MUMTTT2015

Edited by

Gloria Corpas Pastor, Ruslan Mitkov, Johanna Monti and Violeta Seretan

Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT2015)

(2nd edition)

ORGANISING COMMITTEE/COMITÉ ORGANIZADOR

Chair/Presidencia

- Prof. Gloria Corpas Pastor (Universidad de Málaga, Spain)
- Prof. Ruslan Mitkov (University of Wolverhampton, United Kingdom)
- Dr. Johanna Monti (Università degli Studi di Sassari, Italy)
- Dr. Violeta Seretan (Universitè de Genève, Switzerland)

Organisers/Organizadores

- Rosario Bautista Zambrana
- Cristina Castillo Rodríguez
- Hernani Costa
- Isabel Durán Muñoz
- Jorge Leiva Rojo
- Gema Lobillo Mora
- Pablo Pérez Pérez
- Míriam Seghiri Domínguez
- M.ª Cristina Toledo Báez
- Míriam Urbano Mendaña
- Anna Zaretskaya

Secretary/Secretaría

- Míriam Buendía Castro
- Rut Gutiérrez Florido

COMITÉ CIENTÍFICO/ PROGRAMME COMMITTEE

- Iñaki Alegria (Universidad del País Vasco, Spain))
- Giuseppe Attardi (Università di Pisa, Italy)
- Doug Arnold (University of Essex, United Kingdom)
- António Branco (Universidade de Lisboa, Portugal)
- Paul Buitelaar (National University of Ireland, Galway)
- František Čermák (Univerzita Karlova v Praze, Czech Republic)
- Jean-Pierre Colson (Université Catholique de Louvain, Belgium)
- Matthieu Constant (Université Paris-Est, France)
- Gaël Dias (Université de Caen Basse-Normandie, France)
- Mike Dillinger (Association for Machine Translation in the Americas)
- Dmitrij O. Dobrovol'skij (Russian Academy of Sciences, Russia)
- Peter Ďurčo (Univerzita sv. Cyrila a Metoda v Trnave, Slovak Republic)
- Marcello Federico (Fondazione Bruno Kessler, Italy)
- Sabine Fiedler (Universität Leipzig, Germany)
- Natalia Filatkina (Universität Trier, Germany)
- Thierry Fontenelle (Translation Centre for the Bodies of the European Union, Luxembourg)
- Corina Forăscu (Al.I. Cuza University of Iasi, Romania)
- Thomas François (Université Catholique de Louvain, Belgium)
- Ulrich Heid (Universität Hildesheim, Germany)
- Kyo Kageura (University of Tokyo, Japan)
- Cvetana Krstev (University of Belgrade, Serbia)
- Koenraad Kuiper (University of Canterbury, New Zealand)
- Alon Lavie (Carnegie Mellon University, USA)
- Malvina Nissim (Università di Bologna, Italy)
- Preslav Nakov (Qatar Computing Research Institute, Qatar Foundation, Qatar)
- Michael Oakes (University of Wolverhampton, United Kingdom)
- Adriana Orlandi (Università degli studi di Modena e Reggio Emilia, Italy)
- Yannick Parmentier (Université d'Orléans, France)

- Pavel Pecina (Univerzita Karlova v Praze, Czech Republic)
- Carlos Ramisch (Universié de Grenoble, France)
- Victoria Rosén (Universitetet i Bergen, Norway)
- Michael Rosner (University of Malta)
- Manfred Sailer (Goethe Universität, Germany)
- Tanja Samardžić (Universität Zurich, Switzerland)
- Agata Savary (Université François Rabelais Tours, France)
- Gerold Schneider (UniversitätZurich, Switzerland)
- Gilles Sérasset (Université de Grenoble, France)
- Max Silberztein (Universié de Franche-Comté, France)
- Kiril Simov (Bulgarian Academy of Sciences, Bulgaria)
- Kathrin Steyer (Institut für Deutsche Sprache, Germany)
- Joanna Szerszunowicz (University of Bialystok, Poland)
- Marko Tadić (University of Zagreb, Croatia)
- Amalia Todirascu (Université de Strasbourg, France)
- Beata Trawinski (Institut für Deutsche Sprache Mannheim, Germany)
- Dan Tufiş (Romanian Academy, Romania)
- Agnès Tutin (Université de Grenoble, France)
- Lonneke van der Plas (Universität Stuttgart, Germany)
- Veronika Vincze (University of Szeged, Hungary)
- Martin Volk (Universität Zurich, Switzerland)
- Eric Wehrli (Université de Genève, Switzerland)
- Michael Zock (Aix-Marseille Université, France)

INVITED SPEAKER/CONFERENCIANTE PLENARIO

Kathrin Steyer

Research Assistant at the Institute of German Language in Mannheim. She is head of the IDS Project *Usuelle Wortverbindungen* (multi-word expressions). In 2014, she was elected as president of EUROPHRAS.

“Multi word patterns and networks. How corpus-driven approaches have changed our description of language use”

2 July 2015/2 de junio de 2015

Due to the rise of corpus linguistics and the feasibility of studying language data in new quantitative dimensions, it became more and more evident that language use is fundamentally made up by fixed lexical chunks, set phrases, long distance word groups and multiword expressions (MWE). Sinclair’s inductively reconstructed collocations (cf. 1991) and Hausmann’s collocation pairs (cf. 2004) are the two leading concepts in collocation research. Basically, they are merely different ways of looking at the same fundamental principle of language, namely linguistic frozenness and fixedness. Compositional collocations and idioms differ in their degree of lexical fixedness and semantic opacity, their recognisability and prototypicality (Moon 1998, Burger et al. 2007). But they all share the most important characteristic: They are congealed into autonomous units in order to fill a specific role in communication. All these fragments are fixed patterns of language use (cf. Hunston/Francis 2000; cf. Hanks 2013). There is no core and no periphery. The difference is only in the degree of conspicuousness for the observer. These word clusters did not become fixed expressions by chance, but because there was a need of speakers for an economic way to communicate (cf. Steyer 2013).

Two assumptions constitute the basis of my talk:

– MWEs usually have multiple entries in the mental lexicon: on the one hand as more or less specified lexical units (lexemes) and on the other hand as (proto)typical realisations of a more abstract multiword pattern (MW pattern): for example [für ADJ Ohren klingen] (lit. ‘to sound for ADJ ears) ADJ fillers:

deutsche ('German') / westliche ('Western') / heutige ('contemporary') / europäische ('European') / ungeübte ('untrained').

– Independent of their lexical fixedness or variability, MWEs possess a holistic quality in the sense that they fulfil a specific role in communication as autonomous language units. This does not mean that they necessarily have an idiomatic meaning – sometimes they are completely transparent and compositional. The holistic quality can be attached to an abstract pattern and be functional in nature.

MWEs and MW patterns are not clear-cut and distinct entities. On the contrary, fragments and overlapping elements with fuzzy borders are typical for real language use. This means that there really are no MWEs as such. In real communicative situations, some components are focused while others fade into the background.

In my talk, I first discuss the nature of MW patterns that are reconstructed with complex corpus-driven methods. The examples are all taken from the German Reference Corpus (Deutsches Referenzkorpus) (cf.DeReKo) (located at the Institute for the German Language in Mannheim, IDS).

I show how we use an iterative methodology (quantitative - qualitative) to detect the nature of lexical fillers of pattern gaps and to visualize MWE hierarchies and networks. This methodology includes complex phrase searches and reciprocal analysis with COSMAS II (The IDS Corpus Search, Management and Analysis System); collocation analysis (cf. Belica 1995) that not only detects significant word pairs, but also significant syntagmatic cotext patterns; and slot analysis with the help of our UWV Tool that allows us to bundle KWICs. At the end, I will present a vision of a pattern-based lexicographic representation for humans ("MWE fields") (Steyer et al. 2013).

References

BELICA, C. 1995. Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache.

BURGER, H., D. DOBROVOL'SKIJ, P. KÜHN AND NEAL R. NORRICK. 2007 (eds.). *Phraseologie. Ein internationales Handbuch zeitgenössischer Forschung/Phraseology. An international Handbook of Contemporary Research.* (2 Editions). (= HSK 28, 1/2). Berlin/New York: de Gruyter.

DEREKO 2014. *Deutsches Referenzkorpus. / Archiv der Korpora geschriebener Gegenwartssprache 2014-II* (Release 11.09.2014). Mannheim: Institut für

Deutsche Sprache. Accessed on 17 March 2015. http://www.ids-mannheim.de/DeReKo.

HANKS, P. 2013. *Lexical Analysis. Norms and Exploitations*. Cambridge, Massachusetts/London: The MIT Press.

HAUSMANN, F. J. 2004. 'Was sind eigentlich Kollokationen?' In K. Steyer (ed.), *Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. Berlin/New York: de Gruyter, 309-334.

HUNSTON, S. AND G. FRANCIS. 2000. *Pattern Grammar. A corpus-driven approach to the lexical grammar of English.* Amsterdam/Philadelphia: John Benjamins.

MOON, R. 1998. Fixed Expressions and idioms in English. A Corpus-Based Approach. Oxford: Clarendon Press.

SINCLAIR, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

STEYER, K. 2013. *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. (Studien zur Deutschen Sprache 60). Tübingen: Narr.

STEYER, K., A. BRUNNER AND C. ZIMMERMANN 2013. Wortverbindungsfelder Version 3: Grund. Accessed on 31 March 2015. http://wvonline.ids-mannheim.de/wvfelder-v3/index.html.

TABLE OF CONTENTS/ÍNDICE DE CONTENIDOS

AUTOMATIC EXTRACTION OF MULTILINGUAL MWU RESOURCES. EXTRACCIÓN AUTOMÁTICA DE MWU A PARTIR DE RECURSOS MULTILINGÜES.

IDENTIFICATION, ACQUISITION AND EVALUATION OF MULTI-WORD TERMS. IDENTIFICACIÓN, ADQUISICIÓN Y EVALUACIÓN DE MWU Y VARIANTES.

MULTILINGUALISM AND MWU PROCESSING. MWUS AND WORD ALIGNMENT TECHNIQUES.

MULTILINGÜISMO Y PROCESAMIENTO DE MWU. MWU Y TÉCNICAS DE ALINEACIÓN DE PALABRAS.

LEARNING SEMANTIC INFORMATION ABOUT MWUS FROM MONOLINGUAL, PARALLEL OR COMPARABLE CORPORA.

ADQUISICIÓN DE INFORMACIÓN SEMÁNTICA SOBRE MWU A PARTIR CORPUS MONOLINGÜES, PARALELOS Y COMPARABLES.

MWUS IN MACHINE TRANSLATION.

MWU Y TA.

LEXICAL, SYNTACTIC, SEMANTIC AND TRANSLATIONAL ASPECTS IN MWU REPRESENTATION.

REPRESENTACIÓN DE UNIDADES FRASEOLÓGICAS O UNIDADES PLURIVERBALES (MWU): ASPECTOS LÉXICOS, SINTÁCTICOS, SEMÁNTICOS Y DE TRADUCCIÓN.

Automatic extraction of multilingual MWU resources Extracción automática de MWU a partir de recursos multilingües

BUILDING A LEXICAL BUNDLE RESOURCE FOR MT

Natalia Grabar

Université de Lille 3

natalia.grabar@univ-lille3.fr

Marie-Aude Lefer

Marie Haps School of Translation
and Interpreting - Brussels

marie-aude.lefer@ilmh.be

Thanks to the growing availability of large aligned corpora and multilingual resources, such as lexicons and term bases, machine translation has become a well-established and vibrant field of research. However, MT resources are still rather scarce and often need to be adapted and enriched to be applicable to a wide range of specialized domains and text types (see e.g. Arcan et al. 2014). Typically, general language resources are restricted to words (as opposed to MWUs), while term bases, though containing numerous complex terms, fail to include MWUs that are used to express stance (i.e. opinion and degrees of certainty, e.g. it is very important that, it seems to me that) or to structure texts (e.g. and that is why, when it comes to) (see also ten Hacken and Fernández Parra 2008 for similar remarks for CAT).

In this presentation, we try to go some way towards filling this gap by examining the translation of lexical bundles, i.e. “sequences of word forms that commonly go together in natural discourse” (Biber et al. 1999: 990ff), focusing on discourse organizers and stance expressions (see e.g. Biber et al. 2004). In particular, we show how we have compiled a corpus-informed bilingual resource for the English-French language pair that can be used for MT. The method used combines data extracted from comparable and parallel corpora. First (Step 1), lexical bundles in English and French original texts were automatically extracted via the n-gram method. For this, we relied on four corpora, representing four genres: Europarl (Koehn 2005, Cartoni and Meyer 2012), KIAP (research articles; Fløttum et al. 2006), Multed (editorials) and PLECI (news). Together they total 7+ million tokens. We extracted all trigrams (min. 5 occurrences per genre) and the longer n-grams containing them (min. 2 occurrences per genre) and restricted the analysis to the bundles that were found in at least three of the four genres investigated, as they are more likely to be relevant for a rather wide range of text types. This corresponds to 3,251 and 1,600 trigrams in French and English, respectively, with varying numbers of corresponding longer n-grams. Second (Step 2), relying on three parallel corpora (Europarl, PLECI-news and Label France) aligned at word-level with Giza++ (Och and Ney 2000), the discourse organizers and stance expressions identified in Step 1 were automatically matched with their target language equivalents (examples of selected source language bundles include in an attempt to, the result is that, one of the reasons why, made it clear that, it remains to be seen, there is no reason to, now is the time to, it would be wrong to, there is no doubt; il en va de même pour, dans le même temps, de ce point de vue, le cas échéant, pour ne pas dire, ce qui est vrai, d'une certaine façon, il est évident que).

In our presentation, we will describe the method adopted to build the lexical bundle resource and discuss the challenges posed by the discourse organizers and stance expressions in translation (such as the automatic alignment of bundles containing grammatical or highly-frequent words). We will also describe various translation phenomena observed in the parallel corpus data, such as the polyfunctionality of some bundles (e.g. as far as: en ce qui concerne/pour ce qui concerne/s'agissant/concernant/au sujet de vs. aussi loin), the genre-sensitivity of target language equivalents, and categorial changes (e.g. it would be wrong to: ce serait une erreur de).

References

ARCAN, M., GIULIANO, C., TURCHI, M. AND BUITELAAR, P. (2014). Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation. COMPUTERM worskhop at COLING 2014: 22-31.

BIBER, D., CONRAD, S. AND CORTES, V. (2004). *If you look at* … Lexical Bundles in University Lectures and Textbooks. *Applied Linguistics*, 25, 371-405.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. AND FINEGAN, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

CARTONI, B. AND MEYER, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. *8th International Conference on Language Resources and Evaluation (LREC)*. Available at:< https://www.idiap.ch/~tmeyer/res/Cartoni-LREC-2012.pdf>

FLØTTUM, K., DAHL, T. AND KINN, T. (2006). *Academic Voices – across languages and disciplines*. Amsterdam & Philadelphia: John Benjamins.

HACKEN, P. TEN AND FERNÁNDEZ PARRA, M. (2008). Terminology and Formulaic Language in Computer-Assisted Translation. *SKASE Journal of Translation and Interpretation*, 3, 1-16.

KOEHN, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit X*, 79-86. Available at:<http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf>

OCH, F.J. AND NEY, H. (2000). Improved Statistical Alignment Models. In *Proceedings of ACL*. 440-447.

ASSESSING WORDNET FOR BILINGUAL COMPOUND DICTIONARY EXTRACTION

Carla Parra Escartín
University of Bergen
carla.parra@uib.no

Héctor Martínez Alonso
University of Copenhagen
Dennmark alonso@hum.ku.dk

In this paper we present work to explore ways of automatically retrieving compound dictionaries from sentence-aligned corpora using WordNet. More specifically, we focus on the pair of languages German to Spanish and try to retrieve the Spanish translational correspondences of German nominal compounds. German compounds are a challenge because their correspondences into other languages are not straightforward and better methods for aligning them successfully to their translations in parallel corpora are needed. We carried out a pilot experiment to assess whether it is possible to align the formants of a German compound with the words in the Spanish translation which correspond to the main WordNet categories.

As Sag et al. (2001) argue in their seminal ``pain in the neck" paper, Multiword Expressions (MWEs) are a major bottleneck for many Natural Language Processing (NLP) applications. Our research had as a starting point a real problem for human translation and Machine Translation (MT), and therefore is application-driven. Although we focus on compound dictionary extraction, the ultimate aim is to integrate them in Statistical Machine Translation (SMT) tasks.

Our working hypothesis was that the different formants of a compositional compound share semantic features with their corresponding translational equivalents in other languages. Our pilot experiment consisted on semantically tagging the formants of the compounds in German and their Spanish translations using WordNet, and trying to find possible overlappings across languages. To run this experiment, we created a Gold Standard consisting of German compounds and their Spanish translations. The data was extracted from a 261-sentence subset of the TRIS corpus (Parra Escartín, 2012).

We expected to be able to align the split German compound with the Spanish MWE by finding a correlation between the semantic types of their formants. Example 1 below exemplifies this using as an example “Handbremsvorrichtung” (hand brake device).

(1) Hand.BODY PART Bremse.ARTIFACT Vorrichtung.ARTIFACT
mano.BODY PART freno.ARTIFACT dispositivo.ARTIFACT
[DE]: “Handbremsvorrichtung”
[ES]: “Dispositivo de freno de mano”

As can be observed in Example 1, the semantic types of the formants of the compound happen to meet the semantic types of the content words of its translation into Spanish. We tested two approaches, one without setting a limit to the size of the Spanish translation, and another one in which there was a limit. However, the results we obtained were not as positive as we had expected.

As the experiments did not retrieve any results we could analyze, we analyzed our Gold Standard to determine possible sources of error. The accuracy of the Part-of-Speech tagger used for tagging our corpora was particularly damaging for Spanish, and we also faced a WordNet coverage problem. In the light of the results obtained, possible ways of improving the experiment setup are discussed.

References

PARRA ESCARTÌN, C. (2012, May). Design and compilation of a specialized Spanish-German parallel corpus. In N. C. C. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, and S.

Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp. 2199–2206. European Language Re- sources Association (ELRA).

SAG, I. A., T. BALDWIN, F. BOND, A. COPESTAKE, and D. FLICKINGER (2001). Multiword Expressions: A Pain in hte Neck for NLP. In In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), pp. 1–15.

BILINGUAL TERM ALIGNMENT FOR LANGUAGE SERVICE IN COMPARABLE CORPUS

Zhiyan Zhao Xuri Tang

Bilingual term database is an important part in language service. This paper explores a similarity-based method to retrieval bilingual terms from comparable corpus. Comparable corpus differ from aligned bilingual corpus in that it contains the original texts in the source language and the translations of the texts in the target language but the texts are not strictly aligned to each other. Both multi-word expression retrieval and similarity computation are involved in the task. However, this paper is mainly focused on similarity computation.

The data used in the paper is a comparable corpus built from Chinese and English texts. In the initial experiment, about 20 articles in Chinese and their corresponding English texts are collected. For each language, multi-word expression retrieval is then applied to retrieve terms from texts. The remaining task is then to align the terms in the two languages.

Term alignment is conducted with the assumption that term pairs that are the most similar in semantics are potential candidates for alignment. To implement the idea, each Chinese term is split into words and form a set C’term={…Ci…}, and each English term also form a set E’term={…Ej…}，Ej. With the assumption that there is another set D= {…（Ci，Ej）…} in which the Ci and Ej are semantically equivalent, the similarity can be computed using various similarity metrics.

To choose the best candidates, a matrix with similarity values calculated between every Chinese term with every English term is constructed. The bilingual pair with the highest similarity value in the matrix was chosen as the first pair and then the row and column of the chosen pair is eliminated from the matrix. This operation is iterated until no such pair is available.

In our initial experiment, 244 Chinese terms and 191 English terms are retrieved from the comparable corpus. Among them there are 107 bilingual pairs identified. The proposed method currently has a precision of .39 and a recall of .61. Works are under way to improve its performance by incorporating lexicographic knowledge into the system.

Identification, acquisition and evaluation of multi-word terms Identificación, adquisición y evaluación de MWU y variantes

CHUNKING-BASED DETECTION OF ENGLISH NOMINAL COMPOUNDS

Gábor Csernyi
University of Debrecen
csernyi.gabor@gmail.com

Nominal compounds (NCs), a (sub)type of multiword expressions (MWEs) have been widely explored recently. These special linguistic phenomena are lexical items and have been shown that they are quite frequent in any language. They reflect idiosyncratic features: the words making up a nominal compound – while each having their own meaning – together form a single expression functioning as a noun (Sag et al., 2002; Nagy T. et al., 2011).

No matter these expressions are compositional in meaning or not, they are to be treated as a single semantic unit. The importance of their identification in running texts is therefore further underlined by the fields of machine translation as well as information retrieval/extraction, where it is crucial to recognize them each as one semantic whole.

However, it should also be noted that not every nominal combination or co-occurrence functions as a nominal compound (like fat cat, where fat can be an adjectival modifier of cat in one context, while in other cases the whole expression can also mean well-paid executive), and it is generally the context that might help us to decide if the compound candidate is a real compound (Nagy T. and Vincze, 2013). Furthermore, since these expressions are quite productive, they do not constitute a fixed subset of the language; new terms of this type can appear in the language anytime (Nagy T. and Vincze, 2013).

From the nature and features of compounds listed above, taking the natural language processing perspective of automatic detection, it also follows that our focus of interest concerns those cases where the parts of compounds are delimited by a space; hyphenated compounds or those in which the parts are written together do not necessarily pose problems to identifying them as a unit.

This paper presents a supervised learning approach to detecting English nominal compounds (NCs). The method that is based on chunking originates from identifying full noun phrases (as chunks) in POS-tagged text following Bird et al. (2009), and shows that even basic syntactic information (in the form of POS-tags) can be exploited to detect this type of multiword expressions (MWEs) with considerable results compared to dictionary-based and hybrid (the former way combined with machine learning) methods. The results of the experiments presented here also show how the size and the density of the train set (in terms of frequency of the target expressions) as well as of the test set influence the efficiency of the algorithm(s).

References

Bird, S., Klein, E. and Loper, E., 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

NAGY T.I., BEREND, G., AND VINCZE, V., 2011. Noun Compound and Named Entity Recognition and their Usability in Keyphrase Extraction. In *Proceedings of Recent Advances in Natural Lanugage Processing 2011* (RANLP 2011). Hissar, Bulgaria. pp. 162-169.

NAGY, T.I. AND VERONIKA, V., 2013. English Noun Compound Detection With Wikipedia-Based Methods. In Matousek, V., Mautner, P. and Pavelka, T. eds., *Proceedings of the 16th Inter-national Conference on Text, Speech and Dialogue* (TSD 2013*), Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg. pp. 225-232.

SAG, I., BALDWIN, T., BOND, F., COPESTAKE A. AND FLICKINGER, D., 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2002). pp. 1-15

MULTIWORD UNITS TRANSLATION EVALUATION IN MACHINE TRANSLATION: ANOTHER PAIN IN THE NECK?

Johanna Monti
Università degli Studi di Sassari
jmonti@uniss.it

Amalia Todirascu
Université de Strasbourg
todiras@unistra.fr

While MT quality evaluation is a debated topic in MT since its inception, accurate MWU translation evaluation is still a challenge, whatever is the adopted MT approach (statistical, rule-based or example based). The main reason for this being that they display lexical, syntactic, semantic, pragmatic and/or statistical but also translational idiomaticity. Idioms, collocations, verb or nominal compounds, Named Entities or domain specific terms might all be considered as MWUs and in general both Statistical (SMT) or Rule-based Machine Translation (RBMT) fail to translate them correctly for different reasons, as highlighted by several recent contributions (Barreiro et al 2014, Monti 2012, and Ramisch et al. 2013 among others).

MWU translation quality evaluation is not an easy task for several reasons: lack of inter-annotation agreement on the notion of MWU, benchmarking resources and shared assessment methodologies and guidelines. MWU translation quality evaluation has not been discussed so far according to a shared methodological framework and to the best of our knowledge only very few MT quality evaluation metrics consider issues related to MWU translation.

There is the need to develop large data sets, mainly parallel corpora annotated with MWUs, but annotating MWU in parallel texts involves several problems because of the translational asymmetries between languages and because of discontinuity. By translational asymmetries we refer to the differences which occur between an MWU in the source language and its translation. We can have the following cases: many to many (en: kick the bucket – it. tirare le cuoia,) but also, many to one (en: kick the bucket – it. morire) and finally one to many correspondences (fr: dedommager – en: to make good any damage). Besides, each specific MWU category requires different annotation strategies to handle discontinuity or ambiguity in the monolingual part of the corpus.

For these reasons, annotated resources are available only for specific MWU types and they are built to evaluate a specific MWU alignment tool or a specific MWU integration strategy in MT systems (Weller et al, 2014), (Kordoni and Simova, 2014), (Ramisch et al, 2013). Several monolingual (Seddah et al, 2013) or parallel treebanks contain heterogeneous annotations of specific MWU categories. Very few projects annotate all MWU types and they generally annotate continuous MWUs (Flickinger et al, 2012; Bejček et al, 2012).

To sum up, there are only few small-size corpora, containing aligned sentences representative for a specific type of MWU for a limited number of language pairs. (Ramisch et al, 2013, Navlea, 2014, Barreiro et al, 2014 among others)

In this contribution, we analyse the state of the art in the evaluation of MWU translation by MT and we propose to develop a methodologically consistent MWU translation evaluation method by reaching a consensus on (i) the notion of MWU, (ii) inter-annotator agreement guidelines and (ii) consequently by developing benchmarking test sets composed of parallel corpora annotated according to a shared methodology. The idea is to provide linguistic and translational-based evaluation guidelines with high inter-annotation agreement and wide-error coverage.

References

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, *Proceedings LREC 2014*.

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., Žabokrtský, Z. (2012. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)* , Mumbai, India, pp. 231-246,

Flickinger, D. Kordoni, D., Zhang Y. (2012. DeepBank: A Dynamically Annotated Treebank of the Wall Street Journal. In Proceedings of TLT-11, Lisbon, Portugal, 2012.

Kordoni, V. Zhang. Y. (2010. Disambiguating Compound Nouns for a Dynamic HPSG Treebank of Wall Street Journal Texts. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Malta, 2010.

Kordoni, V. AND Simova, I. (2014. Multiword Expressions in Machine Translation. In *Proceedings of the International Conference on Language Resources and Evaluation,* Reykjavik, Iceland: ELRA, pp. 1208-1211.

Seddah; D., Tsarfaty, R., Kübler, S., Candito, M. Choi, J. D., Farkas, R., Foster, J., Goenaga I.; Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M.; Maier, W; Marton, Y., Nivre; J., Przepiórkowski, P. Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A. (2013., Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages, *Proceedings of the Fourth SPMRL Workshop*, Seattle, USA.

Melamed D. I. (1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing,* RI, USA: Providence, pp. 97-108.

MONTI J. (2012, Multi-word unit processing in Machine Translation - Developing and using language resources for Multi-word unit processing in Machine Translation- PhD Thesis University of Salerno, Italy

Navlea, M. (2014. *La traduction automatique statistique factorisée : une application à la paire de langues français - roumain.* Thèse de doctorat, Université de Strasbourg, Strasbourg, 374 pages.

Ramisch, C., Besacier, L., AND Kobzar, A. (2013. How hard is it to automatically translate phrasal verbs from English to French?. In J. Monti, R. Mitkov, G.. Corpas Pastor, V. Seretan (Eds.), *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology,* , Nice (France), pp. 53-61.

Ren, Z, Lü, Cao, J., Liu, Q, AND Huang, Y. (2009. Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009,* pp. 47-54.

Salton; G., Ross, R., Kelleher, J. (2014. Evaluation of a Substitution Method for Idiom Transformation in Statistical Machine Translation, *Proceedings of the 10th Workshop on Multiword Expressions (MWE), EACL 2014*, Göteborg, Sweden, April 2014

Weller, M., Cap, F., Müller, S., Schulte im Walde S. and Fraser, A. (2014. Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComaComa) at COLING,* Dublin, Ireland, August 2014.

PARALLEL SENTENCES FRAGMENTATION

Sergey Potemkin Galina Kedrova

Perfect fragmentation of the parallel bilingual texts represents an essential and first of all step in solving the problem of Example Based Machine Translation (EBMT) (Brown, 1996). Fragmentation of bilingual corpora can be considered with various degrees of details- from the super-sentence unity (paragraph, chapter) to word-by-word matching. We considered fragmentation of a-priori matched sentences of a bilingual text. The steady fragments of sentences sometimes are referred as Phraseologism (Naumova, 2005), occur in the parallel texts more frequently. Necessity for consistent selection of parallel fragments in two sentences one of which is "the perfect translation" of the other one (the perfect translation as a rule is a translation made by a qualified – human – translator, and probably strictly verified, as, for example, Bible translations).

This paper presents a new approach to fragmentation of sentences based on lexical and structural comparison of fragments of the source sentence and the translated one. In the contrary to the known techniques, we use intervals bounded by delimiters (blank) between the words, not the words itself as the elements of proximity matrix in the bilingual space (Melamed, 1999). It enables matching multi word units, not singular words only. Then we derive and process fragments of the source sentence which can be the inverse fragments within the translated sentence. Selection of the best (in some particular sense) fragmentation of each sentence is performed with the use of dynamic programming (Viterby algorithm). It is necessary to choose the best one among all virtual fragmentations according to some criterion connected with:

a) Maximization of weight of all mapping segments included in fragmentation

b) Minimization of total length of interpolating segments

c) Maximization of number of fragments … etc.

The words of two sentences can be matched in different sequence. The same sentence can be translated either in the direct or in the inverse order of words and both translations were right. A more common case - when one groups of words are translated in direct order, other - in inverse and these groups are matched chaotically. However if we consider only monotone mappings (i.e. we assume that order of words of source and target sentence coincide in general), the problem falls into a class of dynamic programming. Really, the sequence of mapping segments in a fragmentation could be considered as a path from point 0 to terminus. Then the path with maximal weight corresponds to the best fragmentation in a sense of criterion a). Each sub-path of the critical path also is a critical path and the problem permits solution within dynamic programming. Tuning the weight factors for each matched interval increases fragmentation quality. The quality was assessed by the experts (Kedrova, Potemkin, 2005)

The algorithm of fragmentation on the corpora without morphology markup is presented. We conclude the paper with presenting and discussing a set of experimental results. Semantic evaluation of the mapping segment weight is offered. Local inversion is effectively processed by inclusion of factious mapping segments in the critical path searching. Obtained fragmentation is evaluated according to structural and semantic criteria (Potemkin, 2004). If either one is violated two subsequent fragments are merged (in extreme case all fragments are merged and form the initial pair of sentences). Our experiments show, that the method improves sentence-level fragmentation of bilingual corpora. Implementation of the procedure enables to build an automatic dictionary of fragments for use in Example Based Machine Translation.

References

BROWN RALF D. (1996). Example-Based Machine Translation in the Pangloss System, *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING-96)*, pp. 169–174 (vol 1).

KEDROVA G.E. AND POTEMKIN S.B. (2005). Automatic evaluation of machine translation based on semantic metrics. *Вестник Луганского НПГУ им. Т.Шевченко* 15(95) pp. 35-41 (in Russian).

MELAMED I. DAN, (1999). Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics 25(1),* pp. 107-130, March

NAUMOVA I.O. (2005). Cognate English and Russian phraseological units as historical traces of European intercultural communication. *Вестник Луганского НПГУ им. Т.Шевченко* 15(95), pp.41-47

POTEMKIN S.B. (2004). Lexical database with superimposed semantic metrics, *II Международный конгресс исследователей русского языка "Русский язык: исторические судьбы и современность", Доклады, М.* (in Russian).

IDENTIFYING MULTI-WORD EXPRESSIONS FROM PARALLEL CORPORA WITH STRING-KERNELS AND ALIGNMENT METHODS

Federico Sangati
Andreas van Cranenburgh
Mihael Arcan
Johanna Monti
Marco Turchi

We propose a new methodology for identifying Multi-Word Expressions (MWEs) from a bilingual parallel corpus (e.g., Cettolo, 2012). Our approach makes use of the non-translatability property of MWEs: they cannot be translated on a word-by-word basis[1] (Sag et al., 2002; Monti, 2012). The methodology envisions a two-stage process. The first phase aims at identifying a list of potential MWEs, while the second filters out those candidates which are not MWEs.

Phase 1: String-Kernel methods. String-kernel algorithms (Lodhi et al., 2002; Rousu and Shawe-Taylor, 2005) are used to efficiently identify matching subsequences (arbitrarily long and with possible gaps) between two sentences. For instance, they can detect that the following two sentences share the expression “as far as [...] are concerned”:

[1] A number of MWEs can be translated literally, such as proper names and universal proverbs. These are excluded from the scope of the current work.

1. She is on a downward slide **as far as** conservatives **are concerned**.
2. We will be there **as far as** development and diplomacy **are concerned**.

In the first phase we extend this method to sentences aligned in parallel corpora. More precisely, the algorithm iteratively detects shared subsequences for each pair of aligned sentences in the two languages. See table 1 for an example.

This methodology is based on the intuition that in order for an expression A to be translatable into B we need to find at least two distinct sentences in the source language containing A which are paired with two sentences of the target language containing B.

English	Italian
I feel we will have to **call it a day** at this point.	Credo che a questo punto dobbiamo **passare oltre**.
He would like to **call it a day** for now.	Il relatore chiede per ora di **passare oltre**.

Table 1: Example of a pair of aligned sentences with a shared subsequence.

Phase 2. Filtering. The first phase is likely to find many NON-MWEs. Many of these cases are accidental co-occurrences of frequent non-related expressions (e.g., EN. “I’d like” → IT. “e non” [*en: and not*]). In the second phase we first make use of co-occurrence counts to filter out those expression pairs whose association is not statistically significant.

Of the remaining cases, associated pairs can be MWEs in both languages, one of them (source or target) or none (see table 2). As we are interested only in the first 3 cases, we adopt a Machine Translation (MT) alignment system trained on the same corpus to distinguish MWEs, for which the system does not produce a coherent alignment (figure 1) from literal translations, which are perfectly aligned (figure 2).

The paper presents the methodology and the resources developed so far.

	English	Italian
1.	MWE bring up to date	NON-MWE modernizzare [*modernize*]
2.	NON-MWE he died	MWE ha tirato le cuoia [*he pulled the leathers*]
3.	MWE can't help but	MWE non poter fare a meno di [*not can do to less of*]
4.	NON-MWE a very complicated thing	NON-MWE una cosa davvero complicata [*a thing really complicated*]

Table 2: All possible outcomes in detecting MWEs from a pair of candidate expressions.

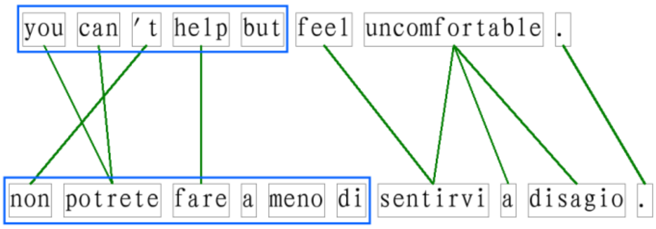

Figure 1: An example of a misalignment between the MWEs (table 2, row 3).

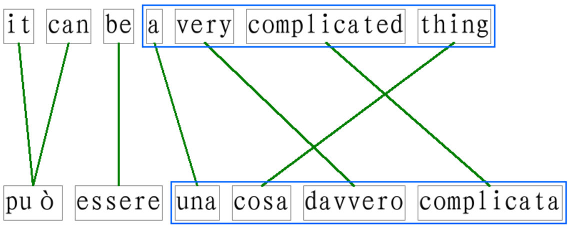

Figure 2: An example of a word-to-word alignment between two NON-MWEs (table 2, row 4).

References

ALESSANDRO, M. (2006). Making Tree Kernels Practical for Natural Language Learning. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics.

CETTOLO, M. GIRARDI, C. AND FEDERICO, M. (2012). Wit3: Web inventory of transcribed and translated talks. In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT), pages 261–268, Trento, Italy.

COLLINS, M. AND DUFFY, N. (2001). Convolution Kernels for Natural Language. In Dietterich, Thomas G., Suzanna Becker, and Zoubin Ghahramani, editors, NIPS, pages 625–632. MIT Press.

LODHI, H. SAUNDERS, C. SHAWE-TAYLOR J. CRISTIANINI N. AND WATKINS C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2:419–444.

MONTI, J. (2012). Multi-word unit processing in Machine Translation - Developing and using language resources for Multi-word unit processing in Machine Translation. PhD thesis, University of Salerno.

ROUSU, J. AND HAWE-TAYLOR, J. (2005). Efficient computation of gapped sub-string kernels on large alphabets. J. Mach. Learn. Res., 6:1323–1344.

SAG, I. A. BALDWIN T. BOND F. COPESTAKE A. AND FLICKINGER D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In Gel- bukh, Alexander, editor, Computational Linguistics and Intelligent Text Processing, volume 2276 of Lecture Notes in Computer Science, pages 1– 15. Springer Berlin Heidelberg.

SANGATI, F. ZUIDEMA AND W. BOD, R. (2010). Efficiently Extract Recurring Tree Fragments from Large Treebanks. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).

VAN CRANENBURGH, A. (2014). Linear average time extraction of phrase-structure fragments. Computational Linguistics in the Netherlands Journal 4:3–16.

Multilingualism and MWU processing.
MWUS and word alignment techniques
Multilingüismo y procesamiento de MWU.
MWU y técnicas de alineación de palabras

STATISTICAL MEASURES TO CHARACTERISE MWES INVOLVING "MORDRE" IN FRENCH OR "BITE" IN ENGLISH

Ismail El Maarouf
University of Wolverhampton
i.el-maarouf@wlv.ac.uk

Michael Oakes
University of Wolverhampton
Michael.Oakes@wlv.ac.uk

A large number of statistical measures exist which measure the collocational strength of MWEs, particularly those which are characterised by two main words (Pecina, 2008). Such measures of collocational strength are useful for discovering new pairs of collocates in corpora. In this paper we will look at statistical measures which have not yet been tested for their ability to discover new collocates, but we have found useful for characterising MWEs containing collocates already found. Smadja (1993) suggested that collocations should be characterised by whether they are flexible (allowing varying numbers of intervening words between the two words in collocation) or rigid (always having exactly the same number of words between them). To characterise flexibility, we suggest the mean and the standard deviation of the distance in words separating the two collocates, taken over all occurrences of the collocation in the corpus. Thus a rigid collocation would have a standard deviation of 0, while a flexible collocation would have a standard deviation above 0 (the higher the value, the more flexible the collocation). For example, the corresponding phrases "mordre la poussière" and "bite the dust" both have standard deviations for their lengths of 0, since in the BNC and Frtenten corpora the verb is always exactly 2 words before the noun. We also suggest Shannon Diversity (originally developed as a measure of ecological diversity) as a measure of diversity within a MWE. Does an MWE always consist of exactly the same set of words, or does it take variant forms? A phrase like “bite the bullet” in English always (in the BNC corpus) occurs as exactly these three words, so its diversity is 0, while “bitten by the bug” takes many forms: “bitten by the acting bug”, “bitten by the travel bug”, “bitten by the golf bug”, and so on. The diversity of “bitten by the bug” is close to its maximum theoretical value of the logarithm to the base 2 of the number of examples in the corpus. Its French counterpart, “mordue de” is also highly diverse, as in the examples “mordue des nuitées en famille sous las tente” (fanatical about nights camping with the family), “mordus des jeux on ligne” (addicted to on-line games) and “mordue d’esperanto” (bitten by the Esperanto bug). The pattern “[[Human]] se mord {les doigts}” rarely takes its literal meaning in French, more often standing for “a person experiencing a bitter time for his past actions”. It usually occurs in the Frtenten corpus as “mordre les doigts”, but sometimes as “mord encore les doigts” (bites his fingers again), “mordrait un peu souvent les doigts” (bit his fingers a bit often) and other variants. This gave a mean and standard deviation of the lengths of 1.19 and 0.15, and a Shannon Diversity of 1.08.

MWE are challenging not for only second language speakers, but also for MT systems. For example, "bite one's fingers" and its apparent French translation "se mordre les doigts" are in stark idiomaticity contrast. While "bite one's fingers" was always found to be literal (5 cases), all instances of “se mordre les doigts” (21) were found to be idiomatic. Systems unaware of this will tend to make two mistakes (as can be checked with Google Translate): when translating from French to English, they will fail to translate the figurative meaning of "se mordre les doigts" with an equivalent idiom like "kick oneself". From English to French, they will fail to translate the literal meaning of "bite one's fingers" and translate it with the frequent idiomatic sequence "se mordre les doigts". For the verbs “mordre” and “bite”, we have shown that the measures of mean and standard deviation of length, Shannon Diversity, and idiomaticity (proportion of occurrences which are idiomatic) give intuitively reasonable results. We propose this measures as parameters in an MT system.

References

Pavel Pecina. (2008). Lexical Association Measures: Collocation Extraction. Ph.D. Thesis, Charles University in Prague.

Frank Smadja. (1993). Retrieving Collocations from Text: Xtract. Computational Linguistics 19: 143-177.

ALIGNING VERB+NOUN COLLOCATIONS TO IMPROVE A FRENCH - ROMANIAN FSMT SYSTEM

Amalia Todirascu
Université de Strasbourg
mirabela_abe@yahoo.com

Mirabela Navlea
Université de Strasbourg
todiras@unistra.fr

We present several MWEs integration and alignment methods using linguistic information, and aiming to improve the results of a French <-> Romanian FSMT system. We focus on a specific class of MWEs: Verb+Noun collocations (prendre des decisions ’make decisions’, tenir compte ’take into account’). Our first strategy aims to extract Verb+Noun collocations from monolingual corpora before aligning them. For this purpose, we combine linguistic information (preference for several morphosyntactic properties, such as definite or indefinite determiners, modifiers or long distance dependencies) and frequent word association criteria. The collocation candidates extracted from one language are transformed into single tokens and aligned to the potential translation equivalents in lexically aligned parallel corpora. Our system uses lemmatized, tagged and sentence-aligned legal parallel corpora. In this paper, we compare this method with the results obtained by applying a French - Romanian Verb+Noun collocation dictionary in the alignment process, as an external resource. We compare our own alignment algorithm with the standard lexical alignment implemented in GIZA++. We evaluate the influence of several collocation alignment methods on the results of the lexical alignment and on the results of the FSMT system.

Learning semantic information about MWUS from monolingual, parallel or comparable corpora. Adquisición de información semántica sobre MWU a partir corpus monolingües, paralelos y comparables

ARANEA: COMPARABLE GIGAWORD WEB CORPORA

Vladimír Benko
Comenius University in Bratislava
& Slovak Academy of Sciences,
Slovakia
vladob@juls.savba.sk

Peter Ďurčo
University of SS. Cyril and
Methodius in Trnava, Slovakia
durco@vronk.ne

Aranea is a family of web corpora intended for use in contrastive linguistic research, multilingual lexicography, as well as for teaching foreign languages and translation studies. The data have been downloaded from Internet at (approximately) the same time, and processed by the same set of open-source and free tools (*SpiderLing and jusText* for crawling and preprocessing, *Unitok* for tokenization, *Onion* for deduplication, and *Tree Tagger* for tagging most languages). We believe that corpora of the same size created in this way (to a large extent) deserve the designation of being "comparable".

As all the "native" tagsets have been mapped to *Araneum Universal Tagset* (*AUT*), this made it possible to create compatible sketch grammars for the Sketch Engine for all the languages involved. The *AUT* contains, besides the traditional 11 word classes (i.e., determiner/article, noun, adjective, pronoun, numeral, verb, adverb, preposition/postposition, conjunction, interjection, and particle), also several tags for other entities typically appearing in corpora (abbreviation/acronym, symbol, number, other (content word), other other (function word), unknown/alien/foreign, punctuation,) and also a special tag for mapping errors.

Compatible sketch grammars (CSG) include common set of rules for all languages, fixed order of tables in word sketches. Rule names represent collocational relationships, i.e. not syntactic relationship. Syntactic functions of keywords and/or collocates are not indicated but just indicated as the "left-hand" and "right-hand" collocates. The CSG uses symbol X for keyword, Y for collocate of any PoS, except for conjunction, preposition and punctuation and Z for collocate of PoS not covered be explicit rules. CSG uses side-sensitive and side-insensitive binary rules, symmetric rules for coordinative relationship, trinary rules for cooccurrences with prepositions and unary rules have been used for PoS categories and PoS subcategories. There are no dual rules used. A collocationally-based sketch grammar has (against a traditional one) several advantages. It can symmetrically cover all relationships between keywords and collocates of all word classes (parts of speech).

At the time of writing this abstract, the *Aranea* family includes corpora for 12 languages (Chinese, Dutch, English, French, Finnish, German, Hungarian, Italian, Polish, Russian, Slovak and Spanish). Currently in preparation there are corpora for Czech, Georgian and Ukrainian. All corpora have "language-neutral" Latin names and come in two sizes: the basic *Maius* ("greater") version has 1.2 billion tokens, i.e., approx. 1 billion words, and a 10% sample *Minus* ("smaller") version is intended for teaching purposes. For some languages, also a *Maximum* ("maximal", as much as we can get) version is being created.

The *Aranea* corpora are accessible via the free web interface at http://ucts.uniba.sk (without word sketches, however) and they are also hosted at http://kontext.korpus.cz (a free registration is required). Users who have account at the Sketch Engine web site can enjoy the full functionality of that system provided by the CSG at http://www.sketchengine.co.uk (a 30-day free trial is available).

In our presentation, we will try to demonstrate that by using large corpora for two languages, consisting of unrelated texts, yet created in a comparable manner, parallel language structures and phenomena can be identified if appropriate tools are involved. With the *Aranea* corpora, the "Bilingual sketch" functionality of the Sketch Engine is one of such tools which provides for analyses of similarities (or differences) of collocation profiles (word sketches) for words and their translation equivalents. As all the *Aranea* sketch grammars are compatible, the respective tables of the Bilingual sketch match for all languages involved and for words of all classes. The Sketch Engine API makes it also possible to analyze the collocational profiles by “non-human” agents, such as components of MT or other NLP systems.

IN-DEPTH STUDY OF THE PHRASEOLOGICAL UNITS IN ISLAMIC AND CHRISTIAN RELIGIONS IN SAMPLES (CORPORA) OF RELIGIOUS TEXTS

Madian Souliman Ali Ahmad

Over the last two decades there has been a great deal of interest in lexical studies, particularly in the combinations of words in natural languages. Conventionalized forms, frames, idioms, and collections have proven to be chiefly appealing in the areas of phraseology. The actuality of present research is conditioned by necessity of studying of the characteristics of the phraseological units in-depth and as they are expressed in the “Holy Bible and Holy Quoran “ which will reveal many methods and approaches in translating them basically in the religious texts and will help us to bind up the whole religious texts (Bible & Quran) in one computerized corpus which will provide us with different translations of those holy texts for a comparative study. We cannot neglect that some of the earliest efforts at grammatical description were based at least in part on corpora of particular religious as the early Arabic grammarians paid particular attention to the language of the Quran which can prove that corpus linguistics adherents have reliable language analysis which best occurs on field-collected samples, in natural contexts and with minimal experimental interference. Within corpus linguistics there are divergent views as to the value of corpus annotation , from John Sinclair advocating minimal annotation and allowing texts to ' speak for themselves', to others, such as the Survey of English Usage team advocating annotation as a path to greater linguistic understanding and rigour. We cannot deny that a computerized corpus of the religious texts was found many years ago. An example is the Andersen-Forbes database of the Hebrew Bible, developed since the 1970s, in which every clause is parsed using graphs representing up to seven levels of syntax, and every segment tagged with seven fields of information. Another example is the Quranic Arabic Corpus which is an annotated corpus for the classical Arabic Language of the Quran. The subject of this presentation is to consider some peculiarities and problems in translating the religious texts especially after taking into consideration that there are many phraseological units in their sacred contexts and the results will be discussed after examining the translation of sixteen examples from both the Holy Qur'an and the Holy Bible. The main goal of the research is to create a computerized corpus for the phraseological units in the religious texts which later on can provide us with not only the state of the language in samples but also different translations with notes on their differences.

References

Ahmad, A. ,Итоговая научно-образовательная конференция студентов Казанского Федерального университета 2014 года:*In-depth study of the paremiological units in Islamic and Christian religions with some contemporary interpretations in their translations*. p. 157.

The observatory of language sciences, 2013, the Dep. of Vernacular Languages & the Graduate Program in Linguistics at the Federal University of Ceará . 1 p. 09/05/2014

Ayupova , 2012 , *English Phraseology* ,Kazan. 7p. 09/05/2014

Арутюнова Н.Д. Дискурс // *Лингвистический энциклопедический словарь*. – М., 1990. – 260с.

http://en.wikipedia.org/wiki/Set_phrase [Accessed 09 May 2014]

Anita Naciscione , *Phraseological units in discourse: towards applied stylistics*, 2001.- 5p. 09/05/2014

Anita Naciscione , *Phraseological units in discourse: towards applied stylistics*, 2001.- 5p. 09/05/2014

http://www.rusnauka.com/10._ENXXIV_2007/Philologia/21605.doc.htm [Accessed 09 May 2014]

Anita Naciscione , *Phraseological units in discourse:* 2001 *towards applied stylistics*,- 5p. 09/05/2014

The Observatory of Language Sciences, 2013 the Dep. of Vernacular Languages & the Graduate Program in Linguistics at the Federal University of Ceará . 1 p. 09/05/2014 http://en.wikipedia.org/wiki/Proverb#Paremiology. [Accessed 17 March 2014]

Wolfgang Mieder. 1990. *Not by bread alone: Proverbs of the Bible*. New England Press, 12p. 17/03/2014

The Holy Bible: *The old and New Testament* No. of books 66 & 1,189 chapters 1,281 ps. 1p. 14/03/2014

Wolfgang Mieder. 1990. *Not by bread alone: Proverbs of the Bible*. New England Press, 12p. 17/03/2014

http://www.americancorpus.org [Accessed 17 March 2014]

http://corpus.byu.edu/bnc/ [Accessed 17 March 2014]

http://corpus.byu.edu/bnc/ [Accessed 17 March 2014]

http://www.americancorpus.org [Accessed 17 March 2014]

http://en.wikipedia.org/wiki/Religious_views_of_Albert_Einstein [Accessed 19 March 2014]

http://www.biographyonline.net/scientists/albert-einstein-quotes.html [Accessed 19 March 2014]

А.В. Кунин. *О переводе английских фразеологизмов в англо-русском фразеологическом словаре*. Тетради переводчика. М.,1964№2. 12/06/2014

Shakespeare —*Much Ado About nothing*||, act 5, scene 1. 12/06/2014

А.В. Кунин. *О переводе английских фразеологизмов в англо-русском фразеологическом словаре*. Тетради переводчика. М.,1964№2. 12/06/2014

The Holy Qur'an , *Surat Al-Baqarah* 15 . 20/06/2014

The Holy Qur'an , *Surat Al-Baqarah* 27 . 20/06/2014

The Holy Qur'an , *Surat Al-Baqarah* 63 . 20/06/2014

The Holy Qur'an , *Surat Al-Imran* 32. 20/06/2014

The Holy Qur'an , *Surat Al-Imran* 92 . 20/06/2014

The Holy Qur'an , *Surat Al-Imran* 103 . 20/06/2014

Imam Zain ul Abideen , *As-Sahifa Al-Kamilah Al-Sajjadiyya, supplication* 53 verse 1. 20/06/2014

The Holy Qur'an , *Al-Anaam* Verse No:78. 20/06/2014

The Holy Bible , *Psalms* 69, 2. 20/06/2014

The Holy Bible , *Proverbs* 16,18. 20/06/2014

The Holy Bible , I *Samuel* 23, 16. 20/06/2014

The Holy Bible , *Proverbs* 1, 14. 20/06/2014

The Holy Bible , *Psalms* 58, 4. 20/06/2014

The Holy Bible , *Luke* 12, 3. 20/06/2014

The Holy Bible , *Proverbs* 25, 13. 20/06/2014

The Holy Bible , *John* 5, 35. 20/06/2014

MWUS in machine translation.

MWU y TA

CHALLENGES ON THE TRANSLATION OF COLLOCATIONS

Angela Costa
CLUNL and INESC-ID
angelampcosta@gmail.com

Luísa Coheur
INESC-ID and IST
luisa.coheur@inesc-id.pt

Teresa Lino
CLUNL, UNL
unl.tlino@mail.telepac.pt

According to Grossmann and Tutin (2002: 3). collocations are defined as "a privileged lexical co-occurrence of two (or more) linguistic elements that establish a syntactic relationship between them." Hausmann (1984, 1985, 1989) observed that the status of the constituents of the collocations are not similar, registering between them an hypotactic relationship. Hausmann calls 'base' to the word that determines the choice of the co-occurring element and 'collocative' the determined constituent. The relationship between base and collocative is, in most cases, unpredictable, and does not demonstrate a particularly clear semantic motivation that can explain it. Collocations are particularly relevant in the context of lexical combinatory, due to their high frequency in languages (Mel'cuk 2001 [1998]: 31). However, their idiosyncratic character and the fact that they cannot yet be considered lexicalized expressions, standing between lexicon and grammar, makes them very complex structures, from the production point of view.

As a linguistic phenomenon, collocations have been the subject of numerous researches both in the fields of theoretical and descriptive linguistics, and, more recently, in automatic Natural Language Processing. Although there are quite a few methods for the extraction of collocations from corpora, the area of post-processing of this structures and their application to various branches of Natural Language Processing is still at the beginning, especially in the area of machine translation (Seretan and Wehrli, 2007). Automatic translation is a task that involves an enormous linguistic knowledge and because collocations cannot be characterized based on syntactic and semantic regularities, not allowing a translation word by word, handling them can be a difficult task.

Having as a starting point our previous work on machine translation error analysis (Costa, 2014). for this article we decided to focus ourselves on the errors present on collocations. For the above mentioned research, we have created a translation corpus composed by three datasets translated from English to Portuguese by two mainstream online translation systems Google Translate (Statistical) and Systran (Hybrid Machine Translation) and two in-house Machine Translation systems.

The collocations wrongly translated were then classified according to the type of error found. Using the location dimension of the error typology from Margarita Alonso Ramos (2011). we marked if the error concerned the collocation as a whole or one of its two elements (base or collocative). For instance, the collocation world fair was literally translated to mundo justo, in this case the whole collocation was wrongly translated. In this other example from our corpus, heart rate was translated to meta cardíaca instead of frequência cardíaca. In this case the error was found on the collocative.

Major translation engines do not handle collocations in the appropriate way and they end up producing literal unsatisfactory translations. To our believe, to have a clear understanding of the difficulties that the collocations represent to the Machine Translations engines, it is necessary a detailed linguistic analysis of their errors.

References

COSTA, A/ Luís, T/ Coheur, L. (2014). Translation Errors from English to Portuguese: an Annotated Corpus, In: *LREC 2014, Ninth International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA). Iceland, 1231 – 1234

GROSSMANN, Francis/Tutin, Agnès (2002). Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif, In; *Revue Française de Linguistique Appliquée*, vol. VII, 2002, 7-25.

HAUSMANN, Franz Josef (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen, In: *Praxis des neusprachlichen Unterrichts* 31, 395-406.

HAUSMANN, Franz Josef (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels", In: Bergenholtz, Henning./Mugdan, Joachim (Hrsg.). *Lexikographie und Grammatik*, Tübingen: Max Niemeyer Verlag, 118-129.

HAUSMANN, Franz Josef (1989). Le dictionnaire de collocations (Artikel 95) In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (1989). *Wörterbücher – Dictionaries – Dictionnaires. Ein internationals Handbuch zur Lexikographie. Erster Teilband*, Berlin/New York: Walter de Gruyter, 1010-1019.

MEL‟CUK, Igor (2001 [1998]). Collocations and Lexical Functions, In: Cowie, A. P. (ed.) (2001[1998]). *Phraseology: Theory, Analysis and Applications*, Oxford: Clarendon Press, 23-53.

RAMOS, M/ Wanner L/ Vinsze O/ Ferraro G/ Nazar R. (2011). Annotation of collocations in a learner corpus for building a learning environment, In: *Learner Corpus Research Conference*, Louvain-la-Neuve, Belgium, 1–10

SERETAN, Violeta/Wehrli, Eric (2007). Collocation translation based on sentence alignment and parsing, In: *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles* (TALN 2007). Toulouse, France, 401–410.

INTEGRATING MULTI-WORD EXPRESSIONS IN STATISTICAL MACHINE TRANSLATION

Zied Elloumi
LIG-GETALP, France
zied.elloumi@imag.fr

Laurent Besacier
University Joseph Fourier ,
LIG, Grenoble-France
laurent.besacier@imag.fr

Olivier Kraif
University Stendhal Grenoble
3, LIDILEM, Grenoble-France
olivier.kraif@u-grenoble3.fr

Multiword expressions (MWEs) processing is a highly important research field in computational linguistics. These expressions are formed by combining two or more words, and they cover a large number of phraseological phenomena: from collocations to phrasal verbs, idioms, etc. These expressions are a challenging issue for statistical machine translation systems, since their meanings cannot be easily predicted from the words they contain.

To deal with these expressions in MT systems, a pre-requisite is to have a corpus containing a lot of MWEs. We propose a semi-automatic method to extract a specific English-French corpus with a high density of MWEs, designed for MT evaluation purpose. This extraction is done by running queries on our syntactic concordancer (Kraif and Diwersy, 2013), starting from a list of MWEs originated from several sources.

Our test-corpus with high density of MWEs includes 500 sentences, and we also built a “control” corpus by selecting randomly 500 sentences from similar sources.

Using the output of our baseline system Moses-LIG (Besacier and al.,2012), we calculated the MT performance using BLEU (Papineni and .al, 2002). The performance on the “control” corpus is 24,87% while we obtain 20,83% on the corpus with higher density of MWEs. This confirms that the test corpus we have designed is more problematic for our MT system, and reinforces our hypothesis that the frequency of MWEs in a corpus has an influence on translation quality.

To improve our system, we considered every phrasal verb (PV) as a single lexical unit, in order to force the segmentation during the alignment. An automatic identification of PV sequences was required. We conducted a parsing on the test corpus and the training corpus using XIP (Aït-Mokhtar et al., 2002) to get linguistic annotation for each form (part of speech, lemma, dependencies).

Then, using the parts of speech and some dependencies (as NUCL_PARTICLE, MOD_POST), we adapted the output of the parser to get an XML version of the corpus compatible with the Moses toolbox, with an additional attribute "EPL = 'verb-id, particle-id'" for the PVs of our test corpus. Then we merged the verb with its particle (as a verb-particle compound). This approach has been applied to both the test and training corpus. We obtained a slight improvement (+0.54 BLEU point) for the test corpus translation quality.

Finally, we handled the idioms expressions using a method of constrained decoding available in the Moses decoder (Koehn, 2014). Using our idiom list, we developed a tool to identify idioms in our English source sentence and to put the correct translation of each expression between the XML tags used by the decoder. Combining this method for idioms with the previous one for PVs, the overall improvement is +4 points (BLEU) on our 500 sentences corpus.

References

Kraif and Diwersy (2013). The Concordancer tools are developed in the LIDILEM laboratory and ILRC, 2013, Stendhal University Grenoble-France and University of Cologne, Germany. *Url: http://emolex.u-grenoble3.fr/emoBase/ .*

Papineni et .al (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation", July 2002, pages 311-318, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia.

Besacier et al., (2012). "The LIG English to French Machine Translation System for IWSLT 2012 ", In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT), 2012.

Koehn (2014). " *Statistical Machine Translation, System User Manual and Code Guide*", modification la plus récente aout 2014, Université de Edinburgh Royaume-Uni.

MULTIWORD EXPRESSIONS IN MACHINE TRANSLATION: THE CASE OF GERMAN COMPOUNDS

Maria Ivanova
Université de Genève
maria.ivanova@unige.ch

Eric Wehrli
Université de Genève
eric.wehrli@unige.ch

Luka Nerima
Université de Genève
luka.nerima@unige.ch

The importance of multiword expressions in translation has been long recognized. Proper identification and well-formed generation are fundamental requirements for high-quality translation. German compounds constitute a particularly important case, due to its high frequency, to its variety and to its highly-productive property, ruling out any hope to simply list them all in the dictionary (Parra Escartín, Peitz and Ney, 2014). Treatment of German needs both a compound analyzer and a compound generator.

This paper presents an on-going research on a compound generation module for the German language, to be integrated into a machine translation system with English, French and Italian as a source language and German as target language. We will restrict ourselves to nominal compounds and furthermore only to bi-nominal compounds.

Two distinct issues arise with respect to German compound generation: when to generate a compound, and how precisely to create it. The first question concerns the determination of what source structure should trigger the generation of a compound. The second question concerns the precise way to combine the constituents, that is to say what connecting morpheme, if any, should be used to glue the two nouns. The main focus of the paper will be the second question.

Regarding the first issue, if we take English as the source language, structures that will trigger compounding are, for example, the noun-noun structure (e.g. chocolate cake), as well as the noun-prep-noun structure (e.g. fall in population, bone of contention) when corresponding to a known collocation (i.e. a collocation listed in the lexical database of the system). Several different structures which can indicate noun compounds are listed in Ziering and van der Plas (2014).

As for the second issue, the majority of all German noun compounds are built without using any additional element to glue the compound constituents (Goldsmith and Reutter, 1998). The rest of the compounds are formed by using a connecting element (Fugenelemente or connecting morpheme) such as –s or -en to merge the constituents. The connecting morpheme is represented by several allomorphs. The relevant linguistic literature considers that the decision about the choice of a connecting element is a non-trivial task (cf. Žepić, 1970; Ortner, et al. (1991; Fuhrhop, 1996, 1998). In some cases the connecting element coincides with inflectional suffixes, e.g. Staat-s-vertrag, Student-en-haus, in others not, e.g. Gesundheit-s-amt, Hahn-en-feder. The first element of the compound is usually responsible for the choice of the connecting morpheme. A number of factors might influence this choice, and often more than one allomorph is possible to apply for the same first element, e.g. (Tag-e-buch, Tag-es-licht or Kind-er-garten, Kind-s-kopf, Kind-es-beine.

Based on the linguistic analyses just mentioned, we have developed a compound generation module using the inflection paradigm, phonetic structure, gender of the first element in order to select the proper compounding rule. To evaluate the quality of the compounding module, we have selected 945 German noun-noun compounds from several different sources (Schulte im Walde, 2013; Fuhrhop, 1998; Henrich and Hinrichs, 2011) that we first analyze using the Fips parser (Wehrli, 2007; Wehrli and Nerima, 2015) in order to retrieve the basic lexemes and then recompose, using the German generation module, thus achieving a true German-to-German translation task.

References

FUHRHOP, N. (1996). Fugenelemente. In: E. Lang and G. Zifonun eds. 1996. *Deutsch – typologisch (IDS Jahrbuch 1995)*. Berlin/New York: de Gruyter. pp.525-550.

FUHRHOP , N. (1998). *Grenzfälle morphologischer Einheiten.* Tübingen: Stauffenburg Verlag

GOLDSMITH, J. AND REUTTER, T. (1998). Automatic Collection and Analysis of German Compounds. In: *Proceedings of the Workshop on Computational Treatment of Nominals (COLING-ACL 1998).* Montreal, Canada. pp.61-69.

HENRICH, V. AND HINRICHS, E. (2011). Determining Immediate Constituents of Compounds in GermaNet. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP). Hissar*, Bulgaria. pp.420-426.

ORTNER, L., MÜLLER-BOLLHAGEN, E., ORTNER, H., WELLMANN, H., PÜMPEL-MADER, M. AND GRTNER, H. (1991). *Deutsche Wortbildung, Haupttl.4, Substantivkomposita: Typen Und Tendenzen In Der Gegenwartssprache: 4. Haupttl (Komposita Und Kompositionsahnliche Strukturen 1)*. Berlin/New York: Gruyter.

PARRA ESCARTÍN, C., PEITZ, S., AND NEY, H. (2014). German Compounds and Statistical Machine Translation. Can they get along? In: *Proceedings of the Tenth Workshop on Multiword Expressions (MWE 2014), EACL 2014.* Gothenburg, Sweden. pp.48-56.

SCHULTE IM WALDE, S., MÜLLER, S. AND ROLLER, S. (2013). Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In: *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM).* Atlanta, GA.

WEHRLI, E. (2007). Fips, a "Deep" Linguistic Multilingual Parser. In: *Proceedings of the Workshop on Deep Linguistic Processing, ACL 2007*. Prague, Czech Republic. pp. 120-127.

WEHRLI, E., AND NERIMA, L. (2015). The Fips Multilingual Parser. In: N. Gala, R. Rapp, and G. Bel-Enguix, eds. 2015. *Language Production, Cognition and The Lexicon. Text, Speech and Language Technology (Volume 48). S*pringer. pp. 473-490.

ŽEPIĆ, S. (1970). *Morphologie und Semantik der deutschen Nominalkomposita.* Zagreb: Izdavač ki zavod Jugoslavenske akademije znanosti i umjetnosti.

ZIERING, P., AND VAN DER PLAS, L. (2014). What good are 'Nominalkomposita' for 'noun compounds': Multilingual Extraction and Structure Analysis of Nominal Compositions using Linguistic Restrictors. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* Dublin, Ireland. pp. 1047-1058.

ANALYSIS OF MULTIWORD EXPRESSION TRANSLATION ERRORS IN STATISTICAL MACHINE TRANSLATION

Natalia Klyueva

Charles University in Prague

kljueva@ufal.mff.cuni.cz

Jeevanthi Liyanapathirana

Copenhagen Business School,
Denmark.

jl.ibc@cbs.dk

In this paper, we are going to evaluate a statistical machine translation (SMT) system, Moses, trained for several language pairs to explore how it cope with multiword expression (MWE) translation. We will experiment with Czech-Russian, English-French and English-Sinhala language pairs to make sure that our conclusions are as language-independent as possible. Multiword Expressions present a sequence of words with non-compositional meaning, they differ from language to language and are highly idiosyncratic. Even for the related languages we can not be sure if the structure of MWE is similar or not to say nothing about typologically different languages.

We translated some frequent MWEs using Moses and checked if they were translated properly. We speculate under which conditions MWEs are translated properly and under which context they got mistranslated. We will distinguish several types of the multiword expressions based on their part of speech and function in a sentence: noun multiword expressions, auxiliary multiword expressions, light verbs, idioms.

Noun multiword expressions. Multiword expressions in our test set are mainly named entities(NE) or belong to domain specific terminology(e.g. english - french : military coup - coup d’etat). They generally contain a noun and some other part of speech. Those terms and NEs get translated properly if they were seen in the training data.

Auxiliary multiword expressions present mainly multiword prepositions (e.g. english - french with regard to / en ce qui concerne) and SMT also does not have a problem to handle them properly because their co-occurence in the data is quite frequent and parts of an expression are not separated by other words.

Light verb constructions (LVC) are generally formed by a verb and a noun where a verb does not bare its initial meaning, so that the whole construction takes the semantics of the noun. Some multiword verbs have identical component words in the languages(cz: hrát úlohu,ru: играть роль – to play role), and some not(cz: dát smysl – give sense vs. ru: иметь смысл – have sense). Generally, multiword expressions are translated properly within SMT when an LVC presents an n-gram, but when a verb is separated from a noun, this LVC is often mistranslated.

Idioms are MWEs that can include words of any part of speech and they generally bear a meaning that has very little to do with any component of MWE. Idiomatic constructions often present a challenge to MT systems because they might be equal in the languages (contain the same words), but that is not always the case. As our data belong to the domain of news, we have not found much idioms in the test set. An example : kick the bucket in English would mean □□□ □□□□ in Sinhalese, which means to die. However, the machine translation system for English to SInhalese has very little resources, so it translates this expression into “□□□□ □□□□□□□ □□□□□” , which just gives the literal meaning of the expression.

We have found out that SMT cope with MWE as soon as a multiword unit fits into a respective bigram or n-gram, which is present and is relatively frequent in the training data. In order to analyse the translation, we use rules to extract “potential” MWEs from the source text. We then investigate how the possible translation errors can be avoided , for example by training the MT system with the extracted MWEs. We also exploited the cases when the languages involved are under-resourced.

References

KOEH, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., AND HERBST, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions* (pp. 177-180). Association for Computational Linguistics.

LIYANAPATHIRANA, J.U. AND WEERASINGHE A.R., (2011). English to Sinhala Machine Translation: Towards Better information access for Sri Lankans . *Conference on Human Language Technology for Development,* Alexandria, Egypt.

GHONEIM, M. AND DIAB, M. (2013). Multiword expressions in the context of statistical machine translation. In *Proc. of IJCNLP* (pp. 1181-1187)

BOUAMOR, D., SEMMAR, N., AND ZWEIGENBAUM, P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *LREC* (pp. 674-679).

MULTI-WORD EXPRESSIONS IN USER-GENERATED CONTENT: HOW MANY AND HOW WELL TRANSLATED? EVIDENCE FROM A POST-EDITING EXPERIMENT

Violeta Seretan
Université de Genève
Violeta.Seretan@unige.ch

According to theoretical claims, multi-word expressions (MWEs) are pervasive in all genres and domains, and, because of their idiosyncratic nature, they are particularly prone to automatic translation errors. The aim of our study is to test this claim empirically, in order to find out, in particular, how frequent MWE translation errors really are, and how much of the total post-editing effort is spent on correcting MWE translation mistakes. We present evidence from a large-scale post-editing experiment carried out in the context of a European project, in which a dataset of 1000 technical forum posts in French have been automatically translated into English using a domain-adapted phrase-based statistical machine translation system, then their translation has been manually corrected by a post-editor.

For the purpose of the present study, a random sample of 500 translation segments – i.e., sentences or sentence fragments like *Bonjour* – have been selected and annotated with MWE information. The annotation results confirmed that MWEs are prevalent in language, even in the technical forum domain which might at first be considered as exhibiting a lesser degree of linguistic richness. Our current investigation, which is to be completed soon, is directed to the evaluation of the translation quality obtained for MWE with our system. We will perform a χ^2 test of goodness of fit in order to find out whether MWEs are indeed often badly translated, as suggested by theoretical work. Moreover, by measuring the difference between the reference translations produced by the post-editor and the versions of the machine translation output with and without MWE post-edition, we will be able to quantify to what extent the segment correction effort is actually devoted to fixing MWE translation errors – for instance, producing the version in (3) instead of (2), which is the output obtained for (1) (MWEs marked in italics):

(1) *Laissez tomber* ..... depuis 5 mois ..... j 'ai *résolu la question* hier

(2) *Let down* ..... for 5 months ..... I 've *resolved the issue* yesterday

(3) *Drop it* ..... after 5 months ..... I *fixed the issue* yesterday.

Our study provides empirical evidence for the prevalence of MWEs in the social media genre represented by forum posts, which is less explored by existing research but is nonetheless one of the biggest challenges for natural language processing for the years to come. In addition to confirming that the technical forum language is as rich in MWEs as the general language, we will be able to find out whether the translation of MWEs is as challenging in the user-generated content domain as in the general domain, and to identify what amount of the total post-editing effort in this domain is devoted to correcting MWE translation errors, as opposed to other kind of errors.

ADDING MULTI-WORD EXPRESSIONS MORE THAN ONCE IN MACHINE TRANSLATION

Liling Tan
Saarland University
liling.tan@uni-saarland.de

Josef van Genabith

Previous studies have shown that adding automatically extracted lexicon or terminology outside of the parallel data prior to machine translation model training improves machine translation performance (Skadins et al. 2013; Meng et al. 2014; Tan and Bond 2014). Alternatively, adding automatically extracted Multi-Word Expressions (MWE) from the parallel data achieves the similar improvements (Tsvetkov and Wintner, 2012; Simova and Kordoni, 2013; lTan and Pal; 2014).

Extending parallel data with lexicon, terminologies or dictionary prior to training a Statistical Machine Translation (SMT) model has shown to improve overall phrase-based MT performace. The primary motivation is to use the addition lexical information for domain adaption (Koehn and Schroeder, 2008; Skadins et al. 2013; Tan and Bond 2014). Theoretically, adding out-of-vocabulary words into the parallel data will always improve SMT performance because the language model do not need to backoff to the unknown word/ngram probability when opitmizing the log-linear likelihood.

Alternative, adding adding automatically extracted Multi-Word Expressions (MWEs) from the parallel data achieves the similar MT improvements. The intuition is that adding extra counts of isolated lexical entries overweighs the alignments between dictionary entries and their translations and minimizes bad word/phrasal alignments with the other context words (Tsvetkov and Wintner, 2012; Simova and Kordoni, 2013; Tan and Pal; 2014). Experimentally, this approach checks out in BLEU or TER improvements as shown in previous studies. Yet emperically, we are unsure of how and why adding these MWEs improves SMT performace.

Our pilot experiment on Japanese-English MT using the ASPEC corpus from the WAT shared task (Nakazawa et al. 2014) has shown that appending the JICST dictionary (JICST 2004) to the parallel corpus has minimal effect on the system (BLEU: 18.57 -> 18.87) and adding the lexicon more than once further improves the overall BLEU scores (18.87 -> 18.91). Howevemr, upon appending the dictionary beyond five times, the performance degrades (<18.57). Our current assumption is that the overweighted lexical items cause alignment model and language model to shift their probability mass too drastically.

This rest paper will describe our attempt to reproduce similar improvements and degradation using MWEs extracted from the parallel data and we discuss the differences between the overweighting effects of the dictionary entries and the extracted MWEs and examine the breaking point where too much lexicon is harming the system.

References

JICST. 2004. *JICST Japanese-English translation dictionaries*. Japan Information Center of Science and Technology.

KOEHN, P., & SCHROEDER, J. 2007.. *Experiments in domain adaptation for statistical machine translation*. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 224-227). Association for Computational Linguistics.

MENG, F., XIONG, D., JIANG, W., & LIU, Q. 2014. *Modeling Term Translation for Document-informed Machine Translation*. In Proceedings of EMNLP 2014.

NAKAZAWA, T., MINO, H., GOTO, I., KUROHASHI, S., & SUMITA, E. (2014). *Overview of the 1st Workshop on Asian Translation*. In Proceedings of the 1st Workshop on Asian Translation (WAT2014).

TAN, L., & BOND, F. 2014. *Manipulating Input Data in Machine Translation*. In Proceedings of the 1st Workshop on Asian Translation (WAT2014).

TAN, L., & PAL, S. 2014. *Manawi: Using multi-word expressions and named entities to improve machine translation*. In Proceedings of ACL 2014, 201.

TSVETKOV, Y., & WINTNER, S. 201). *Extraction of multi-word expressions from small parallel corpora*. Natural Language Engineering, 18(04), 549-573.

SIMOVA, I., & KORDONI, V. 2013, September). *Improving English-Bulgarian statistical machine translation by phrasal verb treatmen*t. In Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, France.

SKADIŅŠ, R., PINNIS, M., GORNOSTAY, T., & VASIĻJEVS, A. 2013. *Application of online terminology services in statistical machine translation*. Proceedings of the XIV Machine Translation Summit, 281-286.

Lexical, syntactic, semantic and translational aspects in MWU representation.

Representación de unidades fraseológicas o unidades pluriverbales (MWU): aspectos léxicos, sintácticos, semánticos y de traducción

TRANSFORMATION AND MWU IN QUECHUA

Maximiliano Durán
Universite de Franche-Comte, Besançon,
France
duran_maximiliano@yahoo.fr

This article presents the process of how with the aid of the transformational engine of the NooJ linguistic development environment we may produce paraphrases and combination of paraphrases for a given Lexicon-Grammar class of Quechua MWU sentences taking into account the grammatical restrictions of the applicability of such transformations. This work is part of our large French-Quechua Machine Translation project. The identification and translation of Quechua MWU has not received much attention up to now.

First of all we had to build the linguistic resource which allows recognizing and annotating a MWU. Searching on a corpus of more than 80000 tokens, we have gathered a dictionary/grammar pair, named QU-MWU, of 500 MWU which is made out of a Lexicon-Grammar bearing their French and Spanish translations and the accompanying columns of the distributional and transformational properties applied to this class.

The Syntactic Grammar which generates paraphrases/transformations takes into account the restrictions on the applicability of transformations like Pronominalization, Cliticization. These transformations are applicable, as a general rule, only to the free constituents.

A phrase like *Rosam Pablopa umanta qoñichin* (Rose has turned Pablo's head) has been analyzed within the Lexicom-Grammar model of M. Gross

(1982). They fallow the structure: N°(m) N2(pa) C1V where N° and N2 represent the free constituents and V, C1 indicate the frozen parts, -m and –pa are nominal suffixes.

The actual syntactic grammar is formed of 12 embedded grammars, it allows the generation/annotation of 9 elementary paraphrases and at least 86 possible combinations of paraphrases. Moreover, the very same grammar also allows the recognition and the annotation of QU-MWU and their paraphrases.

All the agreement constraints are necessary in order to generate only grammatical sentences. If they are not set, NooJ will produce ungrammatical results. After the syntactic grammar is built, it is possible to generate the paraphrases of a given QU-MWU by right clicking on the syntactic grammar, selecting the Produce Paraphrases function and entering the QU-MWU sentences. If we apply one of our sample grammars, to the sentence *Rosam Pablopa umanta qoñichin,* N00J will produce the 86 paraphrases like: *Pablopa umantam quñichin Rosa, Pablotam umanta quñichin Rosaqa, …*

References

Duran, M. (2009). *Diccionario Quechua-Castellano*. Editions HC. Paris.

Gross, M. (1975). *Méthodes en syntaxe*, Paris: Hermann.

—-. (1982). “Une classification des phrases “figées” du francais”, Revue Québécoise de Linguistique 11.2, Montreal: UQAM.

Guardia Mayorga, C. (1973). *Gramatica Kechwa*, Ediciones los Andes, Lima. Peru.

Itier Cesar (2011). *Dictionaire Quechua-Français*, Paris. L’Asiathèque. Paris.

Perroud, Pedro Clemente. 1972. *Gramatica Quechwa Dialecto de Ayacucho*. Lima. 3ª Edicio.

Pino Duran, A. German. Uchuk Runasimi (Jechua – Quechua). *Conversación y vocabulario Castellano-Quechua* Ocopa, Concepción Perú. 1980.

Silberztein, M. (2003). NooJ Manual. htpp://www.nooj4nlp.net (220 pages updated regularly).

Silberztein, M. (2010). *Syntactic parsing with NooJ*”, in Proceedings of the NooJ 2009 International Conference and Workshop, Sfax: Centre de Publication Universitaire, pp. 177-190.

—. (2011). *Automatic Transformational Analysis and Generation*, in Proceedings of the 2010 International Nooj Conference, University of Thrace Ed: Komotini, pp. 221-231

—. (2012). Variable Unification in NooJ v3“ in Same Volume.

Vietri, S. (2010). Building Structural Trees for Frozen Sentences, in Proceedings of the 2009 International Nooj Conference, Sfax: Centre de Publication Universitaire

POPULATING A LEXICON WITH MULTIWORD EXPRESSIONS IN VIEW OF MACHINE TRANSLATION

Voula Gioli
Institute for Language and Speech Processing,
Athena Research and Innovation Centre,
Athens, Greece
voula@ilsp.athena-innovation.gr

Within the NLP community, there is a growing interest in the identification of MWEs and their robust treatment, as this seems to improve parsing accuracy (Nivre and Nilsson, 2004; Arun and Keller, 2005) or MT quality (Ren et al., 2009; Carpuat and Diab 2010). These expressions appear in a continuum of compositionality, which ranges from expressions that are very analysable to others that are partially analysable or ultimately non-analysable (Nunberg et al. 1994). In this respect, the development of large-scale, robust Lexical Resources (LRs) that may be integrated in MT is of paramount importance.

We herby present a LR developed in the context of a lexicographic project that involves the development of a conceptual dictionary of Modern Greek. This LR encompasses MWEs in Greek (EL) and their translational equivalences in English (EN) that belong to specific domains or subject fields, and are mapped onto sets of concepts that are specific to the domains at hand. In this view, the LR developed caters for cross-lingual and inter-lingual alignments that would be valuable for MT. The purpose of the work is two-fold; on the one hand, we aimed at the population of the lexicon with MWEs that pertain to specific domains (namely, transport, education), or semantic fields (emotion, cognition). From another perspective, the study aimed at the identification of cross-lingual correspondences between the EL and EN.

The conceptually organised lexicon that is under development capitalises on two basic notions: (a) the notion of lexical fields, along with (b) the Saussurian notion of sign and its two inseparable facets, namely, the SIGNIFIER and the SIGNIFIED as the building blocks (main classes) of the underlying ontology.

In this sense, the intended language resource is a linguistic ontology in which words are instances in the SIGNIFIER class. At this level, morphological, syntactic and functional information about lemmas is encoded, and instances of the class SIGNIFIER are specified for (a) morphosyntactic properties (PoS, gender); (b) lexical relations (word families, allomorphs); (c) argument structure, (d) lexical semantic relations (synonymy, antonymy), and (e) one or more translational equivalents in English (EN). The latter were obtained from existing parallel corpora that were aligned at the sentence and phrase level. Values for these features are assigned to both single- and multi-word entries in the lexicon.

Similarly, word meanings are instances in the SIGNIFIED class. Each instance in the SIGNIFIER class is mapped onto a concept, that is, an instance in the SIGNIFIED class. Domain-specific features further account for modelling sense (i.e., the features polarity and intensity that are specific to the semantic field of emotions). Furthermore, for each concept, a gloss in the form of a controlled paraphrase is also provided. At the SIGNIFIED level, MWE entries of the lexicon are mapped onto the relevant concepts.

The focus will be on the following aspects: (a) identification and manual extraction of MWEs in the selected domains/semantic fields at the monolingual level from a set of parallel domain-specific EL-EN corpora, (b) alignment MWEs, and (c) encoding of MWEs in the database.

References

ARUN, A. AND F. KELLER. (2005). Lexicalisation in crosslinguistic probablisitic parsing: The case of French. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 306–313. Ann Arbor, MI.

CARPUAT, M., AND DIAB, M. (2010). Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT*

'10). Association for Computational Linguistics, Stroudsburg, PA, USA, 242-245.

NIVRE, J. AND NILSSON, J. (2004). Multiword units in syntactic parsing. *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*.

NUNBERG ,G., SAG I., AND WASOW, T. (1994). Idioms. *Language* 70, pp. 491-538.

REN, Z., LÜ, Y., CAO, J., LIU, Q., AND HUANG, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 47-54.

SAG, I., BALDWIN, T., BOND, F., COPESTAKE, A., AND FLICKINGER, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. F. Gelbukh, ed. 2002. *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02).* Springer-Verlag, London, UK, UK, 1-15*.*

MWES: LIGHT VERB CONSTRUCTIONS VS FROZEN CONSTRUCTIONS IN MODERN GREEK AND FRENCH

Angeliki Fotopoulou
Institute for Language and Speech Processing, Athena Research and Innovation Centre, Athens, Greece
afotop@ilsp.athena-innovation.gr

Voula Giouli
Institute for Language and Speech Processing, Athena Research and Innovation Centre, Athens, Greece
voula@ilsp.athena-innovation.gr

Semi-compositional verb-noun constructions have been the focus of attention in linguistic research from different perspectives. The proposed article is aimed at delineating the boundaries between frozen (or fixed) expressions and *support* or *light* verb constructions (or collocations) in two languages, namely French and Greek.

Traditionally, the multi-word expressions whose idiomatic meaning cannot be deduced from the meaning of their parts (e.g.Bobrow & Bell, 1973; Chomsky, 1980; Fraser, 1970; Swinney & Cutler, 1979; M.Gross, 1982, 1988; Van der Linden, 1992) called Idioms or frozen/fixed expressions. *Light/Support* verb constructions consist of a predicative noun and a support verb or operator verb. However, the distinction between constructions with support verbs and frozen expressions is not always clear

Using French and Greek data, we argue that the phenomenon is not language-specific and can be attested in many languages. Moreover, the delineation of the two types of constructions and their intermediates is crucial not only for linguistic and lexicographic purposes, but also for Natural Language Processing tasks.

Constructions that consist of one of the support verbs *έχω/avoir* (=have), *παίρνω/prendre* (=take) or *χάνω/perdre* (=miss) as well as the operator verb *δίνω/donner* (=give) and a predicative noun, in Greek and French respectively, are light verb constructions or collocations: GR *έχω κουράγιο - παίρνω κουράγιο από – χάνω το κουράγιο μου --δίνω κουράγιο σε* have courage, take courage, loose my courage, give courage to FR *Avoir bcp de courage - prendre du courage – perdre son courage --donner du courage à* (=have courage, take courage, loose my courage, give courage to).

Other verbal expressions contain these verbs, and a number of **variants** are also attested. Therefore, to delineate support verb constructions and frozen constructions (with support verbs) and define their degree of fixedness, a number of linguistic tests have been employed on the basis of their syntactic, lexical and semantic properties. Among others, the following syntactic tests, namely:

a. *substitution* by support verb (έχω=have, είμαι=to be, κάνω=make) or operator verb (δίνω=give):

N0 τρέφει ελπίδες για N1 = N0 έχει ελπίδες για N1 (I cherish /have hopes)

b. *the replacement of the construction* with the *predicative noun*, which maintains the arguments it subcategorizes

H Μαρία έχει ελπίδες για N1

Maria has hopes to N1

= οι ελπίδες της Μαρίας για N1

The hopes of Maria to N

These tests were equally applied to Greek and French with light verbs constructions. However in the frozen expressions with *δίνω*/give like as :

(2) GR *N0 δίνω σάρκα και οστά σε N1*

Give substance to

FR *N0 donne corps à N1*

These tests/criteria are not applicable:

(2a). GR **N1 έχει σάρκα και οστά*

FR **N1 a corps*

(2b). GR * *η σάρκα και οστά του N*

FR * *le corps de N*

despite the fact that these expressions are formed with light verbs and have some of the properties of these constructions (GR. *N1 παίρνει σάρκα και οστά, FR. N1 prend corps).*

In the proposed paper, we will present the syntactic and semantic criteria/tests that have been employed thereof, in view of (a) defining limits between fixed expressions and collocations, and (b) defining intermediate classes between fixed expressions and collocations.

References

BOBROW, S.A. & BELL. S.M. 1973. On Catching on to Idiomatic Expressions. *Memory and Cognition* 1 (3): 343-346.

CHOMSKY, N. 1980. *Rules and Representations*. New York: Columbia University Press.

FRASER, B. 1970. Idioms within a Tranformational Grammar. *Foundations of language* 6 (1): 22-42.

FOTOPOULOU, A., 1992. Dictionnaire électronique des phrases figées : traitement d'un cas particulier : phrases figées - phrases à Vsupport *COMPLEX '92*, Papers in Computational Lexicography, Budapest, Hongrie.

Gross, M. (1981). Les bases empiriques de la notion du prédicat sémantique, Langages 63, Larousse.

Gross, M. (1988). Les limites de la phrase figée. *Langages* 23 (90): 7-22.

Laporte, É., Nakov, P., Ramisch, C., and Villavicencio, A., editors (2010). Proceedings of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010), Beijing, China. ACL.

Nunberg, G., Sag, I., and Wasow, T. (1994).Idioms. Language, 70:491–538.

Ruwet, N., (1983). Du bon usage des expressions idiomatiques dans l'argumentation en syntaxe générative. Revue québécoise de linguistique 13 :1, Presses de l'Université du Québec à Montréal, Montréal.

SWINNEY, D. & Cutler, A. 1979. The Access and Processing of Idiomatic Expressions. *Journal of Verbal Learning and Verbal Behavior* 18 (5): 523-534.

VAN DER LINDEN, E.J. 1992. Incremental Processing and the Hierarchical Lexicon. *Computational Linguistics* 18 (2): 219-238.

Wehrli, E., Seretan, V., and Nerima, L. (2010).Sentence analysis and collocation identification.In [Laporte et al., 2010], pages 27–35