Machine Translation Evaluation Resources and Methods: A Survey

Lifeng Han
ADAPT Centre, Dublin City University, Ireland
lifeng.han@adaptcentre.ie

Abstract

We introduce a Machine Translation (MT) evaluation survey that contains both manual and automatic evaluation methodologies. The traditional human evaluation criteria mainly include intelligibility, fidelity, fluency, adequacy, comprehension, and informativeness. The advanced human assessments include task-oriented measures, post-editing, segment ranking, direct assessment, and other extended criteria. We classify the automatic evaluation methods into two categories including lexical similarity scenario and linguistic features application. The lexical similarity methods contain edit distance, precision, recall, F-measure, and word order. The linguistic features can be divided into syntactic features and semantic features respectively. The syntactic features include part of speech tags, phrase types and sentence structures, and the semantic features include named entity, synonym, textual entailment, paraphrase, semantic role, and language models. The deep learning models for evaluation are very recently proposed due to word embedding popularity. Subsequently, we also introduce the evaluation methodology for MT evaluation including different correlation scores, and the lately quality estimation (QE) tasks for MT.

This paper differs from the previous works (Dorr et al., 2009; EuroMatrix, 2007) from several aspects, by introducing some recent development of MT evaluation measures, the different classifications from manual to automatic evaluation measures, the introduction of lately QE tasks of MT, and the concise construction of the content.

We hope this work will be helpful for MT researchers to easily pick up some metrics that are best suitable for their specific MT model development, and help MT evaluation researchers to get a general clue of how MT evaluation research developed. Furthermore, hopefully, this work can also shine some light on other evaluation tasks, except for translation, in natural language processing (NLP) fields. [1]

1 Introduction

Machine translation (MT) began as early as in the 1950s (Weaver, 1955) , and gained a rapid development since the 1990s (Mariño et al., 2006) due to the development of storage and computing power of computer and the widely available multilingual and bilingual corpora. There are many important works in MT areas, for some to mention by time, IBM Watson research group (Brown et al., 1993) designed five statistical MT models and the ways of how to estimate the parameters in the models given the bilingual translation corpora; (Koehn et al., 2003) proposed statistical phrase-based MT model; Och (Och, 2003) presented Minimum Error Rate Training (MERT) for log-linear statistical machine translation models; (Koehn and Monz, 2005) introduced a Shared task of building statistical machine translation (SMT) systems for four European language pairs; (Chiang, 2005) proposed a hierarchical phrase-based SMT model that is learned from a bitext without syntactic information; (Menezes et al., 2006) introduced a syntactically informed phrasal SMT system for English-to-Spanish translation using a phrase translation model, which was based on global reordering and dependency tree; (Koehn et

[1] Presented in IPRC-2018 (Ireland Postgraduate Research Conference, Nov.8-9, Dublin). Some work was done in NLP2CT-lab/Macau and SLPL-lab/ILLC/UvA/Amsterdam

al., 2007b) developed an open source SMT software toolkit Moses; (Hwang et al., 2007) utilized the shallow linguistic knowledge to improve word alignment and language model quality between linguistically different languages; (Fraser and Marcu, 2007) made a discussion of the relationship between word alignment and the quality of machine translation; (Sánchez-Martínez and Forcada, 2009) described an unsupervised method for the automatic inference of structural transfer rules for a shallow-transfer machine translation system; (Khalilov and Fonollosa, 2011) designed an effective syntax-based reordering approach to address the word ordering problem.

With the fast development of Deep Learning (DL), MT research has evolved from rule-based models to example based models, statistical models, hybrid models, and recent years' Neural models (Nirenburg, 1989; Carl and Way, 2003; Koehn and Knight, 2009; Bahdanau et al., 2014), such as the attention mechanism models, coverage models, multi-modal and multilingual MT models.

Neural MT (NMT) is a recently active topic that conduct the automatic translation workflow very differently with the traditional phrase-based SMT methods. Instead of training the different MT components separately, NMT model utilizes the artificial neural network (ANN) to learn the model jointly to maximize the translation performance through two steps recurrent neural network (RNN) of encoder and decoder (Cho et al., 2014; Bahdanau et al., 2014; **?**). There were far more representative MT works that we havent listed here.

Due to the wide-spread development of MT systems, the MT evaluation became more and more important to tell us how well the MT systems perform and whether they make some progress. However, the MT evaluation is difficult because the natural languages are highly ambiguous and different languages do not always express the same content in the same way (Arnold, 2003).

There are several events that promote the development of MT evaluation research. One of them was the NIST open machine translation Evaluation series (OpenMT), which were very prestigious evaluation campaigns from 2001 to 2009 (LI, 2005).

The innovation of MT and the evaluation methods is also promoted by the annual Workshop on Statistical Machine Translation (WMT) (Koehn and Monz, 2006a; Callison-Burch et al., 2007a; Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011; Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015) organized by the special interest group in machine translation (SIGMT) since 2006. The evaluation campaigns focus on European languages. There are roughly two tracks in the annual WMT workshop including the translation task and evaluation task. The tested language pairs are clearly divided into two directions, i.e., English-to-other and other-to-English, covering French, German, Spanish, Czech, Hungarian, Haitian Creole and Russian.

Another promotion is the international workshop of spoken language translation (IWSLT) that has been organized annually from 2004 (Eck and Hori, 2005; Paul, 2009; Paul et al., 2010; Federico et al., 2011). This campaign has a stronger focus on speech translation including the English and Asian languages, e.g. Chinese, Japanese and Korean.

The better evaluation metrics will be surly helpful to the development of better MT systems (Liu et al., 2011). Due to all the above efforts, the MT evaluation research achieved a rapid development.

This paper is constructed as follow: Section 2 and 3 discuss the human assessment methods and automatic evaluation methods respectively, Section 4 introduces the evaluating methods of the MT evaluation, Section 5 is the advanced MT evaluation, Section 6 is the discussion and conclusion, and the perspective is presented in Section 7.

2 Human Evaluation Methods

This section introduces the traditional human evaluation methods and the advanced methods as shown in Fig. 1.

2.1 Traditional Human Assessment

2.1.1 *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) (Carroll, 1966). The requirement that a translation be 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 sentence would be understandable if originally composed in the translation

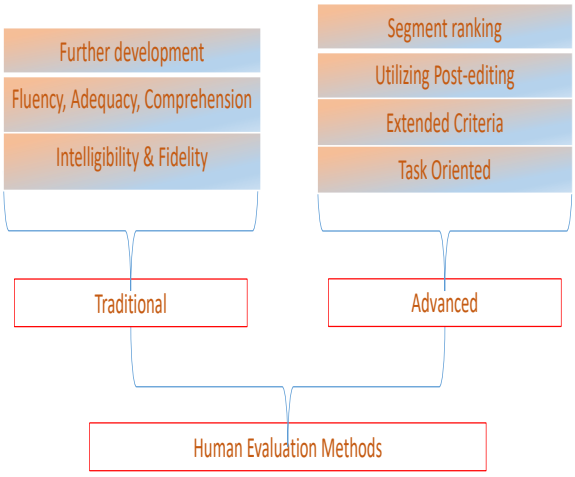

Figure 1: Human Evaluation Methods

language. The requirement that a translation is of high fidelity or accuracy includes that the translation should, as little as possible, twist, distort, or controvert the meaning intended by the original.

2.1.2 *Fluency, Adequacy and Comprehension*

In1990s, the Advanced Research Projects Agency (ARPA) created the methodology to evaluate machine translation systems using the adequacy, fluency and comprehension (Church and Hovy, 1991) in MT evaluation campaigns (White et al., 1994).

$$\text{Comprehension} = \frac{\#\text{Cottect}}{6}, \tag{1}$$

$$\text{Fluency} = \frac{\frac{\text{Judgment point}-1}{\text{S}-1}}{\#\text{Sentences in passage}}, \tag{2}$$

$$\text{Adequacy} = \frac{\frac{\text{Judgment point}-1}{\text{S}-1}}{\#\text{Fragments in passage}}. \tag{3}$$

The evaluator is asked to look at each fragment, delimited by syntactic constituent and containing sufficient information, and judge the adequacy on a scale 1-to-5. The results are computed by averaging the judgments over all of the decisions in the translation set.

The fluency evaluation is compiled with the same manner as that for the adequacy except for that the evaluator is to make intuitive judgments on a sentence by sentence basis for each translation. The evaluators are asked to determine whether the translation is good English without reference to the correct translation. The fluency evaluation is to determine whether the sentence is well-formed and fluent in context.

The modified comprehension develops into the "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. Developed from the reference set of expert translations, six questions have six possible answers respectively including, "none of above" and "cannot be determined".

2.1.3 *Further Development*

(Bangalore et al., 2000) conduct a research developing accuracy into several kinds including simple string accuracy, generation string accuracy, and two corresponding tree-based accuracies.Reeder (2004) shows the correlation between fluency and the number of words it takes to distinguish between human translation and machine translation.

The "Linguistics Data Consortium" (LDC) develops two five-points scales representing fluency and adequacy for the annual NIST machine translation evaluation workshop. The developed scales become the widely used methodology when manually evaluating MT is to assign values. The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation; the second five point scale indicates how fluent the translation is, involving both grammatical correctness and idiomatic word choices.

(Specia et al., 2011) conduct a study of the MT adequacy and design it into four levels, from score 4 to score 1: highly adequate, the translation faithfully conveys the content of the input sentence; fairly adequate, while the translation generally conveys the meaning of the input sentence, there are some problems with word order or tense/voice/number, or there are repeated, added or un-translated words; poorly adequate, the content of the input sentence is not adequately conveyed by the translation; and completely inadequate, the content of the input sentence is not conveyed at all by the translation.

2.2 Advanced Human Assessment

2.2.1 *Task-oriented*

(White and Taylor, 1998) develop 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 evaluation with a

scale of language-dependent tasks, such as scanning, sorting, and topic identification. They develop the MT proficiency metric with a corpus of multiple variants which are usable as a set of controlled samples for user judgments.The principal steps include identifying the user-performed text-handling tasks, discovering the order of text-handling task tolerance, analyzing 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 of task-based MT evaluation work was shown in their later work (Doyon et al., 1999).

Voss and Tate (Voss and Tate, 2006) introduced the tasked-based MT output evaluation by the extraction of who, when, where types elements. They extend the work later into event understanding in (Laoudi et al., 2006).

2.2.2 *Extended Criteria*

(King et al., 2003) extend a large range of manual evaluation methods for MT systems, which, in addition to the early talked accuracy, include suitability, whether even accurate results are suitable in the particular context in which the system is to be used; interoperability, whether with other software or with hardware platforms; reliability, i.e., don't break down all the time or take long time to get running again after breaking down; usability, easy to get the interfaces, easy to learn and operate, and looks pretty; 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 they tend to be improved over time as resources grow and bugs are fixed.

2.2.3 *Utilizing Post-editing*

A measure of quality is to compare translation from scratch and post-edited result of an automatic translation. This type of evaluation is however time consuming and depends on the skills of the translator and post-editor. One example of a metric that is designed in such a manner is the human translation error rate (HTER) (Snover et al., 2006), based on the number of editing steps, computing the editing steps between an automatic translation and a reference translation. Here, a human annotator has to find the minimum number of insertions, deletions, substitutions, and shifts to convert the system output into an acceptable translation. HTER is defined as the number of editing steps divided by the number of words in the acceptable translation.

2.2.4 *Segment Ranking*

In the WMT metrics task, the human assessment based on segment ranking is usually employed. Judgesare frequently asked to provide a complete ranking over all the candidate translations of the same source segment (Callison-Burch et al., 2011; Callison-Burch et al., 2012). In the recent WMT tasks (Bojar et al., 2013),five systems are randomly selected for the judges to rank. Each time, the source segment and the reference translation are presented to the judges together with the candidate translations of five systems. The judges will rank the systems from 1 to 5, allowing tie scores. For each ranking, there is the potential to provide as many as 10 pairwise results if no ties. The collected pairwise rankings can be used to assign a score to each participated system to reflect the quality of the automatic translations.The assigned score can also be utilized to reflect how frequently a system is judged to be better or worse than other systems when they are compared on the same source segment, according to the following formula:

$$\frac{\#\text{better pairwise ranking}}{\#\text{total pairwise comparison} - \#\text{ties comparisons}}. \quad (4)$$

3 Automatic Evaluation Metric

Manual evaluation suffers some disadvantages such as time-consuming, expensive, not tunable, and not reproducible. Due to the weaknesses in human judgments, automatic evaluation metrics have been widely used for machine translation. Typically, they compare the output of machine translation systems against human translations, but there are also some metrics that do not use the reference translation. 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 (Lin and Och, 2004; Han et al., 2012). Common metrics measure the overlap in words and word sequences, as well as word order and edit distance. We classify this kind of metrics as the "Lexical

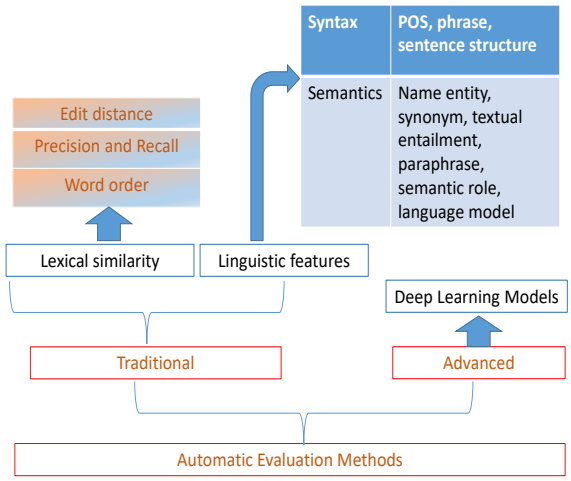

Figure 2: Automatic Evaluation Methods

Similarity" category. Further developed metrics also take linguistic features into account such as syntax and semantics, e.g. POS, sentence structure, textual entailment, paraphrase, synonyms, named entities, semantic roles and language models, etc. We classify these metrics that utilize the linguistic features into "Linguistic Features" category. It is not easy to separate these two categories clearly since sometimes they merge with each other, for instance, some metrics from category one also use certain linguistic features. Further more, we will introduce some advanced researches that apply deep learning into MTE framework, as in Fig. 2.

3.1 Lexical Similarity

3.1.1 *Edit Distance*

By calculating the minimum number of editing steps to transform output to reference, (Su et al., 1992) introduce the word error rate (WER) metric into MT evaluation. This metric 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.

$$\text{WER} = \frac{\text{substitution+insertion+deletion}}{\text{reference}_{\text{length}}}. \quad (5)$$

One of the weak points of the WER is the fact that word ordering is not taken into account appropriately. The WER scores very low when the word order of system output translation is "wrong" according to the reference. 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 expression, 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) (Tillmann et al., 1997) is designed to ignore word order when matching output and reference. Without taking into 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 insertion or deletion ones.

$$\text{PER} = 1 - \frac{\text{correc} - \max(0, \text{output}_{\text{length}} - \text{reference}_{\text{length}})}{\text{reference}_{\text{length}}}. \quad (6)$$

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, (Snover et al., 2006) design the translation edit rate (TER) metric that adds block movement (jumping action) as an editing step.The shift option performs on a contiguous sequence of words within the output sentence. The TER score is calculated as:

$$\text{TER} = \frac{\#\text{of edit}}{\#\text{of average reference words}} \quad (7)$$

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.

3.1.2 *Precision and Recall*

The widely used evaluation metric BLEU (Papineni et al., 2002) is based on the degree of n-gram overlapping between the strings of words produced by the machine and the human translation references at the corpus level. BLEU computes the precision for n-gram of size 1-to-4 with the coefficient of brevity penalty (BP).

$$\text{BLEU} = \text{BP} \times \exp \sum_{n=1}^{N} \lambda_n \log \text{Precision}_\text{n}, \quad (8)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{1-\frac{r}{c}} & \text{if } c <= r. \end{cases} \quad (9)$$

where c is the total length of candidate translation corpus, and r refers to the sum of effective reference sentence length in the corpus. 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 uniform weight. However, the 4-gram precision value is usually very low or even zero when the test corpus is small. To weight more heavily those n-grams that are more informative, (Doddington, 2002) proposes the NIST metric with the information weight added.

$$\text{Info} = \log_2 \left(\frac{\#\text{occurrence of } w_1, \cdots, w_{n-1}}{\#\text{occurrence of } w_1, \cdots, w_n}\right) \quad (10)$$

Furthermore, he replaces the geometric mean of co-occurrences with the arithmetic average of n-gram counts, extend the n-gram into 5-gram ($N = 5$), and select the average length of reference translations instead of the nearest length.

ROUGE (Lin and Hovy, 2003) is a recall-oriented automated evaluation metric, which is initially developed for summaries. Following the adoption by the machine translation community of automatic evaluation using the BLEU/NIST scoring process, Lin conducts a study of a similar idea for evaluating summaries. They also apply the ROUGE into automatic machine translation evaluation work (Lin and Och, 2004).

(Turian et al., 2006) conducted experiments to examine how standard measures such as precision and recall and F-measure can be applied for evaluation of MT and showed the comparisons of these standard measures with some existing alternative evaluation measures. F-measure is the combination of precision (P) and recall (R), which is firstly employed in the information retrieval and latterly has been adopted by the information extraction, MT evaluation and other tasks.

$$F_\beta = (1 + \beta^2) \frac{PR}{R + \beta^2 P} \quad (11)$$

(Banerjee and Lavie, 2005) design a novel evaluation metric METEOR. METEOR is based on general concept of flexible unigram matching, unigram precision and unigram recall, including the match of words that are simple morphological variants of each other by the identical stem 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 by employing the number of matched chunks.

$$\text{Penalty} = 0.5 \times \left(\frac{\#\text{chunks}}{\#\text{matched unigrams}}\right)^3, \quad (12)$$

$$\text{MEREOR} = \frac{10PR}{R + 9P} \times (1 - \text{Penalty}). \quad (13)$$

3.1.3 *Word Order*

The right word order places an important role to ensure a high quality translation output. However, the language diversity also allows different appearances or structures of the sentence. How to successfully achieve the penalty on really wrong word order (wrongly structured sentence) instead of on the "correctly" different order, the candidate sentence that has different word order with the reference is well structured, attracts a lot of interests from researchers in the NLP literature. In fact, the Levenshtein distance and n-gram based measures also contain the word order information.

Featuring the explicit assessment of word order and word choice, (Wong and yu Kit, 2009) develop the evaluation metric ATEC, assessment of text essential characteristics. It is also based on precision and recall criteria but with the designed 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. Combining the precision, order, and recall information together, (Chen et al., 2012) develop an automatic evaluation metric PORT that is initially for the tuning of the MT systems to output higher quality translation. Another evaluation metric LEPOR (Han et al., 2012; Han et al., 2014) is proposed as the combination of many evaluation factors including n-gram based word order penalty in addition to precision, recall,

and sentence-length penalty. The LEPOR metric yields the excellent performance on the English-to-other (Spanish, German, French, Czech and Russian) language pairs in ACL-WMT13 metrics shared tasks at system level evaluation (Han et al., 2013b).

3.2 Linguistic Features

Although some of the previous mentioned metrics employ the linguistic information into consideration, e.g. the semantic information synonyms and stemming in METEOR, the lexical similarity methods mainly focus on the exact matches of the surface words in the output translation. The advantages of the metrics based on lexical similarity are that they perform well in capturing the translation fluency (Lo et al., 2012), and they are very fast and low cost. On the other hand, there are also some weaknesses, for instance, the syntactic information is rarely considered and the underlying assumption that a good translation is one that shares the same lexical choices as the reference translations is not justified semantically. Lexical similarity does not adequately reflect similarity in meaning. Translation evaluation metric that reflects meaning similarity needs to be based on similarity of semantic structure not merely flat lexical similarity.

3.2.1 *Syntactic Similarity*

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

In grammar, a **part of speech (POS)** is a linguistic category of words or lexical items, which is generally defined by the syntactic or morphological behavior of the lexical item. Common linguistic categories of lexical items include noun, verb, adjective, adverb, and preposition, etc. To reflect the syntactic quality of automatically translated sentences, some researchers employ the POS information into their evaluation. Using the IBM model one, (Popović et al., 2011) evaluate the translation quality by calculating the similarity scores of source and target (translated) sentence without using reference translation, based on the morphemes, 4-gram POS and lexicon probabilities. (Dahlmeier et al., 2011) develop the evaluation metrics TESLA, combining the synonyms of bilingual phrase tables and POS information in the matching task. Other similar works using POS information include (Giménez and Márquez, 2007; Popovic and Ney, 2007; Han et al., 2014).

In linguistics, a **phrase** may refer to any group of words that form a constituent and so function as a single unit in the syntax of a sentence. To measure a 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, (Povlsen et al., 1998) perform a research work focusing on the study of English-to-Danish machine-translation system. The syntactic constructions are explored with more complex linguistic knowledge, such as the identifying of fronted adverbial subordinate clauses and prepositional phrases. Assuming that the similar grammatical structures should occur on both source and translations, (Avramidis et al., 2011) perform the evaluation on source (German) and target (English) sentence employing the features of sentence length ratio, unknown words, phrase numbers including noun phrase, verb phrase and prepositional phrase. Other similar works using the phrase similarity include the (Li et al., 2012) that uses noun phrase and verb phrase from chunking and (Echizen-ya and Araki, 2010) that only uses the noun phrase chunking in automatic evaluation and (Han et al., 2013a) that designs a universal phrase tagset for French to English MT evaluation.

Multi-word Expressions (MWEs) set obstacles for MT models due to their richness and complex in presentation (Sag et al., 2002; Han et al., 2020). To investigate the effect of MWEs in MTE, Salehi et al. (2015) focused on the *compositionality* of noun compounds. They identify the **noun compounds** first from the system outputs and reference with Stanford parser, then the matching scores of system outputs and reference sentences are recalculated, added up to Tesla metric, by considering the predicated compositionality of identified noun compound phrases.

Syntax is the study of the principles and processes by which sentences are constructed in particular languages. To address the overall goodness of the translated **sentence's structure**, (Liu and Gildea, 2005) employ constituent labels and head-modifier dependencies from language parser as syntactic features for MT evaluation. They compute the similarity of dependency trees. The overall experiments prove that adding syntactic in-

formation can improve the evaluation performance especially for predicting the fluency of hypothesis translations. Other works that using syntactic information into the evaluation include (Lo and Wu, 2011a) and (Lo et al., 2012) that use an automatic shallow parser and RED metric (Yu et al., 2014) that applies dependency tree, etc.

3.2.2 *Semantic Similarity*

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

To capture the semantic equivalence of sentences or text fragments, the **named entity** knowledge is brought from the literature of named-entity recognition, which is aiming to identify and classify atomic elements in the text into different entity categories (Marsh and Perzanowski, 1998; Guo et al., 2009). The commonly used entity categories include the names of persons, locations, organizations and time. In the MEDAR2011 evaluation campaign,one baseline system based on Moses (Koehn et al., 2007a) utilizes Open NLP toolkit to perform named entity detection, in addition to other packages. The low performances from the perspective of named entities cause a drop in fluency and adequacy.In the quality estimation of machine translation task of WMT 2012, (Buck, 2012) introduces the features including named entity, in addition to discriminative word lexicon, neural networks, back off behavior (Raybaud et al., 2011) and edit distance, etc. The experiments on individual features show that, from the perspective of the increasing the correlation score with human judgments, the feature of named entity contributes nearly the most compared with the contributions of other features.

Synonyms are words with the same or close meanings. One of the widely used synonym database in NLP literature is the WordNet (Miller et al., 1990), which is an English lexical database grouping English words into sets of synonyms. WordNet classifies the words mainly into four kinds of part-of-speech (POS) categories including Noun, Verb, Adjective, and Adverb without prepositions, determiners, etc. Synonymous words or phrases are organized using the unit of synset. Each synset is a hierarchical structure with the words in different levels 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 $\Rightarrow$ TA). Instead of the pure logical or mathematical entailment, the textual entailment in natural language processing (NLP) is usually performed with a relaxed or loose definition (Dagan et al., 2006). 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 $\Rightarrow$ TA also establishes. That the relation is directive also means that the inverse inference (TA $\Rightarrow$ TB) is not ensured to be true (Dagan and Glickman, 2004). Recently, Castillo and Estrella (2012) 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 is to restatement the meaning of a passage or text utilizing other words, which can be seen as bidirectional textual entailment (Androutsopoulos and Malakasiotis, 2010). Instead of the literal translation, word by word and line by line used by metaphrase, paraphrase represents a dynamic equivalent. Further knowledge of paraphrase from the aspect of linguistics is introduced in the works of (McKeown, 1979; Meteer and Shaked, 1988; Barzilayand and Lee, 2003). (Snover et al., 2006) describe a new evaluation metric TER-Plus (TERp). Sequences of words in the reference are considered to be paraphrases of a sequence of words in the hypothesis if that phrase pair occurs in TERp phrase table.

The **semantic roles** are employed by some researchers as linguistic features in the MT evaluation. To utilize the semantic roles, the sentences are usually first shallow parsed and entity tagged. Then the semantic roles used to specify the arguments and adjuncts that occur in both the candidate translation and reference translation. For instance, the semantic roles introduced by (Gim*é*nez and M*á*rquez, 2007; Gim*é*ne and M*á*rquez, 2008) include causative agent, adverbial adjunct, directional adjunct, negation marker, and predication adjunct, etc.In the further development, (Lo and Wu, 2011a; Lo and Wu, 2011b) design the metric MEANT to capture the predicate-argument relations as the structural relations in

semantic frames, which is not reflected by the flat semantic role label features in the work of (Giménez and Márquez, 2007). Furthermore, instead of using uniform weights, (Lo et al., 2012) weight the different types of semantic roles according to their relative importance to the adequate preservation of meaning, which is empirically determined.Generally, the semantic roles account for the semantic structure of a segment and have proved effective to assess adequacy in the above papers.

The **language models** are also utilized by the MT and MT evaluation researchers. A statistical language model usually assigns a probability to a sequence of words by means of a probability distribution. (Gamon et al., 2005) propose LM-SVM, language-model, support vector machine, method 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 dysfluent and illformed sentences.

(Stanojević and Sima'an, 2014a) designed a novel sentence level MT evaluation metric BEER, which has the advantage of incorporate large number of features in a linear model to maximize the correlation with human judgments. To make smoother sentence level scores, they explored two kinds of less sparse features including "character n-grams" (e.g. stem checking) and "abstract ordering patterns" (permutation trees). They further investigated the model with more dense features such as adequacy features, fluency features and features based on **permutation trees** (Stanojević and Sima'an, 2014c). In the latest version, they extended the permutation-tree (Gildea et al., 2006) into permutation-forests model (Stanojević and Sima'an, 2014b), and showed stable good performance on different language pairs in WMT sentence level evaluation task.

Generally, the linguistic features mentioned above, including both syntactic and semantic features, are usually combined in two ways, either by following a machine learning approach (Albrecht and Hwa, 2007; Leusch and Ney, 2009), or trying to combine a wide variety of metrics in a more simple and straightforward way, such as (Giménez and Márquez, 2008; Specia and Giménez, 2010; Comelles et al., 2012),etc.

3.3 DL for MTE

There are researchers applying DL and NNs models for MTE which are promising for further exploration.

For instances, (Guzmán et al., 2015; Guzmn et al., 2017) used neural networks for MTE for pair wise modeling to choose best hypothesis translation by comparing candidate translations with reference, integrating syntactic and semantic information into NNs.

(Gupta et al., 2015b) proposed LSTM networks based on dense vectors to conduct MTE. While (Ma et al., 2016) designed a new metric based on bi-directional LSTM, which is similar with the work of (Guzmán et al., 2015) but with less complexity by allowing the evaluation of single hypothesis with reference, instead of pairwise situation.

4 Evaluating the MT Evaluation

4.1 Statistical Significance

If different MT systems produce translations with different qualities on a data set, how can we ensure that they indeed own different system quality? To explore this problem, (Koehn, 2004) performs a research work on the statistical significance test for machine translation evaluation. The bootstrap resampling method is used to compute the statistical significance intervals for evaluation metrics on small test sets. Statistical significance usually refers to two separate notions, of which one is the p-value, the probability that the observed data will occur by chance in a given single null hypothesis. The other one is the "Type I" error rate of a statistical hypothesis test, which is also named as "false positive" and measured by the probability of incorrectly rejecting a given null hypothesis in favor of a second alternative hypothesis (Hald, 1998).

4.2 Evaluating Human Judgment

Since the human judgments are usually trusted as the golden standards that the automatic evaluation metrics should try to approach, the reliability and coherence of human judgments is very important. Cohen's kappa agreement coefficient is one of the commonly used evaluation methods (Cohen, 1960). For the problem in nominal scale agreement between two judges, there are two relevant quantities p_0 and p_c. The factor 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:

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{14}$$

where $p_0 - p_c$ represents the proportion of the cases in which beyond-chance agreement occurs and is the numerator of the coefficient (Landis and Koch, 1977).

4.3 Correlating Manual and Automatic Score

In this section, we introduce three correlation coefficient algorithms that are commonly used by the recent WMT workshops to measure the closeness of the automatic evaluation and manual judgments. Choosing which correlation algorithm depends on whether the scores or ranks schemes are utilized.

4.3.1 *Pearson Correlation*

Pearson's correlation coefficient (Pearson, 1900) is commonly represented by the Greek letter ρ. The correlation between random variables X and Y denoted as is measured as follow (Montgomery and Runger, 2003).

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{15}$$

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 by:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \tag{16}$$

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

4.3.2 *Spearman rank Correlation*

Spearman rank correlation coefficient, a simplified version of Pearson correlation coefficient , is another algorithm to measure the correlations of automatic evaluation and manual judges, especially in recent years (Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011). When there are no ties, Spearman rank correlation coefficient, which is sometimes specified as (rs) is calculated as:

$$rs_{\varphi(XY)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{17}$$

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.

4.3.3 *Kendall's τ*

Kendall's τ (Kendall, 1938) has been used in recent years for the correlation between automatic order and reference order (Callison-Burch et al., 2010; Callison-Burch et al., 2011; Callison-Burch et al., 2012). It is defined as:

$$\tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \tag{18}$$

The latest version of Kendall's τ is introduced in (Kendall and Gibbons, 1990). (Lebanon and Lafferty, 2002) give an overview work for Kendall's τ showing its application in calculating how much the system orders differ from the reference order. More concretely, (Lapata, 2003) proposes the use of Kendall's τ, a measure of rank correlation, estimating the distance between a system-generated and a human-generated gold-standard order.

4.4 Metrics Comparison Works

There are some researchers who did some work about the comparisons of different kinds of metrics. For example, (Callison-Burch et al., 2006b; Callison-Burch et al., 2007b; Lavie, 2013) mentioned that, through some qualitative analysis on some standard data set, BLEU can not reflect well of the MT systems' performance in many situations, i.e. higher BLEU score cannot ensure better translation outputs. Furthermore, there are some recently developed metrics that can perform much better than the traditional ones especially on the challenging sentence-level evaluation, though they are not popular yet such as nLEPOR and SentBLEU-Moses (Graham et al., 2015; Graham and Liu, 2016). Such kind of comparison works

will help MT researchers to select proper metrics to use for their special tasks.

5 Advanced Quality Estimation

In recent years, some MT evaluation methods that do not use the manually offered golden reference translations are proposed. They are usually called as "Quality Estimation (QE)". Some of the related works have already been mentioned in previous sections. The latest quality estimation tasks of MT can be found from WMT12 to WMT15 (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015). They defined a novel evaluation metric that provides some advantages over the traditional ranking metrics. The designed criterion DeltaAvg assumes that the reference test set has a number associated with each entry that represents its extrinsic value. Given these values, their metric does not need an explicit reference ranking, the way the Spearman ranking correlation does. The goal of the DeltaAvg metric is to measure how valuable a proposed ranking is according to the extrinsic values associated with the test entries.

$$DeltaAvg_v[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S) \quad (19)$$

For the scoring task, they use two task evaluation metrics that have been traditionally used for measuring performance for 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 $s_i, 1 \leqslant i \leqslant |S|$, they denote by $H(s_i)$ the proposed score for entry s_i (hypothesis), and by $V(s_i)$ the reference value for entry s_i (gold-standard value).

$$\text{MAE} = \frac{\sum_{i-1}^{N} |H(s_i) - V(s_i)|}{N} \quad (20)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i-1}^{N} (H(s_i) - V(s_i))^2}{N}} \quad (21)$$

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

6 Discussion and Conclusion

So far, the human judgment scores of MT results are usually considered as the golden standard that the automatic evaluation metrics should try to approach. However, some improper handlings in the process also yield 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 judgments are two important problems.

Automatic evaluation metrics are indirect measures of translation quality, because that they are usually using the various string distance algorithms to measure the closeness between the machine translation system outputs and the manually offered reference translations and they are based on the calculating of correlation score with manual MT evaluation (Moran and Lewis, 2012). Furthermore, the automatic evaluation metrics tend to ignore the relevance of words (Koehn, 2010), for instance, the name entities and core concepts are more important than punctuations and determiners but most automatic evaluation metrics put the same weight on each word of the sentences. Third, automatic evaluation metrics usually yield meaningless score, which is very test set specific and the absolute value is not informative. For instance, what is the meaning of -16094 score by the MTeRater metric (Parton et al., 2011) or 1.98 score by ROSE (Song and Cohn, 2011)? And similar goes to 19.07 by BEER / 28.47 by BLEU / 33.03 by METEOR for a mostly good translation in the paper (see experiments section Table 4) (Maillette de Buy Wenniger and Sima'an, 2015)? Instead, we find one interesting metric family LEPOR and hLEPOR (Han et al., 2012; Han, 2014) that can give a somehow meaningful score for a somehow recognized good translation, e.g. the score can be around 0.60 to 0.80.

The automatic evaluation metrics should try to achieve the goals of low cost, reduce time and money spent on carrying out evaluation; tunable, automatically optimize system performance towards metric;meaningful, score should give intuitive interpretation of translation quality;consistent, repeated use of metric should give

same results;correct, metric must rank better systems higher as mentioned in (Koehn, 2010), of which the low cost, tunable and consistent characteristics are easily achieved by the metric developers, but the rest two goals (meaningful and correct) are usually the challenges in front of the NLP researchers.

There are some related works about MT evaluation survey or literature review before. For instance, in the DARPA GALE report (Dorr et al., 2009), researchers first introduced the automatic and semi-automatic MT evaluation measures, and the task and human in loop measures; then, they gave a description of the MT metrology in GALE program, which focus on the HTER metric as standard method used in GALE; finally, they compared some automatic metrics and explored some other usages of the metric, such as optimization in MT parameter training.

In another research project report EuroMatrix (EuroMatrix, 2007), researchers first gave an introduction of the MT history, then, they introduced human evaluation of MT and objective evaluation of MT as two main sections of the work; finally, they introduced a listed of popular evaluation measures at that time including WER, SER, CDER, X-Score, D-score, NIST, RED, IER and TER etc.

Mrquez (Mrquez, 2013) 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 the chanllenges in the quality estimation task.

Our work differs with the previous ones, by introducing some recent development of MT evaluation models, the different classifications from manual to automatic evaluation measures, the introduction of recent QE tasks, and the concise construction of the content.

7 Perspective

In this section, we mention several aspects that are useful and will attract much attention for the further development of MT evaluation field.

Firstly, it is about the lexical similarity and the linguistic features. Because the natural languages are expressive and ambiguous at different levels (Gimén*e*z and M*á*rquez, 2007), lexical similarity based metrics limit their scope to the lexical dimension and are not sufficient to ensure that two sentences convey the same meaning or not. For instance, the researches of (Callison-Burch et al., 2006a) and (Koehn and Monz, 2006b) report that lexical similarity metrics tend to favor the automatic statistical machine translation systems. If the evaluated systems belong to different types that include rule based, human aided, and statistical systems, then the lexical similarity metrics, such as BLEU, give a strong disagreement between ranking results provided by them and the human evaluators.So the linguistic features are very important in the MT evaluation procedure. However, in-appropriate utilization, or abundant or abused utilization, will result in difficulty in promotion.In the future, how to utilize the linguistic features more accurate, flexible, and simplified, will be one tendency in MT evaluation.Furthermore, the MT evaluation from the aspects of semantic similarity is more reasonable and reaches closer to the human judgments, so it should receive more attention.

Secondly, the Quality Estimation tasks make some difference from the traditional evaluation, such as extracting reference-independent features from input sentences and the translation, obtaining quality score 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 among multiple systems, etc., so they will continuously attract many researchers.

Thirdly, some advanced or challenging technologies that can be tried for the MT evaluation include the deep learning (Gupta et al., 2015a; Zhang and Zong, 2015), semantic logic form, and decipherment model, etc.

8 Aknowledgement

ADAPT Centre for Digital Content Technology funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. Han thanks the funding by NWO VICI under Grant No. 277-89-002 of Netherlands, and by the Research Committee of the University of Macau (Grant No. MYRG2015-00175-FST and MYRG2015-00188-FST) and the Science and Technology Development Fund of Macau (Grant No. 057/2014/A). We thank Ying Shi, Derek F. Wong, Qun Liu, Khalil Sima'an for feedback and

helps.

References

J. Albrecht and R. Hwa. 2007. A re-examination of machine learning approaches for sentence-level mt evaluation. In *The Proceedings of the 45th Annual Meeting of the ACL, Prague, Czech Republic*.

Jon Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

D. Arnold. 2003. *Computers and Translation: A translator's guide-Chap8 Why translation is difficult for computers*. Benjamins Translation Library.

Eleftherios Avramidis, Maja Popovic, David Vilar, and Aljoscha Burchardt. 2011. Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of WMT*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL*.

Srinivas Bangalore, Owen Rambow, and Steven Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG*.

Regina Barzilayand and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings NAACL*.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of WMT*.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Christian Buck. 2012. Black box features for the wmt 2012 quality estimation shared task. In *Proceedings of WMT*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006a. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006b. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL*, volume 2006, pages 249–256.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007a. (meta-) evaluation of machine translation. In *Proceedings of WMT*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007b. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 64–71. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of WMT*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the 4th WMT*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaridan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the WMT*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaridan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of WMT*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of WMT*.

Michael Carl and Andy Way. 2003. Recent advances in example-based machine translation.

John B. Carroll. 1966. An experiment in evaluating the quality of translation. *Mechanical Translation and Computational Linguistics*, 9(3-4):67–75.

Boxing Chen, Roland Kuhn, and Samuel Larkin. 2012. Port: a precision-order-recall mt evaluation metric for tuning. In *Proceedings of the ACL*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Kenneth Church and Eduard Hovy. 1991. Good applications for crummy machine translation. In *Proceedings of the Natural Language Processing Systems Evaluation Workshop*.

Jasob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):3746.

Elisabet Comelles, Jordi Atserias, Victoria Arranz, and Irene Castellón. 2012. Verta: Linguistic features in mt evaluation. In *LREC*, pages 3944–3950.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining workshop*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. *Machine Learning Challenges:LNCS*, 3944:177–190.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. Tesla at wmt2011: Translation evaluation and tunable metric. In *Proceedings of WMT*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT Proceedings*.

Bonnie Dorr, Matt Snover, and etc. Nitin Madnani. 2009. Part 5: Machine translation evaluation. In *Bonnie Dorr edited DARPA GALE program report*.

Jennifer B. Doyon, John S. White, and Kathryn B. Taylor. 1999. Task-based evaluation for machine translation. In *Proceedings of MT Summit 7*.

H. Echizen-ya and K. Araki. 2010. Automatic evaluation method for machine translation using noun-phrase chunking. In *Proceedings of the ACL*.

Matthias Eck and Chiori Hori. 2005. Overview of the iwslt 2005 evaluation campaign. In *In proceeding of International Workshop on Spoken Language Translation (IWSLT)*.

Project EuroMatrix. 2007. 1.3: Survey of machine translation evaluation. In *EuroMatrix Project Report, Statistical and Hybrid MT between All European Languages, co-ordinator: Prof. Hans Uszkoreit*.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the iwslt 2011 evaluation campaign. In *In proceeding of International Workshop on Spoken Language Translation (IWSLT)*.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level mt evaluation without reference translations beyond language modelling. In *Proceedings of EAMT*, pages 103–112.

Daniel Gildea, Giorgio Satta, and Hao Zhang. 2006. Factoring synchronous grammars by sorting. In *Proceedings of ACL*.

Jesús Giménez and Lluís Márquez. 2008. A smorgasbord of features for automatic mt evaluation. In *Proceedings of WMT*, pages 195–198.

Jesús Giménez and Lluís Márquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of WMT*.

Yvette Graham and Qun Liu. 2016. Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1–10.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. 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*, pages 1183–1191.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceeding of SIGIR*.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015a. Machine translation evaluation using recurrent neural networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 380–384, Lisbon, Portugal, September. Association for Computational Linguistics.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015b. 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*, pages 1066–1072. Association for Computational Linguistics, o.A.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. 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)*, pages 805–814, Beijing, China, July. Association for Computational Linguistics.

Francisco Guzmn, Shafiq Joty, Llus Mrquez, and Preslav Nakov. 2017. Machine translation evaluation with neural networks. *Comput. Speech Lang.*, 45(C):180–200, September.

Anders Hald. 1998. *A History of Mathematical Statistics from 1750 to 1930*. ISBN-10: 0471179124. Wiley-Interscience; 1 edition.

Lifeng Han, Derek Fai Wong, and Lidia Sam Chao. 2012. Lepor: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING*.

Aaron Li Feng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, Shuo Li, and Ling Zhu. 2013a. 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*, pages 119–131.

Aaron Li Feng Han, Derek Fai Wong, Lidia Sam Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013b. A description of tunable machine translation evaluation systems in wmt13 metrics task. In *Proceedings of WMT*, pages 414–421.

Aaron Li Feng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, and Yi Lu. 2014. 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*, pages 1–12.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France, May. European Language Resources Association.

Lifeng Han. 2014. *LEPOR: An Augmented Machine Translation Evaluation Metric*. University of Macau, Macao.

Young Sook Hwang, Andrew Finch, and Yutaka Sasaki. 2007. Improving statistical machine translation using shallow linguistic knowledge. *Computer Speech and Language*, 21(2):350–372.

Maurice G. Kendall and Jean Dickinson Gibbons. 1990. *Rank Correlation Methods*. Oxford University Press, New York.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Maxim Khalilov and José A. R. Fonollosa. 2011. Syntax-based reordering for statistical machine translation. *Computer Speech and Language*, 25(4):761–788.

Marrgaret King, Andrei Popescu-Belis, and Eduard Hovy. 2003. Femti: Creating and using a framework for mt evaluation. In *Proceedings of the Machine Translation Summit IX*.

Philipp Koehn and Kevin Knight. 2009. Statistical machine translation, November 24. US Patent 7,624,005.

Philipp Koehn and Christof Monz. 2005. Shared task: Statistical machine translation between european languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

Philipp Koehn and Christof Monz. 2006a. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006b. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of WMT*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. 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*, pages 48–54. Association for Computational Linguistics.

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. 2007a. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007b. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jamal Laoudi, Ra R. Tate, and Clare R. Voss. 2006. Task-based mt evaluation: From who/when/where extraction to event understanding. In *in Proceedings of LREC06*, pages 2048–2053.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*.

Alon Lavie. 2013. Automated metrics for mt evaluation. *Machine Translation*, 11:731.

Guy Lebanon and John Lafferty. 2002. Combining rankings using conditional probability models on permutations. In *Proceeding of the ICML*.

Gregor Leusch and Hermann Ney. 2009. Edit distances with block movements and error rate confidence estimates. *Machine Translation*, 23(2-3).

Liang You Li, Zheng Xian Gong, and Guo Dong Zhou. 2012. Phrase-based evaluation for machine translation. In *Proceedings of COLING*, pages 663–672.

A. LI. 2005. Results of the 2005 nist machine translation evaluation. In *Proceedings of WMT*.

Chin-Yew Lin and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings NAACL*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of ACL*.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedingsof theACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2011. Better evaluation metrics lead to better machine translation. In *Proceedings of EMNLP*.

Chi Kiu Lo and Dekai Wu. 2011a. Meant: An inexpensive, high- accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of ACL*.

Chi Kiu Lo and Dekai Wu. 2011b. Structured vs. flat semantic role representations for machine translation evaluation. In *Proceedings of the 5th Workshop on Syntax and Structure in StatisticalTranslation (SSST-5)*.

Chi Kiu Lo, Anand Karthik Turmuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of WMT*.

Qingsong Ma, Fandong Meng, Daqi Zheng, Mingxuan Wang, Yvette Graham, Wenbin Jiang, and Qun Liu. 2016. 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*, pages 153–161.

Gideon Maillette de Buy Wenniger and Khalil Sima'an. 2015. Labeling hierarchical phrase-based models without linguistic resources. *Machine Translation*, 29(3):225–265.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram based machine translation. *Computational LinguisticsLinguistics*, 32(4):527–549.

Elaine Marsh and Dennis Perzanowski. 1998. Muc-7 evaluation of ie technology: Overview of results. In *Proceedingsof Message Understanding Conference (MUC-7)*.

Kathleen R. McKeown. 1979. Paraphrasing using given and new information in a question-answer system. In *Proceedings of ACL*.

Arul Menezes, Kristina Toutanova, and Chris Quirk. 2006. Microsoft research treelet translation system: Naacl 2006 europarl evaluation. In *Proceedings of WMT*.

Marie Meteer and Varda Shaked. 1988. Microsoft research treelet translation system: Naacl 2006 europarl evaluation. In *Proceedings of COLING*.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Douglas C. Montgomery and George C. Runger. 2003. *Applied statistics and probability for engineers*. John Wiley and Sons, New York, third edition.

John Moran and David Lewis. 2012. Unobtrusive methods for low-cost manual assessment of machine translation. *Tralogy I [Online], Session 5*.

L. Mrquez. 2013. automatic evaluation of machine translation quality. *Dialogue 2013 invited talk, extended*.

Sergei Nirenburg. 1989. Knowledge-based machine translation. *Machine Translation*, 4(1):5–24.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Kristen Parton, Joel Tetreault ans Nitin Madnani, and Martin Chodorow. 2011. E-rating machine translation. In *Proceedings of WMT*.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *Proceeding of IWSLT*.

M. Paul. 2009. Overview of the iwslt 2009 evaluation campaign. In *Proceeding of IWSLT*.

Karl Pearson. 1900. 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. *Philosophical Magazine*, 50(5):157–175.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: Ibm1 scores as evaluation metrics. In *Proceedings of WMT*.

M. Popovic and Hermann Ney. 2007. Word error rates: Decomposition over pos classes and applications for error analysis. In *Proceedings of WMT*.

Claus Povlsen, Nancy Underwood, Bradley Music, and Anne Neville. 1998. Evaluating text-type suitability for machine translation a case study on an english-danish system. In *Proceeding LREC*.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. 2011. "this sentence is wrong." detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

F. Sánchez-Martínez and M. L. Forcada. 2009. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34:605–635.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 54–59, Denver, Colorado, June. Association for Computational Linguistics.

Mattthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceeding of AMTA*.

Xingyi Song and Trevor Cohn. 2011. Regression and ranking based optimisation for sentence level mt evaluation. In *Proceedings of WMT*.

L. Specia and J. Giménez. 2010. Combining confidence estimation and reference-based metrics for segment-level mt evaluation. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Lucia Specia, Naheh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Machine Translation Summit XIII*.

Miloš Stanojević and Khalil Sima'an. 2014a. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Miloš Stanojević and Khalil Sima'an. 2014b. Evaluating word order recursively over permutation-forests. In *Proceedings of the Eight Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Miloš Stanojević and Khalil Sima'an. 2014c. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Keh-Yih Su, Wu Ming-Wen, and Chang Jing-Shin. 1992. A new quantitative quality measure for machine translation systems. In *Proceeding of COLING*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Proceeding of EUROSPEECH*.

Joseph P Turian, Luke Shea, and I Dan Melamed. 2006. Evaluation of machine translation and its evaluation. Technical report, DTIC Document.

Clare R. Voss and Ra R. Tate. 2006. 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*, pages 203–212.

Warren Weaver. 1955. Translation. *Machine Translation of Languages: Fourteen Essays*.

John S. White and Kathryn B. Taylor. 1998. A task-oriented evaluation metric for machine translation. In *Proceeding LREC*.

John S. White, Theresa O' Connell, and Francis O' Mara. 1994. The arpa mt evaluation methodologies: Evolution, lessons, and future approaches. In *Proceeding of AMTA*.

Billy Wong and Chun yu Kit. 2009. Atec: automatic evaluation of machine translation via word choice and word order. *Machine Translation*, 23(2-3):141–155.

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. 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*, pages 2042–2051.

Jiajun Zhang and Chengqing Zong. 2015. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, (5):16–25.