Processing MWE in Machine Translation Systems

Amalia Todirascu, University of Strasbourg, France,

Johanna Monti, University of Sassari, Italy

WG3: WG 3.2 MWE and Translation

Multi-word expressions include a wide list of categories such as idiomatic expressions, proverbs, collocations, compound words, domain-specific terms, named entities. According to Baldwin and Kim (2010) and (Sag, Baldwin, Bond, Copestake and Flickinger 2002), MWE are lexical items composed of several lexemes which display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity. Idiomaticity means deviation from the basic properties of the component items: MWEs have various behaviours, more or less predictable from the properties of the individual items. If some categories of MWEs present a high degree of fixedness in word order and their properties (idioms, domain specific terms), other categories such as collocations are more flexible (accepting modifiers, various senses according to context). These expressions are characterised by specific morpho-syntactic properties (preference for some determiners, numbers or genders), strong lexical preference (*poser une question* but *ask a question*), specific syntactic behaviour and non-compositional sense (Hausmann, 2004). MWE represent major difficulties for MT systems (Hurskainen, 2008). Most of the MT systems fail to produce correct translations of MWE, while word-to-word translations are not appropriate and the sense is more or less compositional. Whatever the method adopted for translation (statistical, rule-based or example based), producing accurate MWE translation is still a challenge.

We present an overview of the existing strategies used to handle MWE in different MT approaches. The use of external resources such as terminological databases, collocation dictionaries or list of named entities might improve the results of MT systems, especially for continuous MWE, but unfortunately these resources are often incomplete or not available for all the languages or domains.

According to the different approaches to MT (rule-based, statistical, example-based, hybrid) alternative strategies to cope with MWEs are used. In this paper, we present an overview of these strategies which are adopted in different stages of the MT process, again according to the different MT approaches, i.e. during the pre or post-processing, the alignment process, etc. Besides, different strategies are used sometimes to process and translate specific categories of MWEs (terms, compound words, NE etc.). Evaluation of MWE alignment and translation requires specific corpora annotating these phenomena.

SMT, which is nowadays the dominant approach in MT, uses, in general, frequency criteria to identify and translate MWE with some differences in the identification of MWEs. Some MT systems identify MWE before starting the lexical alignment and the translation process (Lambert and Banchs, 2006). Other systems include MWE alignment during the processing step (Melamed, 1997).

More recently hybrid methods integrate linguistic knowledge in SMT in order to overcome MWE mistranslations. They use context properties and statistical methods to propose a list of MWE candidates. Statistical systems use frequency criteria to identify MWE expressions which is a valuable strategy for fixed expressions. Phrase-based systems, based on translation tables containing n-grams, might be completed with domain specific bilingual dictionaries (Wu et al, 2008) but other approaches try to integrate syntactic and semantic structures (Chiang, 2005; Marcu et al., 2006; Zollmann & Venugopal, 2006), in order to obtain better translation results.

In SMT, specific attention is devoted to MWE alignment strategies, exploiting linguistic and contextual properties of MWE, or combining frequency criteria and morpho-syntactic properties. Linguistic rules using POS tags are designed to extract MWE from parallel corpora (Bouamor, et al, 2012). MWE alignment is also based on existing simple alignment (Villada Moiron and Tiedemann, 2006).According to their category, specific strategies might be applied to detect noun compounds, verb compounds or named entities. Domain-specific term identification methods could be applied to recognize terms (Dagan and Church, 1994), (Macken et al, 2008), (Kontonatsios, 2014). Some of these methods use the simple lexical alignment as a basis to build new alignments.

Rule-based machine translation systems adopt lexical approaches to identify contiguous MWE, using existing monolingual or bilingual dictionaries. The lexical approach is sometimes integrated by compositional ones, in which specific rules handle non-contiguous, compositional MWE: OpenLogos (Scott, 2003; Scott and Barreiro, 2009; Barreiro *et al.*, 2011).
Finally, example based systems uses examples of possible translations of MWEs, integrated in many cases by linguistic rules (Franz et al. 2000), (Gangadharaiah and Balakrishnan, 2006). They use alignment to extract possible MWE alignments.

(Bouamor *et al*, 2012), (Ren *et al.*, 2009) show that the integration of various MWE detection strategies improves the quality of the MT system.

To evaluate the performance of the MT systems, evaluation corpora containing MWE alignment should be build and specific evaluation strategies should be developed. Annotating MWE in parallel texts involves several problems: these expressions are often discontinuous. Variability in the syntactic structure of these MWE expressions are often sources of ambiguities and of annotator disagreement. Their translation equivalents might be single words but also MWE. Most of the available resources contain alignments of specific collocation classes, for specific languages. (Ramisch *et al*, 2013) builds evaluation corpora annotated by several human annotators for verb compounds for English and French, while (Navlea, 2014) manually builds an evaluation corpora containing aligned, sometimes discontinuous verb+noun collocations, from the law and administrative domain, available for French and Romanian. (Barreiro *et al*, 2014) propose that comparative evaluation tasks among different approaches to MWE translation require the development of specific evaluation corpora for different types of MWEs. In this case, the translation performance of different approaches to each particular type of multiword unit would be adequately evaluated.

References

Baldwin, T. & Kim, (2010) *Multiword expressions in Indurkhya and Damerau Handbook of Natural Language Processing*, Second Edition, pp 267-292

Bouamor, D., Semmar, N., Zweigenbaum, P. (2012) Identifying bilingual multi-word expressions for statistical machine translation. In *LREC 2012*, pages 674-679, Istanbul, Turkey, 2012. ELRA.

Barreiro A., Monti J., Orliac B., Preuß S., Arrieta K., Ling W., Batista F., Trancoso I. (2014) Linguistic Evaluation of Support Verb Construction Translations by OpenLogos and Google Translate, In *Proceedings LREC 2014.*

Barreiro, A., Scott, B., Kasper, W., Kiefer, B. (2011) OpenLogos: Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. In: *Machine Translation* 25, 2011, pp. 107-126

Dagan, I., Church, K. (1994) Termight: Identifying and translating technical terminology. In Proceedings of the fourth conference on Applied natural language processing, pages 34–40. Association for Computational Linguistics.

Hurskainen, A. (2008). *Multiword Expressions and Machine Translation*. Technical Reports in Language Technology Report No 1, 2008

Kontonatsios, G., Korkontzelos, I., Tsujii, J., Ananiadou, S. (2014). Combining String and Context Similarity for Bilingual Term Alignment from Comparable Corpora. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1701-1712, Association for Computational Linguistics

Lambert P. and Banchs R. (2006). Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation. In *Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context*. Trento, Italy.

Macken, L, Lefever, E., Hoste, V. (2008) Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 529–536

Monti J., Barreiro, A., Elia, A., Marano, F., Napoli, A. (2011) Taking on new challenges in multi-word unit processing for Machine Translation, F. Sanchez-Martinez, J.A. Perez-Ortiz (eds.) *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation* Barcelona, Spain, gennaio 2011: 11-19

Monti J. (2012) *Multi-word Unit Processing in Machine Translation. Developing and using language resources for multi-word unit processing in Machine Translation* – PhD dissertation in Computational Linguistics - University of Salerno - Italy, 2012

Navlea, M. (2014) *La traduction automatique statistique factorisée, une application à la paire de langues français – roumain*, Ph.D.Thesis, Université de Strasbourg, France, June 2014.

Ramisch, C., De Araujo, V., Villavicencio, A. (2012) A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions, *Proceedings of the ACL 2012 Student Research Workshop*, Jeju, Republic of Korea, July, 2012.

Ren, Z., Lu, Y., Liu, Q. and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, pages 47–57.

Sag, I., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing (CICLing – 2002)*, volume 2276/2010 de LNCS, pp. 1–15, Mexico City, Mexico.

Scott, B. and Barreiro, A. (2009) OpenLogos MT and the SAL representation language. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation,* Alicante, Spain: Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos. 2–3 November 2009, pp. 19–26

Wu, H., Wang, H., & Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. *Proceedings of Conference on Computational Linguistics (COLING)*: 993-100.