanism.

Similar findings to [39] on the drawbacks of NMT was reported in [268] and [281] that while NMT systems made great progress in fluency, they produced inadequate translations. In this thesis (later chapters) we will make contribution in the adequacy level performance of NMT systems.

Other research work from controlled language study by Marzouk *et al.* [190] carried out the effectiveness analysis of pre-editing/controlled language given different MT paradigms on German-to-English. Controlled (natural) languages (CNL) apply some limitations on natural languages, such as restrictions on vocabulary, grammar, and semantics. These restrictions are expected to make the language easier for computational models to analyse the text for applications such as MT, knowledge representation, user understandability, etc. [163, 267, 26]. [8] In the comparison work by [190], nine individual CNL rules are applied during the examinations, e.g. avoiding participial constructions (pak), avoiding passives (pas), avoiding light-verb construction (fvg), and language specific rules on German avoiding the construction sein + zu + infinitive (comparable to the structure to be+to+base infinitive), etc. Their experimental work shows that CNL rules positively affect rule-based, statistical, and hybrid MT systems, but not the NMT architecture, where the NMT system consistently yields better outputs than when using other models both before and after applying CNL rules. The authors suggest that NMT offers a promising solution CNL rules are no longer required from the aspect of improving MT output [189]. However, this finding may be language specific or data dependent.

For rule-based, example-based and statistical phrase based MT, researchers broke

[8] There has been a regular workshop series since 2008 with the latest event held in 2021 http://www.sigcnl.org/cnl2020.html located in Amsterdam, the Netherlands, called the 7th CNL workshop.

up the sentences, and tried to address the MT output as a comparison to what a human would do, part-by-part of sentences. This is different to earlier MT evaluation where NMT treated sentences as a whole and encoded entire sentence information. Even though, later, when researchers tried to improve NMT, they added linguistically motivated features, such as attention / phrase alignment mechanism, phrase table enhanced NMT, dictionary usage, etc. and also tried to perform explainable NMT e.g. NMT learning representations of middle layer performances. Learning the most transparent representation is still an ongoing topic of investigation.

2.1.2 Summary

Even though the translation quality of MT models has improved hugely, from nonsense to making-sense, and even to "almost correct" in certain situations, domains and language pairs, MT models continue to be criticised in relation to various issues, such as their adequacy level, their fluency level, and stylisation [289, 119, 158]. Adequacy reflects how much of the original sentence meaning has remained or is reflected in the candidate translation, fluency reflects how fluent and natural the candidate translation sounds, and stylisation refers to whether the candidate translation is in a correct style according to the source text and the context, or if it creates its own styles or features (translationese [149, 233, 92]). Professional translators who use MT systems and perform post-editing also complain about restrictions of creativity in MT output, being "trapped by MT", and provide constant negative feedback [29, 148]. In the next section, we present a systematic review on the evaluation of MT systems and their output.

2.2 MT Evaluation Methodologies

In this section, we present a survey of work published on various MT evaluation methodologies. This covers the background knowledge, human evaluation methods, automatic evaluation methods, quality estimation research, and meta-evaluation of

the evaluation methods. We will present the usage of MWEs in MT evaluation in the semantic aspects of automatic methods. We also introduce our own contribution in automatic evaluations, including LEPOR, nLEPOR, hLEPOR, HPPR, and cush-LEPOR models in different sub-categories. Deriving from the drawbacks of current MT evaluation methodologies, we will propose our new evaluation methodology in the later chapter (Chapter 5).

2.2.1 Background

Machine translation (MT) has been one of the main research topics in computational linguistics (CL) and natural language processing (NLP), and has influenced and been influenced by several other language processing tasks such as parsing and language modelling. Starting from rule-based methods to example-based, and then statistical methods [27, 209, 44, 142], to the current paradigm of neural network structures [45, 129, 273, 153], MT quality continue to improve. However, as MT and *translation quality assessment* (TQA) researchers report, MT outputs are still far from reaching *human parity* [159, 158, 108] as measured by MT assessment. Thus we can see that MT quality assessment is an important part of MT research itself, and also for ensuring the quality of downstream applications. TQA remains a challenging and difficult task because of the richness, variety, and ambiguity of natural language itself, e.g. the same concept can be expressed in different word structures and patterns in different languages, even inside one language [6].

In this section, we introduce human judgement and evaluation (HJE) criteria that have been used in standard international shared MT tasks and more broadly. We then introduce automated TQA methods, including the automatic evaluation metrics that were proposed inside these shared tasks and beyond.

Regarding Human Assessment (HA), we categorise these into traditional and advanced sets of methods, with the first set including intelligibility, fidelity, fluency, adequacy, and comprehension, and the second set including task-oriented, extended criteria, utilising post-editing, segment ranking, crowd source intelligence (direct as-

sessment), and revisiting traditional criteria. Regarding automated TQA methods, we present the traditional metrics paradigm that relies on reference translations and the more recent quality estimation paradigm that is reference-free. We refer to Figure 2.1 for a diagram demonstrating this difference. We classify metrics into three categories: simple n-gram based word surface matching, deeper linguistic feature integration such as syntax and semantics, and deep learning (DL) models, with the first two regarded as traditional and the last one regarded as advanced, due to the recent appearance of DL models for NLP. We further divide each of these three categories into sub-branches, each with a different focus. Of course, this classification does not have clear boundaries. For instance some automated metrics involve both n-gram word surface similarity and linguistic features.

We refer to Figure 1.1 for the interpretation of different evaluation paradigms. For instances the combination of (source, reference, output) triplet for Human Evaluation, the (source, output) set for HE and Quality Estimation, and the (reference, output) set for HE and Metrics. Finally, at the end, we introduce an evaluation methodology for evaluation methods, i.e. meta-evaluation, on both human and automatic evaluations.

Related Machine Translation Evaluation Surveys

A systematic survey of work focusing on MT error classifications up to 2005, from translation teaching and translation industry background, is reported in [243]. This is one of the earliest attempts to carry out a systematic review of work on MT evaluation, and offered view points from the translator background. In a research project report from the EuroMatrix project [74], researchers first gave an introduction into MT history and then they introduced human evaluation of MT and objective evaluation of MT as two main sections of the work. Finally, they introduced a list of popular evaluation measures at that time including WER, SER, CDER, X-Score, D-score, NIST, RED, IER and TER etc., up to 2007.

Shortly afterwards in another DARPA GALE project report [65], researchers

first introduced the automatic and semi-automatic MT evaluation measures, and the task and human in loop measures. They then gave a description of the MT metrology used in the GALE program, which focuses on the HTER metric as the standard method used in GALE. Finally, they compared some automatic metrics and explored some other usages of the metric, such as optimisation in MT parameter training.

In an extended invited talk, Màrquez [185] introduced the Asiya online interface developed by their institute for MT output error analysis, where they also briefly mentioned the MT evaluation developments of lexical measures and linguistically motivated measures, and pointed out the challenges in the quality estimation task.

The work in this thesis differs from existing survey works by introducing recent developments in MT evaluation measures, the different classifications from manual to automatic evaluation methodologies, the introduction of more recently developed quality estimation (QE) tasks, the meta-evaluation methods (evaluating the evaluations), and the concise presentation of these concepts.

2.2.2 Human Evaluation Methods

In this section we introduce human judgement methods for MT evaluation, as reflected in Figure 2.6. This categorises human evaluation methods as *Traditional* and *Advanced*. We begin our description with the traditional HE methods, where we introduce the features of Intelligibility and Fidelity; Fluency, Adequacy and Comprehension; and further developments.

Intelligibility and Fidelity

The earliest human assessment methods for MT can be traced back to around 1966. They include the intelligibility and fidelity used by the automatic language processing advisory committee (ALPAC) [38]. The requirement that a translation is intelligible means that, as far as possible, the translation should read like normal, well-edited prose, and be readily understandable in the same way that such a trans-

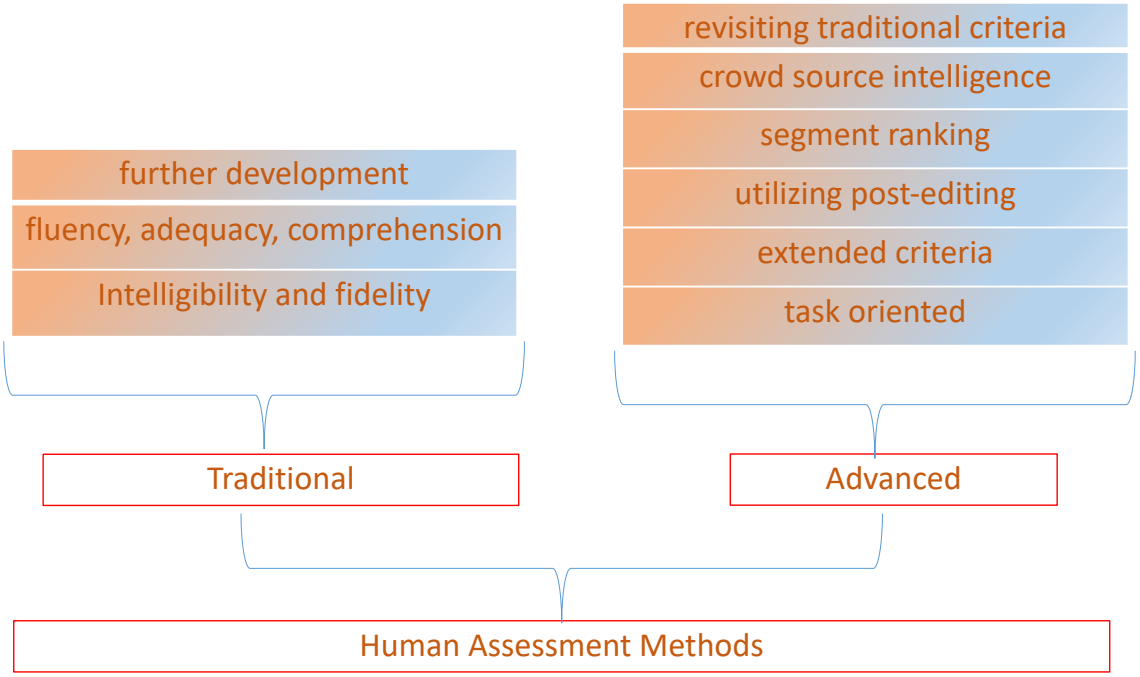

Figure 2.6: Classification of Human Assessment Methods

lation would be understandable if originally composed in the translation language. The requirement that a translation is of high fidelity or accuracy includes the requirement that the translation should, as little as possible, twist, distort, or controvert the meaning intended by the original.

Fluency, Adequacy and Comprehension

In the 1990s, the Advanced Research Projects Agency (ARPA) created a methodology to evaluate MT systems using adequacy, fluency and comprehension of MT output [50], which was subsequently adapted for use in MT evaluation campaigns including [284].

To set up this methodology, a human assessor is asked to look at each fragment, delimited by syntactic constituents and containing sufficient information, and judge its adequacy on a scale of 1-to-5. Results are computed by averaging the judgements over all of the decisions in the translation set.

Fluency evaluation is compiled in the same manner as for adequacy except that the assessor is asked to make intuitive judgements on a sentence-by-sentence basis for each translation. Human assessors are asked to determine whether the translation

is good English without reference to the correct translation. Fluency evaluation determines whether a sentence is well-formed and fluent in context.

Comprehension relates to "Informativeness", whose objective is to measure a system's ability to produce a translation that conveys sufficient information, such that people can gain necessary information from it. The reference set of expert translations is used to create six questions with six possible answers respectively including, "none of the above" and "cannot be determined".

Further Development

In work by Bangalore *et al.* [13] the authors classified accuracy into several categories, including simple string accuracy, generation string accuracy, and two corresponding tree-based accuracy. Other work in [229] found a correlation between fluency and the number of words it takes to distinguish between human translation and MT output.

The Linguistics Data Consortium (LDC) designed two five-point scales representing fluency and adequacy for the annual NIST MT evaluation workshop (referring to *LDC2003T17*) [9]. The developed scales became a widely used methodology when manually evaluating MT by assigning values. The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a translation hypothesis (from None to All); the second five point scale indicates how fluent the translation is, involving both grammatical correctness and idiomatic word choices (from Incomprehensible to Flawless English).

We now examine some of the advanced HE methods and introduce: Task-oriented; Extended Criteria; Utilising Post-editing; Segment-ranking; Crowd-source Intelligence; and Revisiting Traditional Criteria.

[9] archived document available at https://catalog.ldc.upenn.edu/docs/LDC2003T17 *ref.* "TransAssess02.pdf"

Task-oriented

Work by White and Taylor [285] developed a task-oriented evaluation methodology for Japanese-to-English translation to measure MT systems in light of the tasks for which their output might be used. They seek to associate the diagnostic scores assigned to the output used in the DARPA (Defense Advanced Research Projects Agency) evaluation with a scale of language-dependent tasks, such as scanning, sorting, and topic identification. They developed an MT proficiency metric with a corpus of multiple variants which are usable as a set of controlled samples for user judgements. The principal steps include identifying the user-performed text-handling tasks, discovering the order of text-handling task tolerance, analysing the linguistic and non-linguistic translation problems in the corpus used in determining task tolerance, and developing a set of source language patterns which correspond to diagnostic target phenomena. A brief introduction to task-based MT evaluation work was shown in their later work [66].

Another tasked-based MT output evaluation by the extraction of three types of elements namely: *who*, *when*, and *where* was, introduced in [278]. They also later extended their work into event understanding [155].

Our own very recent work on designing a task-oriented MT evaluation framework was introduced in [85], which we named as HOPE evaluation metric. HOPE is based on profession post-editing and annotations using eight pre-defined popular error types. The initial experiments carried out using English-to-Russian translation direction on marketing and business domain corpus demonstrate that HOPE is very effective in reflecting "good enough" translation task evaluation, and in comparing different MT engines regarding how much percents of their outputs fall into "major error", "minor error", and "no need to change" categories.

Extended Criteria

[139] extended a large range of manual evaluation methods for MT systems which, in addition to the earlier mentioned accuracy, include: *suitability*, whether even

accurate results are suitable in the particular context in which the system is to be used; *interoperability*, whether it will operate seamlessly with other software or hardware platforms; *reliability*, i.e., does not break down or take a long time to get running again after breaking down; *usability*, easy to understand the interfaces, easy to learn and operate, and looks well; *efficiency*, when needed, keep up with the flow of dealt documents; *maintainability*, being able to modify the system in order to adapt it to particular users; and *portability*, one version of a system can be replaced by a new version, because MT systems are rarely static and tend to improve over time as resources grow and bugs are fixed.

Utilising Post-editing

One alternative method to assess MT quality is to compare the post-editing correction to the original MT output. This type of evaluation is, however, time consuming and depends on the skills of the human assessor and post-editing performer. One example of a metric that is designed in such a manner is the human-targeted translation edit rate (HTER) [251] which is based on the translation edit rate (TER) metric [212] using the number of editing steps. Here, a human assessor has to find the minimum number of insertions, deletions, substitutions, and shifts to convert the system output into an acceptable translation. HTER calculates the minimum of edits to a *new targeted reference*, i.e. the post-edited translation.

However, it is not clear how to link the number of insertion, deletion and substitution to exactly how good is the quality of MT output. In our own work that we mentioned in the task-oriented evaluation section [85], we designed HOPE a **h**uman-centric evaluation framework based **o**n professional **p**ost-**e**diting and annotations. In HOPE evaluation framework, we designed eight commonly occurring error types including Impact, Required Adaptation Missing, Terminology, Grammar, Mistranslation, Style, Proofreading, and Proper Name error. Each error type can be assigned with different error severity scores to indicate minor error (from 1 to 4) or major error (5+). Post-editing annotators just need to locate each error that they spot from

the MT output into the pre-defined error type and assign a corresponding severity score. HOPE has two variants at the moment, including segment/sentence-level and word-level, which we calculate how many percent of segments/sentences or words fall into different error severity levels. The experiments carried out using two different English-to-Russian MT evaluation tasks using Google MT, a customised MT engine, and DeepL MT showed very effective evaluation results using HOPE from MT output transparency perspective.

Segment Ranking

In the WMT metrics task, human assessment based on segment ranking was often used. Human assessors were frequently asked to provide a complete ranking over all the candidate translations of the same source segment [33]. In the WMT13 shared tasks [20], five systems were randomised for the assessor to give a rank to their output. Each time the source segment and the reference translation were presented together with the candidate translations from the five systems. The assessors ranked the systems from 1 to 5, allowing tied scores. For each ranking, there was the potential to provide as many as 10 pairwise results if there were no ties. The collected pairwise rankings were then used to assign a corresponding score to each participating system to reflect the quality of the automatic translations. The assigned scores could also be used to reflect how frequently a system was judged to be better or worse than other systems when they were compared on the same source segment, according to the following formula:

$$\frac{\#\text{better pairwise ranking}}{\#\text{total pairwise comparison} - \#\text{ties comparisons}} \tag{2.1}$$

Crowd Sourced Intelligence

With the reported very low human inter-agreement scores from the WMT segment ranking task, researchers started to address this issue by exploring new human assessment methods, as well as seeking reliable automatic metrics for segment level ranking [89].

[91] noted that the lower agreements from WMT human assessment might be caused partially by the interval-level scales set up for the human assessor to make a quality judgement of each segment. For instance, the human assessor might be in a situation where neither of the two categories they were forced to choose is preferred. In light of this rationale, they proposed continuous measurement scales (CMS) for human TQA using fluency criteria. This was implemented by introducing the crowd-sourcing platform Amazon Mechanical Turk (MTurk), which has been popular in both NLP and multimedia research tasks [130], with some quality control methods such as the insertion of *bad-reference* and *ask-again*, and statistical significance testing. This methodology reportedly improved both intra-annotator and inter-annotator consistency. Detailed quality control methodologies, including statistical significance testing were documented in direct assessment (DA) [90, 92].

To achieve better human evaluation, moving from crowd-sourced DA to DA by professional linguists was implemented in some very recent shared tasks, e.g. WMT-QE2020 [254], with the expectation that this will produce more reliable reference scores and rankings, but they only implemented this for a small number of language pairs due to the high cost of the manual assessment.

Revisiting Traditional Criteria

There has been some criticism of the traditional human TQA methods because they fail to reflect real problems in translation by assigning scores and ranking several candidates from the same source [222]. Instead, [222] designed a new methodology by asking human assessors to mark all problematic parts of candidate translations, either words, phrases, or sentences. Two questions that were typically asked of the

assessors related to *comprehensibility* and *adequacy*. The first criterion considers whether the translation is understandable, or understandable but with errors; the second criterion measures if the candidate translation has different meaning to the original text, or maintains the meaning but with errors. Both criteria take into account whether parts of the original text are missing in translation. Under a similar experimental setup, [223] also summarised the most frequent error types that the annotators recognised as misleading translations.

2.2.3 Automatic Evaluations of MT

As stated by the Society of Automotive Engineers (SAE) in 2001 when they started to design an automatic reliable evaluation model (SAE metric) by highlighting error categories with different weights, "Consistent use of a metric across the industry will allow tighter control of translation quality. The risks of low-quality translations of service information include erosion of customer confidence, higher warranty costs, and (at an extreme) damage to vehicles or injury to people." [10]

Manual evaluation suffers some disadvantages such as the fact that it is time-consuming, expensive, not tunable, not reproducible, has subjective characteristics, and often shows low agreement levels. Due to these aspects, automatic evaluation metrics have been widely used for MT. Typically, these compare the output of MT systems against human reference translations, but there are also some metrics that do not use reference translations. There are usually two ways to offer the human reference translation, either offering one single reference or offering multiple references for a single source sentence [171, 102].

Automated metrics often measure the overlap in words and word sequences, as well as word order and edit distance. We classify these kinds of metrics as "simple n-gram word surface matching". Further developed metrics also take linguistic features into account such as syntax and semantics, including part-of-speech (POS), sentence structure, textual entailment, paraphrase, synonyms, named entities, multi-word ex-

[10] (available archive files https://www.sae.org/standardsdev/j2450p1.htm and http://www.apex-translations.com/documents/sae_j2450.pdf)

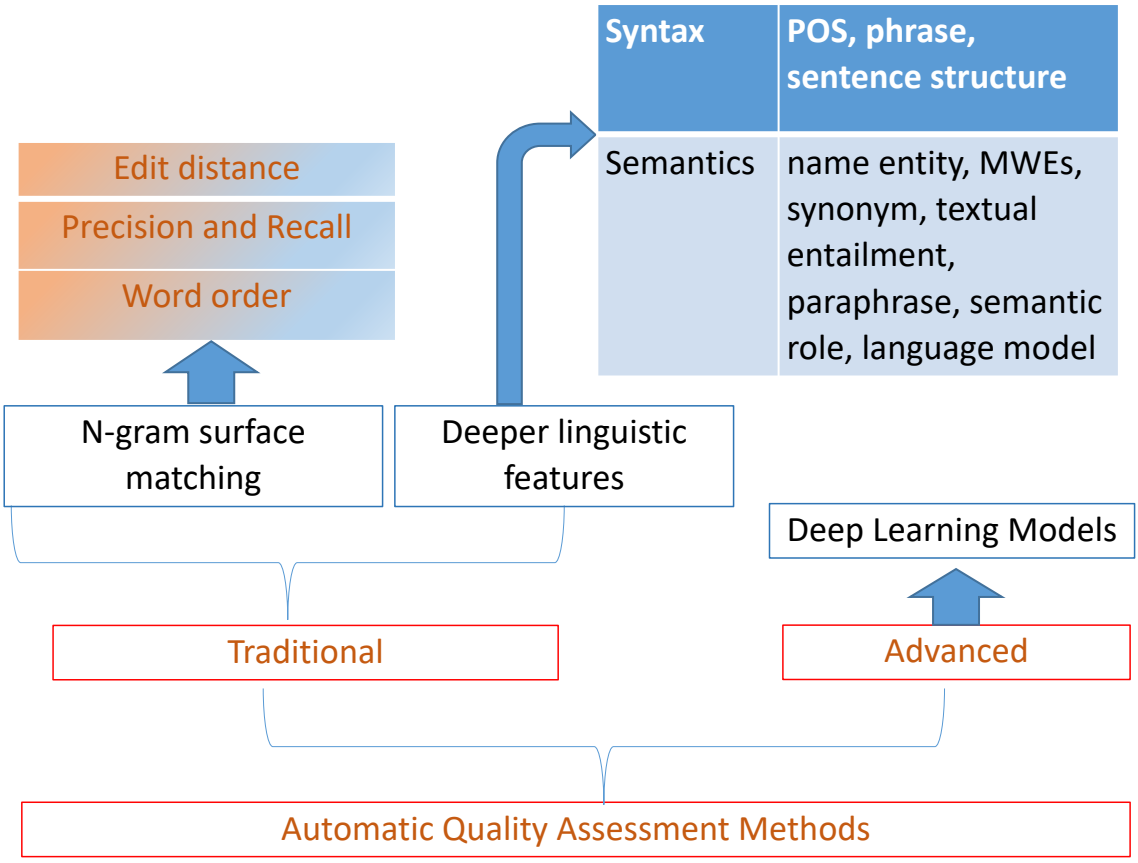

Figure 2.7: Automatic Quality Assessment Methods for MT

pressions (MWEs), semantic roles and language models. We classify these metrics that utilise linguistic features as "Deeper Linguistic Features (aware)". This classification is done purely for easier understanding and better organisation of the content. It is not easy to separate these two categories clearly since sometimes they merge with each other. For instance, some metrics from the first category might also use certain linguistic features. Furthermore, we will introduce some recent models that apply deep learning into the TQA framework, as in Figure 2.7. Due to space limitations, in the appendices we present the MT quality estimation (QE) task which does not rely on reference translations during the automated computing procedure.

We begin with an n-gram word surface match which includes Levenshtein Distance, Precision and Recall, and Revisiting Word Order.

Levenshtein Distance

By calculating the minimum number of editing steps to transform MT output to reference text, [261] introduced the word error rate (WER) metric into MT evaluation. This metric, inspired by Levenshtein Distance (or edit distance), takes word order into account, and the operations include insertion (adding word), deletion (dropping word) and replacement (or substitution, replace one word with another), the minimum number of editing steps needed to match two sequences.

One of the weak points of the WER metric is the fact that word ordering is not treated in an effective way.The WER scores are very low when the word order of system output translation is "wrong" according to the reference text. In the Levenshtein distance, the mismatches in word order require the deletion and re-insertion of the misplaced words. However, due to the diversity of language expressions, some so-called "wrong" order sentences by WER also prove to be good translations. To address this problem, the position-independent word error rate (PER) introduced by [266] is designed to ignore word order when matching output and reference. Without taking account of the word order, PER counts the number of times that identical words appear in both sentences. Depending on whether the translated sentence is longer or shorter than the reference translation, the rest of the words are either insertions or deletions.

Another way to overcome the unconscionable penalty on word order in the Levenshtein distance is adding a novel editing step that allows the movement of word sequences from one part of the output to another. This is something a human post-editor would do with the cut-and-paste function of a word processor. In this light, [212] designed the translation edit rate (TER) metric that adds block movement (jumping action) as an editing step. The shift option is performed on a contiguous sequence of words within the output sentence. Different from HTER we mentioned in human evaluation section (Section 2.2.2), which generates a post-edited reference translation, TER is defined as the sum of the number of editing steps divided by the number of words in a given reference translation.

For the edits, the cost of the block movement, any number of continuous words and any distance, is equal to that of the single word operation, such as insertion, deletion and substitution.

Precision and Recall

The widely used evaluation BLEU metric [213] is based on the degree of n-gram overlap between the strings of words produced by the MT output and the human translation references at the corpus level. BLEU calculates precision scores with n-grams sized from 1-to-4, together multiplied by the coefficient of brevity penalty (BP). If there are multi-references for each candidate sentence, then the nearest length as compared to the candidate sentence is selected as the effective one. In the BLEU metric, the n-gram precision weight λ_n is usually selected as a uniform weight. However, the 4-gram precision value can be very low or even zero when the test corpus is small. To weight more heavily those n-grams that are more informative, [62] proposes the NIST metric with the information weight added. Furthermore, [62] replaces the geometric mean of co-occurrences with the arithmetic average of n-gram counts, extends the n-gram into 5-gram ($N = 5$), and selects the average length of reference translations instead of the nearest length.

ROUGE [170] is a recall-oriented evaluation metric, which was initially developed for summaries, and inspired by BLEU and NIST. ROUGE has also been applied in automated TQA in later work [171].

The F-measure is the combination of precision (P) and recall (R), which was firstly employed in information retrieval (IR) and latterly adopted by the information extraction (IE) community, MT evaluations, and others. [269] carried out experiments to examine how standard measures such as precision, recall and F-measure can be applied to TQA and showed comparisons of these standard measures with some alternative evaluation methodologies.

The METEOR measure was designed in [12] as a novel evaluation metric. METEOR is based on the general concept of flexible unigram matching, whereas preci-

sion and recall include the match between words that are simple morphological variants of each other with identical word stems and words that are synonyms of each other. To measure how well-ordered the matched words in the candidate translation are in relation to the human reference, METEOR introduces a penalty coefficient, different to what is done in BLEU, by employing the number of matched chunks.

Revisiting Word Order

The correct word order plays an important role to ensure a high quality translation output. However, language diversity also allows different appearances or structures of a sentence. How to successfully achieve a penalty on really incorrect word order, i.e. incorrectly structured sentences, instead of on "correct" different order, i.e. the candidate sentence that has different word order to the reference, but is well structured, has attracted a lot of interest from researchers. In fact, the Levenshtein distance (*Section 2.2.3*) and n-gram based measures also contain word order information.

Featuring the explicit assessment of word order and word choice, [286] developed the evaluation metric ATEC (assessment of text essential characteristics). This is also based on precision and recall criteria, but with a position difference penalty coefficient attached. The word choice is assessed by matching word forms at various linguistic levels, including surface form, stem, sound and sense, and further by weighing the informativeness of each word.

Partially inspired by this, our previous work developed LEPOR [102] (from **Le**ngth Penalty, **P**recision, n-gram **Po**sition difference Penalty and **R**ecall). It is designed as a combination of augmented evaluation factors including n-gram based *word order penalty* in addition to *precision*, *recall*, and *enhanced sentence-length penalty*, as shown in the following formulas.

$$\text{LEPOR} = LP \times NPosPenal \times Harmonic(\alpha R, \beta P) \tag{2.2}$$

$$LP = \begin{cases} e^{1-\frac{Length_{ref}}{Length_{hyp}}} \; if \; Length_{hyp} < Length_{ref} \\ 1 \; if \; Length_{hyp} = Length_{ref} \\ e^{1-\frac{Length_{hyp}}{Length_{ref}}} \; if \; Length_{hyp} > Length_{ref} \end{cases} \tag{2.3}$$

$$NPosPenal = e^{-NPD} \tag{2.4}$$

$$NPD = \frac{1}{Length_{hyp}} \sum_{i=1}^{Length_{hyp}} |PD_i| \tag{2.5}$$

$$|PD_i| = |MatchN_{hyp} - MatchN_{ref}| \tag{2.6}$$

where P and R are for precision and recall, LP for length penalty, and $NPosPenal$ for n-gram position difference penalty.

Tested on WMT2011 data, LEPOR outperformed AMBER, METEOR-1.3, BLEU, TER and MP4IBM1 metrics in system-level correlation to human judgements averaged on eight language pairs including Czech/German/Spanish/French<=>English, achieving 0.77 overall score, including the individually highest scores on Czech=>English (0.95), and Spanish=>English (0.96) [102].

LEPOR metric has two variants named *n*LEPOR and *h*LEPOR representing *n*-gram based LEPOR with additional n-gram applied to P and R, and *harmonic* mean of LEPOR with the harmonic applied to three major factors of LEPOR. They are calculated as below [114, 106, 116].

$$n\text{LEPOR} = LP \times NPosPenal \times exp(\sum_{n=1}^{N} w_n log HPR) \tag{2.7}$$

$$h\text{LEPOR} = Harmonic(w_{LP} LP, w_{NPosPenal} NPosPenal, w_{HPR} HPR) \tag{2.8}$$

The factor *HPR* is the harmonic mean of Precision and Recall values.

$$HPR = \frac{(\alpha + \beta)PrecisionxRecall}{\alpha Precision + \beta Recall} \tag{2.9}$$

$$Precision = \frac{Aligned_{num}}{Length_{hypothesis}} \tag{2.10}$$

$$Recall = \frac{Aligned_{num}}{Length_{reference}} \tag{2.11}$$

The n-gram based variant *n*LEPOR using only lexical matching [105, 116] was also verified by MT researchers to be one of the three best performing *segment level* automated metrics (together with METEOR and sentBLEU-MOSES) that correlated with human judgement at a level that was not significantly outperformed by any other metrics, on Spanish-to-English, in addition to an aggregated set of overall tested language pairs (en<=>fr/de/es/cs/ru) on WMT2013 data [89].

The weighted harmonic mean variant of LEPOR, i.e. *h*LEPOR without using any external resources, achieved the *highest* correlations to human judgements on English-to-Russian language pair tested on the shared task data on WMT2013 (0.77 Pearson score) and *Cluster-1* performance on WMT2021 (0.894 Pearson score) at *system level* [105, 113].

We now look at methods which integrate linguistic features including Syntactic Similarity and Semantic Similarity.

Although some of the previously outlined metrics incorporate linguistic information, e.g. synonyms and stemming in METEOR and part of speech (POS) in LEPOR, the simple n-gram word surface matching methods mainly focus on the exact matches of the surface words in the output translation. The advantages of the metrics based on the first category (simple n-gram word matching) are that they perform well in capturing translation fluency, are very fast to compute and have low cost. On the other hand, there are also some weaknesses, for instance, syntactic information is rarely considered and the underlying assumption that a good translation is one that shares the same word surface lexical choices as the reference

translations is not justified semantically. Word surface lexical similarity does not adequately reflect similarity in meaning. Translation evaluation metrics that reflect meaning similarly need to be based on similarity of semantic structure and not merely flat lexical similarity.

Syntactic Similarity

Syntactic similarity methods usually employ the features of morphological POS information, phrase categories, phrase decompositionality or sentence structure generated by linguistic tools such as a language parser or chunker.

In grammar, a **POS** is a linguistic category of words or lexical items, which is generally defined by the syntactic or morphological behaviour of the lexical item. Common linguistic categories of lexical items include noun, verb, adjective, adverb, and preposition. To reflect the syntactic quality of automatically translated sentences, researchers employ POS information into their evaluations. Using the IBM model one [27], [220] evaluate translation quality by calculating the similarity scores of source and target (translated) sentences without using a reference translation, based on the morphemes, 4-gram POS and lexicon probabilities. [57] developed the TESLA evaluation metrics, combining the synonyms of bilingual phrase tables and POS information in the matching task.

In our earlier work [101], we compared the agreement level of POS sequences from MT output candidate translation and the reference translation. Using the hLEPOR algorithm from last section, the model with only POS information included surprisingly achieved the‘ highest correlation score to human judgement on two language pair translations, en-fr and en-de (bi-direction), in comparison to state of the art metrics BLEU, METEOR, MPF, AMBER and MP4IMB1. The experimental testing data was from WMT2011.

The hybrid version of the *h*LEPOR metric, i.e. *h*LEPOR(hybrid) combining both POS as morphology feature and word level similarity, achieved the best performing metric in the WMT2013 shared task on English-to-other (German, French, Spanish,

Czech, and Russian) language pairs in system level Pearson correlation (0.86 in average) to human judgements [114, 106]. This further verified the efficiency of LEPOR and *h*LEPOR metric factors that we designed as augmented models. Other similar work using POS information can be found in [84, 221].

In linguistics, a **phrase** may refer to any group of words that form a constituent, and so it functions as a single unit in the syntax of a sentence. To measure an MT system's performance in translating new text types, such as in what ways the system itself could be extended to deal with new text types, [224] carried out work focusing on the study of an English-to-Danish MT system. The syntactic constructions are explored with more complex linguistic knowledge, such as the identification of fronted adverbial subordinate clauses and prepositional phrases. Assuming that similar grammatical structures should occur in both source and translations, [7] perform evaluation on source (German) and target (English) sentences employing the features of sentence length ratio, unknown words, phrase numbers including noun phrase, verb phrase and prepositional phrase.

In previous work, motivated to facilitate multilingual research and multilingual NLP tasks, we designed a universal phrase tag-set and the corresponding mapping across multiple languages and constituency structure tree-banks [104]. This phrase tagset mapping was then applied into FR-to-EN MT evaluation by comparing the phrase tagset sequence similarity of parsed candidate translation and source text with similarity metric HPPR [115]. HPPR is defined as the Harmonic Mean of n-gram Position penalty, Precision, and Recall. It achieved a strong correlation with human judgements at a level of 0.63 and 0.59 respectively on development (WMT11) and testing (WMT12) sets without relying on reference translation. The calculation algorithms of HPPR is presented as below:

$$HPPR = Harmonic(w_{Ps}N_1 PosPenal, w_{Pr}N_2 Pre, w_{Rc}N_3 Rec) \tag{2.12}$$

where $PosPenal$, Pre, and Rec represent position difference penalty (*ref.* LEPOR

metric factor), precision and recall respectively, and they are calculated using n-gram matching where the "n-gram" can set up as different value for each of these factors. Then, overall, these three main factors are summarised using weighted harmonic mean.

Other similar work using phrase similarity includes [166], which uses noun phrases and verb phrases from chunking, and [70], which only uses noun phrase chunking in automatic evaluation.

Syntax is the study of the principles and processes by which sentences are constructed in particular languages. To address the overall goodness of a translated **sentence structure**, [172] employ constituent labels and head-modifier dependencies from a language parser as syntactic features for MT evaluation. They compute the similarity of dependency trees. Their experiments show that adding syntactic information can improve evaluation performance, especially for predicting the fluency of translation hypotheses. Other work that uses syntactic information in evaluation includes [175] which uses an automatic shallow parser and the RED metric [291], which applies dependency trees.

Semantic Similarity

In contrast to syntactic information, which captures overall grammaticality or sentence structure similarity, the semantic similarity between automatic translations and the source sentences (or references) can be measured by employing semantic features.

To capture the semantic equivalence of sentences or text fragments, **named entity** knowledge is taken from the literature on named-entity recognition, which aims to identify and classify atomic elements in a text into different entity categories [186, 94]. The most commonly used entity categories include the names of persons, locations, organisations and time [100, 117]. Named entity features were proved to be helpful in improving MT outputs from syntactic segmentation and disambiguation points of view [8]. In the MEDAR2011 evaluation campaign, one baseline system

based on Moses [144] utilised an Open NLP toolkit to perform named entity detection, in addition to other packages. The low performances from the perspective of named entities caused a drop in fluency and adequacy. In the quality estimation of the MT task in WMT 2012, [28] introduced features including named entities, in addition to discriminative word lexicon, neural networks, back off behaviour [228] and edit distance. Experiments on individual features showed that from the perspective of increasing the correlation score with human judgements, the named entity feature contributed the greatest impact to the overall performance, in comparison to the impacts of other features.

Multi-word Expressions (MWEs) set challenges for MT models due to their complexity in presentation as well as their idiomaticity [235, 109, 108, 110]. To investigate the effect of MWEs in MT evaluation (MTE), [236] focused on the *compositionality* of noun compounds. They identify the **noun compounds** first from among the MT system outputs and reference this with the Stanford parser. The matching scores of the system outputs and reference sentences are then recalculated, adding up to the Tesla metric, by considering the predicated compositionality of identified noun compound phrases. Our own recent work in this area [108] provides an extensive investigation into various MT errors caused by MWEs which will be detailed in Chapter 5, especially Sections 5.2.4 and 5.2.5.

Synonyms are pairs or larger groups of words with the same or close meanings. One of the most widely used synonym databases in the NLP literature is WordNet [196], which is an English lexical database grouping English words into sets of synonyms. WordNet classifies words mainly into four kinds of part of speech categories; Noun, Verb, Adjective, and Adverb, without prepositions, determiners, etc. Synonymous words or phrases are organised using units called synsets. Each synset is a hierarchical structure with words placed at different levels in the hierarchy according to their semantic relations.

Textual entailment is usually used as a directive relation between text fragments. If the truth of one text fragment TA follows another text fragment TB, then

there is a directional relation between TA and TB (TB ⇒ TA). Instead of pure logical or mathematical entailment, textual entailment in natural language processing (NLP) is usually performed with a relaxed or loose definition [56]. For instance, according to text fragment TB, if it can be inferred that the text fragment TA is *most likely* to be true then the relationship TB ⇒ TA is also established. Since the relation is directive, it means that the inverse inference (TA ⇒ TB) is not ensured to be true [55]. [40] present a new approach for MT evaluation based on the task of "Semantic Textual Similarity". This problem is addressed using a textual entailment engine based on WordNet semantic features.

Paraphrase means to restate the meaning of a passage of text but utilising other words, which can be seen as bidirectional textual entailment [5]. Instead of a literal translation word-by-word and line-by-line used by meta-phrases, a paraphrase represents a dynamic equivalent. Further knowledge of paraphrases from the aspect of linguistics is introduced in works by [191, 192, 14]. [251] describes a new evaluation metric TER-Plus (TERp). Sequences of words in the reference text are considered to be paraphrases of a sequence of words in the hypothesis text if that phrase pair occurs in the TERp phrase table.

Semantic roles can be employed by researchers as linguistic features in MT evaluation. To utilise semantic roles, sentences are usually first shallow parsed and entity tagged. Then the semantic roles are used to specify the arguments and adjuncts that occur in both the candidate translation and reference text. For instance, the semantic roles introduced by [84, 83] include causative agent, adverbial adjunct, directional adjunct, negation marker, and predication adjunct, etc. In a further development, [175] presented the MEANT metric designed to capture predicate-argument relations as structural relations in semantic frames, which are not reflected in the flat semantic role label features in [84]. Furthermore, instead of using uniform weights, the work by [175] weights the different types of semantic roles as empirically determined by their relative importance to the adequate preservation of meaning. Generally, semantic roles account for the semantic structure of a segment and have

proved to be effective in assessing adequacy of translation.

Language models are also utilised by MT evaluation researchers. A statistical language model usually assigns a probability to a sequence of words by means of a probability distribution. [80] propose the LM-SVM language model, and support vector machine methods investigating the possibility of evaluating MT quality and fluency in the absence of reference translations. They evaluate the performance of the system when used as a classifier for identifying highly dis-fluent and ill-formed sentences.

Generally, the linguistic features mentioned above, including both syntactic and semantic features, are combined in two ways, either by following a machine learning approach as in [2, 164], or trying to combine a wide variety of metrics in a more simple and straightforward way, such as [83, 256, 52].

Neural Networks for TQA

We now briefly list some examples of work that have applied deep learning and neural networks for TQA which are promising for further exploration. One example approach is [96] which uses neural networks (NNs) for TQA for pair wise modelling to choose the best hypothetical translation by comparing candidate translations with a reference, integrating syntactic and semantic information into NNs. In an alternative strategy,[95] proposed LSTM networks based on dense vectors to conduct TQA, while [177] designed a new metric based on bi-directional LSTMs, which is similar to the work of [96] but with less complexity by allowing the evaluation of a single hypothesis with a reference, instead of a pairwise situation.

Deep NNs based metrics cost a lot of computational powers, in comparison to traditional lexical based evaluation metrics, for the sake of better performance. To avoid such cost, but also taking advantage of the pre-trained language models, in our own work, we proposed to customise the high performing traditional lexical based metric *h*LEPOR towards pre-trained LMs (PLMs), e.g. LaBSE in a distilled manner [73]. The customised *h*LEPOR (*cush*LEPOR) metric using optimised weighting

parameters trained on German-English and English-Chinese data achieved the first cluster performance in WMT2021 Metrics Task on both these two language pairs in News domain competition out of all submissions from different teams [113], scoring 0.763 in average Pairwise accuracy using MQM human annotation methods [] on en-de, zh-en and en-ru. *cush*LEPOR(LM) also produced the highest system-level Pearson correlation score (0.938) to human judgements on en-de in News domain on WMT2021 metrics task [113].

Furthermore, in addition to the customisation towards PLMs which is cheaper cost named *cush*LEPOR(LM), we also offered an alternative *cush*LEPOR version that is customised towards human labelled data whenever such data is available on the deployed language pair named *cush*LEPOR(pSQM) [11].

2.2.4 Quality Estimation Methods

In recent years, some MT evaluation methods that do not use manually created gold reference translations have been proposed. These are referred to as "Quality Estimation (QE)". Some of the related work in this area has already been introduced in previous sections. The most recent quality estimation tasks can be found at the WMT12 to WMT20 workshops [33, 20, 254] which include our own earlier work [103] where we investigated statistical learning models including conditional random fields (CRFs), Support Vector Machines (SVMs), Naive Bayes classifier (NBs) to learn estimation features such as sentence-length, precision, recall, and part-of-speech (POS) information. These international workshops defined a novel evaluation metric that provides some advantages over the traditional ranking metrics. The DeltaAvg metric assumes that the reference test set has a number associated with each entry that represents its extrinsic value. Given these values, this metric does not need an explicit reference ranking, in the way that Spearman ranking correlation does. The goal of the DeltaAvg metric is to measure how valuable a proposed ranking is

[11] source code and parameters of cushLEPOR is open-source at https://github.com/poethan/cushLEPOR. WMT2021 shared task data: http://www.statmt.org/wmt21/

according to the extrinsic values associated with the test entries.

$$DeltaAvg_v[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S) \tag{2.13}$$

For scoring, two task evaluation metrics are used that have traditionally been used for measuring performance in regression tasks: Mean Absolute Error (MAE) as a primary metric, and Root of Mean Squared Error (RMSE) as a secondary metric. For a given test set S with entries $s_i, 1 \leqslant i \leqslant |S|$, $H(s_i)$ is the proposed score for entry s_i (hypothesis), and $V(s_i)$ is the reference value for entry s_i (gold-standard value).

$$\text{MAE} = \frac{\sum_{i-1}^{N} |H(s_i) - V(s_i)|}{N} \tag{2.14}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i-1}^{N} (H(s_i) - V(s_i))^2}{N}} \tag{2.15}$$

where $N = |S|$. Both these metrics are non-parametric, automatic and deterministic (and therefore consistent), and extrinsically interpretable.

Some further readings on MT QE include the comparison between MT evaluation and QE [257], the QE framework model QuEst [258], the weakly supervised approaches for quality estimation and the limitations analysis of QE Supervised Systems [201], and unsupervised QE models [75].

In recent years, the shared tasks have seen attempts to integrate MT quality estimation and traditional MT evaluation metrics to benefit from knowledge of both measures. For instance, in the WMT2019 shared task, there were 10 reference-less evaluation metrics which were used for the QE task, "QE as a Metric", as well [178].

2.2.5 Meta-evaluation

As shown in Figure 2.8, meta-eval relates to the evaluation methods themselves, which include some categories that we will introduce below, i.e. statistical sig-

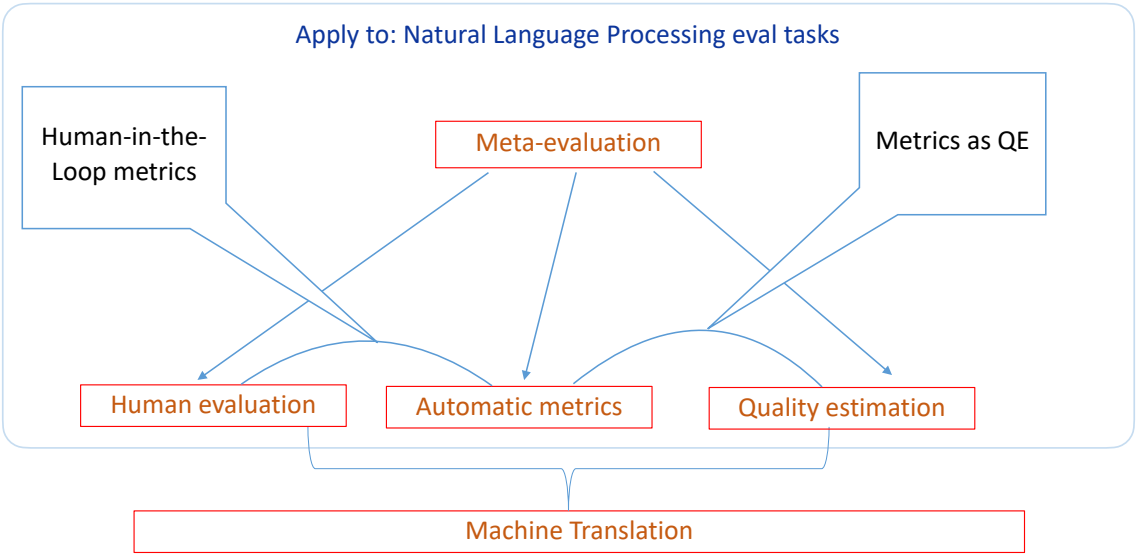

Figure 2.8: Meta-eval of MTE: applying to general NLP tasks

nificance, human judgement agreement levels, correlating manual and automatic evaluations, and metrics comparisons. Figure 2.8 also contains some recent trends in MTE fields including "human-in-the-loop metrics" and "metrics as QE" methods. The concept of upper layer box of Figure 2.8 actually can apply to general NLP evaluation tasks, not limited to MT evaluation tasks.

Statistical Significance

If different MT systems produce translations with different qualities on a dataset, how can we ensure that they indeed have different system quality? To explore this problem, [143] presents an investigation into statistical significance testing for MT evaluation. The bootstrap re-sampling method is used to compute the statistical significance intervals for evaluation metrics on small test sets. Statistical significance usually refers to two separate notions, one of which is the p-value, the probability that the observed data will occur by chance in a given single null hypothesis; the other is the "Type I" error rate of a statistical hypothesis test, which is also called "false positive", and is measured by the probability of incorrectly rejecting a given null hypothesis in favour of a second alternative hypothesis [99].

Evaluating Human Judgement

Since human judgements are usually trusted as gold standards that automatic MT evaluation metrics should try to approach, the reliability and coherence of human judgements is very important. Cohen's kappa agreement coefficient is one of the most commonly used evaluation methods [51]. For the problem of nominal scale agreement between two judges, there are two relevant quantities p_0 and p_c. p_0 is the proportion of units in which the judges agreed and p_c is the proportion of units for which agreement is expected by chance. The coefficient k is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration:

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{2.16}$$

where $p_0 - p_c$ represents the proportion of cases in which beyond-chance agreement occurs and is the numerator of the coefficient [154].

Confidence Sample Size for Evaluation

While automatic machine translation evaluation (MTE) metrics and quality estimation (QE) tools are widely available and easy to access, existing automated tools are not good enough, and human assessment from professional translators (HAP) are often chosen as the golden standard [112]. Human evaluations, however, are often accused of having low reliability and agreement. Is this caused by subjectivity or statistics is at play? How to avoid the entire text to be checked and be more efficient with translation QE from cost and efficiency perspectives, and what is the optimal sample size of the translated text, so as to reliably estimate the translation quality of the entire material? In our work [86], we carried out such motivated research to correctly estimate the confidence intervals depending on the sample size of the translated text, e.g. the amount of words or sentences, that needs to be processed on translation QE workflow step for confident and reliable evaluation of overall translation quality. The methodology we applied for this work is from Bernoulli Statistical

Distribution Modelling (BSDM) and Monte Carlo Sampling Analysis (MCSA). This work addressed the situation where conventional researchers very often took it for granted with a random or un-justified number of sample sentences to judge the overall MT quality.

Correlating Manual and Automatic Score

In this section, we introduce three correlation coefficient methods that have been widely used at recent WMT workshops to measure the closeness between automatic evaluation and manual judgements. The choice of correlation formula depends on whether score or rank schemes are utilised.

Pearson Correlation Pearson's correlation coefficient [216] is commonly represented by the Greek letter ρ. The correlation between random variables X and Y denoted as ρ_{XY} is measured as follows [197]:

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.17}$$

Because the standard deviations of variable X and Y are higher than 0 ($\sigma_X > 0$ and $\sigma_Y > 0$), if the covariance σ_{XY} between X and Y is positive, negative or zero, the correlation score between X and Y will correspondingly result in positive, negative or zero, respectively. Based on a sample of paired data (X, Y) as (x_i, y_i), i=1 to n, the Pearson correlation coefficient is calculated as:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \tag{2.18}$$

where μ_x and μ_y specify the means of discrete random variable X and Y respectively.

Spearman Rank Correlation Spearman rank correlation coefficient, a simplified version of the Pearson correlation coefficient, is another technique to measure the correlations of automatic evaluation and manual judges, e.g. in the WMT metrics task [33]. When there are no ties, the Spearman rank correlation coefficient,

which is sometimes specified as (rs) is calculated as:

$$rs_{\varphi(XY)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{2.19}$$

where d_i is the difference-value (D-value) between the two corresponding rank variables $(x_i - y_i)$ in $\vec{X} = \{x_1, x_2, ..., x_n\}$ and $\vec{Y} = \{y_1, y_2, ..., y_n\}$ describing the system φ.

Kendall's τ Kendall's τ [135] has been used to measure the correlation between automatic order and reference order [33]. This is defined as:

$$\tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \tag{2.20}$$

The latest version of Kendall's τ is introduced in [136]. [161] present an overview of Kendall's τ showing its application in calculating how much the system orders differ from the reference order. More concretely, [156] proposed the use of Kendall's τ, a measure of rank correlation, to estimate the distance between a system-generated and a human-generated gold-standard order.

Metric Comparison

A number of studies have appeared reporting comparisons between different types of metrics. For example, [35, 32, 160] note that, through some qualitative analysis on some standard data set, BLEU cannot reflect MT system performance well in many situations, i.e. higher BLEU score cannot ensure better translation outputs. Some recently developed metrics can perform much better than the traditional ones especially on challenging sentence-level evaluation, although these are not popular as yet. These include: nLEPOR and SentBLEU-Moses [89]. Such comparisons will help MT researchers to select the appropriate metrics to use for specialist tasks.

2.2.6 Discussion

In this section, we examined several topics that can be considered in further development of the MT evaluation field.

The first aspect to be developed should involve both n-gram word surface matching and deeper linguistic features. Because natural languages are expressive and ambiguous at different levels [84], simple n-gram word surface similarity based metrics limit their scope to the lexical dimension and are not sufficient to ensure that two sentences convey the same meaning. For instance, [34] and [146] report that simple n-gram matching metrics tend to favour automatic statistical MT systems. If the evaluated systems belong to different types that include rule-based, human aided, and statistical systems, then simple n-gram matching metrics, such as BLEU, give a strong disagreement between these ranking results and those of human assessors. So deeper linguistic features are very important in the MT evaluation procedure.

However, inappropriate utilisation, or abundant or abused utilisation, of linguistic features will result in limited popularity of measures incorporating linguistic features. In the future, how to use the linguistic features in a more accurate, flexible and simplified way, will be a challenge in MT evaluation. Furthermore, MT evaluation from the aspects of semantic similarity is more reasonable and reaches closer to human judgements, so it should receive more attention.

The second major aspect is MT quality estimation (QE) tasks, which are different to traditional MT evaluation in several ways. These differences include extracting reference-independent features from input sentences and their translation, obtaining quality scores based on models produced from training data, predicting the quality of an unseen translated text at system run-time, filtering out sentences which are not good enough for post processing, and selecting the best translation from among multiple systems. This means that with so many challenges, this topic will continue to attract many researchers.

Thirdly, some advanced or challenging technologies that can be further applied to MT evaluation include the deep learning models [95, 295], semantic logic form,

and decipherment model.

Comparisons of HE vs. AE

So far, human judgement scores of MT results are usually considered as the gold standard against which automatic evaluation metrics should try to match. However, some improper handling in the process can lead to problems. For instance, in the ACL WMT 2011 English-Czech task, the multi-annotator agreement kappa value k is very low and even the exact same string produced by two systems is ranked differently each time by the same annotator. The evaluation results are highly affected by the manual reference translations. How to ensure the quality of reference translations and the agreement level of human judgements are two important problems.

Automatic evaluation metrics are indirect measures of translation quality, because they usually use various string distance algorithms to measure the closeness between the MT system outputs and the manually offered reference translations. They are based on the calculation of the correlation with manual MT evaluation [199]. Furthermore, automatic evaluation metrics tend to ignore the relevance of words [142]. For instance, named entities and core concepts are more important than punctuation and determiners, but most automatic evaluation metrics put the same weight of importance on each word of the sentences. Third, automatic evaluation metrics usually yield scores whose actual values are meaningless, since they are test set specific and the absolute values are not informative. For instance, what is the meaning of -16094 as a score by the MTeRater metric [214] or 1.98 score by ROSE [253] ? Instead, we find our earlier work LEPOR and hLEPOR series including nLEPOR and cushLEPOR [102, 114, 106, 73] can give a somehow meaningful score for a somehow recognised good translation, e.g. the score can be around 0.60 to 0.80.

Automatic evaluation metrics should try to achieve the goals of low cost, reduce time and effort spent on carrying out evaluation, they should be tunable, should automatically optimise system performance towards the metric and generate mean-

ingful scores which give an intuitive interpretation of translation quality. They should be consistent allowing repeated use of the metric to give the same results, and they should be correct in that the metric must rank better systems higher, as mentioned in [142]; of which the low cost, tunable and consistent characteristics are easily achieved by the metric developers. The other two goals (meaningful and correct) are generally challenges in front of NLP researchers. One of the future research direction to address these issues is to customise the automatic metrics using human judgement labelled data, especially the annotated data from professional translators, towards the better performances, such as our recently in-house developed cushLEPOR (customised hLEPOR) [73].

2.2.7 Summary

In this section we have presented a survey of the state-of-the-art in translation quality assessment methodologies from the viewpoints of both manual judgements and automated methods, as well as the meta-evaluation of evaluation methods. The automated methods covered both traditional metrics and advanced quality estimation models without depending on human prepared references. We believe this work can help both MT and NLP researchers and practitioners in identifying appropriate quality assessment methods for their work. We also expect this work could shed some light on evaluation methodologies in other NLP tasks, due to the similarities, tasks such as text summarisation [181, 19], natural language understanding [234], natural language generation [82], as well as programming language (code) generation [169].

2.3 MT Broad Issues and Challenges

In this section, we present a summary of the most important current issues and challenges in MT. Some of these have been mentioned in previous sections, but we list them here in different categories in order to group them. We present MT problems

and challenges from several perspectives: from MT modelling, to corpus issues, low-resource MT, and assessing MT models. This then leads to an introduction of our focus in this thesis.

2.3.1 MT Modelling

If we examine translation models, there are several renowned issues that have not been overcome yet. For instance, from the learning vocabulary perspective, the translation of rare words and out-of-vocabulary (OOV) words is always a challenge. The low frequency of OOV words and phrases that can be learned from a training set leads a model to not being able to acquire the true meaning of such words. This leads to low quality translation of these rare words and phrases, and some MT models just keep these rare words in the source text format without translating them into the output target text, keeping them as foreign words. One effort that has been made to address this issue is sub-word technology using beta-pair encoding (BPE) or similar methods [110, 245, 111].

As mentioned in Section 2.1, the *interpretation* and transparent representation of neural network models and information flow and the transfer from layer-to-layer in NMT are still unclear in most cases. This is an ongoing research topic in all applications of neural networks that attracts many researchers. For instance, [16] investigated the extracted features from the middle layers of an NMT model and their different performances in, for example, POS and semantic tagging tasks. Similarly, [60] investigated the word alignment interpretation in NMT and [174] introduced a representation learning of words, sentences and documents in neural networks for different NLP fields.

MT also faces challenges in *document level* performance since traditional models always treat each sentence separately as a segment during learning and translating, ignoring the coherence between sentences. This will lead to a huge translation quality drop if the sentences that need to be translated are connected and the correct translation needs contextual knowledge. Some recent work that tries to tackle this

challenge uses methods such as topic modelling [151], hierarchical attention networks [194], and there is a recent survey introducing evaluation methods for document-level NMT [187].

Another drawback of NMT is that NMT systems usually produce better fluent output, however the *adequacy* is lower sometimes compared with conventional SMT, e.g. some meaning from the source sentences will be lost in the translation when the sentence is long [268, 268, 145, 204, 45]. One reason for this phenomenon could be due to the above mentioned rare words problem, except for the unclear learning procedure of neural nets.

There is also a huge computational cost from the very *large learning model* issue. As discussed in Section 2.1, to address the word alignment issue in NMT, an attention mechanism can be introduced to model source encoder, target decoder, cross encoder-decoder learning, as well as coverage modelling for overall attention [268, 193, 273]. However, these lead to very large learning models which cost a large amount of resources and computational support. Thus this is not ideal for many language pairs and for carrying out practical research. Model distillations have been a topic of active research and may be a possible solution for this issue. For instance, [79] designed a teacher-student model and a filtering method based on the knowledge of the teacher model to reduce the cost of training time. Knowledge distillation is also receiving attention in other traditional NLP tasks, such as NER [97].

2.3.2 Linguistic Awareness

Linguistic awareness is another problem in state of the art NMT models when they mostly focus on data-driven machine learning methodologies. Linguistic guidance for MT can be traced back to the start of MT as a research topic, from rule-based models to statistical models. Linguistic knowledge can be incorporated into sentence structural level, e.g. grammar and syntax, and semantic level, e.g. disambiguation, idiomatic expression, meaning equivalence, and lexical diversity, as well as language register, e.g. the elegance, style and localisation of translation. Among these chal-

lenges, researchers have defined some clear sub-tasks, such as *translationese*, *gender* and *multi-Word expressions* issues, and the challenge of *parsing for MT*. We expand on some of these concepts below.

Translationese and *gender issues* are not new topics, but have started to draw attention from NMT researchers in recent years. Machine translated outputs are featured with some characteristics that are odd or unnatural according to professional translators. The detection of translationese and addressing this problem is a relatively new topic [149, 233, 92]. Gender is often mistreated by MT models, or biased for example to the male gender by many online translation platforms, including Google Translate. This might be due to the training corpus available for the MT model, but it can also be due to the learning model itself that is not capable of correctly treating gender related lexical translations. Linguistic knowledge integration is one possible solution to addressing this situation [54, 237, 272].

Multi-word expressions will influence both the syntax and the semantic aspects of a sentence and text. Correct recognition of MWEs has an impact on word bundle and overall sentence structure analysis where ambiguous MWEs such as metaphors and idioms present a semantic challenge for MT models [108, 110, 235, 109, 10].

As mentioned in Section 2.1, since the SMT paradigm, researchers have been seeking to incorporate sentence parsing knowledge into MT models, including constituency parsing and dependency parsing structures. However, this is still a challenging research topic [72, 282, 43].

2.3.3 Corpus Issues

There are additional issues regarding the corpora we use for NLP and MT tasks. Firstly, as discussed in [237], the gender imbalance from training data always reduces the performance of an MT model especially when the language has rich morphology and grammatical gender.One effort to address this issue is the creation of the WinoMT challenge set which aims at measuring gender translation quality from MT systems. This was created from English to eight other target languages each of

which have grammatical gender [260].

Register imbalance is another issue which can arise in an MT training and testing corpus. Since the large amount of available corpus data almost always comes from the internet, there are inconsistencies in the register levels. For instance, low register sequences like "haha" and "thx" sometimes appear in corpora that have high register segments [108].

Another issue that presents in many training corpora is the "data sanity" problem. There is research that discusses the test sets that have been used for MT quality evaluation and which have identified these as a problem. Whether the test suite can really reflect MT model differences and model quality is a question that has drawn attention from MT experts. This involves the diversity of the test set, the quality, coverage and the domains of operation [159, 158, 108]. Data sanity checking is also an issue in other NLP tasks such as natural language inference [264].

2.3.4 Language Bias and Low Resources

In view of "valuing diversity & committed to equality", language resources in the past have been very uneven. For data-driven SMT and NMT models, this is a big issue of how to improve model performances in the situation when a large corpus is not available [249].

There are also many large research projects which have focused on language bias and low resources in specific languages. For instance, one EU project, the QT21 consortium, has 5 language pairs as its focus, having English as the source language with the following pairs: English->German, English->Czech, English->Latvian, English->Romanian; and one having English as the target: German->English [12].

Low resource MT modelling has been a long-term research topic and involved in a series of shared tasks and workshops [132]. In WMT2021, there are two shared tasks focusing on low resource MT namely the *Multilingual Low-Resource Translation for Indo-European Languages* and *Unsupervised MT and Very Low Resource Supervised*

[12] http://www.qt21.eu/

MT [13].

Different methodologies have been proposed to tackle the issue of low resources, for instance, back-translation, multilingual modelling, unsupervised monolingual models, knowledge distillation, and NMT optimisation [246, 93, 206, 298].

As suggested by the QT21 project, multilingual embeddings are expected to support transfer learning from a well-resourced language to under-resourced languages. To benefit from multilingual resources for overcoming data shortage, multilingual and zero-shot models are also applied to relevant NLP fields, such as register analysis [232].

2.3.5 Evaluating MT

As discussed in Section 2.2, MT quality is still far from reaching human parity, and MT quality assessment plays an important role in overall research into MT [123, 198, 112, 107]. How to correctly evaluate MT models and assess their translation quality is a very challenging task. Human assessment of MT output quality has high cost and low agreement levels among human assessors. MT evaluation metrics have been used in a feedback loop not just to measure output but also to tune MT model parameters for better performance. As mentioned before, human assessment results have long been regarded as the gold standard reference for automatic metrics to learn from and against which to get higher correlation. However, very recent research from the Google MT group reflected that crowd-sourced human assessors gave very different rankings to MT system, compared to the ranking of professional translators, and get even lower correlation levels to the professional translator than the high performance automatic metrics [77]. The survey work carried out by [130] also revealed many challenges and issues in crowd-sourcing platforms and discussed possible methods on how to improve the quality controls of crowd-sourced workers. However, since the traditional WMT shared task have regarded crowd-source human

[13] available at http://www.statmt.org/wmt21/multilingualHeritage-translation-task.html and http://www.statmt.org/wmt21/unsup_and_very_low_res.html. Multimodal LowRes MT workshop https://sites.google.com/view/mmtlrl-2021/home

assessors as the gold standard reference, this raises a new issue of how to correctly guide MT researchers in pursuing this.

Quality metrics have been a driving force for better MT model development [182]. However, very recent research work reveals that popular automatic evaluation metrics cannot make a correct distinction among state of the art NMT models, especially as they make real progress toward human parity. For instance, the BLEU metric even produced several points drop to better system outputs that human experts preferred, compared to alternative systems that human experts assigned as lower quality [78, 76]. These findings are in consistency with the earlier work by [34, 146] on the mis-leading of automatic BLEU scores. Automatic metrics have favoured n-gram matched "boring translations", instead of really better translations with improved lexical choices. As we discussed previously in Section 2.2.5, statistical significance and power testing is yet another issue in MT evaluation, when we want to know if two system really make a difference to output quality[143, 92].

To evaluate state of the art models in a more efficient way, better test sets and reference sets are also needed. How to better evaluate state of the art translation models, more accurately, is an open issue. Later in this thesis we will propose a new test set and a corresponding new evaluation methodology on this, in Chapter 5.

2.3.6 Summary and Thesis Focus

In this section we have introduced the current challenges with MT from different aspects across MT models, corpora, language resources, and evaluation methods. In this thesis, we focus on **multi-word expressions** in MT with the following rationale. Firstly, MWEs have been a widely recognised bottleneck in NLP for a long time [235, 53, 10]. Secondly, regarding the issues we mentioned in the last section, MWEs have an influence across a broad range of topics. For instance, MWEs present idiomaticity and unexpected syntactical appearances that have a big influence among rare words and idiomatic phrase translations, and consequently also on the adequacy issue in *MT modelling* and *linguistic awareness* challenges.

MWEs are an important indicating factor if the MT models which use them are to achieve real human parity in terms of performance. However, there are currently no available bilingual corpora available for MT testing, something which belongs to the *corpus issue*. Regarding the issue of *evaluating MT*, there is still a lack of translation quality assessment methodology with a focus on MWEs even though relevant knowledge such as named entities and terminologies have been widely used as features in MT evaluation.

In the next section, we give two MT examples with MWEs investigation where we explain how MWEs are a bottleneck in state of the art Neural MT models.

2.4 Multi-word Expressions (MWEs) as a Bottleneck in MT

For more examples on how MWEs play a role in MT, we list two Chinese-to-English MT sentences, one from a popular Chinese poem, and the other a popular Chinese proverb from an ancient essay. For more examples on how MWEs play a role in MT, we list Chinese↔English MT sentences using several state of the art MT systems including DeepL, Google, Bing, and Baidu MT. The Chinese-to-English examples in Section 2.4.1 are from popular Chinese poems and proverbs (Classical Chinese), and the English-to-Chinese examples in Section 2.4.2 are from literature text. These examples will show the different challenges that MWEs present to modern MT systems.

2.4.1 Chinese-to-English Examples of MWEs in MT

ZH-to-EN MT in Poem (詩歌 shī gē)

For the first example in Figure 2.9, we will see how correct understanding of Chinese MWEs can assist disambiguation in machine learning. Conversely, the failure to understand these MWEs can lead to an incorrect translation of the ambiguous Chinese

character even in very well aligned poem sentences. The first example sentence is "年年歲歲花相似，歲歲年年人不同 (phoneticism: nián nián suì suì huā xiāng sì, suì suì nián nián rén bù tóng)" from a poem "< 代悲白頭翁 > (dài bēi bái tóu wēng, translation: *A Song of the Elder*)" of Tang Dynasty by Xiyi Liu[14], shown in Figure 2.9.

The source Chinese sentence contains the MWEs "年年歲歲" (nián nián suì suì) and "歲歲年年" (suì suì nián nián) with clear boundaries. These two MWEs have the same meaning "each/every year", however, the MT model used here translated them in different ways and led to the incorrect translation of the overall sentence. In this example, the *poem* contains two sub-sentences that are well aligned with their inside words. For instance, the MWEs "年年歲歲" (nián nián suì suì, each/every year) and "歲歲年年" (suì suì nián nián, each/every year) are aligned to each other as adverbial phrases specifying the timing; "花" (huā, flower) and "人" (rén, people) are aligned to each other as the subject of the sub-sentences; and "相似" (xiāng sì, similar) and "不同" (bù tóng, different) are aligned to each as the action of the sentences: flowers (are) similar, while people (are) different. Because of the incorrect translation of the first MWE, or failing to recognise them, the MT model did not translate the subsequent character "花" (huā, flower) correctly, and also lost the translation of "人" (rén, people), reflecting the *adequacy loss*. The Chinese character "花" (huā) is an ambiguous word which can be "spend (money, time)" or "flower". The successful recognition and translation of the two MWEs "年年歲歲" (nián nián suì suì, each/every year) and "歲歲年年" (suì suì nián nián, each/every year), and their correct alignment, will help the MT system to translate "花" (huā) as "flower" instead of "spend" since it is aligned to "人" (rén, people) as a noun, instead of a verb. In other situations, such as in this example of Chinese Báihuà, '我花一百，你呢？ (Wǒ huā yībǎi, nǐ ne?)', it means 'I *spend* one hundred, how about you?'.

We assume that the correct recognition and translation of surrounding MWEs,

[14]劉希夷 in Chinese, 651—679, who died at early age due to this famous poem he wrote. https://zh.wikipedia.org/wiki/

ZH source:	年年歲歲花相似，歲歲年年人不同
ZH pinyin:	Nián nián suì suì huā xiāng sì, suì suì nián nián rén bù tóng.
EN reference:	The flowers are similar each year, while people are changing year after year.
EN MT output:	One year spent similar, each year is different

Figure 2.9: Chinese-English MT example from poem (詩歌 shī gē)

in general, can help the MT model to understand the sentences better overall and improve the translation of ambiguous Chinese characters.

ZH-to-EN MT in Proverb (文言 wén yán)

As a second example, similar to example one (Chinese Báihuà), we show how the MT model fails to translate a proverb from Classical Chinese Wényán sentence due to the lack of **Chinese MWE pattern** knowledge. Even though it is still a popular saying, the translation of this Wényán sentence is much worse than the translation of modern Chinese Báihuà. This example also contains the *multi-character named entity* information as one kind of MWE.

This example, shown in Figure 2.10, is a translation of the ancient Chinese 文言 (Wényán) metaphor expression to English: "燕雀安知鴻鵠之志哉？ (phoneticism: yàn què ān zhī hóng hú zhī zhì zāi?)" from the book "《史記》 (shǐ jì, archive of history)" [15] by Sima Qian. This Chinese expression is often used in modern language to express someone's feelings in both verbal and written format. The MWE pattern "A 安知 (ān zhī) B 哉 (zāi)？" is used to express "how can A know B?" or "A does not know B". This metaphor is used to describe that some not serious or very common folks do not know the ambition or great plan of other very motivated ones.

The MT output is poor due to the model not understanding the meaning of the

[15] From Han Dynasty, 206 BC–220 AD https://en.wikipedia.org/wiki/Records_of_the_Grand_Historian

ZH source:	燕雀安知鴻鵠之志哉？
ZH pinyin:	Yàn què ān zhī hóng hú zhī zhì zāi?
EN reference (literal):	How can a finch know the ambition of a big bird (or swan)?
EN MT output:	What is the meaning of Yanque Anzhihong?
EN reference:	Common folk do not know the ambitions of the very motivated people.

Figure 2.10: Chinese-English MT example from proverb (文言 Wényán, Classical Chinese)

entities "燕雀 (yàn què, meaning finch)" and "鴻鵠 (hóng hú, meaning big bird, swan)", the **fixed/patterned expressions** "安知 (ān zhī, meaning 'how to know' or 'do not know')" and "之志 (zhī zhì, meaning *someone*'s ambition)". In the MT output, we can see that it keeps 'Yanque' in the form of the original Chinese pinyin pronunciation. This may be due to the MT system not acquiring this meaning equivalent word from its training data. The MT output also failed by putting 'Anzhihong' together the pinyin pronunciation of the three Chinese characters '安知鴻', which makes no sense at all, since '安知 (ān zhī)' is one term (patterned expression) and '鴻 (hóng)' should be part of another term (named entity) '鴻鵠 (hóng hú)'. The failure to correctly interpret these kinds of expressions presents an obstacle to effective MT.

2.4.2 English-to-Chinese MT Issues with MWEs

In this section, we introduce some broader MT issues related to MWEs that still exist for current state of the art MT models. As a complement to the last section where we used Chinese-to-English examples, here we use the English-to-Chinese direction which will be in line with our corpus construction work to be carried out in Section 5.2. The tested example sentences are from literature domain, e.g. from novels, instead of poems or Classical Chinese (wén yán, 文言) that are used in the

last section. The four MT engines we used are Bing, Google, Baidu, and DeepL. Regarding MWE related or caused issues and phenomena, these include the following situations named entities, *Engli*-Chinese, misusing MWEs, document-MT, MT on Chinese Pinyin, and simplified vs traditional Chinese character sequences in MT.

Named Entities (NEs)

Named entities (NEs) in western languages always match multi-character sequences (MWEs) in the Chinese language, even though it could be a single word in the original form. In some documentary stories with famous names, including the Bible, the messed-up translation of the named entities can lead to very inaccurate history, and cause arguments, e.g. *Absalom* was the son of David, King of Israel, while *Abraham* was the founding father of the Covenant of the pieces, the special relationship between the Hebrews and God [16]. In this example sentence, Fig. 2.11, the named entity "Herodian" was correctly translated by DeepL as "希律王 (xī lǜ wáng)", however, wrongly translated as "hero" by Bing; the named entity "Absalom" was kept as an unknown word by Bing without translation, but was wrongly translated by DeepL into "亚伯拉罕 (yà bó lā hǎn)", which is actually another named entity"Abraham", and the correct translation in Chinese is the multi-character/word sequence"押沙龙 (yā shā lóng)"[17].

Source	The Herodian relics are all that relics should be columns distorted, well worked over by time, Absalom's tomb with its bulbous roof and odd funnel tapering out of it.
DeepL	希律王的遗迹是所有的遗迹都应该是柱子扭曲的，被时间加工过的好的，亚伯拉罕的墓，它的屋顶是球状的，奇怪的漏斗从里面渐渐露出来。(xī lǜ wáng ...,... yà bó lā hǎn)
Bing	英雄遗物都是文物，应该被柱子扭曲，随着时间的推移，Absalom的坟墓与它的球状屋顶和奇怪的漏斗逐渐减少出来。(yīng xióng..., ... Absalom)
Ref.	希律王的遗迹都是柱子成扭曲状的，随着时间的推移精心加工，押沙龙式的坟墓带有球茎状的屋顶和奇怪的漏斗渐渐地从里面露出来。(xī lǜ wáng ...,... yā shā lóng)

Figure 2.11: English-to-Chinese MT: named entity

[16] https://www.thebiblejourney.org
[17] https://biblehub.com http://biblehub.net/searchchin.php?q= 押沙龙

Engli-Chinese

This situation occurs when the MT models face some translation of English phrase patterns, often involving MWEs. We call this *Engli-Chinese* or *EngliChinese*, which means the Chinese sentence sounds like the English pattern based expression. This can happen if the syntax of the Chinese sentence is in English, or there are some English input words (or literally translated words) that apparently make the sentence sound awkward in Chinese.

For the first example sentence, Fig. 2.12 "will find ourselves (forced to) ⋯" is an English patterned MWE that should not be translated just word by word into Chinese. However, DeepL and Google both made the same mistake by the literal translation into Chinese, either dropping "find ourselves" in the translation or replacing "find" as "realise" which would make more sense in Chinese, such as "将(意识到自己) 不得不, jiāng (yì shí dào zì jǐ) bù dé bù". For the second sentence, "see the last of its Christianity" contains one idiom "see the last of (someone)", however, DeepL, Bing and Baidu all translated it into an awkward Chinese sequence, especially Baidu, which literally translated it word by word from English plus a literal moving of "its" to the front.

Misusing MWEs

The current state-of-the-art (SOTA) MT models apparently tried to incorporate target side MWEs into the candidate output. However, some MWEs in the target side were wrongly chosen which resulted in the translation having a different meaning. The output of DeepL with the example sentence in Fig. 2.13 used a Chinese idiom "胸大无脑 (xiōng dà wú nǎo, literally "big chest, no brain") to translate "whose chest is large" which is wrong by virtue of adding some extra meaning that does not exist in the source text. This may be caused by the training corpora that DeepL used which have a higher probability to align "chest ... large ..."to "胸大无脑 (xiōng dà wú nǎo)". In addition, the Google output is not correct either by using "胸大了 (xiōng dà le, chest becomes large)" which is different to the original meaning.

Source	Otherwise we will find ourselves forced to take more serious decisions as regards our trading relations with Israel. At times I suspect that the world would be glad to **see the last of its Christianity**, and that it is the persistency of the Jews that prevents it.
DeepL	否则，我们将发现我们将被迫在与以色列的贸易关系上作出更严肃的决定。(jiāng fā xiàn wǒ men jiāng bèi pò) 有时我怀疑，世界会很高兴**看到其最后的基督教**，而阻止它的是犹太人的顽固不化。(kàn dào qí zuì hòu de jī dū jiào)
Bing	否则，我们将被迫就我们与以色列的贸易关系作出更严肃的决定。(jiāng bèi pò jiù) 有时我怀疑，世界会很高兴**看到其基督教的最后一个**，是犹太人的坚持阻止它。(kàn dào qí jī dū jiào de zuì hòu yī gè)
Google	否则，我们将发现自己被迫就与以色列的贸易关系做出更严肃的决定。(jiāng fā xiàn zì jǐ bèi pò) 有时我怀疑，世界将会很高兴**看到基督教的最后一幕**，而犹太人的顽强阻止了它的发展。(kàn dào jī dū jiào de zuì hòu yī mù)
Baidu	否则，我们将被迫对我们与以色列的贸易关系作出更严肃的决定。(jiāng bèi pò duì) 有时我怀疑，世界会很高兴**看到它的基督教的最后一次**，是犹太人的坚持阻止了它。(kàn dào tā de jī dū jiào de zuì hòu yī cì)
Ref.	否则，我们将(意识到自己)不得不就与以色列的贸易关系作出更严肃的决定。(jiāng *yì shí dào zì jǐ* bù dé bù jiù) 有时我怀疑，世界会很高兴**跟基督教说拜拜**，而阻止它的是犹太人的坚持。(gēn jī dū jiào shuō bái bái)

Figure 2.12: English-to-Chinese MT: *Engli*-Chinese

Source	We step into the street and my friend David Shahar, whose chest is large, takes a deep breath and advises me to do the same.
DeepL	我们步入街头，我的朋友大卫-沙哈尔**胸大无脑**，他深吸一口气，建议我也这样做。(xiōng dà wú nǎo)
Google	我们走进这条街，我的朋友大卫·沙哈（David Shahar）的**胸大了**，深吸了一口气，建议我也这样做。(xiōng dà le)
Ref.	我们步入街头，我的朋友大卫-沙哈胸膛大，他深吸一口气，建议我也这样做。(xiōng táng dà)

Figure 2.13: English-to-Chinese MT: misusing MWEs

Document-level MT

There are some phrases and MWEs that usually need context information to be understood correctly first, before translation. In some cases, this is because of ambiguity, while-others, it is just out of the blue. We present two examples here, one verbal MWE (vMWE) as an idiom and another as a noun phrase (adj+adj+noun) "out of the blue", in Fig. 2.14. The examples prove that the current SOTA MT models can't handle this well yet.

The idiom "swallow one's pride" means "to decide to do something although it will make him/her feel embarrassed or ashamed", and in this context, it means he

Source	Still, it was painful, and he struggled desperately to swallow his pride.
DeepL	sentMT: 但还是很痛苦，他拼命挣扎着吞下自己的骄傲。(tūn xià zì jǐ de jiāo'ào) docMT: 不过，这还是很痛苦的，他拼命挣扎着吞下自己的自尊心。(tūn xià zì jǐ de zì zūn xīn)
Google	sentMT: 仍然，那是痛苦的，他拼命挣扎着吞下自己的骄傲。(tūn xià zì jǐ de jiāo'ào) docMT: 尽管如此，这还是很痛苦的，他拼命挣扎着吞下自己的骄傲。(tūn xià zì jǐ de jiāo'ào)
Ref.	尽管这很痛苦，他拼命挣扎着忍下自己的自尊心（没有做回应）。(rěn xià zì jǐ de zì zūn xīn *méi yǒu zuò huí yīng*)
Context	He was about to tell her who he was, but then he realized that it made no difference. The girl was beyond hope. For five years he had kept William Wilson's identity a secret, and he wasn't about to give it away now, least of all to an imbecile stranger. *Still, it was painful, and he struggled desperately to swallow his pride.* Rather than punch the girl in the face, he abruptly stood up from his seat and walked away.
Source	Shrugging, he gives up and I turn to the twice disagreeable chicken and eat guiltily, my appetite spoiled.
DeepL	sentMT: 耸了耸肩，他放弃了，我转身对着那只两次不欢而散的鸡，内疚地吃着，胃口被破坏了。(nà zhǐ liǎng cì bù huān ér sàn de jī) docMT: 耸耸肩，他放弃了，我转而吃起了两次不喜欢的鸡肉，内疚地吃着，胃口被破坏了。(liǎng cì bù xǐ huān de jī ròu)
Google	sentMT: 耸了耸肩，他放弃了，我转向那两次令人讨厌的鸡肉，内eat地吃了，我的胃口变坏了。(nà liǎng cì lìng rén tǎo yàn de jī ròu) docMT removing # text =: 耸耸肩，他放弃了，我转向那两次令人讨厌的鸡肉，内地吃了，我的胃口变坏了。(nà liǎng cì lìng rén tǎo yàn de jī ròu)
Ref.	他耸了耸肩，放弃了，我转身对着那只引起两次争论的鸡，内疚地吃着，胃口变坏了。(nà zhǐ yǐn qǐ liǎng cì zhēng lùn de jī)
Context	You must never never eat trephena food again." "I can't promise you that. You're asking too much. And just for one sandwich." "I have a duty toward you," he tells me. "Will you listen to a proposition?" "Of course I will." "So let us make a deal. I am prepared to pay you. If you will eat nothing but kosher food, for the rest of your life I will send you fifteen dollars a week." "That's very generous," I say. "Well, you are a Jew," he says. "I must try to save you." "How do you earn your living?" "In a Hasidic sweater factory in New Jersey. We are all Hasidim there. The boss is a Hasid. I came from Israel five years ago to be married in New Jersey. My rabbi is in Jerusalem." "How is it that you don't know English?" "What do I need English for? So, I am asking, will you take my fifteen dollars?" "Kosher food is far more expensive than other kinds," I say. "Fifteen dollars isn't nearly enough." "I can go as far as twenty-five." "I can't accept such a sacrifice from you." *Shrugging, he gives up and I turn to the twice disagreeable chicken and eat guiltily, my appetite spoiled*.

Figure 2.14: English-to-Chinese MT: document-level vs sentence-level

"did not say anything", instead, just walked away. [18] Here we can translate it as "忍下自己的自尊心 (rěn xià zìjǐ de zì zūn xīn)" or "没有做回应 (méi yǒu zuò huí yīng)". DeepL document level MT gave a good try by using the word "自尊心 (zì zūn xīn)", however, "吞下自己的自尊心 (tūn xià zìjǐ de zì zūn xīn)" is not a correct way to say this in Chinese. All other three MT models, including sentence level DeepL, and sentence /document level GoogleMT yielded the same literal translation "吞下自己的骄傲 (tūn xià zìjǐ de jiāo ào)" which makes no sense in Chinese.

From the context, we understand the point that "the twice disagreeable chicken" in the second sentence means that "the chicken caused two disagreements". However, the document level MT of DeepL and Google both failed to translate this meaning

[18] https://dictionary.cambridge.org/dictionary/english/swallow-your-pride

correctly, even though they did make some difference compared with the sentence level model.

MT on Chinese Pinyin

Out of the four MT models we used, two of them offer the Chinese pronunciation in Roman alphabets, i.e., Google and Bing, and this is termed "拼音 (Pīn yīn)". The character "拼 (pin)" indicates writing, and "音 (yin)" indicates sounds/reading. We discuss the issues when MT models offer Pinyin for Chinese characters, which are sometimes related to MWEs.

The Bing Translator just kept each Chinese character's Pinyin separated one by one, without any morphological segmentation. GoogleMT's output of Pinyin tries to perform Chinese word segmentation, i.e., concatenate several characters' Pinyin together if they belong to the same word (in Chinese sense). However, there are several apparent errors in GoogleMT' Pinyin.

1. the Pinyin is wrong itself (totally wrong, or wrong choice in disambiguation situation);

2. the tone is wrong;

3. the word segmentation is wrong which misleads the understanding of the sentence.

In Fig. 2.15, we listed these three types of error with examples. The first error type was presented in two different situations, with one due to wrong choice facing ambiguous pronunciation, and the other is totally wrong.

The characters "希律王 (Xī lǜ **wáng**, Herodian)" in the sample sentence with the named entity issue was wrongly annotated as "Xī lǜ **wán**" by GoogleMT, which belongs to a totally wrong situation [19]. The second sentence of this error, "去地精 (qù dì jīng, *de-gnoming*)" was wrongly annotated as "qù dejīng" which is due

[19]Furthermore, the entity "押沙龙 (yā shā lóng, Absalom)" was wrongly segmented into two pieces "押 (yā, bet)" and "沙龙 (shālóng, salon)". This leads to mis-understanding.

Pinyin Wrong: 100% wrong, Ambiguous wrong	Chiense	**希律王**时代的遗物就是所有的遗物，都应该扭曲成柱状，随着时间的推移经过精心加工，押沙龙的坟墓带有球形的屋顶和奇特的漏斗。(**xī lǜ wáng … yā shā lóng …**) 一知道去地精的事在进行，他们就冲上去观看。 (qù dì jīng)
	Google Pinyin	**xī lǜ wán** shídài de yíwù jiùshì suǒyǒu de yíwù, dōu yīnggāi niǔqū chéng zhùzhuàng, suízhe shíjiān de tuīyí jīngguò jīngxīn jiāgōng, yā shālóng de fénmù dài yǒu qiúxíng de wūdǐng hé qítè de lòudǒu. yī zhīdào qù dejīng de shì zài jìnxíng, tāmen jiù chōng shàngqù guānkàn.
	Ref.	**xī lǜ wáng** shídài de yíwù jiùshì suǒyǒu de yíwù, dōu yīnggāi niǔqū chéng zhùzhuàng, suízhe shíjiān de tuīyí jīngguò jīngxīn jiāgōng, yāshālóng de fénmù dài yǒu qiúxíng de wūdǐng hé qítè de lòudǒu. yī zhīdào qùdìjīng de shì zài jìnxíng, tāmen jiù chōng shàngqù guānkàn.
Tone Wrong	Chiense	椅子很舒服，啤酒已经微微让他上了头。 (wēi wēi) 然后有人说，现在用不了多久，**俄**罗斯人就会让阿拉法特下台。(**é** luó sī)
	Google	yǐzi hěn shūfú, píjiǔ yǐjīng wéiwéi ràng tā shàngle tóu. ránhòu yǒurén shuō, xiànzài yòng bùliǎo duōjiǔ, **èluósī** rén jiù huì ràng ā lā fǎ tè xià tái.
	Ref.	yǐzi hěn shūfú, píjiǔ yǐjīng wēiwēi ràng tā shàngle tóu. ránhòu yǒurén shuō, xiànzài yòng bùliǎo duōjiǔ, **éluósī** rén jiù huì ràng ā lā fǎ tè xià tái.
Segmentation Wrong	Chiense	在72街和**麦迪逊**大道的拐角处，他招手示停了一辆在72街和麦迪逊大道的拐角处，他招手示停了一辆出租车。(**mài dí xùn** … chū zū chē) 在72街和**麦迪逊**大道的拐角处，他招手示停了一辆出租车。(mài dí xùn … guǎi jiǎo … zhāo shǒu shì tíng le)
	Google	zài 72 jiē hé **màidí xùn** dà dào de guǎi jiǎo chù, tā zhāo shǒu shì tíngle yī liàng chūzū chē.
	Ref.	zài 72 jiē hé **màidíxùn** dà dào de guǎijiǎo chù, tā zhāoshǒushìtíng le yī liàng chūzūchē.

Figure 2.15: English-to-Chinese MT: Pinyin on MWEs

to the character "地 (dì)" which has different pronunciations [20]. Due to the model failing to acquire the meaning of this term, it chose an incorrect pronunciation.

As another example, the reference sentence from the "common sense" issue in Fig. 2.15, the named entity "麦迪逊 (mài dí xùn, Madison)" was wrongly segmented as "麦迪 (màidí)" and "逊 (xùn)". These two examples also reflect that the current MT models have issues in recognising or processing *foreign named entities*. In addition, the vMWE "招手示停了 (zhāo shǒu shì tíng le, waved down)" was incorrectly segmented as "招 | 手 | 示 | 停了 (zhāo shǒu shì tíngle)" which should be "招手示停 | 了 (zhāoshǒushìtíng le)". This illustrates the issue in accuracy level of automatic recognition of vMWE. Furthermore, "拐角 (guǎi jiǎo, corner)" and "出租车 (chū zū chē, taxi)" are two concrete concepts and should be presented as "guǎijiǎo" and "chūzūchē" which concatenate their internal characters in word-segmentation format, instead of "guǎi jiǎo" and "chūzū chē" by GoogleMT.

[20] https://baike.baidu.com/item/地/34380 and https://en.wiktionary.org/wiki/地

Simplified vs Traditional Chinese

Traditional Chinese characters have rich linguistic knowledge and have naturally evolved over thousands of years of civilisation from pictographs, while simplified Chinese characters are from recent history with less than one hundred years of usage in mainland China. We think it is valuable to study both the two scripts and their usage in context, especially as the traditional Chinese characters are still being used by Taiwan, Hong Kong, Macau, and other regions, such as in Kanji characters (made up of traditional Chinese) in Japan.

When we choose to translate into traditional Chinese, it usually means simply to replace simplified Chinese characters by traditional ones without changing any grammatical structure or word choices, but using the same Mandarin Chinese. However, there are some issues with the mismatching translations, for instance, some named entities are translated in different ways in simplified or traditional Chinese by mainland China or Hong Kong / Taiwan.

Different to Mandarin, which is the official language of mainland China, Cantonese is another popular dialect of Chinese widely spoken by Hong Kong and Macau. Cantonese was originally from Guangdong Province, the most southern part of China, whose capital city "Guangzhou" was also known as "Canton".

Out of the four MT models, Bing and Baidu offer both traditional Chinese and Cantonese translation outputs. The difference is that Bing offers Cantonese in traditional Chinese characters, while Baidu offers it in simplified characters.

In the example sentence in Fig. 2.16, where 2sim/tra/canto means English to simplified/traditional (Mandarin), and to Cantonese. Regarding the translation of the title <The Sound and the Fury>, there are at least two issues with the Bing Translator at this point: firstly, when it translates into simplified Chinese, it actually uses the named entities from HK/TW translation < 声音与愤怒, shēng yīn yǔ fèn nù>, just the simplified character of < 聲音與憤怒, shēng yīn yǔ fèn nù>, which is a mess-up. Secondly, when it translates into Cantonese in traditional Chinese, it does not translate the term < 聲音與憤怒, shēng yīn yǔ fèn nù> correctly by

Source	Reading The Sound and the Fury last night, I came upon words in Compson's thought that belonged to E E Cummings and the thirties, not to the year 1910.
DeepL	2sim: 昨晚读《声色犬马》，我在康普森的思想中看到了属于E-E-康明斯和三十年代的文字，而不是1910年的文字。(shēng sè quǎn mǎ) 2tra/2canto: not available
Google	2sim: 昨晚阅读《喧Sound与骚动》时，我发现康普森思想中的词语属于卡明斯和三十年代，而不是1910年。(xuān Sound yǔ sāo dòng) 2tra: 昨晚閱讀《喧Sound與騷動》時，我發現康普森思想中的詞語屬於卡明斯和三十年代，而不是1910年。(xuān Sound yǔ sāo dòng) 2canto: not available
Bing	2sim: 昨晚读了《声音与愤怒》时， 我在康普森的念头里发现了属于 E E 卡明斯和三十年代的话， 而不是 1910 年。(shēng yīn yǔ fèn nù) 2tra: 昨晚讀了《聲音與憤怒》時, 我在康普森的念頭里發現了屬於 E E 卡明斯和三十年代的話, 而不是 1910 年。(shēng yīn yǔ fèn nù) 2canto: 琴晚讀咗《聲與憤怒》時,我喺康普森嘅念頭度發現咗屬於E E卡明斯同三十年代嘅話,而唔係1910年。(shēng yīn yǔ fèn nù)
Baidu	2sim: 昨晚读《喧哗与骚动》时，我在康普森的思想里发现了一些属于E E·卡明斯和三十年代的词，而不是1910年的词。(xuān huá yǔ sāo dòng) 2tra: 昨晚讀《喧嘩與騷動》時，我在康普森的思想裏發現了一些屬於E E·卡明斯和三十年代的詞，而不是1910年的詞。(xuān huá yǔ sāo dòng) 2canto: 寻晚睇《嘈吵与骚动》嘅时候，我喺康普森嘅思想度发现咗一啲系属於E E·卡明斯同三十年代嘅字，而唔系1910年嘅字。(cáo chǎo yǔ sāo dòng)
Ref.	2sim: 昨晚读《喧哗与骚动》，我在康普森的思想中间或看到了属于E-E-康明斯和三十年代的文字，而不是1910年的。(xuān huá yǔ sāo dòng) 2tra: 昨晚讀《喧嘩與騷動》，我在康普森的思想中間或看到了屬於E-E-康明斯和三十年代的文字，而不是1910年的。(xuān huá yǔ sāo dòng) 2canto: 琴晚睇《聲音與憤怒》果陣時,我發現喺康普森嘅念頭度，係屬於E E卡明斯同三十年代嘅說話， 而唔係1910年。(shēng yīn yǔ fèn nù)

Figure 2.16: English-to-Chinese MT: Mandarin vs Cantonese, Simplified vs Traditional Chinese.

dropping < 音, yīn>, in addition to that the overall sentence needs to be improved. Baidu's translation of the Cantonese name < 嘈吵与骚动, cáo chǎo yǔ sāo dòng> is incorrect, although the simplified translation < 喧哗与骚动, xuān huá yǔ sāo dòng> gets credit.

Regarding the translation from DeepL, < 声色犬马, shēng sè quǎn mǎ> is a 1974 Chinese movie from HK, which has nothing to do with the tittle <The Sound and the Fury> in the source sentence.

These mis-translations are partially due to the literal translation issue and ambiguity. For instances, "声音与愤怒 (shēng yīn yǔ fèn nù)" is literally from "sound (声音)" and "fury (愤怒)", "嘈吵 (cáo chǎo) is also connected literally to "loud (sound)", and the Chinese movie "声色犬马 shēng sè quǎn mǎ" happens to have "声 shēng (sound, loud, noise)" as part of its title.

This mis-translation of English named entity issue reflects the difference be-

tween mainland China and other Chinese speaking regions (Hong Kong, Macau and Taiwan), also simplified vs. traditional Chinese characters. To train a more precise MT model for English-to-Chinese, we need to make the corresponding changes and clarification when we prepare a simplified or traditional Chinese corpus for MT purpose.

2.5 Summary

In this chapter, we introduced the development of MT from rule-based methods to statistical and neural models, compared their differences and interaction aspects. We then presented the evaluation methodology for MT models covering both human and automated criteria and metrics, as well as quality estimation (QE) models and meta-evaluation of evaluating methods. Following that we summarised the broad issues and challenges remaining in MT research. Finally, we pointed out the focus of this thesis on an investigation into MWEs in MT by presenting: 1) two Chinese-to-English translation examples with MWE issues and explained some of the insights into how the mis-translation of the MWEs causes problems, and 2) detailed MWE related or caused issues in MT on English-to-Chinese translation direction. In the next chapter we introduce the literature on how MWEs are handled in current MT and this leads to the hypotheses and research questions in the thesis on how we propose to address the challenges in a different way, i.e. using decomposition models for MT learning towards better MWE translation and designing new evaluation methodologies incorporating MWEs.

Chapter 3

Literature Review of Addressing Multi-Word Expressions in Machine Translation

In this chapter, we firstly introduce approaches and background about the issue of MWEs in machine translation, and then present a literature review on existing investigations into MT modelling. This is followed by a discussion of current issues, and from there we derive the hypotheses and research questions (RQs) that make up this thesis.

3.1 Activities and Background Related to MWEs in MT

There have been some MWE related activities and investigations in the development of MT, as well as in the development of other computational linguistics and NLP research such as lexicon and word-net studies. The international workshop on MWEs has been organised annually since since 2003 by the SIGLEX-MWE group (Special Interest Group in Lexicon - MWE section), from the Association for Com-

putational Linguistics (ACL)[1]. In 2020 this reached its 16th edition, affiliated to the COLING conference [183][2]. This annual workshop has focused on the analysis, acquisition, identification, discovery, interpretation, disambiguation, annotation and construction of MWEs from both theoretical and practical aspects. Throughout its history, it has been co-organised in collaboration with other relevant research topics including parsing, machine translation, linguistic annotation, and electronic lexicons. A scientific shared task on MWE discovery and identification has been held for several years starting in 2017 [184, 227, 180], where monolingual corpus with annotation of verbal MWEs was offered as challenges across different languages, mainly European ones (the 2020 edition included Chinese, Hindi and other non-European languages). Our recent work concluded that the shared task test set, like many other NLP shared task issues, offered an annotated reference set that does not have good coverage of correct MWEs that system output produced and this in turn lowered the overall evaluation and system ranking accuracy [200]. In other words, some well-performing systems were not given sufficient credit for producing correct MWE annotations.

Another workshop series that has been held every two years is the EUROPHRAS "Workshop on multi-word Units in Machine Translation and Translation Technologies" since 2013, and the 4th edition was held in 2019 in Malaga, Spain [215] [118] [3]. EUROPHRAS conferences have been focusing on computational and corpus-based phraseology, jointly organised by the European Association for Phraseology (EUROPHRAS) and the Association for Computational Linguistics.

Closely affiliated to the MWE annual workshops there are phraseology and multiword expression book series that have encouraged research on MWEs and lexicography and their interactions and applications with other fields, published by Language Science Press [200, 238] [4].

[1] https://multiword.org/

[2] All proceedings are in the ACL anthology https://www.aclweb.org/anthology/venues/mwe/

[3] http://www.lexytrad.es/europhras2019/mumttt-2019-2/

[4] webpage https://langsci-press.org/catalog/series/pmwe and the 2018 edition at https://langsci-press.org/catalog/book/204

A new benchmark task focusing on German verbal idiom disambiguation was held in 2021 with KONVENS2021[5]. For verbal idiom examples, let us examine this sentence "He had a few too many shots of tequila and spent the rest of the night *talking on the big white telephone*." Is the *talking on the big white telephone* in this context literally meaning he is talking on phone or is he is vomiting into a toilet? Sentences like this are a challenge for natural language understanding and generation tasks, including MT [6].

MWE research also includes named entities (institutions, locations, persons, etc.) and terminology studies. In the WMT2021 workshop, there was a new shared task on "Machine Translation using Terminologies" pointing out that in many newly-developed domains, parallel data are often hard to come by. It is our intuition to treat word and multi-word level terminologies in terms of consistency and accuracy for helping to guide professional translators. The shared task aims to develop state of the art methods incorporating terminology knowledge in training and an inference process to help MT systems in new domains on the two language pairs English to French and to Chinese.

From a linguistic background, the MWE phenomenon is related to lexical bundle research as investigated by [299] in a corpus based comparative study covering English and Chinese journal articles, and [173] in a study of native and Chinese scholars' English abstracts taking linguistics and chemistry as an example. Very recent work by linguists also investigated MWEs as discourse markers where they use statistical MT to build a parallel corpus to study the corresponding Hebrew and Lithuanian discourse markers aligned to the pivot English language ones, and explore their corresponding functional and lexical changes [211]. The English source text is taken from TED talks. Another very related work is [61] which identifies structurally fixed discourse marking MWEs in spoken Slovene.

MT is a research task aiming to achieve meaning-preserving transformation from one human language to another. However, there are many customised terms that

[5] https://competitions.codalab.org/competitions/31715
[6] Source of idioms https://idioms.thefreedictionary.com/

native speakers often use in their own language that are not easily captured by other native language speakers or are not common in other languages. The root of the customised terms, such as idioms, fixed expressions, and proverbs is generally connected with the culture and history of the people who speak that language as native speakers. Due to this fact, the expressions / terms are often not exactly the same in different languages. This constitutes a significant challenge for MT, since often literal or surface word translation and such kinds of translations are not enough to make the meaning fully presented in the target language. As mentioned in Section 2.4, MWE is a commonly occuring phenomena that can reflect idioms, fixed expressions (e.g. kick the bucket), customised expression (e.g. once upon a time), proverbs, and more. In light of this, it is natural to think of the combination of MT and MWE, e.g., how to correctly translate MWEs in the MT task to make meaning-equivalent translations. This includes how to combine MWE identification and MWE prepossessing into MT models and systems in order to achieve better translation.

In the next section, we review existing literature that investigates or integrates MWE knowledge into MT models, including both statistical MT (SMT) and neural MT (NMT). We discuss state of the art models, their pros and cons, and current translation quality assessment issues, e.g. lack of indicators on idiomatic expressions as well as test suit narrative, before leading to our hypotheses and research questions.

3.2 Literature Review of MWE Investigations in MT

MWEs play a significant role in language understanding and processing tasks, including MT. This is due to their very frequent appearances and their concept specific presentation. How to recognise MWEs correctly and translate them in a meaning-preserving way, instead of merely surface word translation is a challenging task. This section introduces existing research work in this area.

MWEs in MT are related to word sense **disambiguation** (WSD) [274, 42], phrase **boundary** detection, and semantics [271]. Instead of a single word case in WSD, MWEs are multiple-word expressions, which can be translated in an awkward way if the translation model cannot translate the actual meaning of an MWE in the sentence and context, such as metaphorical MWEs ('apple of someone's eyes', 'kick the bucket', listed as simple examples). Addressing MWE translation also addresses the **semantic** aspects of translation in addition to issues of syntax, e.g. MWE boundary (detection) and its affects on overall sentence understanding. For instance, examples in Figure 2.9 and 2.10 show how the MT model produced very poor output due to not recognising MWE boundaries well. Investigations into WSD have been carried out in the context of research into NMT. [188] shows that despite its general effectiveness, NMT does not provide a full solution to the challenges of WSD. From this result, we have an indication of how challenging it is to find a solution to the issues of *multi-word sense disambiguation* in MT. It is highlighted in [87] that WSD of **rare words** is especially difficult in NMT. The most recent work exploring word senses in NMT, e.g. with the transformer model includes [265].

SMT+MWE

We now introduce research work combining SMT and MWEs. The earliest work that combined MT with MWEs includes [152]. This applied bilingual MWE pairs to modify the word alignment procedure of MT to improve translation quality on an English-Spanish corpus. The modification function on alignment was achieved by grouping the MWEs as one token before training.

Further work includes [230] which integrated bilingual Chinese-English MWEs into the SMT toolkit Moses, [25] which designed models to extract continuous MWEs and integrated them into the Moses system for French-English translation, and [248] which discussed various MWEs in English-Latvian MT. Recent interesting work such as [69] focused on phrasal verb MWEs in Arabic-English phrase-based SMT.

NMT+MWE

This section introduces work on the incorporation of MWEs in NMT. MWEs can appear in different kinds of examples, such as **named entities (NEs)** [100] when the entities appear as a chunk of several words. In [167], the author applied a character-level sequence-to-sequence model to translate named entities and then integrated this into an overall NMT system on a Chinese-to-English task. This model was originally designed to solve the unseen word translation issue, but the results show that NEs in NMT help to improve overall translation effectiveness as measured by BLEU score. It showed the model can derive higher quality named entity alignment in the training corpus.

In recent work [270], researchers designed an additional chunk-level (for NE tags) LSTM layer on the encoder side after the word-level LSTM and embedding layer, trying to acquire knowledge of the NEs based on tagging. This model is inspired by the work of [127], who designed a chunk-based decoder, where the word-level decoder generates target words from the chunk decoder. As a basis for future work, it would be informative to see what the performance would look like if both chunk-level LSTM layers are applied to both encoder and decoder, and if it further improves the performance when integrating a chunk-level layer into the new state of the art transformer NMT structure. This could be further work to investigate.

Similar to the ideas from [230, 25], in [231], the authors showed how enhancing MWE knowledge by adding it into a corpus can improve NMT even with very simple integration. For example, they extracted bilingual MWEs in their corpus and added bilingual MWEs pairs and sentence pairs that included the MWEs into a parallel corpus to train a NMT system in English to Czech and English to Latvian MT. The authors developed an alignment visualisation tool to view the improvement in MWE alignment. The neural network platform they used is from Neural Monkey [120]. [7]

[7] Our first experimental investigation of MWEs in MT for pilot study will take this model's pipeline but with new language pairs English-Chinese and English-German, and new NMT structure.

3.3 Discussion: Revisiting MT Issues and Thesis Hypotheses

We know that MWEs form a bottleneck in advancing MT and NLP technologies. MWEs have been explored for many NLP tasks due to their idiosyncratic nature [235, 53]. As we noted earlier, MWEs can be presented in many different linguistic categories such as idioms, metaphor, customised phrases, named entities, (semi-/) fixed expressions, amongst others. For instance, *kick the bucket*, *apple of my eye*, for idiom and metaphor; *Washington, D.C.*, *National Academy of Science*, and *Donald Trump* for location, institutional or personal names respectively; are all fixed expressions. Knowledge of MWE boundaries and their meaning acquisition is very important for many NLP tasks, including MT and information extraction.

Secondly, as we discussed in the MT issues section (Section 2.3), there are many problems and challenges that still remain. These span from MT modelling (rare words, adequacy) to corpus (register, data sanity), language bias (resources), linguistic awareness (gender, MWEs), and translation quality evaluations (reliability and accuracy).

Thirdly, the investigation of MWEs in MT has a long history, although not much work has been done in this field and there is much scope for further exploration.

From the *MT modelling perspective*, the fluency level has been improved largely with Neural MT structures on high resource language pairs. However, how can the *adequacy level* of translation be further improved, especially for MWEs with idiomatic and metaphorical expressions? Many researchers have re-investigated work that claims to achieve parity of MT models with human translation and discussed the limitations of the "human parity" model on its test suite, domain, and evaluation methodologies, and reveal false claims [159, 108, 198].

From the *MT evaluation perspective*, firstly, well-established automatic metrics such as BLEU do not correlate well to professional translators' judgements, which has been verified by very recent research work such as [77]. Thus, there is a need

to create new and more reliable MTE methods that consider the semantic aspect of evaluations, such as a human in the loop workflow. Secondly, there is a big issue on the test set that has been widely used in WMT community. There are very limited domains in the current practice, such as News, or TED talks. However, as reflected in our earlier MT examples in Section 2.4, MWEs set bottlenecks in MT in the literature text and this huge challenge cannot be recognised adequately without making new efforts by introducing proper multilingual test suites that cover literature domains. Furthermore, better quality test suits also mean that there shall be as little as possible, or no errors existence; however, the current WMT shared task test suites have translation errors in their official reference translation text, which can be mis-leading during evaluation stage, and we will list such examples in our later section of the thesis. Such errors can be reduced by some strict quality control measurements. Finally, it will be helpful to have some linguistically annotated test set, e.g. with highlighted MWEs that we more difficult to translate, and such test set will be very handy for human-targeted evaluation methods.

Following our critique earlier with the aforementioned literature reviews (Section 2.1, 2.2 and 3.2) and discussions (Section 2.3 and 3.3), in this thesis, we aim at tackling the current challenges from both translation models and their evaluation (quality assessment) methods perspectives by looking more closely into MWEs.

This leads to our hypotheses and research questions which we investigate in this thesis as mentioned in Section 1.4.2, where the first part of hypotheses focuses on the possibility of improving and proposing new research methods of addressing MWE challenges in MT modelling; and the second part of hypotheses focuses on the capability of developing a new multilingual test suite that covers literature domain text and has MWE annotations, and a new human-in-the-loop evaluation method looking into MWEs based on the newly developed test suite.

Chapter 4

Experimental Investigations into Multi-Word Expressions in Machine Translation Modelling

This Chapter reports a series of experimental investigations into the use of MWEs in NMT. We re-examine state-of-the-art models that have been introduced in this field in recent years in the first section, Section 4.1, but using different language pairs and a new NMT structure which forms a stronger baseline. The section is designed as a learning section where we learn lessons from peer state-of-the-art work, leading to a discussion on the disadvantages and on how or where we can make any improvements. In Section 4.2 we propose our own methodology for tackling MWE issues in MT with a different approach (decomposition modelling), but with the first section's knowledge as a feature set in our own model to examine their interaction. Both sections include subsections on motivation, model design, experimental setup, and discussion. Finally, we provide a summary in Section 4.3.

Before moving to the first section, we provide a *working definition* of MWEs for our experimental investigation. Conventional researchers have defined MWEs from the perspectives of decompositionality and idiomaticity, as we mentioned in Section 1.4. Following this inspiration, from a multilingual setting, we define an

MWE as *a group of lexemes that are formed into a pattern to carry one concept or expression*. This pattern can be semi-fixed or fixed, and it can be in a presentation of continuous lexemes or discontinuous ones. For instance, the Chinese MWE "洗衣機 (xǐ yī jī)" has three character components "洗 (xǐ)", "衣 (yī)", and "機 (jī)", of which "洗 (xǐ)" can be an independent word meaning "wash (something)", "衣 (yī)" is from the word '衣服 (yī fú)'" meaning "cloth(es)", and "機 (jī)" is from the word "機器 (jī qì)" meaning "machine(s)". All together, these three Chinese characters form a Chinese MWE carrying the concept of a "washing machine" which is a corresponding noun phrase (verb-ing + noun) terminology in English. In German, this is translated into one word "Waschmaschine" which belongs to one kind of MWEs that was listed in Section 1.4 called compound nouns. In Section 4.2, we propose to address MWE translation using Chinese character decomposition models from the following motivations: a) the Chinese MWEs in text can appear as low-frequency words or phrases due to their unique patterns (lexemes combination) being used, b) character decomposition turns lower frequency Chinese MWEs into higher frequency sub-characters such that the MT models can acquire more knowledge from such words at the learning stage.

4.1 Pilot Study with the State-of-the-Art Models

As discussed in our earlier literature review when investigating the use of MWEs in MT, one recent work is from [231] where bilingual MWE terms are automatically extracted and added back into the training corpus to enhance the Neural MT modelling learning, in language pair settings of English to Czech and English to Latvian. Due to the simple concept of the methodology and its surprisingly good performance in their work, we re-investigate this methodology with an automated MWE pairs extraction and alignment procedure, but with higher-performance language pairs of German-English and Chinese-English. We then apply the extracted bilingual MWEs subsequently in the translation task, with a new state-of-the-art attention based NMT structure Transformer [273], to examiner whether this method

helps MT from high-performing language pairs.

4.1.1 Bilingual MWE (BiMWE) Extraction Models

Following the bilingual MWE extraction methodology proposed by [231], where the authors evaluated their model on English to Czech and English to Latvian languages, we re-examine this methodology with new language settings of German-to-English and Chinese-to-English where large amounts of resources are available.

English-to-German

The root parallel corpus that is used here is from the WMT2017 German-English MT training task[1]. This contains 5.8 million German-English sentences. To create a suitable bilingual MWE corpus we adopted the following procedure, shown in Figure 4.1.

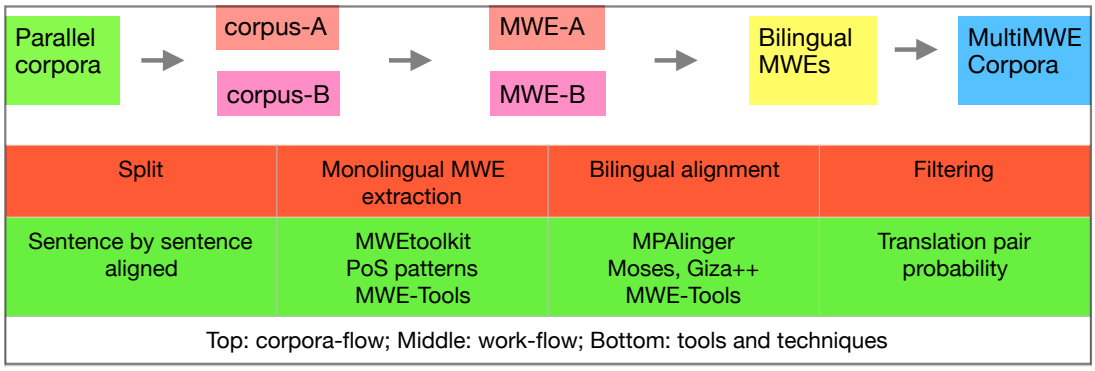

Figure 4.1: MultiMWE corpora extraction workflow.

- Morphological tagging of DE and EN.
- Tagged DE/EN into XML format.
- Design MWE-patterns for DE/EN
- Extract Monolingual MWEs with MWEtoolkit
- Generate DE-EN lexicon translation probability files with Giza++ and Moses

[1] http://data.statmt.org/wmt17/translation-task/preprocessed/

- Align Bilingual MWEs with MPAligner

Firstly, Treetagger[2] [240] was used to tag English and German sentences with morphological information, namely PoS and lemma information. The English and German morphological tag-sets we used were from the British National Corpus (BNC)[3] and STTS[4] corpora. Secondly, we performed a mapping of the English POS patterns for MWE extraction from PENN used in [231] to BNC. We designed the German POS patterns for MWE extraction and then the English and German monolingual candidate MWEs were extracted using MWEtoolkit [226] with the corresponding MWE patterns and the morphological corpus. Thirdly, the MWE-tools[5] from [231] were used to convert the extracted two monolingual candidate MWE files into MPAligner format. Fourthly, we ran word alignment tools Giza++ and SMT platform Moses to get the lexical translation probability files of German-English in both directions [209, 144]. The bilingual MWEs were aligned using MPAligner[6] developed by [219] with the corresponding translation estimation probability. Finally, we examine the use of extracted bilingual MWEs in NMT experiments. We choose a new state-of-the-art attention-based Transformer model for MT experiments, with the open source package THUMT from the Tsinghua University NLP group[7].

Chinese-to-English

To the best of our knowledge, there is no openly available bilingual MWE corpus of Chinese-English (ZH-EN), which means we needed to create our own. For a root parallel training corpora, we use the publicly available WMT2018 Chinese-English pre-processed (word segmented) data. However, due to computational limitations in the follow-up NMT experiment, we chose to use the first 5 million parallel sentences as a training set.

[2] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[3] http://www.natcorp.ox.ac.uk/docs/c5spec.html
[4] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/stts_guide.pdf
[5] https://github.com/M4t1ss/MWE-Tools
[6] https://github.com/pmarcis/mp-aligner
[7] http://thumt.thunlp.org

The monolingual MWE extraction and bilingual MWE alignment procedure was mostly the same with German-English, except for some additional processing that we list as following.

- POS pattern design.
- Stop-word list preparation.
- Chinese-English translation probability files.

MWEs can appear in Chinese in a range of ways, such as fixed (or semi-fixed) expressions, metaphor, idiomatic phrases, and institutional, personal or location names, amongst others. For the part of speech (POS) patterns that MWEtoolkit required for MWE extraction, we performed a POS pattern mapping from English to Chinese and added some other Chinese POS tags that apparently point to *MWEs*, as shown in Table 4.1. These include idioms, fixed expressions, personal names, place names and organisation names. The Chinese tagset is from the Lancaster Corpus of Mandarin Chinese (LCMC)[8].

i	idiom
l	fixed expressions
nr	personal name
ns	place name
nt	orgnisation name

Table 4.1: Added Chinese Patterns for MWEs from the LCMC Tags

In the bilingual MWE alignment step, MPaligner requires a stop word list for each language. We used some open-source packages to build a Chinese stop word file, including the lists from Chinese leading information technology (IT) company Baidu and NLP institutes in the Harbin Institute of Technology (HIT) and Sichuan University. These packages are open source and are available on Github [9]. We

[8] https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
[9] https://github.com/stopwords-iso and https://github.com/goto456/stopwords

removed duplicates when concatenating the word lists together resulting in 2,361 Chinese stop words in the merged file.

Similarly, we ran the Giza++ and Moses toolkits to get the Chinese-English lexicon translation probability files from both directions. These files were used for bilingual MWE alignment by MPaligner.

Bilingual MWE Filtering

We manually examined the extracted bilingual MWEs and found that the MPAligner aligned bilingual MWEs have large amounts of noise, especially for German-English pairs. The output for German-English bilingual MWEs contains many candidates that have very low translation probabilities between 0 and 0.5. For example, the English term 'European Commission' is aligned with German 'Europäische Kommission' with 0.97 translation score, while 'upcoming events' is also aligned to German 'Europäische Kommission' with 0.22 translation probability. This may be due to their co-occurrence and morphological patterns adj+noun, we refer to Figure 4.2 (the higher the estimation scores, the better extracted bilingual candidates) for examples. This also reflects one of the MT issues raised earlier in Section 2.3 where statistical models often fall into a linguistic blind situation. The extracted Chinese-English bilingual MWEs generally have higher translation probability above 0.5 and are of better quality as illustrated in Figure 4.3.

To filter out any low-quality bilingual MWE pairs, we chose to experiment with two thresholds, i.e. 0.70 and 0.85, respectively in our experiments. The initially extracted and subsequently aligned bilingual MWEs are 27,688,373 pairs and 172,900 pairs for German-English and Chinese-English respectively. The package we deployed in automatic processing is very focused on Western languages, e.g. the MPaligner and MWEtoolkit, and this leads to a large proportion of errors in the Chinese-English corpus processing. Thus the extracted pairs in ZH-EN are far smaller sized than in DE-EN. After pruning with an alignment threshold of 0.70 as shown in Figure 4.4 with samples and 0.85, the German-English MWEs moved to

German	English	Estimation
europäische Kommission	european commission	0.970964
upcoming events	european commission	0.071845
upcoming events	upcoming events	1
europäische Kommission	upcoming events	0.22279
lokaler Bürgerforen	hall meetings	0.294691
lokaler Bürgerforen	local fora	0.526792
anstehende Entscheidungen	local fora	0.131985
lokaler Bürgerforen	town hall	0.159491
lokaler Bürgerforen	town hall meetings	0.229983
anstehende Entscheidungen	town hall meetings	0.149191

Figure 4.2: Samples of MPAligner aligned bilingual MWEs (no filtering).

6,518,550 (23.5% of original size) and 3,159,226 (11.4%) pairs. The Chinese-English MWEs moved to 143,042 pairs (82.7% of original size) with alignment threshold 0.85.

From the examples in Figure 4.3 (The extracted pair samples here are from the head of the file, and have good quality.) and in Figure 4.4, we can see that the extracted MWEs include some non-decomposable ones, for instance, the Chinese MWE "簸箕" (bò ji) with two characters together meaning "dustpan", and "電腦" (diàn nǎo) [10] meaning "computer". However, if we split the two characters of any of them, it can not make the same meaning. The Chinese character "電 (diàn)" means "electricity", while "腦 (nǎo)" means "brain". So the combined character sequence "電腦 (diàn nǎo)" is a metaphor to describe "computer". For decomposable MWEs that we extracted, there are "european commission" and "european council" as institutional names in Figure 4.4.

[10]We use traditional Chinese characters overall in the section for consistency, also to solve the character encoding issues

Chinese	English	Estimation
猫 耳	cat ears	0.780979
长 尾巴	long tail	0.820427
小 簸箕	small dustpan	0.856796
艺术 作品	artistic works	0.6281
组 表	group table	0.708438
电脑 专家	computer expert	0.801311
高尔夫球 俱乐部	golf club	0.976473
痘 产品	acne products	0.695547
不同 条件	different conditions	0.887839
常青 植物	evergreen plant	0.610852
note: (电脑)->(電腦) simplified to traditional Chinese character, used in paper content		

Figure 4.3: Extracted Zh-En MWEs without pruning.

4.1.2 Experimental Examinations of MWEs in NMT

To verify the quality of our extracted bilingual MWEs, as one example, we apply them to NMT experiments following [231] as additional knowledge to influence NMT learning, but with these rich training set language settings. This is achieved by concatenating the extracted bilingual MWEs back to the original bilingual training corpus as additional "translation pairs". We call the learning model with the extracted MWEs added to training corpus 'MWE+Base' and call the model with filtered MWEs "MWEpruned (threshold) + Base", e.g. MWEpruned0.7+Base.

The baseline NMT model is a state-of-the-art transformer (THUMT-tensorflow) from [294]. This implements the bidirectional recurrent neural network (RNN) based NMT structures with attention mechanisms developed by the Google Brain team [273]. The sub-word unit translation BPE methodology [245] is applied for the improvement of rare word translation. As a standard setting, the BPE operations

German	English	Estimation
europäische Kommission	european commission	0.970964
upcoming events	upcoming events	1
europäischen Kommission	european commission	0.990844
praktische Informationen	practical information	0.948533
östlichen Teils	eastern part	0.793047
private Konzession	private concession	0.921197
französische Staat	french state	0.853861
europäischen Rat	european council	0.984224
größeren Infrastrukturprojekten	major infrastructure projects	0.853873
zwischengeschalteten Banken	intermediary banks	0.754617

Figure 4.4: Samples of DE-EN MWEs after pruning with threshold 0.70.

size is set to 32k for both German-English and Chinese-English training corpora. The vocabulary-threshold is set to 50, which means any word with frequency less than this threshold will be treated as an (out-of-vocabulary) OOV word. Training set shuffling is applied by randomly relocating the order of each sentence and batch size is set at 6,250. The encoder and decoder are set up with 7+7 layers which can be regarded as a relatively deeper setting in the field.

DE-EN MT

The training corpus for NMT is the same as used for MWE extraction, 5.8 million parallel German-English sentences. The development and testing corpora are 3,003 and 2,169 parallel sentences respectively. To examine the external German-English MWEs that are available, we also set up one experiment where we added the 871

external MWE pairs into the training corpus and this is the only available DE-EN MWE resource we found. We call this ExterMWE871+Base.

After the first 20k learning steps are applied, the evaluation scores are displayed in Table 4.2. This result shows that even though in most n-gram matching the Baseline achieved better scores, the overall BLEU score is lower than the MWE+NMT case. This is due to the Brevity-Penalty (BP) parameter and ratio factors, which give a penalty score if the system output is too short in comparison to the reference translation so that the systems that only produce shorter output for the sake of higher precision scores will receive penalty. The higher BP parameter and ratio value from the MWE+NMT model indicates that it produced more reference-like output than the Baseline model regarding sentence length.

It is unusual to observe that by adding 871 pairs of external DE-EN MWEs into the training set, the ExterMWE871+Base performance score is not higher than the baseline. The reasons could be: 1) due to the added MWEs being too small in size compared with the 5.8 million training set. 2) the external MWEs are kept as one (large) token instead of being split by the BPE model. 3) the external MWE pairs have many metaphor expressions, but such metaphors did not appear often in the training corpus, and can also mislead the learned model. 4) the integration of these MWE pairs indeed improves the DE-EN translation quality, but this is not covered and reflected by reference test suite, and further not reflected by BLEU scores.

	n-gram scores				Params		Combine
models	1-gram	2-gram	3-gram	4-gram	BP	ratio	overall
Baseline	**63.3**	**35.2**	**21.4**	**13.5**	0.942	0.944	26.73
MWEpruned0.7+Base	63.0	35.1	21.3	**13.5**	**0.952**	**0.953**	**26.87**
ExterMWE871+Base	**63.3**	**35.2**	21.2	13.3	0.929	0.932	26.15

Table 4.2: DE-2-EN NMT BLEU Scores with 20k Transformer Learning Steps

ZH-EN MT

For Chinese-English baseline NMT training, we also use the same corpora that were used for MWE extraction, 5 million parallel ZH-EN sentences. The development (newsdev2017) and testing (newstest2017) corpus for NMT were from WMT2017, 2002 and 2001 parallel sentences respectively.

In the evaluation score shown in Table 4.3, model MWEpruned0.85+Base means we pruned the extracted ZH-EN MWEs with a threshold value 0.85, then we used the original BPE operators to encode the pruned MWE pairs and concatenated it to the BPE encoded training set. We used the same vocabulary file from the baseline model. The result shows that the pruned MWE pairs enhanced the model learning by producing improved 3-gram and 4-gram BLEU scores and yielded an overall higher score. This automatic score means that the MWE enhanced model can generally improve the chunk translation, i.e., the MT output sentences include more chunks of 3-gram and 4-grams words that match the reference sentences. Most likely, they are improved MWE translations.

When we look more closely into the translation outputs from the baseline model and the MWE integrated model we found some Chinese MWEs that were not translated by Baseline and were translated properly by the MWEpruned0.85+Base model. Furthermore, some idiomatic MWEs that were translated literally by Baseline, were translated in a meaning-preserved way by MWEpruned0.85+Base. If we look at Figure 4.5 we see that in the first example, Chinese "口水戰 (kǒu shuǐ zhàn)" which means "war of words" was translated into "water fighting" by the Baseline, while it was translated into "oral combat" in a proper way by the MWE enhanced model. The Baseline translation is due to the fact that this is a metaphor expression in Chinese using "口水 + 戰 (kǒu shuǐ + zhàn)" which is a combination of "saliva" and "war".

In the second example, "所謂 朋友 (suǒ wèi péng yǒu)" which means "supposed friend" is translated as "friend" in the Baseline model, and this is lost in the Chinese MWE "所謂 (suǒ wèi)" which is used to express "supposed" or "so-called". The

	n-gram scores				Params		
models	1-gram	2-gram	3-gram	4-gram	BP	Ratio	overall
Baseline	**56.3**	**26.5**	14.3	8.2	**0.9**	**0.905**	18.39
MWE+Base	55.9	26.1	14.3	8.2	0.884	0.89	17.99
MWEpruned0.85+Base	55.9	26.3	**14.5**	**8.4**	0.899	0.903	**18.49**

Table 4.3: Zh-2-En NMT BLEU scores with 20k Transformer learning steps.

MT output yielded a correct translation when we integrated the extracted bilingual MWEs back into the training corpus to enhance the learning.

However, both the two example sentences in Figure 4.5 show that even though the MWE enhanced model produced better MWE translations, the BLEU scores of these two sentences do not improve correspondingly. The reason is that "oral combat" can not match the reference "war of words" in the word surface form as used by the BLEU metric and "so-called" can not match the reference "supposed" either. This reflects the possible explanation that we discussed previously in the DE-EN results (reason 4). This also presents proof of the test suite issues existing in current practice in MT, as we pointed out in the MT problems and challenges section (2.3 and 3.3).

4.1.3 Discussion

In this pilot study, we re-examined one of the state-of-the-art MWE integration models in MT with a different setting of language pairs. We find that even though this methodology gave a boost to performance in English-Latvian and English-Czech NMT by [231], it does not produce much higher scores in BLEU metric with a rich language resource setting, i.e. German-English and Chinese-English.

Furthermore, in the current experiments, the German and Chinese PoS patterns for extracting MWEs are mapped from the English PoS tagset, via meaning equivalent alignment. It could be more reasonable to design German and Chinese patterns specifically for these languages and conduct some linguistic knowledge survey for this.

	Examples of MWE translations in MT outputs
Src	俄罗斯 与 土耳其 领导人 周二 进行 会见 ， 双方 握手 并 宣布 正式 结束 长 达 八个 月 的 口水战 与 经济制裁 。
Ref	the leaders of Russia and Turkey met on Tuesday to shake hands and declare a formal end to an eight - month long war of words and economic sanctions .
Base	Russian and Turkish leaders met Tuesday , shaking hands and declaring the official end of eight months of water fighting and economic sanctions .
B+MWE	Russian and Turkish leaders met on Tuesday and both shook hands and announced a formal end of eight months of oral combat and economic sanctions .
Src	来自 所谓 朋友 的 攻击 更让人 难以 接受
Ref	the offence was even greater , coming from a supposed friend .
Base	attacks from a friend are even harder to accept .
B+MWE	the attack from so-called friends is harder to accept .

Src: source; Ref: reference. B+MWE: Baseline+MWE. Simplified Chinese (战, 谓) mapping into Traditional (戰, 謂), used in paper.

Figure 4.5: Zh-En MT Examples: Baseline vs Base+MWE

Last, when we used MWEtoolkit for Chinese monolingual MWE candidate extraction there were some issues with the toolkit for this language, or ideogram script in general, which meant we had to drop out some parts of the morphological-tagged corpus. This reduced the potential MWE numbers that can be produced by this procedure. We need to further examine this issue and fix the toolkit for the Chinese language or find alternative packages. This will further extend our MWE corpus size for the Chinese-English pair.

4.1.4 Summary

There are some contributions from our work that we list below. Firstly, we presented bilingual MWE corpora for German-English and Chinese-English, two typologically

different languages, which we call *MultiMWE-corpora*. As we discussed previously, the only available bilingual MWE resources available before our work is the DE-EN one with around 800 parallel sentences, while it is zero for ZH-EN case. MultiMWE covers 3,159,226 and 143,042 pairs of German-English and Chinese-English bilingual MWE entries after filtering (with higher quality). These corpora are freely available, and their size is larger than the currently available bilingual MWE corpus. However, this current extraction procedure only generates *continuous* MWEs. In the future, we will design patterns to extract *discontinuous* MWEs or develop new extraction models.

The NMT experiments for German-English and Chinese-English showed one example of the use of extracted bilingual MWEs, where they improved the automated translation evaluation scores slightly by the BLEU metric in quantitative analysis, and assisted better MWE translations in qualitative analysis. According to the BLEU metric, the results are different from one language pair to another. In future work, we will explore more automated metrics and diverse test sets with multiple references that can conduct better meaning-equivalent evaluation such as our own evaluation methods hLEPOR [102, 106], and further investigate the translation output in more detail, such as human in the loop evaluations and looking at MWE translations in general.

We make our extracted bilingual and multilingual MWE corpora openly available. We believe that the MultiMWE corpora can be helpful for other multilingual NLP research tasks such as multi-lingual Information Extraction (IE), Question Answering (QA), and Information Retrieval (IR). For instance, those multi-lingual / cross-lingual tasks can take MultiMWE corpora as external dictionaries/knowledge into their models.

We also plan to extend the MultiMWE corpora to other language pairs, including similar and distant languages, such as Russian-Japanese, English-French, etc. We will use the popular corpus Europarl [11] for this purpose.

[11] https://www.statmt.org/europarl/

Looking back to our hypotheses and research questions in Section 1.4.2, this pilot study verified the opening part of the hypothesis statement namely "MWEs set challenges for MT and investigation into MWEs can help improve MT from translation modelling". In addition we discovered that the improvement varies from language to language (DE-EN vs. ZH-EN). This is an exact examination of the first task in RQ-I. However, this current treatment used a unified MWE extraction pipeline and integration into MT models, without considering language specific knowledge except for the POS pattern design section.

In the next stage of our work, we continue to investigate the first two sub-hypotheses, Hy.a: "current methodology addressing MWEs in MT can be improved to include language-specific characteristics and features (from linguistic point of view)" and Hy.b "(in brief) the low frequency of each MWE can be treated as a rare-words/phrase translation issue (from technical point of view)". To do these, we continue to follow up our RQ-I tasks (Task-II and Task-III), i.e., we focus on the Chinese-English language pair and design new methods to investigate MWEs in MT from an in-depth view of the characteristics of the Chinese language and carry out both quantitative and qualitative validations.

4.2 Chinese Character Decomposition Models Addressing MWEs

In this section we introduce methodologies for tackling MWEs in NMT in Chinese. We firstly introduce the motivation for our work where we re-address the issues and challenges, and introduce some background Chinese character knowledge. Then, we present a literature review of related work in Chinese character decomposition for NLP tasks. Subsequently, we propose our own models and split them into two step experimental investigations. This will be followed by a discussion. This section is based on our published work [111] and [110]

4.2.1 Motivation for this Work

We now introduce the motivation for this work from an MWE and MT perspective, then we introduce relevant background and useful knowledge to help the reader understand Chinese characters, their construction and evolution.

Critique of MWE and MT Literature

Despite Neural Machine Translation (NMT) having recently replaced Statistical Machine Translation (SMT) as the state-of-the-art, research questions still remain, such as how to deal with out-of-vocabulary (OOV) words, how best to integrate linguistic knowledge and how best to correctly translate multi-word expressions (MWEs) [235]. The integration of MWEs into MT generally falls into two categories, firstly combining extracted bilingual MWEs into a training corpus to enhance model learning as additional features, or using MWEs as a linguistically motivated component to change the machine learning structure itself. Of the two past approaches, the first category has been more popular. As we discussed earlier in the literature review on MWE investigation in MT (Section 3.2) and in our pilot study, [230, 25, 231, 109] designed a range of methods to extract bilingual MWEs for Chinese ⇔ English, French ⇔ English, and English → Czech/Latvian translations respectively for statistical as well as neural MT. [152], on the other hand, grouped MWEs as one token prior to SMT training, while [167] added one chunk layer to neural networks focusing on named entities above the word sequence layers in order to guide the NMT learning representation.

However, none of the previous work has investigated the possibility of borrowing methodologies used for tackling *OOV* words to address MWEs in NMT. Some idiomatic and odd-patterned (e.g. POS syntax) MWEs can be treated as *low-frequency* word and phrase translation from the technical side due to their low frequency of occurrence.

For low frequency and OOV word translations for European languages, substantial improvements have been made in terms of rare and unseen words by in-

corporating *sub-word knowledge* using Byte Pair Encoding (BPE) [245]. However, such methods cannot be directly applied to Chinese, Japanese or other ideographic languages.

We design a new methodology to tackle MWE translations as a rare words and phrase translation by decomposing Chinese characters into different degrees gradually and investigate the performance of Chinese-English Neural MT. To address this, for the Chinese-English language pair, we firstly extract Chinese radicals as linguistic features and integrate these into the training set to enhance NMT models as a first step experiment. Then, we investigate the different decomposition levels of Chinese characters and their corresponding performance in NMT, and look at MWE translation quality using quantitative and qualitative analysis. In addition to this, we apply the bilingual MWE pairs extracted from the training set using the methods from the pilot study section as a knowledge base feature set in our new models in the second step.

Chinese Character Knowledge in a Nutshell

Chinese characters often include two pieces of information, with semantics encoded within radicals and a phonetic part. The phonetic part is related to the pronunciation of the overall character, either the same or similar. Figure 4.6 demonstrates the meaning-preservation root of Chinese radicals, where the evolution of the Chinese character 刀 (Dāo), meaning *knife*, evolved from bronze inscription form to contemporary character and radical form, 刂 (named as: tí dāo páng). For instance, Chinese characters with this two-stroke radical, 刂 (tí dāo páng), ordinarily relate to *knife* in meaning, such as the Chinese character 劍 (jiàn, *sword*) and multi-character expression 鋒利 (fēng lì, *sharp*). The radical 刂 (tí dāo páng) preserves the meaning of knife because it is a variation of a drawing of a knife evolving from the original bronze inscription.

Not only can the radical part of a character be decomposed into smaller fragments of stokes but the phonetic part can also be decomposed. Thus there are

Chinese radical 刂 (Dāo, knife) evolution from Pictogram to Regular script					
商 Shang Dynasty (1600-1046BC)		西周 Western-Zhou Dynasty (1045-771BC)	戰國 Warring States period (476-221BC)	漢 Han Dynasty (202BC-220)	東漢 Eastern Han (from 57AD on)
Bronze inscriptions	Oracle bone script	Bronze Inscription	Silk	篆 (on Seal)	Regular script
					

Figure 4.6: Radical 刂 (Dāo): (leftmost) pictogram to regular script (rightmost) containing only two strokes.

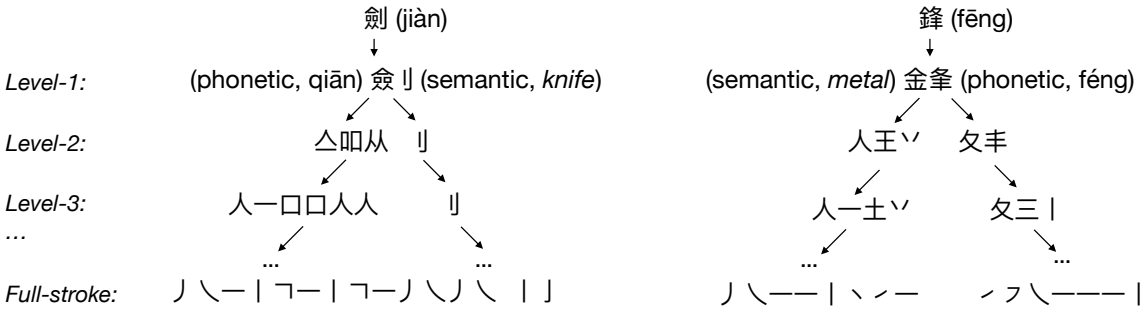

Figure 4.7: Examples of the decomposition of Chinese characters.

often several levels of decomposition that can be applied to Chinese characters by combining different levels of decomposition of each part of the Chinese character. As an example, Figure 4.7 shows the three decomposition levels from our model and the full stroke form of the above-mentioned characters 劍 (jiàn) and 鋒 (fēng). To date, little work has been carried out to investigate the full potential of these alternative levels of decomposition of Chinese characters for the purpose of Machine Translation (MT) and the specific role of MWEs.

4.2.2 Chinese Character Decomposition for NLP

In this section we introduce some related work on Chinese character decomposition investigations for MT and more broadly for other NLP applications. We first introduce Chinese NMT with the consideration of different input units for sequence learning. Then we introduce the Chinese character decomposition work in other related NLP tasks.

Character Decomposition for MT

Neural MT for Asian languages has included translation at the level of phrase, word, and character sequences (see Figure 4.8) where the English reference here is taken from the corpus used for our experiments and pronunciation is Mandarin in Pinyin. Radical knowledge embeddings have been applied as features in both word and character-based neural architectures [205, 111, 297, 296]. using the mapping from Vietnamese to Chinese characters, [205] built a character sequence based NMT system for Vietnamese ⇔ Japanese Kanji with a transformer model [273]. It should be noted that Japanese Kanji is composed of traditional Chinese characters adopted from ancient China and recalled as Kanji, close to Hanzi in pronunciation, meaning characters from the Chinese Han dynasty. This work used Vietnamese → Chinese character alignment knowledge to perform Vietnamese ⇔ Japanese translation. Their models employed character sequences to replace word level information for which experimental results showed improvement in terms of BLEU scores.

Word level	28 / 歲 / 廚師 / 被 / 發現 / 死 / 於 / 舊金山 / 一家 / 商場
Character	28 歲 廚 師 被 發 現 死 於 舊 金 山 一 家 商 場
Pronunciation	èr shí bā Suì chú shī bèi fā xiàn sǐ yú jiù jīn shān yī jiā shāng chǎng
Radical	28 止戌少 广尌 自帀 礻 皮 歹殁 王見 歹匕 方仑 雀臼 人王丷 山 一 宀豕 亠丷冏 土昜
English Ref.	28-Year-Old Chef Found Dead at San Francisco Mall

Figure 4.8: Chinese word to character level MT.

Going deeper than character level units, [111] and [297], considered radical embeddings as additional features for Chinese → English and Japanese ⇔ Chinese NMT.

Furthermore, [296] applied radical embeddings as additional features to character level LSTM-based NMT on Japanese → Chinese translation. None of the aforementioned work has however investigated the performance of decomposed character sequences and the effects of varied decomposition degrees in combination with MWEs.

Subsequently, [297] developed bi-directional English ⇔ Japanese, English ⇔ Chi-

nese and Chinese ⇔ Japanese NMT with word, character, ideograph (the phonetics and semantics parts of character are separated) and stroke levels, with experiments showing that the *ideograph* level was best for ZH→EN MT, while the stroke level was best for JP→EN MT. Although their ideograph and stroke level setting finally removed the original character and word sequences, there was no investigation of intermediate decomposition performance.

Character Decomposition for Broader NLP

Chinese character decomposition has been investigated in previous work for other NLP tasks besides MT. For instance, [247] examined incorporation of radical embeddings into short-text categorisation, Chinese word segmentation, and web search ranking. The authors suggest that radical level decomposition should be the deepest level of analysis in Chinese NLP, since it is difficult to model semantic preservation via deeper Chinese strokes. Results of our experiments however suggest that even deeper Chinese character decomposition to level-3, including strokes, can also achieve good results with the model still able to acquire the meaning of original Chinese word sequences, as shown in Section 4.2.3.

Radical knowledge has also been applied to Chinese word similarity, Chinese named entity recognition, and parsing and semantic role labelling [64, 292, 36, 288]. For instance, [64] concatenated bi-directional LSTM learned radical embeddings and character embeddings together before learning using conditional random fields (CRFs). Their model achieved better performance on the same task than previous CRF models. [292] explored the joint embedding of Chinese words, characters and sub-character components (radicals) in word similarity prediction and word analogy reasoning. In the word analogy task, for instance, the nearest vector representation to vec(意大利, Italy) − vec(罗马, Rome) + vec(柏林, Berlin) is expected to be vec(德国, Germany). Their joint model largely outperformed the baseline word sequence model on the word analogy task.

[288] added glyph-vectors of Chinese character representations into Neural Net-

work learning for improved tagging and labelling.

In [217], sentence-level sentiment analysis with radical embeddings were carried out. One of the four levels of decomposition, full radical embeddings, achieved best performance with a convolutional neural network (CNN) when compared to other traditional classification models such as Naive Bayes (NB). The authors carried out a matching comparison between English and Chinese in composition, in which they aligned English characters with Chinese radicals. However, we believe a better matching for English characters might be the Chinese strokes while the English stems could better match Chinese radicals. The reason for this is that the English characters and the Chinese strokes do not carry any specific meaning independently in general unless we group them together but the English stems and the Chinese radicals carry the meaning of words that they are part of, such as the English stem "trans" indicating some knowledge (across) from the word "transparent", and the Chinese radical "艹" (cǎo zì tóu) indicating some knowledge (grass) from the character "草" (Cǎo, meaning grass). The suggested mapping composition between English and Chinese is shown in Figure 4.9.

English		Chinese		
Hierarchy	Example	Hierarchy	Coresponding	Example
Character	a, b, c, ...	Stroke	筆畫 (bǐhuà)	丶, 一, 丨, 乛, 亅, 丿, 𠂉, 乀, 乚, 𠃌, ...
Stem	trans	Radical	部首 (bùshǒu)	艸, ⺾ (grass), 冖 (lid), 亻(person), 刂 (knife)
Word	transparent	Char/Word	字詞 (zì cí)	草 (grass), 家 (home), 他 (he), 鋒利 (sharp)
Phrase / Multi-word expression	government transparency, neural machine translation	Multi-character/ word expression	組合詞/固定搭配 (zǔhécí / gùdìng dāpèi)	草藥店 (herbal medicine shop), 家庭 (family), 去 ... 了 (went to somewhere), 唯... 是瞻 (only look at ...)

Figure 4.9: A mapping between English and Chinese in composition

Similarly, [134] applied radical embeddings with CNNs and recurrent neural networks (RNNs) for sentiment analysis of Chinese and Japanese languages. Their radical embeddings model achieved similar performance compared to character and word embeddings, but with the advantage of a much smaller vocabulary.

4.2.3 Two-Step Investigations

In the following sections, we introduce the two-step experimental investigations on Chinese character decomposition for MT, expecting that the decomposed methods will address the rare words and OOV words issues and improve low frequency MWE translations. In the first step, we mostly present a quantitative evaluation on the experimental results using the state-of-the-art automatic evaluation metrics hLEPOR, BEER, CharacTER, in addition to BLEU, METEOR, and NIST. The experiments use a sequence-to-sequence Neural MT modelling as a baseline. In the second step, we deploy another state-of-the-art attention based Neural MT model as a new baseline. We carry out both quantitative evaluation with automatic metrics, and qualitative analysis with crowd-sourced human evaluation. Furthermore, in light of the very recent work within this field criticising the incorrectness of crowd-sourced works [77], we validate our model performances with MT experts and native speakers of Chinese.

4.2.4 Model I: Chinese Radicals as NMT Features

For further Chinese radical interpretations, we examine two more examples to illustrate their construction in terms of corresponding characters. One of them can be used as an independent character while the other can only be part of a character (like the two stroke radical for knife). Figure 4.10 shows three Chinese characters (forest, tree, bridge) which contain the same part of the radical (wood) and this radical can be a character used independently. From history we know that a Chinese bridge was built by wood usually, so apparently these three characters carry similar meaning in that they all contain something related to wood.

Figure 4.11 shows three Chinese characters (grass, medicine, tea) which contain the same part of a radical (grass). However, this radical cannot be a character independently in use like a "bound morpheme". This radical means grass in the original development of the Chinese language. In history, Chinese medicine was mostly developed from naturally occurring things like grass and different plants,

Figure 4.10: Example of a Chinese radical as an independent character

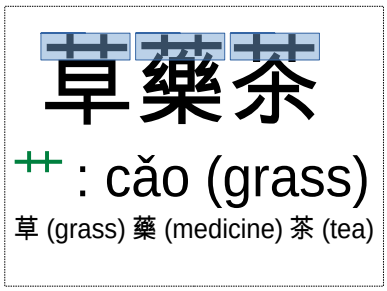

Figure 4.11: Example of a Chinese radical as a non-independent character

and Chinese tea was usually made from the leaves that are related to grass. To the best of our knowledge, our radical integration into NMT for Chinese language was one of the earliest works on this topic [111].

In this first step experiment, we applied one of the earliest encoder-decoder framework NMT models [9, 263] based on recurrent neural networks (RNNs). The encoder apples a bi-directional RNN to encode a source sentence $x = (x_1, x_2, ..., x_{T_x})$ and repeatedly generates the hidden vectors $h = (h_1, h_2, ..., h_{T_x})$ over the source sentence, where T_x is the length of the source sentence. Formally, $h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}]$ is the concatenation of forward RNN hidden state $\overrightarrow{h_j}$ and backward RNN hidden state $\overleftarrow{h_j}$.

The decoder is also an RNN that predicts the next word y_t given the context vector c_t representing the vector of the source sentence, the hidden state of the decoder s_t is the state of the decoder RNN at time step t and the previous predicted word y_{t-1} is the target word.

Traditional NMT models usually use the word-level or character-level information as the inputs of the encoder, which ignores some knowledge of the source sentence, especially for the Chinese language. As mentioned earlier, Chinese words are

usually composed of multiple characters, and characters can be further split into radicals. In this first step experiment, we use the radical, character and word as multiple inputs of NMT and we expect NMT model can learn more useful features based on the different levels of input integration.

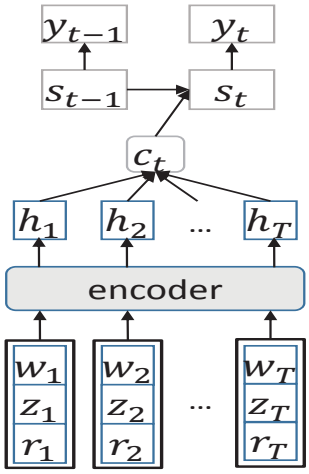

Figure 4.12: Architecture of NMT with multi-embedding.

Figure 4.12 [111] illustrates our designed model. The input embedding x_j consists of three parts: Chinese word embedding w_j, character embedding z_j and radical embedding r_j. We use the character 'z' to represent character, instead of 'c', because we already used 'c' as representation of context vector. The formula is as follows:

$$x_j = [w_j; z_j; r_j] \tag{4.1}$$

where ';' is concatenate operation.

For the word w_j, this can be split into characters $z_j = (z_{j1}, z_{j2}, ..., z_{jm})$ and further split into radicals $r_j = (r_{j1}, r_{j2}, ..., r_{jn})$. In our model, we used simple addition operations to determine the character representation and radical representation of the word, i.e. z_j and r_j can be computed as follows:

$$z_j = \sum_{k=1}^{m} z_{jk} \tag{4.2}$$

$$r_j = \sum_{k=1}^{n} r_{jk} \tag{4.3}$$

Each word can be decomposed into different numbers of characters and radicals, and, by addition operations, we can generate a fixed length representation. In principle our model can handle different levels of input from their combinations. For Chinese character decomposition, i.e. the generation of radicals, we used the HanziJS open source toolkit [12]. For the second step investigation in Section 4.2.6 we will introduce an alternative method for character decomposition. On the usage of target vocabulary [128], we chose 30,000 as the volume size.

This experiment was carried out by deploying the SuDaNMT system developed by the Soochow NLP lab [150] which followed the implementation of attention-based NMT of dl4mt tutorial.[13] Some further information used in the SuDaNMT system settings are listed here. Both the encoder and decoder had 1,000 hidden units. The encoder consisted of a forward (1,000 hidden units) and backward bi-directional RNN. The word embedding dimension was set as 620. The system incorporated a dropout [121] strategy on the output layer. It also used the stochastic descent algorithm with mini-batch and Adadelta [293] to train the model. The parameters ρ and ϵ of Adadelta were set to 0.95 and 10^{-6} respectively. Once the NMT model is trained, beam search was adopted to find possible translations with high probabilities (beam width: 10). The model parameters were selected according to the maximum BLEU score points on the development set. More detailed implementation of this NMT structure can be found in our *Appendix*.

4.2.5 Model I: Experimental Investigations

Experimental Setup

We used 1.25 million parallel Chinese-English sentences for training, which contain 80.9 million Chinese words and 86.4 million English words. The data was mainly from the Linguistic Data Consortium (LDC) parallel corpora, such as LDC2002E18, LDC2003E07, LDC2003E14, LDC2004T07, LDC2004T08, and LDC2005T06. We

[12] HanziJS available at github.com/nieldlr/Hanzi

[13] This is available at github.com/nyu-dl/dl4mt-tutorial/tree/master/session2.

tuned the models with NIST06 as development data using the BLEU metric [213], and used the NIST08 Chinese-English parallel corpus as testing data with four references.

In order to effectively train the baseline model, we limited the maximum sentence length on both source and target sides to 50. We also limited both the source and target vocabularies to the most frequent 30k words and replaced rare words with a special token "UNK" in Chinese and English. In our second step experiments, we keep the rare words as a different strategy. The vocabularies covered approximately 97.7% and 99.3% of the two corpora, respectively.

For our proposed model, all the experimental settings are the same as RNNSearch*, except for the word-embedding dimension and the size of the vocabularies. In our model, we set the word, character and radical to have the same dimensions, each being 620. The vocabulary sizes of word, character and radical are set to 30k, 2.5k and 1k respectively.

To integrate the character radicals into an NMT system, we designed several different settings as demonstrated in Table 4.4 . Both the baseline and our settings used the attention-based NMT structure.

Settings	Description	Abbreviation
Baseline	Words	W
Setting1	Word+Character+Radical	W+C+R
Setting2	Word+Character	W+C
Setting3	Word+Radical	W+R
Setting4	Character+Radical	C+R

Table 4.4: Model Settings for Radical as MT Features

Automatic Evaluation Metrics

Firstly, there are many works reflecting the insufficiency of the BLEU metric which we have covered earlier, and in light of such work, we try to validate our work in a deeper and broader evaluation setting and from more aspects. We use a wide range

of state-of-the-art MT evaluation metrics, which have been developed in recent years, to do a more comprehensive evaluation, including hLEPOR [114, 106], CharacTER [280], BEER [259] in addition to BLEU and NIST [213].

The hLEPOR metric is augmented with multiple factors including sentence length, precision, recall, positional difference, and word n-grams. CharacTER is a character level editing distance rate metric. BEER uses permutation trees and character n-grams integrating many features such as paraphrase and syntax. They have shown top performances in recent years' WMT[14] shared tasks [179, 259, 89, 24]. Furthermore, both CharacTER and BEER metrics achieved the parallel top performance in correlation scores with human judgements on Chinese-to-English MT evaluation in WMT17 shared tasks [23]. The hLEPOR metric series are evaluated by MT researchers as one of the best-performing metrics that are not significantly outperformed by others in the metrics comparison work by [89] using WMT data.

Evaluations on the Development Set

On the development set NIST06, we obtained the evaluation scores shown in Table 4.5 and Table 4.6. The cumulative N-gram scoring of BLEU and NIST metric, with bold case as the highlight of the winner in each n-gram column situation, is shown in the table respectively. Researchers usually report their 4-gram BLEU while 5-gram NIST metric scores, so we also followed this tradition.

	1-gram	2-gram	3-gram	4-gram
Baseline	0.7211	0.5663	.4480	0.3556
Setting1: W+C+R	**0.7420**	**0.5783**	**0.4534**	**0.3562**
Setting2: W+C	0.7362	0.5762	0.4524	0.3555
Setting3: W+R	0.7346	0.5730	0.4491	0.3529
Setting4: C+R	0.7089	0.5415	0.4164	0.3219

Table 4.5: BLEU Scores on the NIST06 Development Data

[14] http://www.statmt.org/wmt17/metrics-task.html

	1-gram	2-gram	3-gram	4-gram	5-gram
Baseline	5.8467	7.7916	8.3381	8.4796	8.5289
Setting1: W+C+R	**6.0047**	**7.9942**	**8.5473**	**8.6875**	**8.7346**
Setting2: W+C	5.9531	7.9438	8.5127	8.6526	8.6984
Setting3: W+R	5.9372	7.9021	8.4573	8.5950	8.6432
Setting4: C+R	5.6385	7.4379	7.9401	8.0662	8.1082

Table 4.6: NIST Scores on the NIST06 Development Data

From the automatic scoring results, we can see that the model setting one, i.e. W+C+R, was the best of the baseline models in all categories from uni-gram to 4-gram BLEU and to 5-gram NIST scores. Furthermore we can see that by adding characters and/or radicals to the words, the model settings two and three also outperformed the baseline models. However, setting 4 that only used character and radical information in the model lost both BLEU and NIST scores compared with the word-level baseline. This means that for Chinese NMT the word segmentation knowledge is important to show some guiding in learning for Chinese translation models.

For uni-gram BLEU score, our Model one gets 2.1 higher score than the baseline model which means that by combining W+C+R, the model can yield higher adequacy level translation, though the fluency score (4-gram) does not show much difference. This is exactly the point that we want to improve about neural models, an issue highlighted by many researchers as discussed in Sections 2.1 and 2.3).

The evaluation scores with broader state-of-the-art metrics are shown in Table 4.2.5. Since CharacTER is an edit distance based metric, the lower score means better translation result.

From the broader evaluation metrics, we can see that our designed models also outperformed the baseline system in all the metrics. Our model setting one, i.e. the W+C+R model, was better using both BEER and CharacTER scores, while our model two, i.e. the W+C, was better with the hLEPOR metric score, though setting four continues to be the poorest performance, which is consistent with the BLEU

	Metrics on Single Reference		
Models	hLEPOR	BEER	CharacTER
Baseline	0.5890	0.5112	0.9225
Setting1: W+C+R	0.5972	**0.5167**	**0.9169**
Setting2: W+C	**0.5988**	0.5164	0.9779
Setting3: W+R	0.5942	0.5146	0.9568
Setting4: C+R	0.5779	0.4998	1.336

Table 4.7: Broader Metric Scores on the NIST06 Development Data

and NIST metrics. Interestingly, we find that the CharacTER score of settings two and three are both worse than the baseline, which means that by adding character and radical information separately the output translation needs more editing effort. However, if we add both the character and radical information into the model, i.e. setting one, then the editing effort required became less than the baseline.

Evaluation on Test Sets

The evaluation results on the NIST08 Chinese-to-English test data are presented in this section. Firstly, we show the evaluation scores on BLEU and NIST metrics, with four reference translations and a case-insensitive setting. The tables show the cumulative N-gram scores of BLEU and NIST, with bold case as the best-performing of each n-gram situation in each column.

	1-gram	2-gram	3-gram	4-gram
Baseline	0.6451	0.4732	0.3508	0.2630
Setting1: W+C+R	**0.6609**	**0.4839**	**0.3572**	**0.2655**
Setting2: W+C	0.6391	0.4663	0.3412	0.2527
Setting3: W+R	0.6474	0.4736	0.3503	0.2607
Setting4: C+R	0.6378	0.4573	0.3296	0.2410

Table 4.8: BLEU Scores on NIST08 Test Data

The results show that our model setting one was best performing for both BLEU and NIST scores on each n-gram evaluation scheme, while model setting three,

	1-gram	2-gram	3-gram	4-gram	5-gram
Baseline	5.1288	6.6648	7.0387	7.1149	7.1387
Setting1: W+C+R	**5.2858**	**6.8689**	**7.2520**	**7.3308**	**7.3535**
Setting2: W+C	5.0850	6.5977	6.9552	7.0250	7.0467
Setting3: W+R	5.1122	6.6509	7.0289	7.1062	7.1291
Setting4: C+R	5.0140	6.4731	6.8187	6.8873	6.9063

Table 4.9: NIST Scores on NIST08 Test Data

i.e. the W+R model, was best performing for the uni-gram and bi-gram BLEU scores, and obtained a very close score to the baseline model in the NIST metric. Furthermore, model setting four, i.e. the C+R one, continued showing the worst performance, which verifies again that word segmentation information and word boundaries are indeed helpful to Chinese translation models.

What is worth mentioning here is that the detailed evaluation scores from BLEU reflect that our Model one yields higher BLEU score (+1.58) on uni-gram, similar with the results on development data, while a little bit higher performance on 4-gram (+0.25). These mean that in the fluency level our translation is similar to the state-of-the-art baseline. However, our model yields much better adequacy level translation in NMT since uni-gram BLEU reflects the adequacy aspect. This verifies the value of our model in the original problem we want to address.

The evaluation results on recent years' advanced metrics are shown below. The scores are also evaluated on the four references scheme. We calculate the average score of each metric from 4 references as the final evaluation score. Bold case means the best performing system, as usual.

From the broader evaluations we can see that our model setting one was best performing for both the LEPOR and BEER metrics. Though the baseline model was best performing, the CharacTER metric, the margin between the two scores from baseline (0.9846) and our model three, i.e. W+R, (0.9882) is very small (around 0.0036). The setting four with C+R performed the worst and is consistent with our previous findings.

	Metrics Evaluated on 4-references		
Models	hLEPOR	BEER	CharacTER
Baseline	0.5519	0.4748	**0.9846**
Setting1: W+C+R	**0.5530**	**0.4778**	1.3514
Setting2: W+C	0.5444	0.4712	1.1416
Setting3: W+R	0.5458	0.4717	0.9882
Setting4: C+R	0.5353	0.4634	1.1888

Table 4.10: Broader Metrics Scores on NIST08 Test Data

Discussion

We investigated the performance of the multiple model settings by integrating Chinese characters and especially radicals into what were in 2017, state-of-the-art attention-based BiRNN NMT systems, which we find can be helpful information for other researchers to look inside and gain general clues about how the radical works. This reflected our hypotheses in MT model learning and our RQ-I, as well as an effort of integrating language-specific linguistic knowledge and features into MT.

Our models showed that only using character+radical is not sufficient or suitable for Chinese language translation, which is a different finding to the work on Western languages such as [48]. Our model results showed that the word segmentation and word boundaries, which probably also includes the boundary of MWEs, are helpful knowledge for Chinese translation systems. In light of this, our step-II investigation in the next section will keep word segmentation i.e. word borders, in our processing as shown in Section 4.2.6.

Although the combination of words, characters and radicals mostly yielded the best performance scores, the detailed evaluations also showed that the model setting W+R, i.e. using both words and radicals, was generally better than the model setting W+C, i.e. words plus characters without radicals, which verified the value of this work by exploring radicals in Chinese NMT. Our Models yielded much better adequacy level translation output according to uni-gram BLEU scores compared to

the baseline systems, which showed that this work is important in exploring how to improve the adequacy aspect of neural MT models.

Even though our model settings perform well in both the traditional BLEU and NIST metrics, in recent years we have seen the development of advanced metrics which showed some differences and interesting phenomena, especially at the character-level translation error rate metric CharacTER. This encourages MT researchers to use state-of-the-art metrics to find potentially useful insights into their models. We will also include human assessment as a more credible evaluation method in the next set of experiments.

In this work, we aimed at exploring the effectiveness of Chinese radicals, so we did not use the BPE technique for English side splitting. However, to achieve a new state-of-the-art Chinese-English translation, in the next section, we will apply the splitting on both the Chinese and English sides.

In this experiment, rare words are treated by replacing them with a special token "UNK" in Chinese and English, and in our next investigation we will keep these rare words and OOV words for a different solution.

It will be interesting to see if fully decomposed Chinese characters with original word boundary information included can represent the original word and character sequence in meaning preservation, if the different decomposition levels of Chinese characters (as shown earlier in Figure 4.7 in Section 4.2.1) can have different performance regarding the translation of MWEs and context in general, and if bilingual MWE terms extracted from the pilot study in Section 4.1 can boost the model as a knowledge based feature. To answer these questions, in the next investigation we carry out step-II, in the next section.

4.2.6 Model II: Chinese Character Pieces for NMT

In this step-II investigation into decomposed models, instead of using radical sequence as enhancing features, we test if each fine-gained decomposition level of character pieces can directly represent the original word and character sequences in

meaning representation for NMT.

We firstly investigate translation at increasing levels of decomposition of Chinese characters using underlying radicals, as well as the additional Chinese character strokes corresponding to ever-smaller units, breaking down characters into component parts as this is likely to reduce the number of unknown words, rare words, and low frequency MWEs. Then, in order to better deal with MWEs which have a common occurrence in the general context [235], and working in the opposing direction in terms of meaning representation, we investigate translating larger units of Chinese text, with the aim of restricting translation of larger groups of Chinese characters that should be translated together as one unit. In addition to investigating the effects of decomposing characters we simultaneously apply methods of incorporating bilingual MWEs into translation that were extracted with the methodology mentioned in the last pilot study section.

Character Decomposition Methods

We now introduce the character decomposition approach and the extraction tools which we apply in this work. These tools are publicly available at: https://github.com/poethan/MWE4MT/tree/master/radical4mt. We employ the open source IDS dictionary which was derived from the CHISE (CHaracter Information Service Environment) project [15]. It is comprised of 88,940 Chinese characters from CJK (Chinese, Japanese, Korean script), unified ideographs and the corresponding decomposition sequences of each character. Most characters are decomposed as a single sequence, but characters can have up to four possible decomposed representations. The reason for this is that the character can come from different resources, such as Chinese Hanzi (G, H, T for Mainland, Hong Kong, and Taiwan), Japanese Kanji (J), Korean Hanja (K), and Vietnamese ChuNom (V), etc.[16] Even though they have the same root of Hanzi, the historical development of languages and writing systems in

[15] http://www.chise.org/

[16] Universal Coded Character Set (10646:2017) standards.iso.org/ittf/PubliclyAvailableStandards/

different territories has resulted in certain degrees of variation in the appearance and stroke order, for instances, (且 vs 目) and (了 vs 丂) from the second and third examples in Figure 4.13 where the grey parts of decomposition graphs represent the construction structure of the character.

Figure 4.13 shows example characters that have two different decomposition sequences. In our experiments, when there is more than one decomposed representation of a given character, we choose the Chinese mainland decomposition standard for the model, since the corpora we use correspond best to simplified Chinese as used in mainland China. The examples in Figure 4.13 also show the general construction and corresponding decomposition styles of Chinese characters, such as *left-right*, *up-down*, *inside-outside*, and *embedded* amongst others. To obtain a decomposition level L representation of Chinese character α, we go through the IDS file L times. Each time, we search the IDS file character list to match the newly-generated smaller-sized characters in the character pieces and replace them with a decomposed representation recursively.

Unicode	Character	Decomposition	Decomposition
U+4E3D	丽 (lì)	⿱一⿰⿰冂丶⿰冂丶[G]	⿰⿱一⿰冂丶⿱一⿰冂丶[T]
U+5177	具 (jù)	⿱⿻且一八[GTKV]	⿱目一八[J]
U+51FD	函 (hán)	⿴凵⿻了⿱丷八[GTV]	⿴凵⿻丂⿱丷八[JK]
U+52C7	勇 (yǒng)	⿱甬力[GTV]	⿱⿱マ田力[JK]
Character construction: ⿱: up-down, ⿰: left-right, ⿰⿴⿵: inside-outside, ⿻: embedded			

Figure 4.13: Chinese character examples from the IDS dictionary.

4.2.7 Model II: Experimental Settings

We now test the various levels of decomposed Chinese and Chinese MWEs using publicly available data from the WMT-2018 shared tasks Chinese to English, using the pre-processed (word segmented) data as training data [22]. The original word

boundaries were preserved in decomposition sequences, as advised by the step-I experiment findings shown in Section 4.2.4. To obtain better generalisability of our decomposition model, we use a large enough size namely the first 5 million parallel sentences for training across all learning steps which is even larger than our last experimental setting in Section 4.2.4 where we used 1.25 million sentences. The corpora "newsdev2017" used for development and "newstest2017" for testing are from the WMT-2017 MT shared task [23]. These include 2002 and 2001 parallel Chinese ⇔ English respectively. We use the THUMT [294] toolkit developed by the Tsinghua University NLP group which is an implementation of several attention-based transformer architectures for NMT and set up the encoder-decoder as 7+7 layers [273], a similar setting to the pilot study experiments in Section 4.1. Batch size is set as 6,250. For sub-word encoding BPE technology, we use 32K BPE operations that are learned from the bilingual training set. Experiments are carried out on Google Colab [17].

In order to assess the performance of each model employing a different meaning representation in terms of decomposition and MWEs, we carried out both automatic BLEU and human evaluation and direct assessment (DA) using crowd-source workers [90] of the outputs of the system.

Bilingual MWE Terms as Features

We further prepare bilingual MWEs that are comprised of radical and stroke level decomposition from the word level, and examine their effect on MT system performance as well as adopting the MWE extraction pipeline from our pilot study [231, 109] in Section 4.1). We briefly summarise the procedure there for refreshment in Figure 4.1.

Treetagger [240] is used to attach Part-of-Speech (POS) tag and lemma information to English and Chinese sentences. Subsequently, we designed Chinese POS patterns to extract MWEs by mapping from the English patterns used by [231,

[17] colab.google.com

	20k	100k	120k	140k	160k	180k
baseline	18.39	**21.56**	21.45	21.31	21.29	21.42
base+MWE	**18.49**	21.39	**21.67**	**21.83**	**21.42**	**21.86**
RXD3	16.48	20.75	20.73	20.93	20.98	21.14
RXD3+MWE	17.82	21.36	21.50	21.31	**21.42**	21.47
RXD2	11.84	13.26	12.88	13.02	13.38	12.86
RXD1/ideograph	15.52	20.67	20.61	21.26	20.76	21.00

Table 4.11: Chinese → English NMT BLEU scores for increasing learning steps.

109] and manual addition of some new tags for idioms, fixed expressions, personal names, place names and organisation names by native speakers. We then extract the English and Chinese monolingual MWEs using MWEtoolkit [226]. Next, MWE converting tools and MPaligner [231, 219] are deployed to extract bilingual MWE candidates. Filtering is applied to remove low quality bilingual MWE pairs using the threshold score which is estimated according to the statistical translation scores of the phrases. This is set to a threshold of 0.85 in all experiments.

Finally, we decompose the Chinese MWEs (from bilingual pairs) into radical and stroke sequences to different degrees and add these new bilingual pairs back into the training corpus.

4.2.8 Model II: Automatic Evaluation Using BLEU

The evaluation scores of word sequence baseline NMT, character decomposed NMT and MWE-NMT according to the BLEU metric are presented in Table 4.11 where RXD1/2/3 represents the decomposition level of Chinese characters. The RXD1 model, decomposition level 1, is the ideograph model [297] used for their experiments where the phonetic (声旁 shēng páng) and semantic (形旁 xíng páng) parts of characters are separated initially. Since decomposition level 3 yields generally higher scores than the other two levels, we also applied decomposition of MWEs to level 3 and concatenated the bilingual glossaries to the training data.

From the automated evaluation results, we see that decomposition model RXD3 has very close BLEU scores to the baseline word sequence model, even though the original Chinese word sequences are all removed in learning. This is interesting since the level 3 Chinese decomposition is typically very difficult for even native language human speakers to read and understand which is like a cipher code. Furthermore, by adding the decomposed MWEs back into the learning corpus, "rxd3+MWE" (RXD3MWE) yields higher BLEU scores in some learning steps than the baseline word model. To gain further insight, we show the learning curve with the learning steps and corresponding automated-scores in Figure 4.14. Importantly, our experiments revealed that decomposition of level-2 (RXD2) is not a good choice of representation of the original word sequence for the MT in our settings according to BLEU metric.

We also discovered that the decomposed models generated fewer system parameters for the neural nets to learn, which potentially reduces computational complexity. This addresses the MT modelling issue on the model sizes issue raised earlier in Section 2.3. For instance, the total trainable variable size (as reported by Colab) of the word-based baseline model is 89,456,896, while this number decreased to 80,288,000 and 80,591,104 respectively for RXD3 and RXD2 models (10.25% from RXD3). As mentioned by [88], in NLP tasks the total number of possible words is so large that the word sequence models have to operate on an extremely high-dimensional and sparse discrete space. The decomposition model reduced the overall size of possible tokens for the model to learn, which is more space efficient.

The BLEU score increasing ratio in decomposed models (from RXD3 to RXD3MWE) is larger than the ratio in word sequence models (from BASE to BASEMWE) by adding MWE pairs in general. Furthermore, the increase in performance is consistent by adding MWEs from the decomposed model, compared to the conventional word-based model. For instance, the performance has a surprising drop at 100K learning steps for BASEMWE.

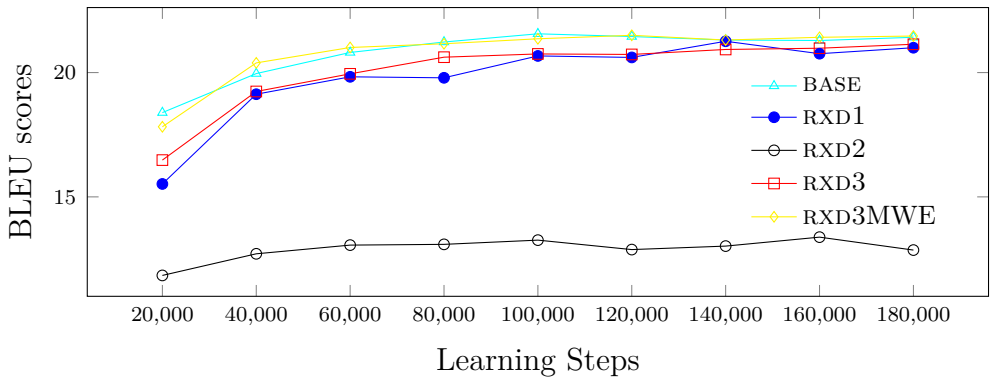

Figure 4.14: Learning curves from different models with BLEU metric

4.2.9 Model II: Crowd-sourced Human Evaluation

Due to the inconsistency in BLEU scores observed in the above experiments, and the criticism of an automated metric by some MT researchers, we additionally conducted human evaluation using direct assessment (DA) [90]. We used the models with the most learning steps, 180K, and ran human evaluation on the Amazon Mechanical Turk crowd-sourcing platform,[18] including the strict quality control measures of [90].

Human evaluations were collected on Mechanical Turk after quality filtering was applied. Direct Assessment scores for systems were calculated as in [92] by firstly computing an average score per translation before calculating the overall average for a system from its average scores for translations. Significance tests in the form of Wilcoxon Rank-Sum test were then applied to score distributions of the latter to identify systems that significantly outperform other systems according to the human evaluation.

Results of the Direct Assessment human evaluation are shown in Table 4.12 where similarly performing systems are clustered together (denoted by horizontal lines in the table). Systems in a given lower ranked cluster are significantly outperformed by all systems in a higher ranked cluster and Avg. raw = the average score for translations calculated from raw Direct Assessment scores for translations, Avg. z = the average score for translations after score standardisation per human assessor mean and standard deviation score, n is the number of distinct translations included

[18] https://www.mturk.com

in the human evaluation i.e. the sample size used in significance testing, N is the number of human assessments including repeat assessment, horizontal lines denote systems belonging to distinct clusters in which all systems in a lower ranked cluster are significantly outperformed by all systems in a higher ranked cluster. Amongst the six models included in the human evaluation, the first five form a cluster with very similar performance according to crowd-sourced human assessors, including the baseline, MWE, RXD1, RXD3MWE, and RXD3. RXD2, on the other hand, is far behind the other models in terms of performance according to human judges, performing significantly worse than all other runs (at $p< 0.05$). Furthermore, human evaluation results in Table 4.12 show that the top five models all achieve high performance on-par with state-of-the-art in Chinese to English MT.

Avg. raw	Avg. z	n	N	
73.2	0.161	1,232	1,639	BASE
71.6	0.125	1,262	1,659	MWE
71.6	0.113	1,257	1,672	RXD1
71.3	0.109	1,214	1,593	RXD3MWE
70.2	0.073	1,260	1,626	RXD3
53.9	−0.533	1,227	1,620	RXD2

Table 4.12: Human evaluation results for systems using Direct Assessment

4.2.10 Model II: Expert Validation - Phase One

Very recent investigations on evaluation of MT outputs has presented the issues from crowd-sourced human evaluations [77]. The crowd-sourced human evaluation achieved lower correlation to expert translator ranking, in comparison to state-of-the-art automatic metrics. At the same time, BLEU has been criticised for not reflecting the real improvement of state-of-the-art MT models especially when single reference translation is used in practice. BLEU has been favouring "boring translations" instead of the improved one with higher lexical diversity and more

natural outputs according to human experts [78]. According to translation post editing research, BLEU can have an up to 6 point score drop for human preferred translations [76].

We carried out expert validations of our MT models using ourselves, the MT experts and Chinese native speakers.

We select typical sentences from the test data (100K learning step models) to provide insights into the kind of errors made by each model. Even though the automated BLEU metric gives the baseline model a higher score 21.56 than the RXD3 model (20.75), the translation of some Chinese MWE terms is better with the RXD3 model. For instance, in Figure 4.15, RXD3 is the Chinese decomposition with level 3, BASE is the word sequence model, and src/ref represents source/reference. The Chinese MWE 商场 (shāng chǎng) in the first sentence is correctly translated as *mall* by RXD3 model but translated as *shop* by the baseline word sequence model. The MWE 楼梯间 (lóu tī jiān) in the second sentence is correctly translated as *stairwell* by the RXD3 model while translated as *stairs* by baseline. Furthermore, the MWE 近日 (jìn rì) meaning *recently* is totally missed out by the word sequence model, which results in a misleading ambiguous translation of an even larger content, i.e., did the chief move to San Francisco (SF) *recently* or *this week*. We will not get this clearly from the word base sequence model, however, the MWE 近日 (jìn rì) is correctly translated by the RXD3 model and the overall meaning of the sentence is clear that the chef moved to SF *recently* and was found dead *this week*.

We also attach the translations of these two sentences by four other models. With regard to the first sentence MWEs, all the four models translate San Francisco mall correctly as REF and RXD3 beating BASE model. In terms of the second sentence MWEs, BASEMWE and RXD2 drop out the MWE 近日 (jìn rì, *recently*) as BASE model, and all the four models drop out the translation of MWE 楼梯间 (lóu tī jiān, *stairwell*).

For the automatic and human evaluation results, where the decomposition level 2 achieved a surprisingly lower score than the other levels, error analysis revealed

src	28 岁 厨师 被 发现 死 于 **旧金山** 一家 **商场** 近日 刚 搬 至 旧金山 的 一位 28 岁 厨师 本周 被 发现 死 于 当地 一家 商场 的 楼梯间 。
ref	28 @-@ Year @-@ Old Chef Found Dead at **San Francisco Mall** a 28 @-@ year @-@ old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week .
rxd3	the 28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef who recently moved to San Francisco has been found dead on a stairwell in a local mall this week .
base	the 28 @-@ year @-@ old chef was found dead in a **shop** in **San Francisco** a 28 @-@ year @-@ old chef who has moved to San Francisco this week was found dead on the stairs of a local mall .
base MWE	28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef who recently moved to San Francisco was found dead this week at a local mall .
rxd3 MWE	28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef recently moved to San Francisco was found dead this week at a local mall .
rxd1	the 28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef recently moved to San Francisco was found dead in a local shopping mall this week .
rxd2	the 28 @-@ year @-@ old chef was found dead in a **San Francisco mall** a 28 @-@ year @-@ old San Francisco chef was found dead in a local mall this week .

Figure 4.15: Samples of the ZH-EN MT output at 100K learning steps.

an important insight. While level-1 decomposition encoded the original character sequences into radical representations, and this typically contains semantic and phonetic parts of the character, level-3 gives a deeper decomposition of the character such as the stroke level pieces with sequence order. In contrast, however, level-2 decomposition appears to introduce some intermediate characters that mislead model learning. These intermediate level characters are usually constructed from fewer strokes than the original root character, but can be decomposed from it. As in Figure 4.7, from decomposition level-2, we get new characters 从 (cóng) and 王 (wáng) respectively from 劍 (jiàn, *sword*) and 鋒 (fēng, *edge/sharp point*), but they have no direct meaning from their father characters, instead meaning "from" and "king" respectively.

In summary, decomposition level-2 tends to generate some intermediate characters that do not preserve the meaning of the original root character's radical, nor those of the strokes, but rather smaller sized independent characters with fewer strokes that result in other meanings.

4.2.11 Model II: Expert Validation - Phase Two

In Phase-two expert validation, we extended the sample size of validation sentences, and report evaluation results from typical sentences that reveal the mutual enhancement of BiMWE features and our decomposition models, e.g. via rxd3MWE (decomposition level-3 with BiMWE features), in Figure 4.16.

Figure 4.16 includes the sample sentences of source, reference, and MT outputs. We use underline to indicate the best overall translation at sentence level, and use bold case to indicate the correctly translated terms/MWEs highlighted in colored font.

The sample sentences for experts' validation are translation outputs from 180k learning steps, i.e., the most trained translation models. We use these sample sentences to acquire an in-depth view of the errors made by different models, i.e., word sequence based models including *base* and BiMWE enhanced word sequence models *base+BiMWE*, and decomposition models including *rxd1*; *rxd2*; *rxd3*; and BiMWE enhanced decomposition model *rxd3+BiMWE*.

Findings from expert validation Phase-two verify that the BiMWE enhanced models and decomposition models generate better translation outputs than word sequence based strong baseline models despite having very similar BLEU scores, with the exception of the rxd2 model. This tells us that expert validation can reveal interesting findings that cannot be revealed by the BLEU score or crowd-sourced human evaluations. We explain this in detail as bellow.

Sentence 1, "裁定为他杀 (ruled as a homicide)" is translated as "found dead" by the baseline model, which overlooks the meaning of the original MWE "他杀 (homicide)". The reason is the word "他 (him/he)", which has an ambiguous meaning in this context. Instead of the commonly used meaning of "him or he", in this context, it indicates "killed (杀) by *somebody else*". All the models fail to correctly translate this ambiguous MWE term, except for the rxd3+BiMWE model, which successfully translates it into a meaning equivalent phrase "convicted of killing him", but this cannot be reflected by n-gram matching BLEU scores due to the usage

src	1. 旧金山 警察局 称该 起 死亡 案件 被 裁定 为 他杀 ， 并 正在 进行 调查 。 2. 该 发言人 称 ， “ 我们 是 相处 得 像 亲密无间 的 一家人 的 一个 小 团队 ， 我们 将 深深 怀念 他 。 ” 3. Sons & Daughters 餐馆 的 一位 发言人 表示 ， 他们 对于 Frank 的 死 感到 “ 非常 震惊 ” 。 4. 但 受害人 的 哥哥 表示 想不出 有 谁 会 想要 加害于 他 ， 并称 “ 一切 终于 好 起来 了 。 ”
ref	1. the San Francisco Police Department said the death was ruled a homicide and an investigation is ongoing . 2. " we are a small team that operates like a close knit family and he will be dearly missed , " the spokesperson said . 3. a spokesperson for Sons & Daughters said they were " shocked and devastated " by his death . 3.correction: a spokesperson for Sons & Daughters restaurant said they were " shocked and devastated " by Frank's death . 4. but the victim 's brother says he can 't think of anyone who would want to hurt him , saying , " Things were finally going well for him . "
Base	1. the San Francisco police station says he was found dead and is investigating the death case . 2. " we are a small team that looks like a close @-@ knit family and we will miss him deeply , " the spokesman said . 3. a spokeswoman for Sons & Daughers said they were " shocked " by Frank 's death . 4. but the victim 's brother said he could not think of anyone who would harm him and said , " Everything is all right . "
base MWE	San Francisco police said the death was found to be a death case for him , and is under investigation . " we are a small team that is **as intimate as a family** , and we will remember him deeply , " the spokesman said . a Sons & Daughers spokesman said they were " shocked " at Frank 's death . but the victim 's brother said he couldn 't think of anyone who would want to hurt him and said , " **It 's all right at last** . "
Rxd1	the San Francisco police station said the death was ruled by him and was investigating . " we are a small team of people who live like a close family , and we will miss him deeply , " the spokesman said . a spokesman at the Sons & Daughers **restaurant** said they were " shocked " by Frank 's death . but the victim 's brother said he did not know who would want to harm him and said , " It 's all right . "
Rxd2	the San Francisco police station is investigating the death toll from a triangle . " we are a small team of families like the one with the same relationship , and we will miss him deeply , " the spokesman said . a spokesman for the Conversation &amp ; Conversation **Restaurant** said they were " very shocked " by the death of Frank . but the **elder brother** said that who would want to have the one who would , " all **is well at last** . "
Rxd3	the San Francisco police station said the death case was ruled by him and is under investigation . " we are a small group of close @-@ knit families and we will miss him deeply , " the spokesman said . a spokesman for Sons & Dauders said they were " shocked " by Frank 's death . but the victim 's brother said he could not imagine anyone trying to hurt him and said , " All is well . "
rxd3 MWE	San Francisco police said the death was convicted of **killing him** and was conducting an investigation . the spokesman said , " We are **as close as a family** of a small team , we will miss him deeply . " a spokesman for the Sons & Daughers **restaurant** said they were " shocked " by Frank 's death . but the victim 's brother said he could not think of anyone who would want to harm him , and said " **everything is better at last** . "

Figure 4.16: Samples of the ZH-EN MT output at 180K learning steps.

of different words from the reference translation “ruled a homicide”.

Sentence 2, the phrase “相处得像亲密无间的一家人 (operates like a close knit family)” is translated by the baseline as “that looks like a close-knit family”, which

is not very accurate. The translations by base+BiMWE, rxd1 and rxd3+BiMWE have better meaning equivalent outputs to interpret the MWE "相处得 (operates like, instead of *looks like*)" using terms "as intimate as", "live like", and "as close as".

For sentence 3, the reference translation in the WMT offered data is actually plainly wrong, and we corrected it in the figure presentation as "a spokesperson for Sons and Daughters restaurant said they were shocked and devastated by Frank's death". The baseline model surprisingly translated the MWE "发言人(spokesperson)" into a female gender "spokeswoman", even though other models produce "spokesman". The baseline model also drops out the MWE term "餐馆(restaurant)", while the decomposition models rxd1, rxd2 and rxd3+BiMWE all translate this term correctly.

For sentence 4, the MWE "哥哥(elder brother)" is only translated correctly by rxd2, and all other models translate it into "brother". The phrase "终于好起来了(finally going well (for him))" is translated as "is all right" by the baseline model, which drops the meaning that "things were not good in the beginning" which is reflected by MWEs "终于(finally)" and "好起来 (getting better)". This phrase is better translated by base+BiMWE, rxd2, and rxd3+BiMWE by "is all right at last", "is well at last", and "is better at last".

In summary: we can see the baseline model performs more poorly than other models in all four sample sentences.

We also discover that the rxd2 model actually can translate some MWE terms better than the baseline model even though its overall sentence-level translation quality is lower. For instance, rxd2 translates "哥哥(gē gē)" correctly into "elder brother" which is the only correct translation from all competitors; and it also translates "餐馆(cān guǎn)" into "restaurant" correctly unlike word based baseline and base+BiMWE models.

The rxd3+BiMWE model is the best performing model, with the rxd1 model being next, and closely followed by base+BiMWE. This indicates that both BiMWEs

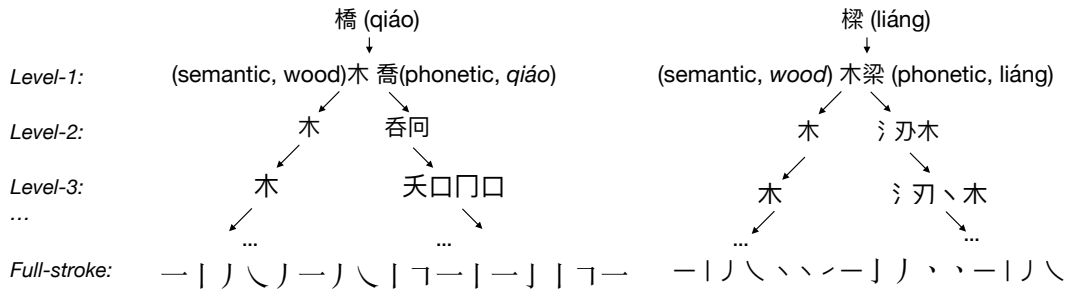

Figure 4.17: Character decomposition examples of qiáo and liáng

and character decomposition contribute to translation model learning, and that their combination contributes to translation accuracy in a larger degree. This verifies the *assumption* of our main focus, i.e., the if the decompositionality of Chinese characters helps the word meaning representation for MT task, and if it affects the BiMWE integrated NMT learning structure towards more accurate translation of MWEs.

Our other findings include that the testing corpus needs to be improved, e.g., sentence 3 where the offered testing reference has two errors: missing "restaurant" entity translation; and error translation from "Frank's" into "his". The reference of sentence 4 also has one error of "brother''instead of "elder brother". This indicates the issues in the testing suit that WMT (the Annual Workshop of MT) and MT communities have, and the need of higher quality alternative testing-suit creation, e.g. reducing errors and introducing multiple reference semantic equivalent translations.

As in Phase-one validation, we have interesting examples regarding level-2 decomposition, where the mis-leading smaller sized characters can be generated for model learning, as in Figure 4.17. For instance, 吞 (tūn) and 囧 (jiōng) are produced in Level 2 of character 橋 (qiáo, bridge), however they have very different meanings, from bridge, of "swallow" and "the face of a person when he/she does not know if he/she shall laugh or cry by the situation, e.g. a person open his/her mouth and becomes numb as indicated by the presentation of this character". Similarly, 办(chuāng) is produced in Level 2 from character 梁 (liáng, beam, a structural element to resist loads), however, it means "cut, wound, trauma" which is distinct

from the meaning of its root character. Level 3 decomposition will further split this smaller sized character 刃小 (chuāng) into a partial sequence with strokes.

4.2.12 Discussion

In this section, one step further to step-I in Section 4.2.4, we investigated varying the level of Chinese character decomposition for MT and also adding MWEs into NMT, including decomposed MWEs. We conducted experiments for decomposition levels 1 to 3. We conducted automated evaluation with the BLEU metric, crowd sourced human evaluation with direct assessment (DA), as well as expert validations. Our conclusion is that the Chinese character decomposition levels 1 and 3 can be used to represent or replace the original word sequence in an MT task, and that this achieves similar performance to the original word sequence model in NMT according to BLEU and crowd-sourced workers, but the expert validation shows that decomposed models can improve MWE translations. However, decomposition level 2 is not suitable to represent the original character sequence in meaning at least for MT. We leave it to future work to explore the performance of different decomposition levels in other NLP tasks.

Another finding from our experiments is that while adding bilingual MWE terms can both increase word and decomposed level MT score according to the automatic metric BLEU, the crowd-sourced human evaluation shows no statistical significance between them. This might be due to the high volume of training set we deployed, 5 million sentences, which covers all the test set.

We also report a deep analysis of the learning steps, to provide more information about how many learning steps will start to over-fit the model or decrease performance. We suggest that with 5 million training sentences and 7+7 layers of encoder-decoder neural nets, the Transformer model becomes too flat in its learning rate curve with 100K learning steps, and this applies to both original word sequence model and decomposition models.

Regarding the difference of each decomposition degree and the full stroke se-

quence of original characters, we can have a look at one example in Fig. 4.7, the first character 劍 (jiàn), where the level-3 decomposition output sequence is "人一口口人人刂" while full stroke is like "丿乀一丨一丨一丿乀丿乀丨丨". To reach the full stroke level, the decomposition number is uncertain for different characters due to their different level of complexity. The IDS file that we applied in this work from the CHISE project limited the decomposition to full stroke sequences. However, we aim to keep developing this aspect, as part of our future research.

In summary, in this section, we investigated: (i) the degree to which Chinese radical and stroke sequences represent the original word and character sequences that they are composed of; (ii) the difference in performance achieved by each decomposition level regarding MWE translations; (iii) the effect of radical and stroke representations in MWEs for MT.

Looking back to the MT challenges, our hypothesis and research questions, these sections (from 4.2.3) continued our pilot study in Section 4.1 to address current MT challenges in Section 2.3 by investigating the first half of our hypothesis "MWEs set challenges for MT and investigation into MWEs can help improve MT from translation modelling". Further to the pilot study, we move on to the investigation of Hy.a "including language specific features to address some weak points of existing methods" and Hy.b "MWEs in MT can be tackled as rare words challenges from the statistical perspective due to their low frequencies" by two step modelling. The first step used Chinese radicals as the semantic part of the Chinese character to enhance the MT learning model, thus to boost the translation model performance, the step two decomposition models treated MWEs as rare words and low frequency phrases by splitting the Chinese character into character pieces, to see if the character piece information can be used to estimate the low frequency MWE translation. Furthermore, we integrated the bilingual MWE terms extracted in the pilot study into the decomposition models in various degrees. These investigations on Hy.a and Hy.b also addressed general MT challenges in the areas of OOV words, model size, linguistic awareness and adequacy issues as in Section 2.3. We strictly followed our

research question I (RQ-I) to "do list" continuously by validating state-of-the-art MWE investigation models in MT (via a pilot study in Section 4.1), design of new methods focusing on specific languages (via step-I model in Section 4.2.4), and validating the proposed model with quantitative and qualitative analysis (via step-II model in Section 4.2.6).

4.3 Summary

In this chapter, we first made one pilot study re-examining one of the recent models integrating MWEs in MT using alternative language pairs (German-English and Chinese-English) as well as a new NMT structure, attention based BiRNN, with a large amount of language recourse. We found that bilingual MWE extraction and integration (BiMWEs) into the training can improve the automatic translation evaluation BLEU score, but with lower increasing rate, in this high-resource and high-performance baseline scenario, compared to the lower performance baseline used by [231] on English-Latvian and English-Czech. In addition to this finding, we also contributed a multilingual parallel MWE glossary called MultiMWE, covering the examined languages, German-English and Chinese-English, with 3,159,226 and 143,042 pairs of entries respectively. This is a relatively large amount of bilingual MWE resource to date, which is extracted automatically from a bilingual corpus with quality control by filtering, considering the bilingual MWE we found before: DE-EN 800 sentences and ZH-EN zero.

An analysis of this kind of method treats MWEs as auxiliary bilingual terms to improve model learning by enhancing the training corpus. We assume this is a similar approach to translation of rare words and OOV words. Focusing on the Chinese language, we propose to decompose Chinese characters into radicals where the meaning preservation part remains and to integrate the extracted radicals into the neural network training structure. We expect that this integration will improve the rare words and phrase translation accuracy, hence improving the adequacy of MT. The step-I experiments indeed produced positive evaluations using a wide range

of metrics and multi-reference test suites.

In the step-II investigation, we further decomposed the Chinese characters into different degrees (fine-gained) and examined their corresponding performances by replacing the original character sequences in Neural MT learning and translating with a new state-of-the-art baseline Transformer. We also use the bilingual MWE terms extracted in the pilot study as a knowledge feature by integrating them into the training corpus in Step-II. While the automatic evaluation using single-referenced BLEU and the crowd-sourced human assessment did not find a significant difference between the decomposed models (level 1 and 3) and the character sequence models, the expert validation Phase-one reflected that the decomposed models (level 3 and rxd3MWE) improved the MWE translations. Moreover, the step-II investigation using character pieces instead of character sequences produced a smaller number of parameters for Neural MT learning, which is more space efficient.

Our further expert assessment Phase-two with larger size sample data reveals that the BiMWE models and our character decomposition models can achieve a mutual enhancement to each other, e.g. "rxd3MWE" the decomposition level-3 plus BiMWE features, boosting the translating performances of both MWE terms and the overall sentence level accuracy, as shown in Figure 4.16.

In summary, in this chapter, our pilot study and proposed Chinese character decomposition models addressed some MT issues introduced in Section 2.3 on MT modelling (OOV and low-frequency words, model size), linguistic awareness (MWEs, Chinese radicals), and MT adequacy, and investigated our hypotheses **Hy.a** (including language specific features to improve current MT models investigating MWEs) and **Hy.b** (MWEs in MT can be tackled as a low-frequency or OOV words and phrases challenge from a statistical perspective), and exactly followed our **RQ-I** proposed "to do" list. In the next chapter, we move on to our **Hy.c** "investigating translation quality assessment (TQA) looking into MWEs" and the corresponding **RQ-II** "to do" list, namely a critique on MT issues with current test sets being used, corpus creation, and new TQA methods with the knowledge and component

of MWEs.

Chapter 5

Machine Translation Quality Assessments: Looking into Multi-Word Expressions

Following the last chapter summary, in this chapter, we delve more deeply into some of the machine translation (MT) issues and challenges as discussed in Section 2.3. We further investigate the second part of hypotheses, i.e. "investigation into MWEs can help to improve MT from the translation evaluation perspective", and our **Hy.c** that "current MT quality assessment models and test suites have disadvantages and we can tackle these by introducing better test suites from the data sanity point of view and the design of new evaluation methodologies that deploy MWEs as important features". These will guide our further experiments on multilingual parallel corpus creation with MWEs annotation, and new TQA methodology design incorporating the knowledge of MWEs, which follows our **RQ-II** listed research questions: i) critique on TQA, ii) better test suite covering literature domain, and iii) human-in-the-loop MT evaluation looking into MWEs.

We firstly revisit the MT issues and challenges in translation quality assessment (TQA) and the findings from our own experimental investigations. Then, based on these findings, we create a multilingual corpus with MWEs annotated in the

corpus which is called AlphaMWE. The corpus is sentence-by-sentence aligned, as well as MWE term aligned, across different languages. Our *methodology* of the creation of AlphaMWE in a multilingual setting is also part of the contribution of this thesis chapter. We discuss the issues in the current state-of-the art MT models in handling MWE related translations, revealing that they are far from achieving human (expert) parity. We then classify the errors into different categories expecting this to be helpful for future researchers focusing on each or some of them to tackle as *challenges*. After this, based on the AlphaMWE corpus created, we then introduce a new MT evaluation methodology, with the human-in-the-loop design and we examine MWEs in Section 5.3 called **HiLMeMe**. We hope this new evaluation methodology will help the community to achieve better and more reliable evaluation of MT system outputs. Our methodology is proposed based on an MT evaluation task, but researchers can integrate this model into their corresponding other NLP tasks easily, such as text summarization and image captioning evaluation, etc.

5.1 Critique of MT Issues in Assessment Methods

In Chapter 2, Section 2.3, we mentioned that MT is still far from reaching human parity and that MT evaluation plays one of the crucial roles in advancing its further development [63]. MT challenges include idiomatic and multiword expressions, as reflected in the two Chinese-English translation examples in Section 2.4. MT evaluation has a long history since the beginning of MT and reached a high level of popularity in the SMT and NMT eras with automatic metrics. These metrics are often based on test suites using reference translations, and try to incorporate syntax or semantic knowledge into the evaluation. On the one hand, the test suites popularly used in WMT shared tasks often come from the news domain without much variation, for instance, without including text from the literature domain [137]. Secondly, the test suites being used often lack linguistic driven annotations, e.g. metaphorical and idiomatic expressions. Thus the usage of these test suites for assessment involves a lack of linguistic awareness or guidance.

For better evaluation of state-of-the-art MT systems, firstly, varied domain data is needed, including literature text to verify the MT model quality in the translation of idiomatic and ambiguous expressions. Secondly, these expressions should be highlighted or labelled in the test suite so that they can be deployed easily by the emulators and assessors to pay attention to these components in the test suites, in addition to the overall sentence and segment translation quality, in scoring and ranking the systems. Neither of these two aspects are achieved by current research and practice.

From the experimental investigations that we carried out in our modelling evaluations in the last chapter in Sections 4.1 and 4.2.3, we also observe that the current popular practice in MT model evaluation is that they are only deployed on news domain test suites and the popular evaluation metrics (BLEU, LEPOR, BEER, CharacTER) only generate an overall scores for MT systems, based on different factors, e.g. n-gram matching or trying to capture word order issues, without detailed error classifications for the evaluated systems. It would help MT researchers to find out exactly what issues exist in their models, and use the evaluation results as a guide to determine the next steps. Thus, it is highly plausible to create a corresponding evaluation methodology that produces error categorisations in addition to quantitative scores.

Another issue as we discussed previously, is that automatic metrics tend to give higher scores to "seemingly fluent translations" which have more n-gram matching to reference translations without reflecting whether the idiomatic and ambiguous expressions are correctly interpreted in meaning equivalence. Thus, a corpus with ambiguous expression annotation would help assessors look into this situation. For instance, the n-gram match scores can be very high but have a totally opposite meaning to the source or reference text.

However, we realise that there is in fact a lack of a corresponding corpus for this, especially, a test suite that has MWEs annotated in a bilingual setting, not to say multilingual.

5.2 AlphaMWE: A Multilingual Corpus with MWEs

Following the critique of last section, in this section, we describe how a multilingual parallel corpus with MWE annotations was constructed to investigate the performance of different NMT models on MWE related content.

5.2.1 Motivations

Multi-word Expressions (MWEs) have long been of interest to both natural language processing (NLP) researchers and linguists [235], [53], [225]. The automatic processing of MWEs has posed significant challenges for some tasks in computational linguistics (CL), such as word sense disambiguation (WSD), parsing and (automated) translation [152], [25], [248], [167], [109], which is verified by our own NMT experiments in the last chapter as well as in Chapter 2 with example sentences that have MWEs as a challenge (Section 2.4). This is caused by both the variety and the richness of MWEs as they are used in language. However, as noted by other NLP researchers for example in [53], there is a lack of bilingual or even multilingual parallel corpora with MWE annotations available for cross-lingual NLP research and for downstream applications such as machine translation (MT) and its evaluation [129, 109].

We have introduced the definition of MWEs in Chapter 1 (Section 1.4) and the working definition of MWEs for our MT experiments in the beginning of Chapter 4, and in this section, we introduce a sub-category of MWE research, namely "verbal MWEs (vMWEs)" for our corpus creation.

The working definition of vMWEs was introduced by PARSEME shared task on vMWE identification and discovery that has been organised almost each year since 2017 [180, 200]. It is defined as following: a vMWE has a verb as the head of the studied MWE term and functions as a verbal phrase, with examples such as "kick the bucket". We also have more examples from our corpus creation some of which will be explained in Section 5.2.4 including "cutting capers" and "(beer) gone to his head", etc.

In the work reported here, to improve our evaluation of MT systems as well as facilitate multilingual MWE research, we present the construction of a multilingual corpus with vMWE annotations as a first step, including English-Chinese, English-German, English-Polish and English-Italian language pairs. We started with the same source monolingual corpus in English with its vMWE tags from the shared task affiliated with the SIGLEX-MWE workshop in 2018 [279, 227]. Several state-of-the-art (SOTA) MT models were used to perform an automated translation, and then human post-editing and annotation for the target languages was conducted with cross-validation to ensure the quality, i.e., with each sentence receiving post-editing and manual rechecking by at least two people.

In order to get deeper insights into the difficulties of processing MWEs we carried out a categorisation of the errors made by MT models when processing MWEs. From this we conclude that current state-of-the-art MT models are far from reaching parity with human experts in terms of translation performance, especially on idiomatic MWEs, and even for sentence level translation, although some researchers sometimes claim otherwise [289, 119].

A brief guide to the sections that follow is below. In the next section (5.2.2) we present related work and then detail the corpus preparation stages including selection of MT models and the resulting AlphaMWE in Section 5.2.3. We then look at various issues that MT has with regard to MWEs in Section 5.2.4, followed by a broader discussion section (5.2.5).

The error analysis and discussion of MT models on MWEs once again reflects and verifies one of our hypothesis that "MWEs set challenges for MT and investigation into MWEs can help improve MT from both translation modelling and translation evaluation perspectives". Furthermore, the AlphaMWE multilingual parallel corpora will serve as a test bed for our new MT evaluation method that we will propose to address the *Hy.c* "(in brief) current MT quality test suites have disadvantages and we can tackle these by introducing better test suites and deploying a new evaluation method taking MWEs into account". And these strictly follow the 2nd and 3rd task

lists in our **RQ-II**.

5.2.2 Related Work on Corpus Construction

There are a number of published studies which focus on the creation of monolingual corpora with vMWE annotations, such as the PARSEME shared task corpora [239, 227, 238]. The 2017 edition of this task included 18 European languages, and the 2020 edition of the task covers 14 languages including Chinese, Hindi, and Turkish as non-European languages. [276] prepared a manual annotation of a Hungarian corpus with Light Verb Constructions (LVCs) and indicated its usability for machine translation and information extraction. Some work from monolingual English corpora includes the MWE aware "English Dependency Corpus" from the Linguistic Data Consortium (LDC2017T01), which covers *compound words* used to train parsing models, and Wiki50 by [277], which consists of 50 Wikipedia articles (4,350 sentences) with the annotation of MWEs (compound, verb-particle constructions, idiom, Light verb construction, multi-word verbs) and named entities (PER, LOC, ORG and MISC). Also related to this are English MWEs from "web review data" by [241] that covers noun, verb and preposition super-senses and English verbal MWEs from [279] and [133] that covers PARSEME shared task defined vMWE categories. However, all these works were performed in monolingual settings, independently by different language speakers without any bilingual alignment. These corpora are helpful for monolingual MWE research and tasks such as discovery or identification, however it would be difficult to use these corpora for bilingual or multilingual research such as MT or cross-lingual information extraction.

The work most related to that presented in this thesis is from [275], who created an English-Hungarian parallel corpus with annotations for light verb constructions (LVCs). A total of 703 LVCs for Hungarian and 727 for English were annotated in this work, and a comparison between English and Hungarian data was carried out. However, the work did not cover other types of vMWEs, for instance inherently adpositional verbs, verbal idioms, or verb-particle constructions, and it was

not extended to any other language pairs. In our work, we annotate in a multilingual setting including far distance languages such as English, German, Polish, Italian and Chinese, in addition to the extension of vMWE categories. In our pilot study presented in Section 4.1 [109], we performed an automatic construction of bilingual MWE terms based on a parallel corpus, in this case English-Chinese and English-German referred to as MultiMWE. We first conducted automated extraction of monolingual MWEs based on part-of-speech (POS) patterns and then aligned the two side monolingual MWEs into bilingual terms based on statistical lexical translation probability. However, due to the automated procedure, the extracted bilingual "MWE terms" contain not only MWEs but also normal phrases. Part of the reason for this is due to the POS pattern design which is a challenging task for each language and needs to be further refined [248], [231], [109].

Similarly on the use of POS patterns, [81] carried out an experimental investigation into multilingual collocation extraction based on English to Spanish and to Portuguese. Both bilingual word embedding and bilingual translation dictionaries (handcrafted and statistical) were applied as well to some comparable corpora. The extraction was based on five different syntactic patterns: adjective-noun, noun-noun, verb-object, verb-subject, and verb-adverb.

Another restriction from both our pilot study work in Section 4.1 [109] and [81] is that the POS patterns we used as well as by [81] can only extract continuous MWEs leaving discontinuous MWEs as a topic for future possible exploration. Our AlphaMWE corpus, however, includes both continuous and discontinuous MWEs.

5.2.3 AlphaMWE Corpus Construction

We now describe our corpus preparation method, selection of the MT models used in our investigation, and the open-source corpus AlphaMWE.

Preparations and Workflow

To construct a well aligned multilingual parallel corpus, our approach is to take a monolingual corpus from the PARSEME *vMWE discovery and identification shared task* as our root corpus. Our rationale is that this shared task is well established and its process of tagging and categorisation is clear. Furthermore, as we plan to extend the MWE categories in future, we enrich the PARSEME shared task corpus with potential for other downstream research and applications, including bilingual and multilingual NLP models. The English corpus [279] we used from the PARSEME shared task (V1.1) follows the annotation guidelines having a broad range of vMWE categories tagged. These include inherently adpositional verbs (IAV), light verb constructions (LVC.cause and LVC.full), multi-verb constructions (MVC), verbal idioms (VID), and verb-particle constructions (VPC.semi and VPC.full). The PARSEME English corpus is from the Universal Dependencies English treebanks [1] including English Web Treebank, LinES Parallel Treebank, and Parallel Universal Dependencies treebanks.

The English corpus contains sentences from several different topics, such as news, literature, and IT documents. For the IT document domain, vMWEs are usually easier or more straightforward to translate, with a high chance of repetition, e.g. "apply filter" and "based on". For the literature annotations, the vMWEs include richer samples with many idiomatic or metaphor expressions, such as "cutting capers" and "gone slightly to someone's head" that cause MT challenges.

Figure 5.1 shows our workflow. This first uses MT models to perform automated translation for the target language direction, then human post editing of the output hypotheses with annotation of the corresponding target side vMWEs which are aligned with the source English ones. Finally, to avoid human introduced errors, we apply a cross validation strategy, where each sentence receives at least a second person's quality checking after the first post-editing. Tagging errors are more likely to occur if only one human has seen each sentence and we discuss some error samples

[1] http://universaldependencies.org/

from English source corpora in later sections.

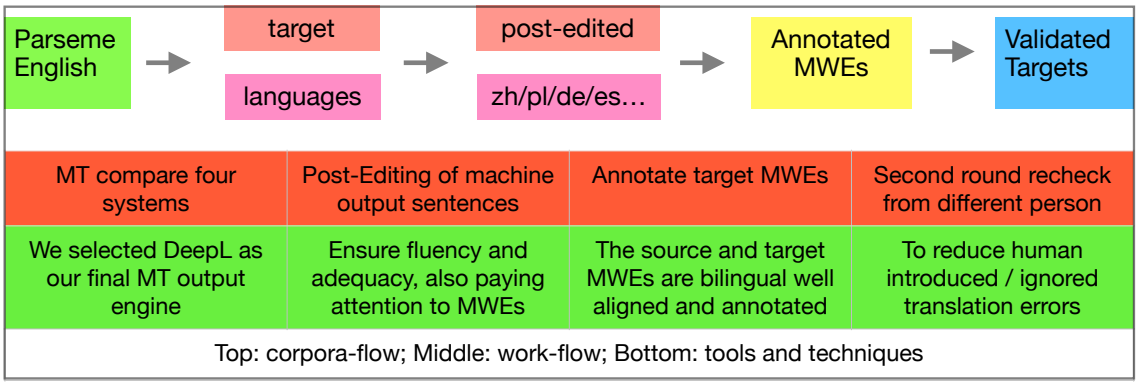

Figure 5.1: Workflow to prepare the AlphaMWE corpus.

MT Model Selection

We tested a number of example sentences from the English test set to compare state-of-the-art MT from Microsoft Bing [47], GoogleMT [273], Baidu Fanyi [262], and DeepL[2], as in Figure 5.2. We illustrate the comparative performances with two worked example translations. As a first example sentence, GoogleMT and Bing Translator have very similar outputs, where the MT output sentences try to capture and produce as much information as possible, but make the sentences redundant or awkward to read, such as the phrase "验证... 是否验证了 (yàn zhèng ... Shì fǒu yàn zhèng le)" where they use a repeated word "验证" (yàn zhèng, *verify*). Although the DeepL Translator does not produce a perfect translation since it drops the source word "validated" which should be translated as "有效性 (yǒu xiào xìng)" (as one candidate translation), the overall output is fluent and the source sentence meaning is mostly preserved. Baidu translator yields the worst output in this example. It produces some words that were not in the source sentence (或者, huò zhě, *or*), loses some important terms' translation from the source sentence ("SQL Server", the subject of the sentence), and the reordering of the sentence fails resulting in an incorrect meaning ("在没有密码的情况下, zài méi yǒu mì mǎ de qíng kuàng xià" is

[2] https://www.deepl.com/en/translator (All testing was performed in July 2020)

moved from the end of the sentence to the front and made as a condition). So, for this case, DeepL performed best.

As a second example sentence, GoogleMT confused the original term TSQL as SQL. Bing MT had a similar issue with the last example, i.e. it produced redundant information "有关 (yǒu guān)" (about/on). In addition it concatenated the website address and the normal phrase "了解有关 (liǎo jiě yǒu guān)" together with a hyperlink. GoogleMT and Bing both translate half of the source term/MWE "Microsoft Developer Network Web" as "Microsoft 开发人员网络网站" (kāi fā rén yuán wǎng luò wǎng zhàn) where they kept "Microsoft" but translated "Developer Network Web". Although this is one reasonable output since Microsoft is a general popular named entity while "Developer Network Web" consists of common words, we interpret "Microsoft Developer Network Web" as a named entity/MWE in the source sentence that consists of all capitalised words which would be better translated overall as "微软开发人员网络网站 (wēi ruǎn kāi fā rén yuán wǎng luò wǎng zhàn)" or be kept as the original capitalised words as a foreign term in the output, which is how DeepL outputs this expression. However, Baidu totally drops out this MWE translation and another word translation is not accurate, "more" into 详细 (xiáng xì). These samples illustrate why we chose to use DeepL as the provider of our MT translations.

AlphaMWE: a Glimpse of the Corpus

Regarding the size of the corpus, we extracted all 750 English sentences which have vMWE tags included. The target languages include Chinese, German, Polish and Italian with sample sentences in Figure 5.3. There are several situations and decisions that are worth noting:

- when the original English vMWEs are translated into a general phrase in the target language but not choosing sequence of MWEs, we tried to offer two different references, with one of them being revised in a vMWE/MWE presentation in the target;

Two sample sentences' MT outputs comparison from head of test file	
Source	# text = SQL Server verifies that the account name and password were validated when the user logged on to the system and grants access to the database, without requiring a separate logon name or password.
DeepL	# text = SQL Server 会在用户登录系统时验证账户名和密码，并授予对数据库的访问权限，而不要求单独的登录名或密码。
Google	# text = SQL Server验证用户登录系统时是否验证了帐户名和密码，并授予对数据库的访问权限，而无需单独的登录名或密码。(yàn zhèng … shì fǒu yàn zhèng le)
Bing	[text] SQL Server 验证用户登录到系统时是否验证了帐户名称和密码，并授予对数据库的访问权限，而无需单独的登录名称或密码。(yàn zhèng … shì fǒu yàn zhèng le)
Baidu	#或者，在没有密码的情况下验证用户名和用户名是否被登录到数据库中，并且系统是否授予登录权限。(huò zhě, zài méi yǒu mì mǎ de qíng kuàng xià)
Ref.	# 文本 = SQL Server 会在用户登录系统时验证账户名和密码的有效性，并授予对数据库的访问权限，而不要求单独的登录名或密码。
Source	# text = See the http://officeupdate.microsoft.com/, Microsoft Developer Network Web site for more information on TSQL.
DeepL	# text = 有关 TSQL 的更多信息，请参见 http://officeupdate.microsoft.com/，Microsoft Developer Network Web 站点。
Google	# text =有关Microsoft SQL的更多信息，请参见http://officeupdate.microsoft.com/，Microsoft开发人员网络网站。
Bing	[文本] 有关 TSQL http://officeupdate.microsoft.com/了解有关 TSQL 的相关信息，请参阅 Microsoft 开发人员网络网站。
Baidu	#text=参见http://officeupdate.microsoft.com/，以获取有关TSQL的详细信息。
Ref.	# text = 有关 TSQL 的更多信息，请参见 http://officeupdate.microsoft.com/，微软开发人员网络互联网站点。
Blue: redundancy; green: adding error; pink: reordering error; yellow: dropping error.	

Figure 5.2: Sample comparison of outputs from four MT models.

- when the original English sentence terms were translated into the correct target language but in a different register, e.g. the source language has low register (thx, for instance), we offer two reference sentences, with one of them using the same low register and the other with (formal) full word spelling;

- for the situations where a single English word or normal phrase is translated into a typical vMWE in the target language, or both source and target sentences include vMWEs, but the source vMWE was not annotated in the original English corpus, we made some *additions* to include such vMWE (pairs) into AlphaMWE;

- for some wrong/incorrect annotation in the source English corpus, or some mis-spelling of words, we corrected them in AlphaMWE;

- we chose English as root/source corpus, since the post-editing and annotation of target languages requires the human annotators to be fluent/native in both-side languages, and all editors were fluent in English as well as being native

speakers in the specific target languages respectively.

We examined the development and test data sets from the annual Workshop of MT (WMT) [21] and also from the NIST MT challenges where they offered approximately 2,000 sentences for development/testing over several years. This means that our bilingual/multilingual corpora with 750 sentences is comparable to such standard shared task usage. Our confidence estimation work in [86] on translation quality assessment of testing sample size also verifies that the AlphaMWE corpus is large enough to be deployed to validate and test the MT system output quality.

As shown in the examples in Figure 5.3 from Chinese, German, Polish and Italian, the alignment across languages is at sentence-level. In addition, the MWEs in the sentences are annotated in the end of each sentence, and are concatenated by semi-colon in the order of their occurrences. AlphaMWE also includes statistics of the annotated vMWEs, and a multilingual vMWEs glossary. The extracted 750 sentences from both training and testing files of EN corpus have a total number of vMWEs 832 (331+501), IAV 60 (16+44), LVC.cause 43 (7+36), LVC.full 244 (78+166), MVC 4 (0+4), VID 139 (60+79), VPC.full 297 (151+146), VPC.semi 45 (19+26) [3]. The AlphaMWE corpora are divided evenly into five portions which were designed in the post-editing and annotation stage. As a result, it is convenient for researchers to use them for testing NLP models, choosing any subset portion or combination, or cross validation usage.

5.2.4 MT Issues with MWEs during AlphaMWE Construction

We performed an analysis of the behaviour of various MT systems when required to translate MWEs or MWE-related content. In this analysis we focus on the English → Chinese language pair. We also highlight some issues on English → German and English → Polish in different sections, but leave the detailed analysis of other

[3]This indicated that the training file size is smaller than the testing file size, also the training file has fewer number of annotations of PARSEME EN v1.1.

Plain English Corpus	The chair was comfortable, and the beer had gone slightly to his head. I was smoking my pipe quietly by my dismantled steamer, and saw them all cutting capers in the light, with their arms lifted high, when the stout man with mustaches came *tearing down* to the river, a tin pail in his hand, assured me that everybody was 'behaving splendidly, splendidly, dipped about a quart of water and *tore back* again. (*the italic was not annotated in source English*)
English MWEs	gone (slightly) to his head, cutting capers, tearing down, tore back
Target Chiense Corpus	椅子很舒服，啤酒已经微微让他上了头。[sourceVMWE: gone (slightly) to his head][targetVMWE: (微微)让他上了头] 我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下欢呼雀跃，高举双臂，这时，那个留着胡子的大块头，手里拿着一个铁皮桶，快速来到河边，向我确保大家都"表现得很精彩，很精彩"，他浸了大约一夸脱的水，又快速回去了。[sourceVMWE: cutting capers; tearing down; tore back][targetVMWE: 欢呼雀跃; 快速到; 快速回去]
Target German Corpus	Der Stuhl war bequem, und das Bier war ihm leicht zu Kopf gestiegen. [sourceVMWE: gone (slightly) to his head][targetVMWE: (leicht) zu Kopf gestiegen] Ich rauchte leise meine Pfeife an meinem zerlegten Dampfer und sah, wie sie alle im Licht mit hoch erhobenen Armen Luftsprünge machten, als der stämmige Mann mit Schnurrbart mit einem Blecheimer in der Hand zum Fluss hinunterkam und mir versicherte, dass sich alle "prächtig, prächtig benahmen, etwa einen Liter Wasser eintauchte und wieder zurückwankte". [sourceVMWE: cutting capers; tearing down; tore back][targetVMWE: Luftsprünge machten; hinunterkam; zurückwankte]
Target Polish Corpus	Krzesło było wygodne, a piwo lekko uderzyło mu do głowy. [sourceVMWE: gone (slightly) to his head][targetVMWE: (lekko) uderzyło mu do głowy] Cicho paliłem swoją fajkę przy zdemontowanym parowcu i widziałem, jak wszyscy pląsają w świetle, z podniesionymi wysoko ramionami, gdy twardziel z wąsami przyszedł szybkim krokiem do rzeki, blaszany wiaderko w dłoni, zapewnił mnie, że wszyscy "zachowują się wspaniale, wspaniale, nabrał około ćwiartkę wody i zawrócił szybkim krokiem". [sourceVMWE: cutting capers; tearing down; tore back][targetVMWE: pląsają; przyszedł szybkim krokiem; zawrócił szybkim krokiem]
Target Italian Corpus	La sedia era comoda, e la birra gli aveva leggermente dato alla testa. [sourceVMWE: gone (slightly) to his head][targetVMWE: aveva (leggermente) dato alla testa] Stavo fumando tranquillamente la pipa vicino al mio piroscafo smontato, e li ho visti tutti giocare gioiosamente alla luce, con le braccia alzate, quando l'uomo robusto con i baffi è venuto giù al fiume alacremente, un secchio di latta in mano, mi ha assicurato che tutti si stavano "comportando splendidamente, splendidamente, ha preso circa un litro d'acqua ed è tornato indietro velocemente." [sourceVMWE: cutting capers; tearing down; tore back] [targetVMWE: giocare gioiosamente; venuto giù alacremente; tornato indietro velocemente]
AlphaMWE corpora examples from multilingual parallel files. "cutting capers" was annotated as VID type of MWEs, while "tearing down" and "tore back" were not annotated in the source English corpus. We added them into AlphaMWE multilingual corpora since they do cause translation errors for most state-of-the-art MT models. The bilingual MWEs are aligned with their appearance order from sentence inside the afterwards attached bracket-pairs.	

Figure 5.3: AlphaMWE corpora samples with two sentences

language pairs for future work.

English to Chinese

When MT produces incorrect or awkward translations this can fall into different categories, and from our analysis we **classify** them as: common sense, super sense, abstract phrase, idiom, metaphor, and ambiguity with ambiguity further sub-divided. These classifications are to be further refined in future, e.g. the differences between metaphor and idiom are sometimes unclear. We now list each of these with exam-

ples to support future MT research on improving the quality of MT when handling MWEs.

Category-I: Common Sense (CS)

Source	At the corner of 72nd Street and Madison Avenue, he waved down a cab.
DeepL	在72街和麦迪逊大道的拐角处，他向一辆出租车招手。 Zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù, tā xiàng yī liàng chū zū chē zhāo shǒu.
Bing	在72街和麦迪逊大道的拐角处，他挥手示意一辆出租车。 zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù , tā huī shǒu shì yì yí liàng chū zū chē.
Google	在第72街和麦迪逊大街的拐角处，他挥舞着出租车。 Zài dì 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā huī wǔ zhe chū zū chē.
Baidu	在72街和麦迪逊大街的拐角处，他挥手叫了一辆出租车。 zài 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā huī shǒu jiào le yí liàng chū zū chē.
Ref.	在72街和麦迪逊大道的拐角处，他招手示停了一辆出租车。 Zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù, tā zhāo shǒu shì tíng le yī liàng chū zū chē.

Figure 5.4: MT issues with MWEs: common sense. Pinyin offered by GoogleMT with post-editing.

The first error category is called *common sense*. As an example, the sentence in Figure 5.4 includes the vMWE "waved down" which in general understanding indicates that "he succeeded in getting the cab" and not only "waved his hand". However, in the translation by DeepL and Bing this vMWE was incorrectly translated as "he waved his hand to the cab" missing part of the original meaning. The MT output by GoogleMT is also incorrect, saying "he waves with the cab in hand". The Baidu translation of this sentence is semantically correct that "he waved and got one cab" though it does not use a corresponding Chinese side vMWE "招手示停 (zhāo shǒu shì tíng)".[4]

Category-II: Super Sense (SS)

This category of translation issue is related to a form of state of mind and we need to make a logical prediction to guess the positiveness or negativeness of some words, in the choice of Chinese characters. As in Figure 5.5, the MT systems each have

[4]We give full sentence pronunciation (in Pinyin) of Chinese characters in this figure, for the following examples, we only present the Chinese Pinyin for MWEs and studied words of the sentences to save space.

Source	Each time he took a walk, he felt as though he were leaving himself behind, and by giving himself up to the movement of the streets, by reducing himself to a seeing eye, he was able to escape the obligation to think, and this, more than anything else, brought him a measure of peace, a salutatory emptiness within.
DeepL	每一次散步，他都觉得自己好像把自己抛在了脑后，把自己交给了街道的运动，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这比什么都重要，这给他带来了一定程度的平静，内心的空虚让人敬佩。(jiāo gěi; nèi xīn de kōng xū ràng rén jìng pèi)
Bing	每次散步时，他都觉得自己好像要把自己抛在脑后，通过向街道运动自在，通过把自己留眼，他能够逃避思考的义务，这比什么都重要，给他带来了一定程度的和平，一种有益的空虚。(zì zài; yǒu yì de kōng xū)
Google	每次散步时，他都会感觉自己好像在抛弃自己，投身于大街小巷，睁大了眼睛，这使他摆脱了思考的义务，而这， 最重要的是，他带给他一种和平的感觉，一种内在的称呼空虚。(tóu shēn yú; nèi zài de chēng hu kōng xū)
Baidu	每次他散步，他都觉得自己好像是在离开自己，把自己交给街道的流动，把自己变成一个有眼光的人，他就可以逃避思考的义务，而这比其他任何事情都能给他带来某种程度的平和，一种内在的致意的空虚。(nèi zài de zhì yì de kōng xū)
Ref.	每一次散步，他都觉得自己好像把自己抛在了脑后，投身于大街小巷，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这给他带来了某种程度的平静和内心悦纳的空无，远胜于其他。(nèi xīn yuè nà de kōng wú)

Figure 5.5: MT issues with MWEs: super sense.

advantages for different parts of this long sentence. However, none of them is perfect. For instance, for the translation of vMWE "giving (himself) up (to)", the DeepL and Baidu outputs give very literal translations by saying "he gives himself to", while the Bing translator drops the vMWE, and GoogleMT preserves the correct meaning in the translation "投身于 (tóu shēn yú)" from the reference indicating "he devoted himself". However, GoogleMT's output for the phrase "salutatory emptiness within" is very poor and makes no sense; the reference is "the emptiness that he welcomes" for which Baidu has a closer translation "内在的致意的空虚 (nèi zài de zhì yì de kōng xū)". All four MT outputs also use the same Chinese words "空虚 (kōng xū)" which is a term with negative meaning, however, the sentence indicates that he is welcoming this emptiness, which should be the corresponding Chinese words "空无 (kōng wú)", an unbiased or positive meaning.

Category-III: Abstract Phrases (APs)

Abstract phrases can have different exact meanings and we usually need some background information from the sentence or paragraph to select the correct word choices in the target language. This sometimes belongs to the context-unaware ambiguity (CUA) that we will mention later. However, CUA does not necessarily mean "abstract phrase", and usually needs paragraph information, not only at the sentence level. Furthermore, in some situations, we just do not know how to interpret

Source	Quinn had his doubts, but this was all he had to **go on**, his only bridge to the present.
DeepL	奎恩有他的疑虑，但这是他唯一的依靠，是他通往现在的唯一桥梁。(wéi yī de yī kào)
Bing	奎因有他的怀疑，但这就是他必须**去**的，他唯一的桥梁到现在。 (bì xū qù de)
Google	奎因有他的疑惑，但这就是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de)
Baidu	奎因有他的疑虑，但这是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de)
Ref.	奎恩曾有他的疑虑，但这是他**开展工作**的所有依据，是他通往现在的唯一桥梁。(**kāi zhǎn gōng zuò** de suǒ yǒu yī jù)

Figure 5.6: MT issues with MWEs: abstract phrases

"abstract phrase", i.e. the candidate interpretations are unknown without context, and this is different from ambiguity. With the example sentence in Figure 5.6, from the context, we know that "go on" in this sentence means "to work from" using all the information he had. The phrase "this was all he had to go on" is then to be interpreted as "this is all the information he had to work from". At the end of the sentence, "the present" is the "present person" he needs to look for (with the picture of this person's younger age portrait). However, Bing translated it as "this is (where) he had to go" which is an incorrect interpretation of "had to go"; furthermore, Bing's translation of the second half of the sentence kept the English order, without any reordering between the words, which is grammatically incorrect in Chinese, i.e. "他唯一的桥梁到现在 (tā wéi yī de qiáo liáng dào xiàn zài)". GoogleMT and Baidu translated it as "what he needs to do" which is also far from correct, while DeepL successfully translated the part "his only thing to rely on" but dropped the phrase "go on", i.e., *to do what*. Abstract Phrase can include Super Sense as a sub-category, however, it does not necessarily relate to a state of mind.

Category-IV: Idioms

The use of idioms often causes wrongly translated sentences, mostly resulting in humorous output due to literal translation. For example, in the sentence in Figure 5.7, the vMWEs "*cutting capers*" and "tore back" are never translated correctly at the same time by any of the four MT models we used. The idiom "cutting capers" indicates to frolic or romp, to "act in the manner of a young goat clumsily frolicking about" and here it means "they are in a happy mood, playful and lively movement"

Source	I was smoking my pipe quietly by my dismantled steamer, and saw them all cutting capers in the light, with their arms lifted high, when the stout man with mustaches came tearing down to the river, a tin pail in his hand, assured me that everybody was 'behaving splendidly, splendidly, dipped about a quart of water and tore back again.
DeepL	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下高举双臂切着荠菜，这时，那个留着胡子的壮汉撕扯着来到河边，手里拿着一个铁皮桶，向我保证说，大家都'表现得很精彩，很精彩，浸了大约一夸脱的水，又撕扯回来了。 (qiē zhe jì cài; sī chě huí lái)
Bing	我悄悄地用我拆除的蒸笼抽我的烟斗，看到他们所有的切割披肩在光中，他们的手臂抬起来高，当粗壮的人用胡子来到河中，一个锡桶在他的手中，向我保证，每个人都'表现辉煌，辉煌，浸约夸脱的水，并再次撕毁。(qiē gē pī jiān; zài cì sī huǐ)
Google	我正用拆卸的蒸锅安静地抽着烟斗，看到他们都在灯光下割下刺山柑，双臂高高举起，当时那留着小胡子的矮壮男人向河边撕下，手里拿着一个铁桶，向我保证 每个人都表现得非常出色，表现得非常出色，蘸了一夸脱的水，然后又撕了回来。(gē xià cì shān gān; sī le huí lái)
Baidu	我正静静地在我拆掉的汽船旁抽着烟斗，看见他们都高举着胳膊，在灯光下割着山柑，这时那个长着胡子的胖男人手里拿着一个锡桶，朝河里跑来，向我保证每个人都表现得很好，很漂亮，蘸了一夸脱水，然后又往回跑。(gē zhe shān gān; wǎng huí pǎo)
Ref.	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下欢呼雀跃，高举双臂，这时，那个留着胡子的大块头，手里拿着一个铁皮桶，快速来到河边，向我确保大家都"表现得很精彩，很精彩"，他浸了大约一夸脱的水，又快速回去了。(huān hū què yuè; kuài sù huí qù)

Figure 5.7: MT issues with MWEs: idioms

which should properly be translated as the corresponding Chinese idiom "欢呼雀跃 (huān hū què yuè, happily jumping around like sparrows)". However, all four MT models translated it literally into "cutting" actions just with different subjects, i.e., what they cut. The idiom (slang) "tore back" means the stout man walked back rapidly, which the Baidu translation gives the closest translation as "往回跑 (wǎng huí pǎo, run back)" but the other three models translated into an action "tear something (to be broken)" which is incorrect.

Category-V: Metaphors

The first sentence vMWE "blown to bits" in Figure 5.8 is a metaphor to indicate "everything is gone", instead of the physical "blowing action". However, the three MT models DeepL, GoogleMT and Baidu translate it as "exploded into pieces (by bombs)", while BingMT translates it even more literally into "blown to (computer) bits". There is a corresponding Chinese vMWE "化为乌有 (huà wéi wū yǒu, vanish into nothing)" which would be a proper choice for this source vMWE translation.

The second sentence vMWE "gone (slightly) to his head" is a metaphor to indicate "got slightly drunk". However, all four MT models translate it as physically "beer moved to his head" but by slightly different means such as *flow* or *flutter*. The corresponding translation as a MWE should be "微微让他上了头 (wéi wéi ràng tā

Source	The what? Auster laughed, and in that laugh everything was suddenly blown to bits. The chair was comfortable, and the beer had gone slightly to his head.
DeepL	那个什么？奥斯特笑了，在这笑声中，一切突然被炸得粉碎。(bèi zhà dé fěn suì) 椅子很舒服，啤酒已经微微到了他的头上。 (wēi wēi dào le tā de tóu shàng)
Bing	什么？奥斯特笑了，在笑，一切都突然被吹成位。(bèi chuī chéng wèi) 椅子很舒服，啤酒稍微到他的头去了。 (shāo wēi dào tā de tóu qù le)
Google	什么啊 Auster笑了起来，在那笑声中，一切突然被炸碎了。 (bèi zhà suì le) 椅子很舒服，啤酒微微飘到他的头上。(wēi wēi piāo dào tā de tóu shàng)
Baidu	什么？奥斯特笑了，在那笑声中，一切都突然被炸成碎片。(bèi zhà chéng suì piàn) 椅子很舒服，啤酒已经稍稍流到他的头上了。(shāo shāo liú dào tā de tóu shàng le)
Ref.	那个什么？奥斯特笑了，在这笑声中，一切突然化为乌有。(huà wéi wū yǒu) 椅子很舒服，啤酒已经微微让他上了头。 (wēi wēi ràng tā shàng le tóu)

Figure 5.8: MT issues with MWEs: metaphor

shàng le tóu)", using the same characters, but the character order here makes so much difference, meaning "slightly drunk".

Category-VI: Ambiguities

We encountered different kinds of situation that cause ambiguity in the resulting translation when it meets MWEs or named entities, so we further divide ambiguity into three sub-classes.

Category-VI: Context-Unaware Ambiguity (CUA)

In this case, the context, i.e. the background information, is needed for correct translation of the sentence. For instance, see Figure 5.9. DeepL gives the translation "it did not give me time though", while Bing and GoogleMT give the same translation "it/this did not give me one day's time" and Baidu outputs a grammatically incorrect sentence. From the pre-context, we understand that it means the speaker "did not feel that is special to him" or "did not have affection of that" after *all the Mormon missionary's effort towards him.* Interestingly, there is a popular Chinese idiom (slang) that matches this meaning very well "不是我的菜 (bù shì wǒ de cài, literally *not my dish*)". From this point of view, the context based MT model deserves some more attention, instead of only focusing on the sentence level. When we tried to put all background context information as shown in Figure 5.9 into the four MT models, they produce as the same output for this studied sentence, as for sentence

level MT. This indicates that current MT models still focus on sentence-by-sentence translation when meeting paragraphs, instead of using context inference.

Source	But it did not give me the time of day.
DeepL	但它并没有给我时间。 (gěi wǒ shí jiān)
Bing	但它没有给我一天的时间。 (gěi wǒ yī tiān de shí jiān)
Google	但这没有给我一天的时间。 (gěi wǒ yī tiān de shí jiān)
Baidu	但它没有给我一天中的时间。 (gěi wǒ yī tiān zhōng de shí jiān)
Ref.	但我没有感到这个对于我特殊 / 但这不是我的菜。(gǎn dào zhè ge duì yú wǒ tè shū / … wǒ de cài)
Context	An old Mormon missionary in Nauvoo once gripped my knee hard as we sat side by side, and he put his arm about me and called me "Brother." We'd only met ten minutes before. He took me to his good bosom. His eyes began to mist. I was a prospect, an exotic prospect in old tennis shoes and a sweatshirt. His heart opened to me. It opened like a cuckoo clock. But it did not …

Figure 5.9: MT issues with MWEs: context-unaware ambiguity

Category-VI: Social/Literature-Unaware Ambiguity (SLUA)

In this case, social knowledge of current affairs from news, or literature knowledge about some newly invented entities / phrases is required in order to get correct translation output. For instance, Figure 5.10 includes two sentences, one from politics and another from literature.

Source	The moment they know the de-gnoming's going on they storm up to have a look. Then someone says that it can't be long now before the Russians write Arafat off.
DeepL	他们一知道去核的事，就会冲上去看一看。 (qù hé) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**注销**。(**bǎ** ā lā fǎ tè **zhù xiāo**)
Bing	当他们知道去诺格明是怎么回事， 他们冲了起来看看。 (qù nuò gé míng) 然后有人说，现在俄罗斯人要不长了，就**把**阿拉法特**注销**了。(**bǎ** ā lā fǎ tè **zhù xiāo** le)
Google	当他们知道正在逐渐消失的那一刻，他们便冲上去看看。 (zhèng zài zhú jiàn xiāo shī) 然后有人说，不久之后俄罗斯人**将**阿拉法特**注销**。 (**jiāng** ā lā fǎ tè **zhù xiāo**)
Baidu	他们一知道德格诺明正在进行，就冲上去看一看。 (dé gé nuò míng) 然后有人说，俄国人很快就会把阿拉法特**一笔勾销**了。 (bǎ ā lā fǎ tè **yī bǐ gōu xiāo** le)
Ref.	一知道去地精的事在进行，他们就冲上去观看。 (qù dì jīng) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**下課 /** 让…下台。 (**bǎ** ā lā fǎ tè **xià kè;** ràng…xià tái)

Figure 5.10: MT issues with MWEs: social/literature-unaware ambiguity

In the first sentence, “de-gnoming” is a word from Harry Potter, invented by its author, to refer to the process of ridding a garden of gnomes, a small magical beast. Without this literary knowledge it is not possible to translate the sentence

correctly. For instance, even though this sentence is from a very popular novel that has been translated into many languages, DeepL translated it as "去核 (qù hé, de-nuclear)", Bing translated it as "去诺格明 (qù nuò gé míng, *de*-nuògémíng" where "nuògémíng" is a simulation of the pronunciation of "gnoming" in a Chinese way, Baidu translated it as "德格诺明 (dé gé nuò míng)" which is the simulation of the pronunciation of the overall term "de-gnoming".

In the second sentence, "write Arafat off" is to dismiss "Yasser Arafat", former Chairman of the Palestine Liberation Organisation, who is a person of historical significance. However, all three models DeepL, Bing, and GoogleMT translated it into "把/将阿拉法特注销 (bǎ/jiāng ā lā fǎ tè zhù xiāo, *deregister Arafat*)" which treated "Arafat" as the title of a certain policy/proceeding, not being able to recognise it as a personal named entity, while Baidu made the effort to use the Chinese idiom "一笔勾销 (yī bǐ gōu xiāo, *cancel everything*, or *never mention historical conflicts*)" for "write off", but it is not a correct translation. Interestingly, if we put these two sentences into a web search engine it retrieves the correct web pages as context in the top rank position of the search result. This may indicate that future MT models could consider including web search results as part of their knowledge of the background or for disambiguation, for translation purposes.

Category-VI: Coherence-unaware Ambiguity (CohUA)

Source	Two months ago I had to have an operation for a serious ***complaint***.
DeepL	两个月前，我因为一次严重的**投诉**不得不做手术。(**tóu sù** … zuò shǒu shù)
Bing	两个月前，我不得不做一个严重的**投诉**手术。 (zuò … **tóu sù** shǒu shù)
Google	两个月前，我不得不接受一次手术以应对严重的**投诉**。 (jiē shòu yī cì shǒu shù … **tóu sù**)
Baidu	两个月前，我因为严重的**投诉**不得不动手术。 (**tóu sù** … dòng shǒu shù)
Ref.	两个月前，我因为一次严重的**症状**不得不做手术。(**zhèng zhuàng** … zuò shǒu shù)

Figure 5.11: MT issues with MWEs: coherence-unaware ambiguity

This kind of MWE ambiguity can be solved by the coherence of the sentence itself, for instance, the example in Figure 5.11. The four MT models all translated the vMWE itself "have an operation" correctly in meaning preservation by "做/接

受/动手术 (zuò/jiē shòu/dòng shǒu shù)" just with different Chinese word choices. However, none of the MT models translated the "reason of the operation", i.e., "complaint" correctly. The word complaint has two most commonly used meanings "a statement that something is unsatisfactory or unacceptable" or "an illness or medical condition" and all four models chose the first one. According to simple logic of social life, people do not need to "have an operation" due to "a statement", instead their "medical condition" should have been chosen to translate the word "complaint". Because of the incorrectly chosen candidate translation of the word "complaint", Bing's output even invented a new term in Chinese "投诉手术 (tóu sù shǒu shù, *a surgery of complaint statement kind*)" which makes no sense.

English to German

In the case of English-to-German MWEs, there are some cases where the corresponding German translation of English MWEs can be one word. This is partially because German has compound verbs. For instance, the vMWE "woke up" in the sentence "An old woman with crinkly grey hair woke up at her post outside the lavatory and opened the door, smiling and grasping a filthy cleaning rag." has corresponding German aligned word "erwachte" with a suitable translation "Eine alte Frau mit krausem, grauem Haar erwachte auf ihrem Posten vor der Toilette und öffnete die Tür, lächelte und griff nach einem schmutzigen Putzlappen.".

This also occurs in the English-to-Chinese translation, such as an English verb+particle MWE being aligned to one single Chinese character/word. For example, in this sentence "The fact that my name has been mixed up in this.", the vMWE (VPC) *mixed up* gets aligned to a single character word "混 (hùn)" in a suitable translation "事实上，我的名字已经被混在这里面了。(shì shí shàng, wǒ de míng zì yǐ jīng bèi *hùn* zài zhè lǐ miàn le)".

A second issue is that the automatic translation to German can be biased towards choosing the polite or formal form, see the examples such as "Sie"instead of the second form singular "du"for "you", "auf Basis von" instead of "basierend auf" for

"based on". Achieving a higher accuracy level of MT will depend on the context of usage to decide which form is more suitable.

A third issue is that for the English verbal multi-word expressions that are often not translated as verbal multi-word expressions to German, this indicates some further work to explore by MT researchers to develop better models to have the machine produce corresponding German MWEs.

English to Polish

The MT output issues in English-to-Polish mostly fall into the categories of coherence-unaware error, literal translation, and context unaware issue.

Regarding the MT output, issues in English-to-Polish that fall into coherence-unaware error, for instance, the vMWE "write off" in the sentence "Then someone says that it can't be long now before the Russians write Arafat off." was translated as "Wypiszą" (Potem ktoś mówi, że już niedługo Rosjanie wypiszą Arafata.) which means "prescribe", instead of the correct "spiszą na straty (Arafata)". This error can be avoided by the coherence of the sentence itself in meaning preservation models.

For a literal translation, we can see the example vMWE "gave (him) a look" in the sentence "She ruffled her feathers and gave him a look of deep disgust." which was literally translated as "dała mu spojrzenie", however, in Polish, people use "**throw** a look" as "rzuciła (mu) spojrzenie" instead of "gave (dała, a female form)" and a proper translation would be *Nastroszyła sobie pióra i rzuciła mu spojrzenie głębokiego obrzydzenia.* Also the MT output word for "Nastroszyła" was "Zdruzgotała" which has an incorrect meaning. . Another example of literal translation leading to errors is the vMWE "turn the tables" from the sentence "Now Iran wants to turn the tables and is inviting cartoonists to do their best by depicting the Holocaust." which is translated as "odwrócić stoliki (turn tables)", however, it should be "odwrócić sytuację (turn the situation)" or "odwrócić rolę (turn role)" with a proper translation "*Teraz Iran chce odwrócić sytuację i zachęca rysowników, by zrobili wszystko, co w ich mocy, przedstawiając Holocaust.*" These two examples

illustrate the localisation issue in the target language.

For the context unaware issue, we can look back to the example sentence "But it did not give me the time of day." from Figure 5.9. This was literally translated word by word into "Ale nie dało mi to pory dnia." which is in the sense of hour/time. However, it should be "Nie sądzę aby to było coś wyjątkowo/szczególnie dla mnie. (I do not think this is special to me.)" based on the context, or "Ale to nie moja bajka" as an idiomatic expression which means "not my fairy tale" (indicating *not my cup of tea*).

5.2.5 Discussion

In this section, we discuss some issues discovered during the corpus construction process, which are related to MWEs in MT.

MWE Issues: Source Corpus

Some problems occurred in the source English corpus which impact the sentences we extracted from the monolingual training and test data with vMWE annotation tags.

Firstly, there is an accuracy issue with the tagging, which may have been unavoidably introduced as a result of the tagging task being carried out by one single person for each case, i.e. without a cross-validation quality checking stage [5].

Some error annotations of vMWEs in the source monolingual corpus will thus have some impact on the accuracy level of the vMWE discovery and identification shared task, but also affect the bilingual usage of AlphaMWE. We tried to address all these cases. For instance, for the example sentence in Figure 5.6, the English corpus annotated wrongly the sequence "had to go on" as a verbal idiom (VIDs) which is not accurate. The verb "had" here is affiliated with "all he had" instead of "to go on". So either we annotate "go on" as vMWE in the sentence or the overall clause "*all he had to go on*" as a studied term.

[5] https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.1/EN

Another example with a different type of vMWE is the sentence "He put them on in a kind of trance." where the source English corpus tagged "put" and "trance" as Light-verb construction (LVC.cause). However, the phrase is with "put...on" instead of "put...trance". "put someone into a trance" is a phrase to express "make someone into a half-conscious state". However, for this sentence, if we check back a bit further of the context, it means "he put on his cloth in a kind of trance". The word "trance" is affiliated with the phrase "*in a kind of trance*" instead of "put".

Source	The pressure had been building up in him since Stillman's disappearance that morning, and *it came out of him now as* ***a torrent of words***.
DeepL	自从那天早上斯蒂尔曼失踪后，他身上的压力就一直在积聚，现在却如**滔滔江水**般涌了出来。(jī jù, ... què rú **tāo tāo jiāng shuǐ** bān yǒng le chū lái)
Bing	自从斯莱特曼那天早上失踪后，他的压力就一直在增加，现在他突然**大言不出**。(zēng jiā, ... tā tú rán **dà yán bù chū**)
Google	自从斯蒂尔曼（Stillman）那天早上失踪以来，他身上的压力一直在增加，现在，他的**言语如潮水**般涌出。(zēng jiā, ... tā de **yán yǔ rú cháo shuǐ** bān yǒng chū)
Baidu	自从斯蒂尔曼那天早上失踪后，他的压力就一直在增加，现在他身上的压力就像**滔滔不绝的话**。(zēng jiā, ... tā shēn shang de yā lì jiù xiàng **tāo tāo bù jué de huà**)
Ref.	自从那天早上斯蒂尔曼失踪后，他身上的压力就一直在积聚，现在这些压力变成**滔滔不绝的话**倾诉而出。(jī jù, ... zhè xiē yā lì biàn chéng **tāo tāo bù jué de huà** qīng sù ér chū)

Figure 5.12: Additional vMWEs or MWEs

A second issue we discovered with the English language corpus is that there are some interesting sentences in the corpus that include non-decomposable MWEs, but these MWEs are not annotated. Future work could include adding further annotation on this aspect and extend this kind of bilingual pairs.

For instance, in Figure 5.12, the vMWE category verb-particle constructions (VPC.semi) is tagged to the phrase "building up", which may be an interesting case for vMWE discovery and identification. However, for cross-lingual research, as the initial aim of our corpus construction, such as MT, the ending part of this sentence "came out of him now as a torrent of words" poses more challenges, and would draw more attention from researchers.

We test this with four MT models and found the following outputs: DeepL literally translated it into "a strong and fast-moving stream of water" and dropped

"words"; Bing gave a translation with opposite meaning "does not say a word"; Google and Baidu produced much better translation covering "torrent of words" even though the sentence level translation contains errors and can be improved, i.e., "it" meaning the "pressure"in the source sentence was dropped out by Google; "came out of him" was dropped out by Baidu.

5.2.6 Summary of AlphaMWE

In this section, we presented the construction of a multilingual parallel corpus, AlphaMWE, with vMWEs as pioneer annotations by native speakers of the corresponding languages. This is inspired by a need for better MT evaluation test suites in assessing whether MT models reach human parity. We described the procedure of MT model selection, human post editing and annotation, and compared different state-of-the-art MT models and classified the MT errors from vMWEs related sentence/context translations. We characterised the errors into different categories to help MT research to focus on one or more of them to improve the performance of MT. The vast majority of error categories verified our assumption that MT is far from human parity, and that MWEs are a bottleneck.

We performed the same process as described here for English → Chinese, English → German and English → Polish and similarly categorised the MT issues when handling MWEs. We also included a section on corpus development.

We will use AlphaMWE for our new evaluation methodology that we will present in the next section. This is inspired by the fact that current translation quality estimation (TQA) does not reflect differences in state-of-the-art MT models. We offer a new gold standard in evaluation.

5.3 HiLMeMe: *H*uman-*i*n-the-*L*oop *M*T *E*valuation with *MWE*s

As discussed previously on MT challenges and MT evaluation issues (Sections 2.3 and 2.2), current evaluation methods using automatic metrics fail to correctly reflect the differences among state-of-the-art MT models, and to distinguish them from the real human parity. Meanwhile crowd-sourced human evaluations suffer from a lack of clear guidance on linguistic inputs and of well-trained workers, leading to low correlation to experts' judgements [77, 112]. Following the creation of the AlphaMWE multilingual parallel corpora with MWEs annotation (Section 5.2), we design a *h*uman-*i*n-the-*l*oop *M*T *e*valuation method looking into *m*ulti-word *e*xpressions (*HiLMeMe*). This methodology takes MWEs as one important factor in the assessment procedure, in addition to judging the overall contextual sentence or segment level translation performances. HiLMeMe is designed for experts or well-trained native speakers with guidelines to carry out translation evaluations. It asks assessors to do certain level classification of the error types regarding MWE translations, e.g. if it is translated correctly or not, using reference MWEs or alternatives or common phrases, etc. These classification behaviours are saved in our toolkit and can be exported for researchers to carry out further analysis of their system outputs. With this in mind, we explore whether HiLMeMe can have a positive influence for MT modelling research.

5.3.1 HiLMeMe Model Design

If we look back at Section 2.2, our survey of MT evaluation, HiLMeMe can be placed into the area of semantic features integrated human assessment, with MWEs as the lexical terms featured, in addition to fluency and adequacy criteria being used. It is also connected to task-oriented evaluations. There is a three-step assessment task designed in HiLMeMe, including general text (sentences, segments) level fluency and adequacy score $General(fluency, adequacy)$, highlighted MWEs translation quality

score $MWE(\alpha, \beta, \gamma, \theta)$, and a weighting parameter for MWEs on overall text Φ. We describe these separately below before we get to the computation of the overall HiLMeMe score.

I. How good is the MT output text in general?

- Look at the two factors when scoring.
- Fluency: Is the candidate translation fluent, e.g. grammatically correct
- Adequacy: Does the candidate translation cover all the meaning in the source / reference text?
- Give a score 0 to 10 $\rightarrow General(fluency, adequacy)$.

In this interface, we give a scoring range 0 to 10.

II. Look into the highlighted MWEs and classify if they are translated, and if so then how?

- Correctly translated using reference MWEs ($\alpha++$, score:10)
- Correctly translated using alternative MWEs ($\beta++$, score:10)
- Translated using other words, non-MWEs ($\gamma++$, score:0 to 10)
- Not translated, lost, NULL (θ, score:0)
- Score $\rightarrow MWE(\alpha, \beta, \gamma, \theta)$

In this interface, we give four choices ref-MWE, alt-MWE, non-MWE, and NULL, in addition to a scoring range 0 to 10. The triple set (α, β, γ) stores how often the MWEs are translated using the reference MWE, alternative MWE, or other words, and θ stores how often the source MWEs are left in a loss in the translation or kept as foreign words without any translation.

III. From which aspects do the MWEs present difficulty, affect the translation, and to what degrees?

- Semantics: word meanings and relations between them

- Grammar: syntax and morphology
- Idiomaticity: a group of words established by usage as having a meaning not deducible from those of the individual words
- Ambiguity: the quality of being open to more than one interpretation
- Degrees (ϕ, score:0 to 1)
- Output $\rightarrow \Phi(Sem, Gra, Idi, Amb, \phi)$

where the parameters Sem, Gra, Idi, Amb represent semantics, grammar, idiomaticity and ambiguity respectively. The classification of different situations in step III is to facilitate further analysis on the MWEs appearing in our test set (corpus), as well as the possible extension in future to cover more labelled data with broader aspects. This is a multiple choice classification where the assessors can tick more than one of the categories.

Finally, the overall score of HiLMeMe, i.e. $HiLMeMe(General, MWE, \Phi)$, is the weighted sum of the general text score $General(fluency, adequacy)$ and MWE score $MWE(\alpha, \beta, \gamma, \theta)$ with the weighting parameter from the third step ϕ on the influence of MWEs on overall text.

The scoring function is as below and is based on the three step judgements where we use $HiLMeMe(\bullet)$ to indicate $HiLMeMe(General, MWE, \Phi)$.

$$HiLMeMe(\bullet) = General(fluency, adequacy) + \phi \times MWE(\alpha, \beta, \gamma, \theta) \tag{5.1}$$

$$HiLMeMe_{norm} = HiLMeMe/Point_{Max} \tag{5.2}$$

where $Point_{Max}$ is the maximum point that step-I and step-II can generate, and $HiLMeMe_{norm}$ is the normalised score propagating the HiLMeMe score into the interval (0, 1).

The overall score is the combination of step-I and step-II with a weighting parameter attached to the second point. The normalised score of HiLMeMe is the raw score divided by the highest potential score they can get, such that the normalised score ranges from 0 to 1. The normalisation is to give the user a more straightforward instinct on how much the assessors judge the translation text quality in a 0-to-1 (0 to 100) range. Another benefit of the normalisation is that it can be used for automatic evaluation metrics tuning by calculating their correlation to HiLMeMe, e.g. Spearman, Pearson, or Kendall Tau correlation methods at system level or segment level [112].

This platform can also be used to create new resources. For example, the step-II MWE question where we ask if the translation uses alternative MWEs and if so then which, here we can set a further storing option to save the alternative MWEs that are correct translations of the source MWEs. We can also store plain phrases that are correct translations of source MWEs. In this way, we generate more bilingual parallel MWE terms, including paraphrasing at single side at MWE level. These resources can be important linguistic driven knowledge base features for popular automatic evaluation metrics such as METEOR, which depends on high quality paraphrase data to achieve better evaluations. From the translation modelling perspective, the extracted and stored multilingual paraphrased MWEs can be integrated into MT modelling learning and translating to generate alternative high quality translation outputs with lexical diversity. Furthermore, paraphrase databases are widely used in NLP communities in different tasks such as natural language inference, natural language understanding, text entailment, searching, etc.

5.3.2 HiLMeMe Implementation and Platform

HiLMeMe is implemented via the PsychoPy3 platform, relying on Python3 packages. PsychoPy[6] has been a popular platform for researchers to carry out experiments, especially in the situations where human interactive or assessments are needed. These

[6] https://www.psychopy.org

```
HiLMeMe: Algorithms

Input: src(i) as source sentence(i), tar(i) as candidate MT ouput for target
sentence, ref(i) as reference translation of src(i), MWE_src(i.n) as MWEs in
src(i), tran_MWE(i.m) as translations of MWE_src(i.n), MWE_ref(i.p) as MWEs in
ref(i), alt_MWEs(i.q) as alternative correct target MWEs not included in
MWE_ref(i.p), non_MWEs(i) as common words and correct translation but not using
MWEs in tar(i), NULL indecating MWE_src(i.n) not transalted or lost in tar(i).

if tar(i) matches ref(i) in any degrees
    Point(i.I) = General(fluency, adequacy)
    Point(i.II) = MWE(α, β, γ, θ) where
       α++ if tran_MWE(i.m) = MWE_ref(i.p)
       β++ if tran_MWE(i.m) = alt_MWEs(i.q)
       γ++ if tran_MWE(i.m) = non_MWEs(i)
       θ++ if tran_MWE(i.m) = NULL
    Weight(i) = Φ(Sem,Gra,Idi,Amb,ϕ) where
       Sem++ if (MWE_src(i.n), MWE_ref(i.p)) meet semantics factor
       Gra++ if (MWE_src(i.n), MWE_ref(i.p)) meet grammar factor
       Idi++ if (MWE_src(i.n), MWE_ref(i.p)) meet idiomaticity factor
       Amb++ if (MWE_src(i.n), MWE_ref(i.p)) meet ambiguity factor
       ϕ = the value of MWE_src(i.n) weighting for src(i)

HiLMeME(i) = Point(i.I) + Weight(i) x Point(i.II)
```

Figure 5.13: HiLMeMe Algorithms

behavioural sciences include neuroscience, psychology, psychophysics, and linguistics. It can easily accommodate our human-in-the-loop evaluation methodology by offering a straightforward interface and storing all the classification data during the assessments. The HiLMeMe initial PsychoPy3 platform is available and will be open source and publicly available at git.HiLMeMe.

In the implemented platform, we designed the following HiLMeMe workflow as shown in Figure 5.14: with the following sequence of steps ... consent form -> task introduction -> sample practice with three questions -> real assessment of MT results -> stored assessment data.

5.3.3 Summary of HiLMeMe Evaluation Methodology

To achieve more reliable MT quality assessment and advance the state-of-the-art in MT modelling, we designed a new evaluation methodology, having a human-in-the-loop and looking specifically into MWEs. We introduced three-step assessment models with corresponding questions and error classifications. We presented the

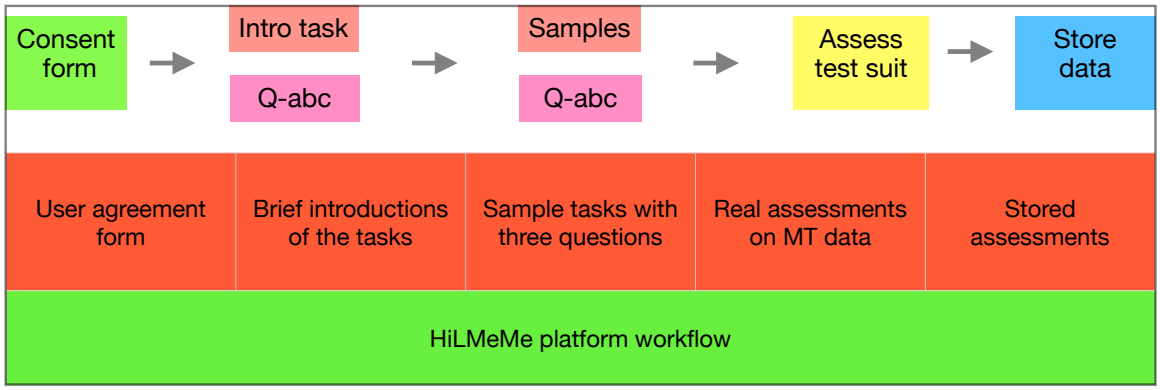

Figure 5.14: HiLMeMe Platform Workflow

scoring functions based on the three main questions and the implemented platform. We discussed the potential impact of HiLMeMe and its output data.

HiLMeMe is based on our multilingual parallel corpus AlphaMWE. We expect this new TQA method to reflect the differences of SOTA MT models in performance towards human parity, e.g. translation in idiomatic and ambiguous situations. We will open source our platform to MT researchers and the MT evaluation community for better evaluating their MT models, and for better correlating their evaluation metrics with expert evaluations.

5.4 Chapter Summary

In this chapter we first presented a brief critique on the issues in MT quality assessment and especially the corpus issues (e.g. corpus sanity), as we had discussed earlier in the MT evaluation and MT challenges Sections 2.2 and 2.3.

Subsequently, we carried out the corpus construction work of AlphaMWE, a multilingual parallel corpus with MWE annotations, where we introduced the motivation, related corpus work, and corpus construction workflow, as well as extensive error categorisations we observed on translating MWEs and MWEs related context from state of the art MT models.

Finally, we proposed a new evaluation methodology, human-in-the-loop MT eval-

uation looking into MWEs, which we call as HiLMeMe. We described the three-step evaluations with corresponding questions designed, and the overall scoring formula. We also discussed the value of HiLMeMe outputs, such as new multilingual MWE corpus construction, high quality paraphrase database, as well as the detailed error classification of MT models which can be used to advise MT researchers to further improve the MT models.

This chapter further addressed our research hypotheses (*Hy.c*) on the improvement of current MT evaluation models and corresponding MT test suites in practice by looking into MWEs and using MWEs as an important factor. We also fully investigated our **RQ-II** *to-do-list*, by revisiting the MT assessment issues, creating a new corpus with MWE annotations, and developing a human-in-the-loop new evaluation method and open source platforms looking into MWEs.

Chapter 6

Conclusions

"What can be said at all can be said clearly, and what we cannot talk about we must pass over in silence."

–L. Wittgenstein (1918, Vienna)

We first (in Section 6.1) revisit the Hypotheses and RQ-s that we proposed for this thesis, having a review of how we addressed them with our experimental investigations, to what levels, and what are our findings. Then, in Section 6.2, we conclude our thesis with discussions and future work.

6.1 Revisiting Hypotheses and Research Questions (RQs)

In each chapter we revisited how our *Hypotheses* and *RQ-s* were proposed and addressed in the thesis.

In Chapter 1, we firstly introduced the background of computer development and its application in various AI fields, such as robots, speech recognition, computer vision, gaming, as well as automatic processing of human languages (c.f. NLP). Subsequently, we introduced the first topic of the thesis focus, i.e. machine translation (MT), as one of the main tasks in NLP, and the broad applications of MT.

In Chapter 2, we introduced the history of the development of MT from rule-based methods since the 1950s to neural models in the mid-2010s and afterwards. We presented various MT evaluation methods from human manual work and automatic categories, and the strategies for evaluation methods (meta-eval). Then, we discussed the broad issues and challenges existing in current state-of-the-art MT models and in MT evaluation. Finally, we derived another research focus for the thesis, i.e. multi-word expressions (MWEs), as one of the bottlenecks in MT and given their broader influence on MT problems and challenges, where we listed illustrative examples.

In Chapter 3, we presented another literature review on MWE investigations in MT, from both statistical MT and neural MT paradigms. We revisited the MT challenges and problems in Chapter 2, gave a critical analysis on current MWE investigations in MT, and proposed our hypotheses and research questions for this thesis, i.e., the investigation into MWEs in MT from both MT modelling and MT evaluation perspectives.

In Chapter 4, we firstly carried out a pilot study investigating one of the very recent published models on MWE integration into neural MT, where we examined different language pairs and extended the training set into a rich-resourced and high-performance scenario. We carried out a critical analysis on the deployed model, and proposed a very different method to address MWE problems in MT as a rare words issue from the technical point of view. We then carried out two-step investigations, with the first one incorporating partial Chinese characters (semantic radicals) as a linguistic feature in an attention based BiRNN neural model expecting this can enrich the model learning to tackle rare words and MWE translations. In the second step investigations we replaced the neural MT input from words and characters into decomposed character sequences in various degrees in a new state-of-the-art Transformer structure NMT model. Furthermore we used the bilingual MWE terms (BiMWEs) extracted in the pilot study as a knowledge feature for our second step experiments to examine their interactive functions, and the two-phases expert vali-

dations demonstrated the *positive engagement* of these two models in a combination setting.

Both the pilot study and the two-step character decomposition experiments are an investigation into our hypotheses from the MWE for MT model learning point of view, i.e., *Hy.a* and *Hy.b*, which followed the task list we designed for **RQ-I**. Our experimental results verify the first half of our hypotheses that addressing MWE issues can improve MT quality from the adequacy and semantic representation aspects, thus to improve overall MT performance.

In chapter 5, we further investigated our hypotheses, *Hy.c*, on the aspect of improving MT evaluation (or quality assessment) using MWEs. We followed the investigation list in **RQ-II**, by critique analysis on current TQA methods and corpus issues, followed by the creation of a new multilingual parallel corpus with MWE annotations where we made an extensive error classification when MT meets MWEs. Subsequently, based on our new corpus, we proposed a human-in-the-loop MT evaluation model HiLMeMe looking into MWEs, and we analysed the potential impact of the developed toolkit in a broader view. The creation of AlphaMWE and the HiLMeMe evaluation methodology addressed the second half of our hypotheses and research questions, i.e. the introduction of MWEs as semantic knowledge or feature set can be used to improve current MT evaluation methods in practice.

6.2 Conclusions and Future Work

In conclusion, among all the fields in AI and NLP, this thesis made one investigation into MWEs in MT. This was motivated by the survey of state-of-the-art MT research and its evaluations and critique analysis on current problems and challenges as presented in Chapter 2, Section 2.3. MWEs are one of the bottlenecks in MT (and NLP) due to their very common usage, low frequency, ambiguity and idiomaticity. We addressed our hypotheses on investigating MWEs from both the MT modelling and MT evaluation points of view in Chapter 4 and Chapter 5 following RQ-I and RQ-II respectively.

On the one hand, we re-examined one of the very recent MWE integration models into neural MT as a pilot study with larger sizes of training corpus and different language pairs. We carried out a critical analysis based on our findings and proposed to tackle MWE translations from a new angle as a low-frequency words challenge and corresponding solutions focusing on the Chinese language, by decomposing Chinese characters into character-pieces as neural MT sequential inputs. The first decomposition model extracted Chinese radicals as feature set to enhance character-sequence NMT learning, while the second step investigation decomposed the Chinese characters into different degrees gradually and replaced the character-sequence input. The bilingual MWEs extracted in the pilot study were used as a new feature set to enhance the step two decomposition models.

On the other hand, based on the critique of MT evaluation as used in current practice, we created a new resource i.e. a multilingual parallel test suite with MWE annotations, and designed a new evaluation methodology with a human in the loop and looking into MWEs as an important assessment factor. This plays a role in MT error analysis and assessing MT development towards human parity.

All the tool-kits and resources we developed along with the thesis investigation are open-source and publicly available to the NLP community, including the Chinese character decomposition tools (radical4mt), automatically extracted bilingual MWE terms from WMT data (MultiMWE, German-English and Chinese-English) in our pilot study and decomposed Chinese-English bilingual MWE terms from decomposition models, created multilingual parallel corpus with manual post-editing and annotation of MWEs (AlphaMWE, covering English, Chinese, German, Polish, and Italian), and the new MT evaluation models and platform using human-in-the-loop looking into MWEs (HiLMeMe, language independent). These open source resources can be used by both MT researchers, to further investigate MWE translations, and other NLP task practitioners, to use as a knowledge base.

In the future work, we have interests in further investigation of our proposed decomposition models as well as the usage of Chinese character decomposition for

MT evaluations. We plan to carry out human assessment on our step-I decomposition model where we applied Chinese radicals as a feature set to enhance NMT learning. We also plan to further analyse the step-II decomposition model output using multi-referenced test suites for automatic evaluation and sampling part of the test suites for carrying out expert validations in scale.

The training corpus we used for decomposition model investigations is very large using 5 million sentence pairs (compared to 1.25 million in step-I), which could have a 100% coverage of the total vocabularies in the development and test set (compared to 97.7 and 99.3 coverage in step-I). In future work we plan to further investigate the model performance in lower resource settings to address rare words and phrases, as well as alternative test set usage (currently 2,000 test sentences). The 1.25 million corpus of the step-I setting covered most of the testing words.

Chinese radicals preserve the meaning part of Chinese characters, and from this point of view we expect they can be applied in MT evaluation (e.g. metrics, Section 2.2.3) as a semantic feature set to improve automated evaluation accuracy.

In the event that automatic metrics can have an improved correlation to experts' assessment by using multiple high quality and lexically diverse references, we plan to create such resources and test suite based on AlphaMWE using paraphrase technology.

Regarding the HiLMeMe MT evaluation model, we plan to further empower this methodology and open-source platform by extensive experimental exploration and testing, such as its usage for our own model evaluations from the pilot study and proposed decomposition methods.

We also have interests in investigating other MT problems and challenges that we discussed in Section 2.3 that are not addressed in this thesis. In connection to the second and third examples on Chinese-English MT in Section 2.4 Figures 2.9 and 2.10 on poems and Wenyan (classic) Chinese, we are interested in working on these two challenging domains, which is more difficult than the translation of commonly spoken Mandarin (Baihua, Putonghua), modern Chinese.

Another topic we want to carry out experimental investigation into is the combined neural model on MWE identification and MT. As the neural network leanings have been successfully applied to both these two fields, it is highly possible to explore the possibility of integrating these two tasks together into one neural model, with further references [98, 109].

Appendices

Appendix A

Further Details on NN and NMT Models

Some further details on Neural Networks (NNs) and Neural MT (NMT) background models are introduced in Appendix A.1. We also introduce some helpful knowledge on the Soochow University NMT (or SuDaNMT) systems in 2017 that we used for the Step-I investigation for decomposition model in Appendix A.2.

A.1 NMT structures

First, let us see the encoder-decoder NMT models in a mathematical way as follows. The first NMT architecture used convolutional neural networks (CNNs) for the encoder and Recurrent NNs (RNNs) for the decoder [131]. Let $P(f|e)$ be the probability of a source sentence $e = e_i, ...e_k$ translating to a target sentence $f = f_1, ..., f_m$, then, $P(f|e)$ is calculated as:

$$P(f \mid e) = \Pi_{i=1}^{m} P(f_i | f_{1:i-1}, e) \tag{A.1}$$

where f_i means the ith word in the target sentence and $f_{1:i-1}$ means the the preceding words. Here we can see that the generation probability of each target word f_i is conditioned on both the source sentence e and the already generated preceding words

$f_{1:i-1}$.

The language model (decoding part) is displayed in Figure A.1, where I is an input vocabulary transformation, R as a recurrent transformation, and O as an output vocabulary transformation. The equations are displayed as:

$$h_1 = \sigma(I \cdot v(f_1)) \tag{A.2}$$

$$h_{i+1} = \sigma(R \cdot h_i + I \cdot v(f_{i+1})) \tag{A.3}$$

$$o_{i+1} = O \cdot h_i \tag{A.4}$$

where σ is a nonlinear function, and v is the fixed-dimensional representation of an input sequence.

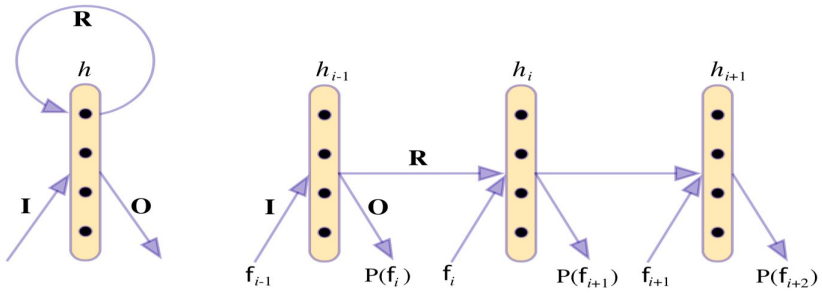

Figure A.1: Recurrent Continuous Translation Models (RCTM) and RLM model from [131].

The encoding was a convolutional NNs shown in Figure A.2 where K and L are denoted as weight matrices, and M as a generic matrix.

Because the CNNs encoding can not keep the ordering information of the source words, RNNs (especially LSTM and LSTM-like Gated Recurrent Units) were applied to replace the CNNs [263, 45], as shown in Figure A.3 which was a fully RNN encoder-decoder, although it still used fixed-dimensional vectors for sentence representations.

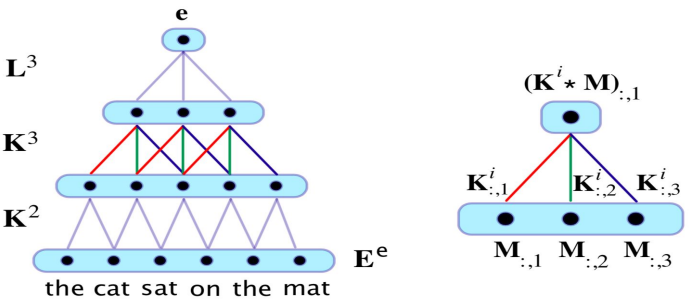

Figure A.2: Convolutional sentence embedding example from [131].

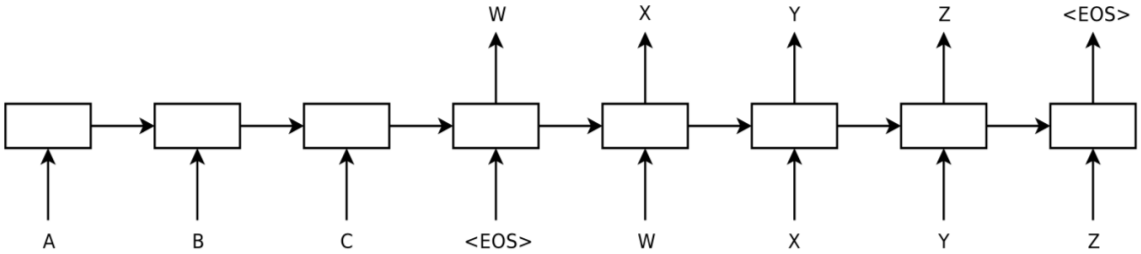

Figure A.3: Fully RNNs encoder-decoder NMT from [263].

$$h_t = u_t \odot \tilde{h}_t + (1 - u_t) \odot h_{t-1}$$
$$\tilde{h} = \tanh(W[x_t] + U(r_t \odot h_{t-1}) + b)$$
$$u_t = \sigma(W_u[x_t] + U_u h_{t-1} + b_u)$$
$$r_t = \sigma(W_r[x_t] + U_r h_{t-1} + b_r)$$

$$h_t = o_t \odot \tanh(c_t)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$\tilde{c}_t = \tanh(W_c[x_t] + U_c h_{t-1} + b_c)$$
$$o_t = \sigma(W_o[x_t] + U_o h_{t-1} + b_o)$$
$$i_t = \sigma(W_i[x_t] + U_i h_{t-1} + b_i)$$
$$f_t = \sigma(W_f[x_t] + U_f h_{t-1} + b_f)$$

Figure A.4: GRU (left) [45] vs LSTM (right) [122].

Compared with general RNNs, Gated Recurrent Unit (GRU) and LSTM have an advantage to address the long-sentence poor translation issue, because the gated parameter settings can selectively remember or forget some previous hidden state information (see Figure A.4)[1]. Hidden state h_t in GRU is like the cell c_t in LSTM. Both of these have forgetting gates u_t and f_t respectively. Then r_t and i_t are reset gates to decide how much previous information to keep and input gate respectively. LSTM has an additional output gate O_t compared with GRU.

Alignment information is introduced to the NMT structures in order to put more

[1]image credit: Prof Chis Manning, from Stanford NLP course

attention on the surrounding words of the source and target translation token. This is also designed to address long sentence translation, together changing the fixed-length (dimensions) vector to various lengths by [9], as illustrated in Figure A.5 where a as the alignment weights to reflect the importance of surrounding words. The encoder is updated as bi-directional RNNs (BiRNN), i.e., the annotation h_j contains the summary information of both the preceding words and the following words $[\overrightarrow{h_j}; \overleftarrow{h_j}]$.

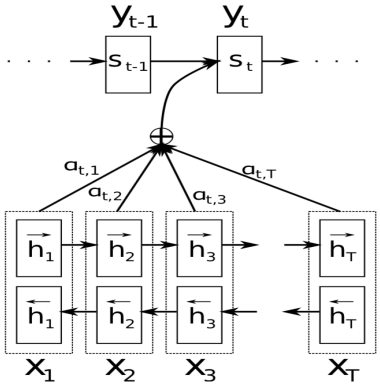

Figure A.5: BiRNN for encoder and RNN for decoding with attention (soft-alignment) from [9].

The alignment formula is described below with c_i as the context vector depending on a sequence of annotations $(h_1, ..., h_{T_x})$, a_{ij} as weights and e_{ij} as alignment model, s_{i-1} as hidden state before emitting y_i.

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{A.5}$$

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{A.6}$$

$$e_{ij} = a(s_{i-1,h_j}) \tag{A.7}$$

The decoder generates y_t with the probability formula (σ is a nonlinear function):

$$p(y_t \mid y_1, ..., y_{t-1}, c) = \sigma(y_{t-1}, s_t, c) \tag{A.8}$$

The later important change includes the introduction of pre-processing models BPE and Word-piece for OOV word translation on Western alphabet languages, the full-attention NMT without RNNs to reduce computational complexity and cost, and pre-trained language model BERT to pre-tune the model parameters for follow-up downstream sub-tasks. Here we recall one of our contributions in this thesis is a novel pre-processing model focusing on Chinese character decomposition and integrating this pre-processing model to an attention based NMT model to handle MWE translation.

For the attention based Transformer model, it includes three forms of attention: Encoder-Decoder Attention, Encoder Self-attention, and Masked Decoder Self-attention. Encoder-Decoder attention connects the encoder-decoder, Encoder self-attention connects all input information, and Masked decoder self-attention only attends some information just before current (a single matrix multiply with mask). In Transformer, every encoded word information is transferred in parallel with feed-forward nets so it reduced the computational perplexity compared with RNN models.

A.2 SuDaNMT structure applied in decomposition models

In Soochow University's NMT (SuDaNMT) model in 2017 [150, 111] the encoder apples a bidirectional RNN to encode a source sentence $x = (x_1, x_2, ..., x_{T_x})$ and repeatedly generates the hidden vectors $h = (h_1, h_2, ..., h_{T_x})$ over the source sentence, where T_x is the length of the source sentence. Formally, $h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}]$ is the concatenation of forward RNN hidden state $\overrightarrow{h_j}$ and backward RNN hidden state $\overleftarrow{h_j}$, and $\overrightarrow{h_j}$ can be computed as follows:

$$\overrightarrow{h_j} = f(\overrightarrow{h_{j-1}}, x_j) \tag{A.9}$$

where function f is defined as a Gated Recurrent Unit (GRU) [49].

The decoder is also an RNN that predicts the next word y_t given the context

vector c_t, the hidden state of the decoder s_t and the previous predicted word y_{t-1}.

$$p(y_t|y_{<t}, x) = softmax(g(s_t, y_{t-1}, c_t)) \tag{A.10}$$

where g is a non-linear function. and s_t is the state of decoder RNN at time step t, which is calculated by:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{A.11}$$

where c_t is the context represent vector of source sentence.

Usually c_t can be obtained by attention model and calculated as follows:

$$c_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j \tag{A.12}$$

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{T_x} e_{tk}} \tag{A.13}$$

$$e_{tj} = v_a^T tanh(s_{t-1}, h_j) \tag{A.14}$$

SuDaNMT also followed the implementation of attention-based NMT of the "dl4mt" tutorial [2], which enhances the attention model by feeding the previous word y_{t-1} to it, therefore the e_{tj} is calculated by:

$$e_{tj} = v_a^T tanh(\widetilde{s}_{t-1}, h_j) \tag{A.15}$$

where $\widetilde{s}_{t-1} = f(s_{t-1}, y_{t-1})$, and f is a GRU function. The hidden state of the decoder is updated as following:

$$s_t = f(\widetilde{s}_{t-1}, c_t) \tag{A.16}$$

[2] available on github.com/nyu-dl/dl4mt-tutorial/tree/master/ session2

Appendix B

Publications on Work from this Thesis

All the publications listed below have been peer-reviewed excepted for pre-prints in Section B.6, and wherever clearly stated "under-review".

B.1 Journal Article Publications/Submissions

1. "Towards a resource for multilingual lexicons: a multilingual parallel corpus with MWE annotation". 27 pages. Extended work on AlphaMWE. Submitted to Journal of Language Resource and Evaluation. Publisher: Springer. *Received first round review, conditional acceptance.*

2. "Chinese Character Decomposition for Multiword Expressions Aware Neural Machine Translation: an empirical investigation". Frontiers in Artificial Intelligence: Language and Computation. In Special issue: Word Senses and Multiword Expressions in the Lexicon. 22 pages. *Under-review.*

3. "Meta-Evaluation of Machine Translation Evaluation Methods" To be submitted. 20 pages. *Invited submission*, to Journal of Data Science and Informetrics (DSI). Extended work from Metrics2021 Workshop presentation, organised by SIG-MET: The ASIS&T Special Interest Group for Metrics.

B.2 Book Chapters Published

4. Moreau, Erwan, Alsulaimani, Ashjan, Maldonado, Alfredo , **Han, Lifeng**, Vogel, Carl and Dutta Chowdhury, Koel (2018) Semantic reranking of CRF label sequences for verbal multiword expression identification. In: Markantonatou, Stella, Ramisch, Carlos , Savary, Agata and Vincze, Veronika , (eds.)

Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop. Language Science Press, Berlin, pp. 177-207. ISBN 978-3-96110-124-5 https://hal.archives-ouvertes.fr/hal-01930987

B.3 Conference Paper Publications

5. *Maldonado, Alfredo* , ***Han, Lifeng***, *Moreau, Erwan* , Alsulaimani, Ashjan, Chowdhury, Koel, Vogel, Carl and Liu, Qun (2017) Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In: 13th Workshop on Multiword Expressions (MWE 2017)@EACL, pages 114–120, Apr 2017, Valencia, Spain. Publisher: Association for Computational Linguistics. (***three-main contributors***) https://aclanthology.org/W17-1715.pdf

6. ***Han, Lifeng*** *and Kuang, Shaohui* (2018) Incorporating Chinese radicals into neural machine translation: deeper than character level. In: Proceedings of the ESSLLI 2018 Student Session (Language and Computation), pp 54-65. 30th European Summer School in Logic, Language & Information. 6-17 Aug 2018, Sofia, Bulgaria. Publisher: Association for Logic, Language and Information (FoLLI). http://doras.dcu.ie/24732/

7. **Han, Lifeng**, Jones, Gareth J.F., and Smeaton, Alan F. (2020) MultiMWE: building a multi-lingual multi-word expression (MWE) parallel corpora. In: 12th International Conference on Language Resources and Evaluation (LREC), pages 2970–2979, 11-16 May, 2020, Marseille, France. (Virtual). Publisher: Association for Computational Linguistics & European Language Resource Association (ELRA). https://www.aclweb.org/anthology/2020.lrec-1.363

8. **Han, Lifeng**, Jones, Gareth J.F., and Smeaton, Alan F. (2020) AlphaMWE: construction of multilingual parallel corpora with MWE annotations. In: Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020)@COLING, Pages 44–57, 13 Dec 2020, Barcelona, Spain (Online). Publisher: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.mwe-1.6

9. **Han, Lifeng** , Jones, Gareth J.F., Smeaton, Alan F., and Bolzoni, Paolo (2021) Chinese Character Decomposition for Neural MT with Multi-Word Expressions. In NoDaLiDa2021: The 23rd Nordic Conference on Computational Linguistics, pages 336–344. Publisher: Association for Computational Linguistics. https://aclanthology.org/2021.nodalida-main.35

10. **Han, Lifeng** , Jones, Gareth J.F., and Smeaton, Alan F. (2021) Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. In Proceedings of MoTra21: Workshop on Modelling Translation: Translatology in the Digital Age @NoDaLiDa2021 - The 23rd Nordic Conference on Computational Linguistics. Pages 15–33. Publisher: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.motra-1.3.pdf

11. Erofeev, Gleb, Sorokina, Irina, **Han, Lifeng**, and Gladkoff, Serge (2021) cushLEPOR uses LABSE distilled knowledge to improve correlation with human translation evaluations. In Proceedings of Machine Translation Summit XVIII: Users and Providers Track. pages 421–439. August 16-20, Online. Publisher: Association for Computational Linguistics & AMTA. https://aclanthology.org/2021.mtsummit-up.28/

12. **Han, Lifeng**, Sorokina, Irina, Erofeev, Gleb, and Gladkoff, Serge (2021) cushLEPOR: Customised hLEPOR Using LABSE Distilled Knowledge Model to Improve Agreement with Human Judgements. In Proceedings of EMNLP 2021 Sixth Conference on Machine Translation - WMT2021 (Forthcoming). November 10-11, Online. Publisher: Association for Computational Linguistics. https://arxiv.org/abs/2108.09484

B.4 Conference Tutorials Presented/Accepted

13. **Han, Lifeng** (2021) Meta-Evaluation of Machine Translation Evaluation Methods. Tutorial Research Presentation in Metrics 2021: Workshop on Informetric and Scientometric Research (SIG-MET), Affiliated with 84th ASIS&T Annual Meeting. Association for Information Science and Technology. October 23-24. http://doras.dcu.ie/26280/

14. **Han, Lifeng** (2021) Meta-Evaluation of Translation Evaluation Methods: a systematic up-to-date overview. Tutorial Proposal accepted to LREC2022: The International Conference on Language Resources and Evaluation. (A half-day Main conference tutorial) https://lrec2022.lrec-conf.org/

B.5 Conference Abstract Presentations

15. **Han, Lifeng** (2018) Apply Chinese Radicals Into Neural Machine Translation: Deeper Than Character Level. Research Presentation In: LPRC - Limerick Postgraduate Research Conference, May 24, Limerick, Ireland. http://doras.dcu.ie/26279/

16. **Han, Lifeng** (2018) Machine translation evaluation resources and methods: a survey. Research Presentation In: IPRC - Irish Postgraduate Research Conference, 8-9 Nov 2018, Dublin, Ireland. http://doras.dcu.ie/24493/

17. Alekseeva, Alexandra, Gladkoff, Serge, Sorokina, Irina , and **Han, Lifeng** (2021) Monte Carlo Modelling of Confidence Intervals in Translation Quality Evaluation (TQE) and Post-Editing Distance (PED) Measurement. Research Presentation in Metrics 2021: Workshop on Informetric and Scientometric Research (SIG-MET). Affiliated with 84th ASIS&T Annual Meeting. Association for Information Science and Technology (**corresponding author**). October 23-24. http://doras.dcu.ie/26281/

B.6 *e*-Preprints

18. **Han, Lifeng** (2016) Machine Translation Evaluation: A Survey. arXiv:1605.04515v6 [cs.CL] https://arxiv.org/abs/1605.04515v6

19. Gladkoff, Serge, Sorokina, Irina , **Han, Lifeng**, and Alekseeva, Alexandra (2021) Measuring Uncertainty in Translation Quality Evaluation (TQE). arXiv:2111.07699 [cs.CL] (**corresponding author**) https://arxiv.org/abs/2111.07699

20. Gladkoff, Serge, and **Han, Lifeng** (2021) HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation. *arXiv:2112.13833 [cs.CL]* https://arxiv.org/abs/2112.13833.

Appendix C

Created Open-sourced Resources and Toolkits

C.1 Corpus Resources

C.1.1 MultiMWE

MultiMWE is a multilingual MWE corpus including automatically extracted parallel MWE glossaries from 5 millions parallel sentences in German-English and Chinese-English. The number of MWE pairs is 3,159,226 and 143,042 respectively for DE-EN and ZH-EN after filtering. The open-source website: https://github.com/poethan/MWE4MT.

C.1.2 AlphaMWE

AlphaMWE is a multilingual parallel corpus including 750 sentences each language with annotations on MWEs. The source English corpus was from PARSEME shared task on MWEs, and the other languages we offered covering Chinese, Polish, German, and Italian. We carried out automated translation plus manual post-editing and annotation by native speakers. AlphaMWE is maintained under the page: https://github.com/poethan/AlphaMWE.

C.2 Data-Processing Toolkits/Platforms

C.2.1 Chinese Character Decomposition Tools

The "radical4mt" toolkit can be used to decompose Chinese characters into different degrees from shallow to deeper, currently using the IDS dictionary files from CHISE (CHaracter Information Service Environment) project. This toolkit is available at: https://github.com/poethan/MWE4MT/tree/master/radical4mt.

C.2.2 HiLMeMe: Human-in-the-Loop MT Evaluation with MWEs

HiLMeMe initial platform is an open source toolkit for researchers carrying out human-in-the-loop MT evaluations which takes Multi-word Expressions as an impor-

tant assessment feature. It involves segment/sentence level translation fluency, adequacy, and the translation qualities of highlighted MWE terms. It is implemented via PsyChoPy3.0 and will be available at https://github.com/poethan/HilMeMe.

C.2.3 cushLEPOR: Customised hLEPOR Metric for MT Evaluation

Co-developed with LSP company Logrus Global, cushLEPOR takes advantage of huge language models, but uses them as a light weight fashion; also uses human evaluation labelled scores whenever they are available, tuning the automatic metric. Overall: towards high quality human level eval via auto-metric using optimised and distilled knowledge models! Two variants: cushLEPOR(LM): tuned on language model LaBSE (Language-agnostic BERT Sentence Embedding); cushLEPOR(pSQM): tuned on human professional annotated score labels (Scalar Quality Metric) using WMT20 data. cushLEPOR(LM) ranked *Cluster 1* in WMT2021 Metrics shared task on **En-De**, **Zh-En** and **En-Ru** in News domain data https://github.com/lHan87/cushLEPOR.

C.2.4 HOPE: Human-Centric Evaluation Framework Based on Professional Post-editing Annotations

HOPE is a human-centric and task-oriented MT evaluation framework, co-developed with LSP company Logrus Global. The experimental evaluation on Russian-English MT data demonstrated that HOPE is more efficient and effective on MT engine quality evaluation with transparent feedback. https://github.com/lHan87/HOPE.

References

[1] Roee Aharoni and Yoav Goldberg. "Towards String-to-Tree Neural Machine Translation". In: *CoRR* abs/1704.04743 (2017). URL: http://arxiv.org/abs/1704.04743.

[2] Joshua Albrecht and Rebecca Hwa. "A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 880–887. URL: https://aclanthology.org/P07-1111.

[3] James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, Wessel Kraaij, John Lafferty, Victor Lavrenko, David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal, Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu, and ChengXiang Zhai. "Challenges in Information Retrieval and Language Modeling: Report of a Workshop Held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002". In: *SIGIR Forum* 37.1 (Apr. 2003), pp. 31–47. ISSN: 0163-5840. DOI: 10.1145/945546.945549. URL: https://doi.org/10.1145/945546.945549.

[4] ALPAC. *Languages and machines: computers in translation and linguistics*. Tech. rep. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council., Feb. 1966, 124pp.

[5] Jon Androutsopoulos and Prodromos Malakasiotis. "A Survey of Paraphrasing and Textual Entailment Methods". In: *Journal of Artificial Intelligence Research* 38 (2010), pp. 135–187.

[6] Doug Arnold. *Computers and Translation: A translator's guide-Chap8 Why translation is difficult for computers*. Benjamins Translation Library, 2003.

[7] Eleftherios Avramidis, Maja Popovic, David Vilar, and Aljoscha Burchardt. "Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features". In: *Proceedings of WMT*. 2011.

[8] Bogdan Babych and Anthony Hartley. "Improving Machine Translation Quality with Automatic Named Entity Recognition". In: *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*. 2003. URL: https://aclanthology.org/W03-2201.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473 (2014). URL: http://arxiv.org/abs/1409.0473.

[10] Timothy Baldwin and Su Nam Kim. "Multiword Expressions". In: *Handbook of Natural Language Processing, Second Edition*. Taylor and Francis Group, 2010, pp. 267–292. URL: http://www.crcnetbase.com/doi/abs/10.1201/9781420085938-c12.

[11] Timothy Baldwin and Su Nam Kim. "Multiword Expressions". In: *Handbook of Natural Language Processing, Second Edition*. Chapman and Hall, 2010,

pp. 267–292. URL: http://www.crcnetbase.com/doi/abs/10.1201/9781420085938-c12.

[12] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In: *Proceedings of the ACL*. 2005.

[13] Srinivas Bangalore, Owen Rambow, and Steven Whittaker. "Evaluation metrics for generation". In: *Proceedings of INLG*. 2000.

[14] Regina Barzilayand and Lillian Lee. "Learning to paraphrase: an unsupervised approach using multiple-sequence alignment". In: *Proceedings NAACL*. 2003.

[15] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. "Graph Convolutional Encoders for Syntax-aware Neural Machine Translation". In: *arXiv preprint https://arxiv.org/abs/1704.04675* (2017).

[16] Yonatan Belinkov, Llúis Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. "Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 1–10. URL: https://www.aclweb.org/anthology/I17-1001.

[17] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. "A Neural Probabilistic Language Model". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944966.

[18] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. "Neural versus Phrase-Based Machine Translation Quality: a Case Study". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguis-

tics, Nov. 2016, pp. 257–267. DOI: 10.18653/v1/D16-1025. URL: https://aclanthology.org/D16-1025.

[19] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. "Re-evaluating Evaluation in Text Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9347–9359. DOI: 10.18653/v1/2020.emnlp-main.751. URL: https://www.aclweb.org/anthology/2020.emnlp-main.751.

[20] Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. "Findings of the 2013 Workshop on Statistical Machine Translation". In: *Proceedings of WMT*. 2013.

[21] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. "Findings of the 2017 Conference on Machine Translation (WMT17)". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 169–214. DOI: 10.18653/v1/W17-4717. URL: https://www.aclweb.org/anthology/W17-4717.

[22] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. "Findings of the 2018 Conference on Machine Translation (WMT18)". In: *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 272–307. URL: http://www.aclweb.org/anthology/W18-6401.

[23] Ondřej Bojar, Yvette Graham, and Amir Kamran. "Results of the WMT17 Metrics Shared Task". In: *Proceedings of the Second Conference on Machine*

Translation, Volume 2: Shared Tasks Papers. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017.

[24] Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. "Results of the WMT16 Metrics Shared Task". In: *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 199–231. URL: http://www.aclweb.org/anthology/W/W16/W16-2302.

[25] Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. "Identifying bilingual Multi-Word Expressions for Statistical Machine Translation". In: *Conference on Language Resources and Evaluation*. 2012.

[26] Pierrette Bouillon, Liliana Gaspar, Johanna Gerlach, Victoria Porro Rodriguez, and Johann Roturier. "Pre-editing by Forum Users: a Case Study". fre. In: Workshop (W2) on Controlled Natural Language Simplifying Language Use - 9th International Conference on Language Resources and Evaluation (LREC). ID: unige:39603. 2014, pp. 3–10. URL: https://archive-ouverte.unige.ch/unige:39603.

[27] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. "The Mathematics of Statistical Machine Translation: Parameter Estimation". In: *Computational Linguistics* 19.2 (1993), pp. 263–311. URL: https://www.aclweb.org/anthology/J93-2003.

[28] Christian Buck. "Black box features for the WMT 2012 quality estimation shared task". In: *Proceedings of WMT*. 2012.

[29] Kristine Bundgaard. "Translator Attitudes towards Translator-Computer Interaction - Findings from a Workplace Study". In: *HERMES - Journal of Language and Communication in Business* 56 (Oct. 2017), pp. 125–144. DOI: 10.7146/hjlcb.v0i56.97228. URL: https://tidsskrift.dk/her/article/view/97228.

[30] Ozan Caglayan, Loïc Barrault, and Fethi Bougares. "Multimodal Attention for Neural Machine Translation". In: *CoRR* abs/1609.03976 (2016). URL: http://arxiv.org/abs/1609.03976.

[31] Iacer Calixto, Qun Liu, and Nick Campbell. "Multilingual Multi-modal Embeddings for Natural Language Processing". In: *CoRR* abs/1702.01101 (2017). URL: http://arxiv.org/abs/1702.01101.

[32] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. "(Meta-) Evaluation of Machine Translation". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics. 2007, pp. 64–71.

[33] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. "Findings of the 2012 Workshop on Statistical Machine Translation". In: *Proceedings of WMT*. 2012.

[34] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. "Improved statistical machine translation using paraphrases". In: *Proceedings of HLT-NAACL*. 2006.

[35] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. "Re-evaluating the role of BLEU in machine translation research". In: *Proceedings of EACL*. Vol. 2006. 2006, pp. 249–256.

[36] Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. "cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[37] Michael Carl and Andy Way. *Recent advances in example-based machine translation*. Vol. 21. Springer Science & Business Media, 2003.

[38] John B. Carroll. "An experiment in evaluating the quality of translation". In: *Mechanical Translation and Computational Linguistics* 9.3-4 (1966), pp. 67–75.

[39] Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Antonio Valerio Miceli Barone, and Maria Gialama. "A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators". In: *MT Summit 2017*. Sept. 2017.

[40] Julio Castillo and Paula Estrella. "Semantic Textual Similarity for MT evaluation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 52–58. URL: https://www.aclweb.org/anthology/W12-3103.

[41] Daniel Cer, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. "Phrasal: A Toolkit for Statistical Machine Translation with Facilities for Extraction and Incorporation of Arbitrary Model Features". In: *Proceedings of the NAACL HLT 2010 Demonstration Session*. HLT-DEMO '10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 9–12. URL: http://dl.acm.org/citation.cfm?id=1855450.1855453.

[42] Yee Seng Chan, Hwee Tou Ng, and David Chiang. "Word Sense Disambiguation Improves Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic*. 2007. URL: http://aclweb.org/anthology/P07-1005.

[43] Xinyun Chen, Chang Liu, and Dawn Song. "Tree-to-tree Neural Networks for Program Translation". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*. 2018, pp. 2552–2562. URL: http://papers.nips.cc/paper/7521-tree-to-tree-neural-networks-for-program-translation.

[44] David Chiang. "A Hierarchical Phrase-Based Model for Statistical Machine Translation". In: *Proceedings of the 43rd Annual Meeting of the Associa-*

tion for Computational Linguistics (ACL'05). Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 263–270. DOI: 10.3115/1219840.1219873. URL: https://www.aclweb.org/anthology/P05-1033.

[45] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". In: *CoRR* abs/1409.1259 (2014). arXiv: 1409.1259. URL: http://arxiv.org/abs/1409.1259.

[46] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014.

[47] Vishal Chowdhary and Scott Greenwood. "EMT: End To End Model Training for MSR Machine Translation". In: *Proceedings of the 1st Workshop on Data Management for End-to-End Machine Learning*. 2017. ISBN: 978-1-4503-5026-6. URL: https://www.microsoft.com/en-us/research/publication/emt-end-end-model-training-msr-machine-translation/.

[48] Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. "A Character-level Decoder without Explicit Segmentation for Neural Machine Translation". In: *ACL*. 2016.

[49] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *Presented in NIPS 2014 Deep Learning and Representation Learning Workshop* (2014).

[50] Kenneth Church and Eduard Hovy. "Good Applications for Crummy Machine Translation". In: *Proceedings of the Natural Language Processing Systems Evaluation Workshop*. 1991.

[51] Jasob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.

[52] Elisabet Comelles, Jordi Atserias, Victoria Arranz, and Irene Castellón. "VERTa: Linguistic Features in MT Evaluation". In: *LREC*. 2012, pp. 3944–3950.

[53] Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. "Survey: Multiword Expression Processing: A Survey". In: *Computational Linguistics* 43.4 (2017), pp. 837–892. DOI: 10.1162/COLI_a_00302. URL: https://www.aclweb.org/anthology/J17-4005.

[54] Marta R. Costa-jussà and Adrià de Jorge. "Fine-tuning Neural Machine Translation on Gender-Balanced Datasets". In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 26–34. URL: https://www.aclweb.org/anthology/2020.gebnlp-1.3.

[55] Ido Dagan and Oren Glickman. "Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability". In: *Learning Methods for Text Understanding and Mining workshop*. 2004.

[56] Ido Dagan, Oren Glickman, and Bernardo Magnini. "The PASCAL recognising textual entailment challenge". In: *Machine Learning Challenges:LNCS* 3944 (2006), pp. 177–190.

[57] Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. "TESLA at WMT2011: Translation evaluation and tunable metric". In: *Proceedings of WMT*. 2011.

[58] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. "Learning where to Attend with Deep Architectures for Image Tracking". In: *CoRR* abs/1109.3737 (2011). URL: http://arxiv.org/abs/1109.3737.

[59] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[60] Shuoyang Ding, Hainan Xu, and Philipp Koehn. "Saliency-driven Word Alignment Interpretation for Neural Machine Translation". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 1–12. DOI: 10.18653/v1/W19-5201. URL: https://www.aclweb.org/anthology/W19-5201.

[61] Kaja Dobrovoljc. "Multi-word discourse markers and their corpus-driven identification: The case of MWDM extraction from the reference corpus of spoken Slovene". In: *International Journal of Corpus Linguistics* 22.4 (2017), pp. 551–582. ISSN: 1384-6655. DOI: https://doi.org/10.1075/ijcl.16127.dob. URL: https://www.jbe-platform.com/content/journals/10.1075/ijcl.16127.dob.

[62] George Doddington. "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics". In: *HLT Proceedings*. 2002.

[63] Stephen Doherty. "Issues in human and automatic translation quality assessment". In: ed. by D. Kenny. Human issues in translation technology. Routledge, 2017, pp. 131–148. ISBN: 9781315648934.

[64] Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. "Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition". In: *NLPCC/ICCPOL*. 2016.

[65] Bonnie Dorr, Matt Snover, and etc. Nitin Madnani. "Part 5: Machine Translation Evaluation". In: *Bonnie Dorr edited DARPA GALE program report*. 2009.

[66] Jennifer B. Doyon, John S. White, and Kathryn B. Taylor. "Task-Based Evaluation for Machine Translation". In: *Proceedings of MT Summit 7*. 1999.

[67] Quan Duong, Mika Hämäläinen, and Simon Hengchen. "An Unsupervised method for OCR Post-Correction and Spelling Normalisation for Finnish". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics*

(NoDaLiDa). Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 240–248. URL: https://aclanthology.org/2021.nodalida-main.24.

[68] Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. "Cdec: A Decoder, Alignment, and Learning Framework for Finite-state and Context-free Translation Models". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics2010 System Demonstrations*. ACLDemos '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 7–12. URL: http://dl.acm.org/citation.cfm?id=1858933.1858935.

[69] Sara Ebrahim, Doaa Hegazy, Mostafa Gadal-Haqq M. Mostafa, and Samhaa R. El-Beltagy. "Detecting and Integrating Multiword Expression into English-Arabic Statistical Machine Translation". In: *Procedia Computer Science* 117 (2017). Arabic Computational Linguistics, pp. 111–118. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.10.099. URL: http://www.sciencedirect.com/science/article/pii/S1877050917321567.

[70] Hiroshi Echizen-ya and Kenji Araki. "Automatic Evaluation Method for Machine Translation Using Noun-Phrase Chunking". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 108–117. URL: https://aclanthology.org/P10-1012.

[71] Desmond Elliott and Ákos Kádár. "Imagination improves Multimodal Translation". In: *CoRR* abs/1705.04350 (2017). URL: http://arxiv.org/abs/1705.04350.

[72] Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. "Tree-to-Sequence Attentional Neural Machine Translation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*

2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. 2016. URL: http://aclweb.org/anthology/P/P16/P16-1078.pdf.

[73] Gleb Erofeev, Irina Sorokina, Lifeng Han, and Serge Gladkoff. "cushLEPOR uses LABSE distilled knowledge to improve correlation with human translations". In: *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*. Virtual: Association for Machine Translation in the Americas, Aug. 2021, pp. 421–439. URL: https://aclanthology.org/2021.mtsummit-up.28.

[74] Project EuroMatrix. "1.3: Survey of Machine Translation Evaluation". In: *EuroMatrix Project Report, Statistical and Hybrid MT between All European Languages, co-ordinator: Prof. Hans Uszkoreit*. 2007.

[75] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. "Unsupervised Quality Estimation for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 539–555. DOI: 10.1162/tacl_a_00330. URL: https://www.aclweb.org/anthology/2020.tacl-1.35.

[76] Markus Freitag, Isaac Caswell, and Scott Roy. "APE at Scale and Its Implications on MT Evaluation Biases". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 34–44. DOI: 10.18653/v1/W19-5204. URL: https://www.aclweb.org/anthology/W19-5204.

[77] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation". In: *arXiv e-prints*, arXiv:2104.14478 (Apr. 2021), arXiv:2104.14478. arXiv: 2104.14478 [cs.CL].

[78] Markus Freitag, David Grangier, and Isaac Caswell. "BLEU might be Guilty but References are not Innocent". In: *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020, pp. 61–71. DOI: 10.18653/v1/2020.emnlp-main.5. URL: https://www.aclweb.org/anthology/2020.emnlp-main.5.

[79] Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. "Ensemble Distillation for Neural Machine Translation". In: *CoRR* abs/1702.01802 (2017). arXiv: 1702.01802. URL: http://arxiv.org/abs/1702.01802.

[80] Michael Gamon, Anthony Aue, and Martine Smets. "Sentence-level MT evaluation without reference translations Beyond language modelling". In: *Proceedings of EAMT*. 2005, pp. 103–112.

[81] Marcos Garcia. "Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Ed. by Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze. Berlin: Language Science Press., 2018, pp. 319–342. DOI: 10.5281/zenodo.1469571.

[82] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt,

Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. "The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics". In: *arXiv e-prints*, arXiv:2102.01672 (Feb. 2021), arXiv:2102.01672. arXiv: 2102.01672 [cs.CL].

[83] Jesús Giméne and Lluís Márquez. "A Smorgasbord of Features for Automatic MT Evaluation". In: *Proceedings of WMT*. 2008, pp. 195–198.

[84] Jesús Giménez and Lluís Márquez. "Linguistic Features for Automatic Evaluation of Heterogenous MT Systems". In: *Proceedings of WMT*. 2007.

[85] Serge Gladkoff and Lifeng Han. *HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation*. 2021. arXiv: 2112.13833 [cs.CL].

[86] Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. "Measuring Uncertainty in Translation Quality Evaluation (TQE)". In: *CoRR* abs/2111.07699 (2021). arXiv: 2111.07699. URL: https://arxiv.org/abs/2111.07699.

[87] Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. "Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings". In: *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*. 2017, pp. 11–19. URL: https://aclanthology.info/papers/W17-4702/w17-4702.

[88] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[89] Yvette Graham, Timothy Baldwin, and Nitika Mathur. "Accurate Evaluation of Segment-level Machine Translation Metrics". In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. 2015, pp. 1183–1191. URL: http://aclweb.org/anthology/N/N15/N15-1124.pdf.

[90] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. "Can machine translation systems be evaluated by the crowd alone". In: *Natural Language Engineering* 23.1 (2017), pp. 3–30. DOI: 10.1017/S1351324915000339.

[91] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. "Continuous Measurement Scales in Human Evaluation of Machine Translation". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 33–41. URL: https://www.aclweb.org/anthology/W13-2305.

[92] Yvette Graham, Barry Haddow, and Philipp Koehn. "Statistical Power and Translationese in Machine Translation Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 72–81. DOI: 10.18653/v1/2020.emnlp-main.6. URL: https://www.aclweb.org/anthology/2020.emnlp-main.6.

[93] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. "Universal Neural Machine Translation for Extremely Low Resource Languages". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 344–354. DOI: 10.18653/v1/N18-1032. URL: https://www.aclweb.org/anthology/N18-1032.

[94] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. "Named entity recognition in query". In: *Proceeding of SIGIR*. 2009.

[95] Rohit Gupta, Constantin Orasan, and Josef van Genabith. "ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks". In: *Proceedings of the 2015 Conference on Emperical*

Methods in Natural Language Processing. Association for Computational Linguistics. o.A., 2015, pp. 1066–1072.

[96] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. "Pairwise Neural Machine Translation Evaluation". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL'15)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 805–814. URL: http://www.aclweb.org/anthology/P15-1078.

[97] Lovisa Hagström and Richard Johansson. "Knowledge Distillation for Swedish NER models: A Search for Performance and Efficiency". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 124–134. URL: https://www.aclweb.org/anthology/2021.nodalida-main.13.

[98] Ali Hakimi Parizi and Paul Cook. "Do Character-Level Neural Network Language Models Capture Knowledge of Multiword Expression Compositionality?" In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 185–192. URL: http://aclweb.org/anthology/W18-4920.

[99] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. ISBN-10: 0471179124. Wiley-Interscience; 1 edition, 1998.

[100] Aaron L-F Han, Derek F Wong, and Lidia S Chao. "Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics". In: *Language Processing and Intelligent Information Systems*. Springer, 2013, pp. 57–68.

[101] Aaron L. -F. Han, Derek F. Wong, Lidia S. Chao, and Liangye He. "Automatic Machine Translation Evaluation with Part-of-Speech Information". In: *Text, Speech, and Dialogue*. Ed. by Ivan Habernal and Václav Matoušek. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 121–128. ISBN: 978-3-642-40585-3.

[102] Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. "LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors". In: *Proceedings of COLING 2012: Posters*. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 441–450. URL: https://aclanthology.org/C12-2044.

[103] Aaron Li-Feng Han, Yi Lu, Derek F. Wong, Lidia S. Chao, Liangye He, and Junwen Xing. "Quality Estimation for Machine Translation Using the Joint Method of Evaluation Criteria and Statistical Modeling". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 365–372. URL: https://aclanthology.org/W13-2245.

[104] Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Yi Lu, Liangye He, and Liang Tian. "A Universal Phrase Tagset for Multilingual Treebanks". In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Ed. by Maosong Sun, Yang Liu, and Jun Zhao. Cham: Springer International Publishing, 2014, pp. 247–258. ISBN: 978-3-319-12277-9.

[105] Li Feng Han, Fai Wong, Sam Chao, Yi Lu, Liang Ye He, Yi Ming Wang, and Jia Ji Zhou. "A description of tunable machine translation evaluation systems in WMT13 metrics task". In: *The eighth workshop on statistical machine translation, Aug. 2013, p. 414-421*. 2013.

[106] Lifeng Han. "LEPOR: An Augmented Machine Translation Evaluation Metric". University of Macau, Macao, 2014.

[107] Lifeng Han. "Machine Translation Evaluation Resources and Methods: A Survey". In: *arXiv e-prints*, arXiv:1605.04515 (May 2016), arXiv:1605.04515. arXiv: 1605.04515 [cs.CL].

[108] Lifeng Han, Gareth Jones, and Alan Smeaton. "AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations". In: *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. online: Association for Computational Linguistics, Dec. 2020, pp. 44–57. URL: https://www.aclweb.org/anthology/2020.mwe-1.6.

[109] Lifeng Han, Gareth Jones, and Alan Smeaton. "MultiMWE: Building a Multilingual Multi-Word Expression (MWE) Parallel Corpora". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 2970–2979. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.363.

[110] Lifeng Han, Gareth Jones, Alan Smeaton, and Paolo Bolzoni. "Chinese Character Decomposition for Neural MT with Multi-Word Expressions". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 336–344. URL: https://aclanthology.org/2021.nodalida-main.35.

[111] Lifeng Han and Shaohui Kuang. "Incorperating Chinese Radicals Into Neural Machine Translation: Deeper Than Character Level". In: *Proceedings of ESSLLI-2018* abs/1805.01565 (2018). arXiv: 1805.01565. URL: http://arxiv.org/abs/1805.01565.

[112] Lifeng Han, Alan Smeaton, and Gareth Jones. "Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods". In: *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital*

Age. online: Association for Computational Linguistics, May 2021, pp. 15–33. URL: https://www.aclweb.org/anthology/2021.motra-1.3.

[113] Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. "cushLEPOR: customising hLEPOR metric using Optuna for higher agreement with human judgments or pre-trained language model LaBSE". In: *Proceedings of Six Conference on Machine Translation (WMT2021)*. Association for Computational Linguistics, 2021. URL: https://github.com/poethan/cushLEPOR.

[114] Lifeng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. "Language-independent Model for Machine Translation Evaluation with Reinforced Factors". In: *Machine Translation Summit XIV*. International Association for Machine Translation. 2013, pp. 215–222.

[115] Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, Shuo Li, and Ling Zhu. "Phrase Tagset Mapping for French and English Treebanks and Its Application in Machine Translation Evaluation". In: *International Conference of the German Society for Computational Linguistics and Language Technology, LNAI Vol. 8105*. 2013, pp. 119–131.

[116] Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, and Yi Lu. "Unsupervised Quality Estimation Model for English to German Translation and Its Application in Extensive Supervised Evaluation". In: *The Scientific World Journal. Issue: Recent Advances in Information Technology*. 2014, pp. 1–12.

[117] Lifeng Han, Xiaodong Zeng, Derek F Wong, and Lidia S Chao. "Chinese Named Entity Recognition with Graph-based Semi-supervised Learning Model". In: *SIGHAN workshop with ACL-IJCNLP 2015*. 2015, p. 15.

[118] Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. "Multiword units in machine translation and translation technology". In: *Machine Translation* (Aug. 2019). ISSN: 1573-0573. DOI: 10.1007/s10590-019-09239-4. URL: https://doi.org/10.1007/s10590-019-09239-4.

[119] Hany Hassan, Anthony Aue, C. Chen, Vishal Chowdhary, J. Clark, C. Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, W. Lewis, M. Li, Shujie Liu, T. Liu, Renqian Luo, Arul Menezes, Tao Qin, F. Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and M. Zhou. "Achieving Human Parity on Automatic Chinese to English News Translation". In: *ArXiv* abs/1803.05567 (2018).

[120] Jindřich Helcl and Jindřich Libovický. "Neural Monkey: An Open-source Tool for Sequence Learning". In: *The Prague Bulletin of Mathematical Linguistics* 107.1 (2017), pp. 5–17. ISSN: 0032-6585. DOI: 10.1515/pralin-2017-0001. URL: http://ufal.mff.cuni.cz/pbml/107/art-helcl-libovicky.pdf.

[121] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012).

[122] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[123] Juliane House. "Translation Quality Assessment: Past and Present". In: *Translation: A Multidisciplinary Approach*. Ed. by Juliane House. London: Palgrave Macmillan UK, 2014, pp. 241–264. ISBN: 978-1-137-02548-7. DOI: 10.1057/9781137025487_13. URL: https://doi.org/10.1057/9781137025487_13.

[124] Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. "Attention-based Multimodal Neural Machine Translation". In: *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 639–645. DOI: 10.18653/v1/W16-2360. URL: http://aclanthology.coli.uni-saarland.de/pdf/W/W16/W16-2360.pdf.

[125] Matthias Hüning and Barbara Schlücker. "Multi-word expressions". In: *Müller et al. eds.: Word formation, An International Handbook of the Languages of Europe*. 1 (2015), pp. 450–467.

[126] John Hutchins. *Recent Applications of Machine Translation*. Oxford: The Oxford Handbook of Translation Studies, Edited by Kirsten Malmkjær, Kevin Windle, 2012.

[127] Shonosuke Ishiwatari, Jingtao Yao, Shujie Liu, Mu Li, Ming Zhou, Naoki Yoshinaga, Masaru Kitsuregawa, and Weijia Jia. "Chunk-based Decoder for Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1901–1912. DOI: 10.18653/v1/P17-1174. URL: https://www.aclweb.org/anthology/P17-1174.

[128] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. "On Using Very Large Target Vocabulary for Neural Machine Translation". In: *ACL 2015*. 2014.

[129] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". In: *CoRR* abs/1611.04558 (2016). URL: http://arxiv.org/abs/1611.04558.

[130] Gareth J. F. Jones. "An Introduction to Crowdsourcing for Language and Multimedia Technology Research". In: *Information Retrieval Meets Information Visualization: PROMISE Winter School 2012, Zinal, Switzerland, January 23-27, 2012, Revised Tutorial Lectures*. Ed. by Maristella Agosti, Nicola Ferro, Pamela Forner, Henning Müller, and Giuseppe Santucci. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 132–154. ISBN: 978-3-642-

36415-0. DOI: 10.1007/978-3-642-36415-0_9. URL: https://doi.org/10.1007/978-3-642-36415-0_9.

[131] Nal Kalchbrenner and Phil Blunsom. "Recurrent Continuous Translation Models". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, USA: Association for Computational Linguistics, Oct. 2013.

[132] Alina Karakanta, Atul Kr. Ojha, Chao-Hong Liu, Jonathan Washington, Nathaniel Oco, Surafel Melaku Lakew, Valentin Malykh, and Xiaobing Zhao, eds. *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. Dublin, Ireland: European Association for Machine Translation, Aug. 2019. URL: https://www.aclweb.org/anthology/W19-6800.

[133] Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. "Construction of Large-scale English Verbal Multiword Expression Annotated Corpus". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. ISBN: 979-10-95546-00-9.

[134] Yuanzhi Ke and Masafumi Hagiwara. "Radical-level Ideograph Encoder for RNN-based Sentiment Analysis of Chinese and Japanese". In: *CoRR* abs/1708.03312 (2017). arXiv: 1708.03312. URL: http://arxiv.org/abs/1708.03312.

[135] Maurice G. Kendall. "A new measure of rank correlation". In: *Biometrika* 30 (1938), pp. 81–93.

[136] Maurice G. Kendall and Jean Dickinson Gibbons. *Rank Correlation Methods*. New York: Oxford University Press, 1990.

[137] Dorothy Kenny. *Lexis and Creativity in Translation - A Corpus Based Approach*. Routledge, 2001. ISBN: 9781900650397.

[138] Dorothy Kenny. "Machine Translation". In: *The Routledge Handbook of Translation and Philosophy*. Ed. by P. Rawling and P. Wilson. Routledge, 2018, pp. 430–447. ISBN: 9781315678481. URL: https://doi.org/10.4324/9781315678481.

[139] Marrgaret King, Andrei Popescu-Belis, and Eduard Hovy. "FEMTI: Creating and Using a Framework for MT Evaluation". In: *Proceedings of the Machine Translation Summit IX*. 2003.

[140] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *CoRR* abs/1701.02810 (2017). URL: http://arxiv.org/abs/1701.02810.

[141] Philipp Koehn. "Europarl: A parallel corpus for statistical machine translation". In: *MT summit*. Vol. 5. 2005, pp. 79–86.

[142] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

[143] Philipp Koehn. "Statistical significance tests for machine translation evaluation". In: *Proceedings of EMNLP*. 2004.

[144] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of Conference on Association of Computational Linguistics*. 2007.

[145] Philipp Koehn and Rebecca Knowles. "Six Challenges for Neural Machine Translation". In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, Aug. 2017,

pp. 28–39. DOI: 10.18653/v1/W17-3204. URL: https://aclanthology.org/W17-3204.

[146] Philipp Koehn and Christof Monz. "Manual and Automatic Evaluation of Machine Translation between European Languages". In: *Proceedings of WMT*. 2006.

[147] Philipp Koehn, Franz Josef Och, and Daniel Marcu. "Statistical Phrase-based Translation". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. NAACL '03. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 48–54. DOI: 10.3115/1073445.1073462. URL: https://doi.org/10.3115/1073445.1073462.

[148] Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. "MT for Subtitling: Investigating professional translators' user experience and feedback". In: *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*. Virtual: Association for Machine Translation in the Americas, Oct. 2020, pp. 79–92. URL: https://www.aclweb.org/anthology/2020.amta-pemdt.6.

[149] Moshe Koppel and Noam Ordan. "Translationese and Its Dialects". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 1318–1326. URL: https://www.aclweb.org/anthology/P11-1132.

[150] Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong. "Attention Focusing for Neural Machine Translation by Bridging Source and Target Embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1767–

1776. DOI: 10.18653/v1/P18-1164. URL: https://www.aclweb.org/anthology/P18-1164.

[151] Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. "Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 596–606. URL: https://www.aclweb.org/anthology/C18-1050.

[152] Patrik Lambert and Rafael E. Banchs. "Data Inferred Multi-word Expressions for Statistical Machine Translation". In: *Proceedings of Machine Translation Summit X*. Phuket, Thailand, Sept. 2005, pp. 396–403. URL: https://hal.archives-ouvertes.fr/hal-00705007.

[153] Guillaume Lample and Alexis Conneau. "Cross-lingual Language Model Pretraining". In: *CoRR* abs/1901.07291 (2019). arXiv: 1901.07291. URL: http://arxiv.org/abs/1901.07291.

[154] J Richard Landis and Gary G. Koch. "The measurement of observer agreement for categorical data." In: *Biometrics* 33 1 (1977), pp. 159–74.

[155] Jamal Laoudi, Ra R. Tate, and Clare R. Voss. "Task-based MT Evaluation: From Who/When/Where Extraction to Event Understanding". In: *in Proceedings of LREC 06*. 2006, pp. 2048–2053.

[156] Mirella Lapata. "Probabilistic text structuring: Experiments with sentence ordering". In: *Proceedings of ACL*. 2003.

[157] Hugo Larochelle and Geoffrey E Hinton. "Learning to combine foveal glimpses with a third-order Boltzmann machine". In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta. Curran Associates, Inc., 2010, pp. 1243–1251. URL: http://papers.nips.cc/paper/4089-learning-to-combine-foveal-glimpses-with-a-third-order-boltzmann-machine.pdf.

[158] Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. "A Set of Recommendations for Assessing Human–Machine Parity in Language Translation". In: *Journal of Artificial Intelligence Research* 67 (Mar. 2020). ISSN: 1076-9757. DOI: 10.1613/jair.1.11371. URL: http://dx.doi.org/10.1613/jair.1.11371.

[159] Samuel Läubli, Rico Sennrich, and Martin Volk. "Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4791–4796. DOI: 10.18653/v1/D18-1512. URL: https://www.aclweb.org/anthology/D18-1512.

[160] Alon Lavie. "Automated Metrics for MT Evaluation". In: *Machine Translation* 11 (2013), p. 731.

[161] Guy Lebanon and John Lafferty. "Combining rankings using conditional probability models on permutations". In: *Proceeding of the ICML*. 2002.

[162] Sangmin-Michelle Lee. "The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis". In: *Computer Assisted Language Learning* 0.0 (2021), pp. 1–23. DOI: 10.1080/09588221.2021.1901745. eprint: https://www.tandfonline.com/doi/pdf/10.1080/09588221.2021.1901745. URL: https://www.tandfonline.com/doi/abs/10.1080/09588221.2021.1901745.

[163] Sabine Lehmann, Ben Gottesman, Robert Grabowski, Mayo Kudo, Siu Kei Pepe Lo, Melanie Siegel, and Frederik Fouvry. "Applying CNL Authoring Support to Improve Machine Translation of Forum Data". In: *Controlled Natural Language*. Ed. by Tobias Kuhn and Norbert E. Fuchs. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–10. ISBN: 978-3-642-32612-7.

[164] Gregor Leusch and Hermann Ney. "Edit distances with block movements and error rate confidence estimates". In: *Machine Translation* 23.2-3 (2009).

[165] Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang, and Guodong Zhou. "Modeling Source Syntax for Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 688–697. DOI: 10.18653/v1/P17-1064. URL: https://aclanthology.org/P17-1064.

[166] Liang You Li, Zheng Xian Gong, and Guo Dong Zhou. "Phrase-Based Evaluation for Machine Translation". In: *Proceedings of COLING*. 2012, pp. 663–672.

[167] Xiaoqing Li, Jinghui Yan, Jiajun Zhang, and Chengqing Zong. "Neural Name Translation Improves Neural Machine Translation". In: *Machine Translation*. Ed. by Jiajun Chen and Jiajun Zhang. Singapore: Springer Singapore, 2019, pp. 93–100. ISBN: 978-981-13-3083-4.

[168] Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Jonathan Weese, and Omar F. Zaidan. "Demonstration of Joshua: An Open Source Toolkit for Parsing-based Machine Translation". In: *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*. ACLDemos '09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 25–28. URL: http://dl.acm.org/citation.cfm?id=1667872.1667879.

[169] Pietro Liguori, Erfan Al-Hossami, Domenico Cotroneo, Roberto Natella, Bojan Cukic, and Samira Shaikh. "Shellcode_IA32: A Dataset for Automatic Shellcode Generation". In: *arXiv e-prints*, arXiv:2104.13100 (Apr. 2021), arXiv:2104.13100. arXiv: 2104.13100 [cs.SE].

[170] Chin-Yew Lin and E. H. Hovy. "Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics". In: *Proceedings NAACL*. 2003.

[171] Chin-Yew Lin and Franz Josef Och. "Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics". In: *Proceedings of ACL*. 2004.

[172] Ding Liu and Daniel Gildea. "Syntactic features for evaluation of machine translation". In: *Proceedingsof theACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 2005.

[173] Jia Liu and Yuqing Lu. "A Corpus-Based Comparative Study on Lexical Bundles in Native and Chinese Scholars'English Abstracts—Taking Linguistics and Chemistry as an Example". In: *Chinese Journal of Applied Linguistics* 42.4 (2019), pp. 488–502. DOI: doi:10.1515/CJAL-2019-0029. URL: https://doi.org/10.1515/CJAL-2019-0029.

[174] Zhiyuan Liu, Yankai Lin, and Maosong Sun. "Representation Learning for Natural Language Processing". In: *CoRR* abs/2102.03732 (2021). arXiv: 2102.03732. URL: https://arxiv.org/abs/2102.03732.

[175] Chi Kiu Lo and Dekai Wu. "MEANT: An Inexpensive, High- Accuracy, Semi-Automatic Metric for Evaluating Translation Utility based on Semantic Roles". In: *Proceedings of ACL*. 2011.

[176] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *CoRR* abs/1508.04025 (2015). URL: http://arxiv.org/abs/1508.04025.

[177] Qingsong Ma, Fandong Meng, Daqi Zheng, Mingxuan Wang, Yvette Graham, Wenbin Jiang, and Qun Liu. "MaxSD: A Neural Machine Translation Evaluation Metric Optimized by Maximizing Similarity Distance". In: *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*.

2016, pp. 153–161. DOI: 10.1007/978-3-319-50496-4_13. URL: http://dx.doi.org/10.1007/978-3-319-50496-4_13.

[178] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. "Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 62–90. DOI: 10.18653/v1/W19-5302. URL: https://www.aclweb.org/anthology/W19-5302.

[179] Matouš Macháček and Ondřej Bojar. "Results of the WMT13 Metrics Shared Task". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 45–51. URL: http://www.aclweb.org/anthology/W13-2202.

[180] Alfredo Maldonado, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel, and Qun Liu. "Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking". In: *The 13th Workshop on Multiword Expressions @ EACL 2017*. ACL. 2017.

[181] Inderjeet Mani. "Summarization Evaluation: An Overview". In: *NTCIR*. 2001.

[182] Katrin Marheinecke. "Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective". In: *Proceedings of the LREC 2016 Workshop Translation Evaluation –From Fragmented Tools and Data Sets to an Integrated Ecosystem, Georg Rehm, Aljoscha Burchardt et al. (eds.)* 2016. URL: http://www.cracking-the-language-barrier.eu/wp-content/uploads/Marheinecke.pdf.

[183] Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova, and Agata Savary, eds. *Proceedings of the Joint Workshop on Multiword Expressions and Electronic*

Lexicons. online: Association for Computational Linguistics, Dec. 2020. URL: https://www.aclweb.org/anthology/2020.mwe-1.0.

[184] Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, eds. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017. DOI: 10.18653/v1/W17-17. URL: https://www.aclweb.org/anthology/W17-1700.

[185] Lluís Màrquez. "automatic evaluation of machine translation quality". In: *Dialogue 2013 invited talk, extended* (2013).

[186] Elaine Marsh and Dennis Perzanowski. "MUC-7 Evaluation of IE Technology: Overview of Results". In: *Proceedingsof Message Understanding Conference (MUC-7)*. 1998.

[187] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. "A Survey on Document-Level Neural Machine Translation: Methods and Evaluation". In: *ACM Comput. Surv.* 54.2 (Mar. 2021). ISSN: 0360-0300. DOI: 10.1145/3441691. URL: https://doi.org/10.1145/3441691.

[188] Rebecca Marvin and Philipp Koehn. "Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems (Non-archival Extended Abstract)". In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Boston, MA: Association for Machine Translation in the Americas, 2018, pp. 125–131. URL: https://www.aclweb.org/anthology/W18-1812.

[189] Shaimaa Marzouk. "An in-depth analysis of the individual impact of controlled language rules on machine translation output: a mixed-methods approach". In: *Machine Translation* (2021). DOI: 10.1007/s10590-021-09266-0. URL: https://doi.org/10.1007/s10590-021-09266-0.

[190] Shaimaa Marzouk and Silvia Hansen-Schirra. "Evaluation of the Impact of Controlled Language on Neural Machine Translation Compared to Other

MT Architectures". In: *Machine Translation* 33.1–2 (June 2019), pp. 179–203. ISSN: 0922-6567. DOI: 10.1007/s10590-019-09233-w. URL: https://doi.org/10.1007/s10590-019-09233-w.

[191] Kathleen R. McKeown. "Paraphrasing using given and new information in a question-answer system". In: *Proceedings of ACL*. 1979.

[192] Marie Meteer and Varda Shaked. "Microsoft Research Treelet Translation System: NAACL 2006 Europarl Evaluation". In: *Proceedings of COLING*. 1988.

[193] Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. "A Coverage Embedding Model for Neural Machine Translation". In: *CoRR* abs/1605.03148 (2016). URL: http://arxiv.org/abs/1605.03148.

[194] Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. "Document-Level Neural Machine Translation with Hierarchical Attention Networks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2947–2954. DOI: 10.18653/v1/D18-1325. URL: https://www.aclweb.org/anthology/D18-1325.

[195] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781 (2013). URL: http://arxiv.org/abs/1301.3781.

[196] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. "WordNet: an on-line lexical database". In: *International Journal of Lexicography* 3.4 (1990), pp. 235–244.

[197] Douglas C. Montgomery and George C. Runger. *Applied statistics and probability for engineers*. third. New York: John Wiley and Sons, 2003.

[198] Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. "Introduction". In: *Translation Quality Assessment: From Principles to Practice*. Ed. by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen

Doherty. Cham: Springer International Publishing, 2018, pp. 1–6. ISBN: 978-3-319-91241-7. DOI: 10.1007/978-3-319-91241-7_1. URL: https://doi.org/10.1007/978-3-319-91241-7_1.

[199] John Moran and David Lewis. "Unobtrusive methods for low-cost manual assessment of machine translation". In: *Tralogy I [Online], Session 5* (2012).

[200] Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel, and Koel Dutta Chowdhury. "Semantic reranking of CRF label sequences for verbal multiword expression identification". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Language Science Press, 2018, pp. 177–207. DOI: 10.5281/zenodo.1469559. URL: https://hal.archives-ouvertes.fr/hal-01930987.

[201] Erwan Moreau and Carl Vogel. "Limitations of MT Quality Estimation Supervised Systems: The Tails Prediction Problem". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 2205–2216. URL: https://www.aclweb.org/anthology/C14-1208.

[202] Jonathan Mutal, Lise Volkart, Pierrette Bouillon, Sabrina Girletti, and Paula Estrella. "Differences between SMT and NMT Output - a Translators' Point of View". In: *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*. Varna, Bulgaria: Incoma Ltd., Shoumen, Bulgaria, Sept. 2019, pp. 75–81. DOI: 10.26615/issn.2683-0078.2019_009. URL: https://aclanthology.org/W19-8709.

[203] Ramón P. Neco and Mikel L. Forcada. "Asynchronous translations with recurrent neural nets". In: *Neural Networks,1997., International Conference on*. Vol. 4. June 1997, 2535–2540 vol.4. DOI: 10.1109/ICNN.1997.614693.

[204] Graham Neubig. "Neural Machine Translation and Sequence-to-sequence Models: A Tutorial". In: *arXiv preprint arXiv:1703.01619* (2017).

[205] Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. *How Transformer Revitalizes Character-based Neural Machine Translation: An Investigation on Japanese-Vietnamese Translation Systems*. 2019. arXiv: 1910.02238 [cs.CL].

[206] Toan Q. Nguyen and David Chiang. "Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 296–301. URL: https://www.aclweb.org/anthology/I17-2050.

[207] Sergei Nirenburg. "Knowledge-Based Machine Translation". In: *Machine Translation* 4.1 (1989), pp. 5–24. ISSN: 09226567, 15730573. URL: http://www.jstor.org/stable/40008396.

[208] Franz Josef Och. "Minimum Error Rate Training for Statistical Machine Translation". In: *Proceedings of ACL*. 2003.

[209] Franz Josef Och and Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models". In: *Computational Linguistics* 29.1 (2003), pp. 19–51.

[210] Franz Josef Och and Hermann Ney. "Improved statistical alignment models". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2000, pp. 440–447.

[211] Giedre Valunaite Oleskeviciene and Chaya Liebeskind. "Multiword expressions as discourse markers in Hebrew and Lithuanian". In: *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*. online: Association for Computational Linguistics, May 2021, pp. 46–56. URL: https://www.aclweb.org/anthology/2021.motra-1.5.

[212] Joseph Olive. "Global Autonomous Language Exploitation (GALE)". In: *DARPA/IPTO Proposer Information Pamphlet*. 2005.

[213] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: 2002, pp. 311–318.

[214] Kristen Parton, Joel Tetreault ans Nitin Madnani, and Martin Chodorow. "E-rating Machine Translation". In: *Proceedings of WMT*. 2011.

[215] Gloria Corpas Pastor and Ruslan Mitkov, eds. *Computational and Corpus-Based Phraseology - Third International Conference, Europhras 2019, Malaga, Spain, September 25-27, 2019, Proceedings*. Vol. 11755. Lecture Notes in Computer Science. Springer, 2019. ISBN: 978-3-030-30134-7. DOI: 10.1007/978-3-030-30135-4. URL: https://doi.org/10.1007/978-3-030-30135-4.

[216] Karl Pearson. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *Philosophical Magazine* 50.5 (1900), pp. 157–175.

[217] Haiyun Peng, Erik Cambria, and Xiaomei Zou. "Radical-Based Hierarchical Embeddings for Chinese Sentiment Analysis at Sentence Level". In: *Florida Artificial Intelligence Research Society Conference*. 2017. URL: https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15460.

[218] Jan-Thorsten Peter, Arne Nix, and Hermann Ney. "Generating Alignments Using Target Foresight in Attention-Based Neural Machine Translation". In: *The Prague Bulletin of Mathematical Linguistics* 108.1 (2017), pp. 27–36.

[219] Mārcis Pinnis. "Context Independent Term Mapper for European Languages". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sept. 2013, pp. 562–570. URL: https://www.aclweb.org/anthology/R13-1074.

[220] Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. "Evaluation without references: IBM1 scores as evaluation metrics". In: *Proceedings of WMT*. 2011.

[221] Maja Popovic and Hermann Ney. "Word Error Rates: Decomposition over POS Classes and Applications for Error Analysis". In: *Proceedings of WMT*. 2007.

[222] Maja Popović. "Informative Manual Evaluation of Machine Translation Output". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5059–5069. DOI: 10.18653/v1/2020.coling-main.444. URL: https://www.aclweb.org/anthology/2020.coling-main.444.

[223] Maja Popović. "Relations between comprehensibility and adequacy errors in machine translation output". In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, Nov. 2020, pp. 256–264. DOI: 10.18653/v1/2020.conll-1.19. URL: https://www.aclweb.org/anthology/2020.conll-1.19.

[224] Claus Povlsen, Nancy Underwood, Bradley Music, and Anne Neville. "Evaluating Text-Type Suitability for Machine Translation a Case Study on an English-Danish System". In: *Proceeding LREC*. 1998.

[225] Virginia Pulcini. "English-derived Multi-word and Phraseological Units across Languages in the Global Anglicism Database". en. In: *Textus, English Studies in Italy* 1/2020 (2020), pp. 127–143. ISSN: 1824-3967. DOI: 10.7370/97355. URL: https://www.rivisteweb.it/doi/10.7370/97355.

[226] Carlos Ramisch. *Multiword Expressions Acquisition: A Generic and Open Framework*. Vol. XIV. Theory and Applications of Natural Language Processing. Springer, 2015, p. 230. ISBN: 978-3-319-09206-5. DOI: 10.1007/978-3-319-09207-2.

[227] Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. "Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions". In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 222–240. URL: https://www.aclweb.org/anthology/W18-4925.

[228] Sylvain Raybaud, David Langlois, and Kamel Smaïli. ""this sentence is wrong." detecting errors in machine-translated sentences." In: *Machine Translation* 25.1 (2011), pp. 1–34.

[229] Florence Reeder. "Investigation of Intelligibility Judgments". In: *Machine Translation: From Real Users to Research*. Ed. by Robert E. Frederking and Kathryn B. Taylor. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 227–235. ISBN: 978-3-540-30194-3.

[230] Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. "Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions". In: *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. MWE '09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 47–54. ISBN: 978-1-932432-60-2. URL: http://dl.acm.org/citation.cfm?id=1698239.1698249.

[231] Matīss Rikters and Ondřej Bojar. "Paying Attention to Multi-Word Expressions in Neural Machine Translation". In: *Proceedings of the 16th Machine Translation Summit (MT Summit 2017)*. Nagoya, Japan, 2017.

[232] Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. "Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 157–165. URL: https://www.aclweb.org/anthology/2021.nodalida-main.16.

[233] Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. "Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 960–970. DOI: 10.18653/v1/N16-1110. URL: https://www.aclweb.org/anthology/N16-1110.

[234] Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. "XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation". In: *arXiv e-prints*, arXiv:2104.07412 (Apr. 2021), arXiv:2104.07412. arXiv: 2104.07412 [cs.CL].

[235] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. "Multiword Expressions: A Pain in the Neck for NLP". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 1–15. ISBN: 978-3-540-45715-2.

[236] Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. "The Impact of Multiword Expression Compositionality on Machine Translation Evaluation". In: *Proceedings of the 11th Workshop on Multiword Expressions*. Denver, Colorado: Association for Computational Linguistics, June 2015,

pp. 54–59. DOI: 10.3115/v1/W15-0909. URL: https://www.aclweb.org/anthology/W15-0909.

[237] Danielle Saunders and Bill Byrne. "Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7724–7736. DOI: 10.18653/v1/2020.acl-main.690. URL: https://www.aclweb.org/anthology/2020.acl-main.690.

[238] Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. "PARSEME multilingual corpus of verbal multiword expressions". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Berlin: Language Science Press., 2018, pp. 87–147. DOI: 10.5281/zenodo.1469555.

[239] Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. "The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions". In: *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain, 2017, pp. 31–47. DOI: 10.18653/v1/W17-1704. URL: https://www.aclweb.org/anthology/W17-1704.

[240] Helmut Schmid. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, 1994.

[241] Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. "Comprehensive Annotation of Multiword Expressions in a Social Web Corpus". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Languages Resources Association (ELRA), May 2014, pp. 455–461. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf.

[242] Holger Schwenk, Daniel Dchelotte, and Jean-Luc Gauvain. "Continuous Space Language Models for Statistical Machine Translation". In: *Proceedings of the COLING/ACL on Main Conference Poster Sessions*. COLING-ACL '06. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 723–730. URL: http://dl.acm.org/citation.cfm?id=1273073.1273166.

[243] Alina Secară. "Translation Evaluation - a State of the Art Survey". In: *Proceedings of the eCoLoRe/MeLLANGE workshop*. Leeds, 2005, pp. 39–44.

[244] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. "Nematus: a Toolkit for Neural Machine Translation". In: *Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, 2017.

[245] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://www.aclweb.org/anthology/P16-1162.

[246] Rico Sennrich and Biao Zhang. "Revisiting Low-Resource Neural Machine Translation: A Case Study". In: *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 211–221. DOI: 10.18653/v1/P19-1021. URL: https://www.aclweb.org/anthology/P19-1021.

[247] Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. "Radical Embedding: Delving Deeper to Chinese Radicals". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, 2015, pp. 594–598. DOI: 10.3115/v1/P15-2098. URL: http://aclanthology.coli.uni-saarland.de/pdf/P/P15/P15-2098.pdf.

[248] Inguna Skadina. "Multi-word Expressions in English-Latvian Machine Translation". In: *Baltic J. Modern Computing* 4 (2016), pp. 811–825.

[249] Inguna Skadina, Robert Gaizauskas, Bogdan Babych, Nikola Ljubesic, Dan Tufis, and Andrejs Vasiljevs. *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*. Springer, 2019. DOI: 10.1007/978-3-319-99004-0.

[250] Alan F. Smeaton, Paul Over, and Wessel Kraaij. "Evaluation Campaigns and TRECVid". In: *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. MIR '06. Santa Barbara, California, USA: Association for Computing Machinery, 2006, pp. 321–330. ISBN: 1595934952. DOI: 10.1145/1178677.1178722. URL: https://doi.org/10.1145/1178677.1178722.

[251] Mattthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceeding of AMTA*. 2006.

[252] Michaela Socolof, Jackie Chi Kit Cheung, Michael Wagner, and Timothy J. O'Donnell. "Characterizing Idioms: Conventionality and Contingency".

In: *arXiv e-prints*, arXiv:2104.08664 (Apr. 2021), arXiv:2104.08664. arXiv: 2104.08664 [cs.CL].

[253] Xingyi Song and Trevor Cohn. "Regression and ranking based optimisation for sentence level MT evaluation". In: *Proceedings of WMT*. 2011.

[254] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. "Findings of the WMT 2020 Shared Task on Quality Estimation". In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 743–764. URL: https://www.aclweb.org/anthology/2020.wmt-1.79.

[255] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. "A Shared Task on Multimodal Machine Translation and Crosslingual Image Description". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 543–553. DOI: 10.18653/v1/W16-2346. URL: https://www.aclweb.org/anthology/W16-2346.

[256] Lucia Specia and Jesús Giménez. "Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation". In: *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*. Denver, Colorado, USA: Association for Machine Translation in the Americas, Oct. 2010. URL: https://aclanthology.org/2010.amta-papers.3.

[257] Lucia Specia, Dhwaj Raj, and Marco Turchi. "Machine translation evaluation versus quality estimation". English. In: *Machine translation* (2010). ISSN: 0922-6567.

[258] Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. "QuEst - A translation quality estimation framework". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System*

Demonstrations. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 79–84. URL: https://www.aclweb.org/anthology/P13-4014.

[259] Miloš Stanojević and Khalil Sima'an. "BEER: BEtter Evaluation as Ranking". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. 2014.

[260] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. "Evaluating Gender Bias in Machine Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1679–1684. DOI: 10.18653/v1/P19-1164. URL: https://www.aclweb.org/anthology/P19-1164.

[261] Keh-Yih Su, Wu Ming-Wen, and Chang Jing-Shin. "A New Quantitative Quality Measure for Machine Translation Systems". In: *Proceeding of COLING*. 1992.

[262] Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. "Baidu Neural Machine Translation Systems for WMT19". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 374–381. DOI: 10.18653/v1/W19-5341. URL: https://www.aclweb.org/anthology/W19-5341.

[263] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[264] Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. "NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online):

Linköping University Electronic Press, Sweden, May 2021, pp. 276–287. URL: https://www.aclweb.org/anthology/2021.nodalida-main.28.

[265] Gongbo Tang, Rico Sennrich, and Joakim Nivre. "An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation". In: *CoRR* abs/1810.07595 (2018). arXiv: 1810.07595. URL: http://arxiv.org/abs/1810.07595.

[266] Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. "Accelerated DP Based Search For Statistical Translation". In: *Proceeding of EUROSPEECH*. 1997.

[267] Norbert E. Fuchs Tobias Kuhn. *Controlled Natural Language - Third International Workshop, CNL 2012, Zurich, Switzerland, August 29-31, 2012. Proceedings*. Springer, 2012.

[268] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. "Modeling Coverage for Neural Machine Translation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. 2016. URL: https://www.aclweb.org/anthology/P16-1008/.

[269] Joseph P Turian, Luke Shea, and I Dan Melamed. *Evaluation of machine translation and its evaluation*. Tech. rep. DTIC Document, 2006.

[270] Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. "Neural Machine Translation Incorporating Named Entity". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3240–3250. URL: https://www.aclweb.org/anthology/C18-1274.

[271] Tim Van de Cruys and Begoña Villada Moirón. "Semantics-based Multiword Expression Extraction". In: *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Prague, Czech Republic: Association

for Computational Linguistics, June 2007, pp. 25–32. URL: https://www.aclweb.org/anthology/W07-1104.

[272] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. "Getting Gender Right in Neural Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3003–3008. DOI: 10.18653/v1/D18-1334. URL: https://www.aclweb.org/anthology/D18-1334.

[273] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Conference on Neural Information Processing System*. 2017, pp. 6000–6010.

[274] David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. "Word-sense Disambiguation for Machine Translation". In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005, pp. 771–778. DOI: 10.3115/1220575.1220672. URL: https://doi.org/10.3115/1220575.1220672.

[275] Veronika Vincze. "Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association, May 2012, pp. 2381–2388. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/177_Paper.pdf.

[276] Veronika Vincze and János Csirik. "Hungarian Corpus of Light Verb Constructions". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing

Committee, Aug. 2010, pp. 1110–1118. URL: https://www.aclweb.org/anthology/C10-1125.

[277] Veronika Vincze, István Nagy T., and Gábor Berend. "Multiword Expressions and Named Entities in the Wiki50 Corpus". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria: Association for Computational Linguistics, Sept. 2011, pp. 289–295. URL: https://www.aclweb.org/anthology/R11-1040.

[278] Clare R. Voss and Ra R. Tate. "Task-based Evaluation of Machine Translation (MT) Engines: Measuring How Well People Extract Who, When, Where-Type Elements in MT Output". In: *In Proceedings of 11th Annual Conference of the European Association for Machine Translation (EAMT-2006*. 2006, pp. 203–212.

[279] Abigail Walsh, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider, and Clarissa Somers. "Constructing an Annotated Corpus of Verbal MWEs for English". In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 193–200. URL: https://www.aclweb.org/anthology/W18-4921.

[280] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. "CharacTer: Translation Edit Rate on Character Level." In: *WMT*. 2016, pp. 505–510.

[281] Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. "Neural Machine Translation Advised by Statistical Machine Translation". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 3330–3336.

[282] Xinyi Wang, Hieu Pham, Pengcheng Yin, and Graham Neubig. "A Tree-based Decoder for Neural Machine Translation". In: *CoRR* abs/1808.09374 (2018). arXiv: 1808.09374. URL: http://arxiv.org/abs/1808.09374.

[283] Warren Weaver. "Translation". In: *Machine Translation of Languages: Fourteen Essays* (1955).

[284] John S. White, Theresa O' Connell, and Francis O' Mara. "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches". In: *Proceeding of AMTA*. 1994.

[285] John S. White and Kathryn B. Taylor. "A Task-Oriented Evaluation Metric for Machine Translation". In: *Proceeding LREC*. 1998.

[286] Billy Wong and Chun yu Kit. "ATEC: automatic evaluation of machine translation via word choice and word order". In: *Machine Translation* 23.2-3 (2009), pp. 141–155.

[287] Shuangzhi Wu, Ming Zhou, and Dongdong Zhang. "Improved Neural Machine Translation with Source Syntax". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 2017, pp. 4179–4185. DOI: 10.24963/ijcai.2017/584. URL: https://doi.org/10.24963/ijcai.2017/584.

[288] Wei Wu, Yuxian Meng, Qinghong Han, Muyu Li, Xiaoya Li, Jie Mei, Ping Nie, Xiaofei Sun, and Jiwei Li. "Glyce: Glyph-vectors for Chinese Character Representations". In: *Annual Conference on Neural Information Processing Systems* (2019).

[289] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and

Jeffrey Dean. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). URL: http://arxiv.org/abs/1609.08144.

[290] Tong Xiao, Jingbo Zhu, Hao Zhang, and Qiang Li. “NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics2012 System Demonstrations*. ACL ’12. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 19–24. URL: http://dl.acm.org/citation.cfm?id=2390470.2390474.

[291] Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. “RED: A Reference Dependency Based MT Evaluation Metric”. In: *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. 2014, pp. 2042–2051. URL: http://aclweb.org/anthology/C/C14/C14-1193.pdf.

[292] Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. “Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 286–291. DOI: 10.18653/v1/D17-1027. URL: https://www.aclweb.org/anthology/D17-1027.

[293] Matthew D Zeiler. “ADADELTA: an adaptive learning rate method”. In: *arXiv preprint arXiv:1212.5701* (2012).

[294] Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huan-Bo Luan, and Yang Liu. “THUMT: An Open Source Toolkit for Neural Machine Translation”. In: *ArXiv* abs/1706.06415 (2017).

[295] Jiajun Zhang and Chengqing Zong. “Deep Neural Networks in Machine Translation: An Overview”. In: *IEEE Intelligent Systems* 5 (2015), pp. 16–25.

[296] Jinyi Zhang and Tadahiro Matsumoto. "Improving Character-level Japanese-Chinese Neural Machine Translation with Radicals as an Additional Input Feature". In: *CoRR* abs/1805.02937 (2018). arXiv: 1805.02937. URL: http://arxiv.org/abs/1805.02937.

[297] Longtu Zhang and Mamoru Komachi. "Neural Machine Translation of Logographic Language Using Sub-character Level Information". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 17–25. DOI: 10.18653/v1/W18-6303. URL: https://www.aclweb.org/anthology/W18-6303.

[298] Xinlu Zhang, Xiao Li, Yating Yang, and Rui Dong. "Improving Low-Resource Neural Machine Translation With Teacher-Free Knowledge Distillation". In: *IEEE Access* 8 (2020), pp. 206638–206645. DOI: 10.1109/ACCESS.2020.3037821.

[299] Yi-tao Zhou. "Four-Word Bundles in English Abstracts of Chinese and English Linguistics Journal Articles: A Corpus-based Comparative Study". In: *Journal of Literature and Art Studies* 11.2 (2021), pp. 116–122. DOI: 10.17265/2159-5836/2021.02.009.