Machine Translation
https://doi.org/10.1007/s10590-019-09239-4

BOOK REVIEW

Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor, and Violeta Seretan (eds): Multiword units in machine translation and translation technology

Current Issues in Linguistic Theory, Volume 341, John Benjamin Publishing Company, Amsterdam & Philadelphia, 2018, ix+259 pp, ISBN 978-90-272-0060-0 (HB), ISBN 978-90-272-6420-6 (e-book)

Rejwanul Haque[1] · **Mohammed Hasanuzzaman**[2] · **Andy Way**[1]

Received: 9 August 2019 / Accepted: 20 August 2019

The book *Multiword Units in Machine Translation and Translation Technology* demonstrates the importance of multiword units (MWUs) in natural language processing (NLP) applications, and explores computational treatments of how they can be handled in NLP, particularly in machine translation (MT) and translation technology. The book was edited by Ruslan Mitkov (University of Wolverhampton), Johanna Monti (L'Orientale University of Naples), Gloria Corpas Pastor (University of Málaga), and Violeta Seretan (University of Geneva) who are renowned researchers in NLP and computational linguistics. The book contains twelve chapters including an introductory chapter compiled by the editors of the book themselves. The remaining eleven chapters were contributed by NLP researchers and experts in the field, and each of these chapters investigates specific problems relating to MWUs.

The introductory chapter (Chap. 1) has same title as the the book itself (*Multiword Units in Machine Translation and Translation Technology*), and presents a survey of the field with particular attention to MT and translation technology. At the start of the chapter, the authors present a comprehensive definition of MWUs with an example: "*Multiword units or multiword expressions are meaningful lexical units made of two or more words in which at least one of them is restricted by linguistic conventions in the sense that it is not freely chosen. For example, in the expression to 'smell a rat' the*

✉ Rejwanul Haque
rejwanul.haque@adaptcentre.ie

Mohammed Hasanuzzaman
mohammed.hasanuzzaman@adaptcentre.ie

Andy Way
andy.way@adaptcentre.ie

1 School of Computing, ADAPT Centre, Dublin City University, Dublin, Ireland

2 Department of Computer Science, Cork Institute of Technology, Cork, Ireland

Published online: 28 August 2019

word rat cannot be replaced with similar words such as mouse or rodent". The authors also provide definitions of multiword expressions (MWEs) from other papers (Firth 1957; Sag et al. 2002; Baldwin and Kim 2010; Ramisch 2015). The first section of the article highlights statistics relating to MWUs (e.g. frequency of occurrences of 'multiword', 'collocations', and 'idiom' in research papers collected from the ACL anthology)[1] from their own experiments and previous studies (Jackendoff 1997; Biber et al. 1999; Sag et al. 2002; Ramisch et al. 2013). Integration of MWU knowledge in NLP tasks and applications is a known problem and arguably one of the most challenging and complex processes in the field. The second section of the article presents a brief historical overview of computational treatments of MWUs in the following NLP areas: POS tagging, parsing, word sense disambiguation, information retrieval, information extraction, question answering, sentiment analysis, and text mining. As far as computational treatment of MWUs in MT is concerned, there are generally two major issues to be addressed: (a) the identification of the MWUs in the source language, and (b) their transfer into and correct generation in the target language. The third section of the article reviews some of the seminal works that illustrate the main approaches to MWU processing in different MT paradigms: rule-based MT, example-based MT, statistical MT (SMT) and neural MT (NMT). The final section of the paper discusses how MWUs are dealt with in Translation Memory systems and other support tools for translators.

The second article (Chap. 2: *Analysing linguistic information about word combinations for a Spanish-Basque rule-based machine translation system*), by Iñurrieta et al. describes an in-depth analysis with a particular type of MWU (i.e. *noun + verb combinations*) in Spanish-to-Basque translations. The first section of the paper claims that one has to take the level of *syntactic fixedness* of MWUs into account as far as their processing is concerned, citing the fact that a large number of MWUs can be separated by other words, and sometimes even the word order can be changed, which cannot be identified by applying the most basic and widely used method *words-with-spaces* strategy (Alegria et al. 2004; Zhang et al. 2006). The paper first undertakes an in-depth linguistic analysis of Basque and Spanish *noun + verb* combinations, and draws a conclusion, namely that MWUs cannot usually be translated word for word and morpheme for morpheme, as such expressions vary considerably from language to language. The authors select a tiny set of frequent *noun + verb* combinations in Spanish, and analyse them further and classify them according to their *syntactic fixedness* and their *semantic compositionality*, which helps to determine the kind of treatment that each MWU needs. In short, the article presents a study of the identification of MWUs and reports interesting observations with Spanish-to-Basque translation, using rule-based MT.

The third article of the book (Chap. 3: *How do students cope with machine translation output of multiword units? An exploratory study*), by Daems et al. investigates students' post-editing (PE) of MWUs from English-to-Dutch. The authors study the translation of MWUs from English to Dutch on the basis of their *degree of contrast* with the target language. In short, they divide English MWUs in two ways: by *category* (compound, collocation, multiword verb) and by *contrastiveness*: if a direct

[1] https://www.aclweb.org/anthology/.

translation of the English MWU would be correct in Dutch, the unit is classified as '*non-contrastive*', whereas MWUs that cannot be translated literally into Dutch are classified as '*contrastive*'. Their results indicate that contrastive MWUs are harder to translate for the MT system, and harder to correct by student post-editors than non-contrastive MWUs. The authors suggest that the difference between contrastive and non-contrastive MWUs is a useful new way of classifying MWUs with regard to MT and subsequent PE.

The fourth article (Chap. 4: *Aligning verb + noun collocations to improve a French–Romanian FSMT system*) was written by Todiraşcu et al. who aim to improve French–Romanian factored SMT (FSMT) (Koehn and Hoang 2007) by employing different collocation integration methods, namely (a) collocation extraction from monolingual corpora, using a hybrid method which combines morphosyntactic properties and frequency criteria, (b) use of a bilingual collocation dictionary to identify collocations, and (c) applying a specific alignment algorithm for identification of collocations. In their first strategy, given a large monolingual corpus, Todiraşcu et al. apply the well-known statistical measure log-likelihood ratio (Dunning 1993) in order to extract collocations. In particular, the authors restrict their investigations to a specific type of MWEs (i.e. *verb + noun* class of collocations), which are highly productive, vary in their syntactic structures and provide various degrees of compositionality. As far as MT is concerned, a word-for-word translation strategy fails to handle such MWEs, due to their strong lexical preferences.

Chapter 5 (*Multiword expressions in multilingual information extraction*), by Gregor Thurmair, presents computational treatments of MWUs in the context of *multilingual indexing* and cross-lingual information extraction. In short, MWUs [e.g. key terms, named entities (NE)] are translated and presented in the user's native language, to help determine whether or not the document is relevant with regard to a given profile of interest. The task also involves NE recognition and key term identification, and the languages involved are both European and Middle Eastern (Persian, Turkish, Arabic, Pashto). In addition to the translation and generation of the MWUs, the paper presents MWU processing as three subsequent sub-tasks: (a) *MWU extraction*, (b) *MWU representation*, and (c) *MWU analysis and identification*. The major portion of the article deals with the MWU representation sub-task (i.e. lexical representation of MWUs, annotation schemes, MWU extension). The MWU analysis sub-task treats MWUs as an integral part of the analysis itself, while abandoning pre-processing [(e.g. the *words-with-spaces* approach (Samaridi and Markantonatou 2014)] and post-analysis steps due to many drawbacks.

Chapter 6 (*A multilingual gold standard for translation spotting of German compounds and their corresponding multiword units in English, French, Italian and Spanish*), by Clematide et al. describes a multilingual gold standard for German nominal compounds and their multiword translation equivalents in English, French, Italian, and Spanish. The gold standard was created for the evaluation of translation spotting[2] of German compounds. The paper first talks about selection and preprocessing of the gold standard material and automatic linguistic annotations and the alignment pro-

[2] The term "translation spotting" refers to the task of identifying the target-language words that correspond to a given set of source-language words in a pair of text segments known to be mutual translations (Simard 2003).

cess for sentences and words. Then, the chapter explains sampling criteria for German compounds, annotation guidelines and processes, and finally presents the performance of their proposed model measured against the gold standard.

Chapter 7 (*Dutch compound splitting for bilingual terminology extraction*) by Macken & Tezcan describes a compound splitter for Dutch that makes use of corpus frequency information and linguistic knowledge. The authors show that the compound splitter combined with a proposed word alignment technique considerably improves bilingual terminology extraction results. In theory, the *union* and the *grow-diag-final* heuristics of Moses (Koehn et al. 2007) result in an overall word alignment with a gain in recall but which causes a substantial loss in precision. The authors describe why this poses a problem for applications such as bilingual terminology extraction in which precision is important. In addition to this problem, the authors explain why Dutch compound words lead to data sparseness. The authors propose a solution to overcome the problems of data sparseness and word alignment. In a nutshell, the authors demonstrate how their proposed compound splitter combined with a word alignment technique improves bilingual terminology extraction results.

Chapter 8 (*A flexible framework for collocation retrieval and translation from parallel and comparable corpora*) by Rivera et al. presents a framework for collocation retrieval and translation from parallel and comparable corpora, developed with translators and language learners in mind. Collocations are compositional and statistically idiomatic MWEs (Baldwin and Kim 2010); their translations are crucial in many NLP applications, especially in MT and CAT tools. Notably, translators and language learners encounter difficulties in translating collocations. The authors describe the development, implementation and exploitation of language-independent computational tools, e.g. Tree Tagger (Schmid 1994) and the MWE Toolkit (Ramisch 2012) for collocation extraction according to specific POS-patterns, and Hunalign (Varga et al. 2007) for collocation translation. The paper also describes to what extent the proposed framework aids language learners and translators to retrieve collocations in a source language and their translations in a target language from bilingual parallel and comparable corpora.

Chapter 9 (*On identification of bilingual lexical bundles for translation purposes*), by Łukasz Grabowski, explores the use of bilingual *lexical bundles* (Biber et al. 1999), a peculiar type of MWUs,[3] to improve the *degree of naturalness* and *textual fit* of translated texts. In particular, the authors present a framework that extracts bilingual lexical bundles semi-automatically from an English–Polish comparable corpus of patient information leaflets for integration into PB-SMT systems or CAT software.

With an end goal identical to that of the article of Chapters 9 and 10 (*The quest for Croatian idioms as multiword units*) by Kocijan and Librenjak presents Nooj, a rule-based automated text processing tool, which can recognise idiomatic expressions in text as *continuous* and *discontinuous* MWUs by using grammatical rules. The authors differentiate five basic syntactic types of idiomatic expressions, and the major portion of the article explains the creation of syntactic grammar rules for each of the five types.

[3] Definition of lexical bundle from Wikipedia: "a sequence of two or more words that occur in language with high frequency but are not idiomatic; a bundle, chunk, or cluster".

Finally, the article shows that Nooj can detect idioms in all kinds of Croatian texts, regardless of their genre and the inflection or inversion of the idiom in case.

Chapter 11 (*Corpus analysis of Croatian constructions with the verb doći 'to come'*) by Bartolec and Ivanković demonstrates a corpus-based analysis of constructions consisting of the verb *doći* 'to come' followed by the prepositions *do* 'to' or *na* 'onto' and a noun. The authors present an analysis of the lexicographic description of the verb doći in three contemporary Croatian dictionaries, followed by a lexicographic analysis of the same in three Croatian corpora.

As stated above while discussing Chap. 8, collocation identification is viewed as a crucial task in many NLP applications, and particularly in MT and CAT environments. The same statement can also be applicable for *anaphora resolution* (Mitkov 2002). The final chapter of the book (Chap. 12: *Anaphora resolution, collocations and translation*) by Wehrli and Nerima focuses on the intersection of these two problems: *collocation identification* and *anaphora resolution*. In particular, the article presents treatments for the translation from English to French of collocations of the type verb-direct object, in which the object has been pronominalised. First, the authors explain translation problems with respect to both collocations and anaphors, and show how current MT systems fail to handle such cases. Then, they try to address the problems with their in-house rule-based MT system, Its-2 (Wehrli et al. 2009).

In summary, the book represents many interesting topics in the area of computational treatment of multiword expressions, with a special focus on MT and translation technology. The book covers the treatments of different types of MWUs (e.g. idioms, lexical bundles, collocations, compounds) in translation (i.e. by MT system or CAT software) and a number of use-cases (e.g. translation between European and Middle Eastern languages, impact on bilingual terminology extraction, impact on post editing of MT, use of parallel and comparable corpora). The first chapter by the editors of the book presents an extensive background study and survey on the computational treatment of MWUs in NLP applications (particularly with respect to different MT approaches). This book can essentially be viewed as an important contribution to a specialised area (i.e. computational treatment of MWUs) of interest, which will be a great help to NLP researchers, and MT researchers and users in particular.

Acknowledgements The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. This project has partially received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 713567, and the publication has emanated from research supported in part by a research Grant from SFI under Grant No. 13/RC/2077.

References

Alegria I, Ansa O, Artola X, Ezeiza N, Gojenola K, Urizar R (2004) Representation and treatment of multiword expressions in Basque. In: Proceedings of the workshop on multiword expressions: integrating processing. Barcelona, pp 48–55

Baldwin T, Kim SN (2010) Multiword expressions. In: Indurkhya N, Damerau FJ (eds) Handbook of natural language processing, 2nd edn. CRC Press, Boca Raton, pp 267–292

Biber D, Johansson S, Leech G, Conrad S, Finegan E (1999) Longman grammar of written and spoken english. Longman, London

Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. Comput Linguist 19(1):61–74

Firth JR (1957) Papers in linguistics 1934–1951. Oxford University Press, Oxford

Jackendoff R (1997) The architecture of the language faculty. MIT Press, Cambridge

Koehn P, Hoang H (2007) Factored translation models. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). Prague, pp 868–876

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, College W, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: ACL 2007, proceedings of the interactive poster and demonstration sessions. Prague, pp 177–180

Mitkov R (2002) Anaphora resolution. Longman, London

Ramisch C (2012) A generic framework for multiword expressions treatment: from acquisition to applications. In: Proceedings of ACL 2012 student research workshop. Jeju Island, pp 61–66

Ramisch C (2015) Multiword expressions acquisition: a generic and open framework. Springer, New York

Ramisch C, Villavicencio A, Kordoni V (2013) Introduction to the special issue on multiword expressions: from theory to practice and use. ACM Trans Speech Lang Process 10(2):3

Sag IA, Baldwin T, Bond F, Copestake A, Flickinger D (2002) Multiword expressions: a pain in the neck for NLP. In: Gelbukh A (eds) International conference on intelligent text processing and computational linguistics. Lecture Notes in computer science. Springer, Berlin, pp 1–15

Samaridi N, Markantonatou S, (2014) Parsing modern greek verb mwes with lfg/xle grammars. In: Proceedings of the 10th workshop on multiword expressions (MWE). Gothenburg, pp 33–37

Schmid H (1994) Probabilistic part-of-speech tagging using decision trees. In: Proceedings of international conference on new methods in language processing (NEMLAP), vol 12. Manchester, pp 44–49

Simard M (2003) Translation spotting for translation memories. In: Proceedings of the HLT-NAACL 2003 workshop on building and using parallel texts: data driven machine translation and beyond. Edmonton, pp 65–72

Varga D, Halácsy P, Kornai A, Viktor NB, Laszlo N, Laszlo, N, Viktor T (2007) Parallel corpora for medium density languages. In: Proceedings of proceedings of RANLP 2007: recent advances in natural language processing. Borovets, pp 590–596

Wehrli E, Seretan V, Nerima L, Russo L (2009) Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In: Proceedings of the 13th annual meeting of the european association for machine translation. Barcelona, pp 128–135

Zhang Y, Kordoni V, Villavicencio A, Idiart M (2006) Automated multiword expression prediction for grammar engineering. In: Proceedings of the workshop on multiword expressions: identifying and exploiting underlying properties. Sydney, pp 36–44