Discovery of MWEs
WG3 Report on MWE Processing

Tomas Krilavičius, Justina Mandravickaitė, Michael Oakes,
Carlos Ramisch, Federico Sangati, Veronika Vincze

1 Introduction

This poster proposal summarizes the WG3 survey on tools and techniques for automatic MWE discovery. It serves as a ground for the State of the Art report of WG3 (Hybrid & Multilingual Processing of MWEs). Given length limitations, this abstract provides keywords and pointers that are further detailed in our shared document[1].

2 Lexical Association Measures

Lexical measures that estimate the association strength between words are one of the main tools employed in unsupervised discovery of MWEs in corpora. They are often based on the statistical distribution of the expression and of the words composing it. There are different ways of measuring this strength of word association:

- Measures based on raw frequency of the word combination [49, 33, 50].
- Measures based on information theory, e.g. pointwise mutual information [17, 46].
- Measures based on the contingency tables, e.g. chi-square [17].
- Statistical significance [46].
- Measures of association between three or more words, e.g. specific correlation [38, 2, 44].
- Measures which use linguistic information in addition to word frequencies, e.g. collostructional strength is the affinity of a word to a syntactic pattern [48].

Although much has been discussed about association measures, there is no consensus yet about the best type of measure to use in each case.

[1] https://docs.google.com/document/d/118iNh1bUWmODa3ChT3T9fVgt7zMEaw_oY7aH9uN1i1o/edit?usp=sharing

3 Supervised Machine Learning

A number of succesful approaches based on supervised machine learning have been proposed for MWE discovery.

Most of them rely on thoroughly developed lexical resources, i.e., corpora, treebanks, dictionaries, lexicons, etc. Therefore, even if the method per se is language independent, dependency on certain resources makes supervised machine learning approaches partly language-dependent [15, 25, 11, 4, 5, 41, 43, 34, 32]. Primary lexical resources can be complemented with additional or secondary ones, like web dictionaries [35, 53, 14, 15], and WordNet [3, 24, 35].

A number of approaches apply a variety of features ranging from shallow to deep ones, i.e. frequencies of n-grams [30], lemmas [43], orthographic variations [41], scores of association measures [53, 32], morphosyntactic patterns [11, 10]. Common techniques for complementing supervised machine learning include: filtering [11, 30], pregrouping [15], re-ranking [15], thresholds for MWE candidates [20, 30], combination of simple methods (i.e. combination of association measures [32, 53], POS tags, chunk tags and chunk sequences [4, 5]), manual annotation [23, 24], and evaluation to some degree [25].

Application of genetic algorithms show promising results [1, 9], i.e. for evolving new association measures that perform at least on the same level as already known ones [39].

4 Methods based on Semantic Properties

A number of works have used the *non-decomposability* property of MWEs to identify them: the meaning of an MWE cannot be derived from the meanings of its component words. As suggested in [28], the methods can be classified between those based in *context distributions* and those based on *substitutions*.

Context distribution methods use distributional semantic measures (verctor-based distance, e.g. Latent Semantic Analysis) to determine the distance between a MWE candidate to that of one or more of its component words [6, 27, 45, 12].

Substitution methods assess the degree of rigidity of a MWE by evaluating whether replacing a component word by a similar word gives rise to a valid expression (e.g. emotional baggage vs. emotional luggage) [29, 31, 7, 21].

5 Parallel corpora

Parallel corpora are of high importance in the automatic identification of MWEs. Usually, one-to-many correspondences are exploited when designing methods for detecting multiword expressions. On the other hand, aligned parallel corpora can also enhance the identification of multiword expressions in different languages: if an algorithm is implemented for one language, data from other languages can also be gathered with the help of aligned units. Related work in multilingual MWE discovery has been carried out using:

- Word alignments in parallel corpora [18, 43, 42]
- Dependency-parsed word-aligned sentences [52]
- Alignment mismatches [37, 40]
- Statistical machine translation systems [11]
- Decision trees in parallel corpora [47]

6 Other Methods

Other techniques have been proposed for MWE discovery, for instance, using Wikipedia [8], using terminology extraction methods based on linguistic pattens [26] or using syntactic parsing [36].

7 Tools for MWE discovery

Most tools for automatic discovery of MWEs take as input a textual corpus. Sometimes, they require prior automatic or manual linguistic analysis such as sentence segmentation, tokenisation, POS-tagging or even full parsing. Most tools listed below implement one of the techniques described above:

- Tools for corpus searches and concordancers (Sketch engine, AntConc)
- Association measures and patterns (UCS, Text::NSP, mwetoolkit, LocalMaxs, ACCURAT Toolkit, Xtract Dragon toolkit, bgMWE).
- Token-based MWE identification (jMWE, AMALGr, FIPS-Co, StringNet)
- Find and extract recurring tree fragments from syntax trees (FragmentSeeker, DiscoDOP, Varro).

8 Evaluation

One of the open challenges in MWE discovery is evaluation. Some works present the results of their methods by showing a list of the top-k MWEs returned according to some ranking criterion [16]. It is possible to manually annotate these top-k candidates, obtaining an estimation of the method's precision [36]. Traditional information retrieval measures report precision and recall with respect to a gold standard dictionary [53]. In order to avoid setting a hard threshold, it is possible to average precision over all recall points through mean average precision [19]. Given one or more objective evaluation measures, it is possible to perform a simultaneous comparative evaluation of a set of methods [31]. Finally, the use of the acquired MWEs in an NLP application (like a parser) can give an indirect usefulness measure of the MWE discovery method by the performance improvement of the application [22, 51, 13].

References

[1] Araujo, Lourdes. How evolutionary algorithms are applied to statistical natural language processing. *Artificial Intelligence Review*, 28(4):275–303, 2007.

[2] Aroonmanakun, Wirote. Extracting thai compounds using collocations and POS bigram probabilities without a POS tagger. In *International Conference on Asian Language Processing, IALP 2009, Singapore*, pages 118–122, 2009.

[3] Attia, Mohammed, Lamia Tounsi, Pavel Pecina, Josef van Genabith, and Antonio Toral. Automatic extraction of arabic multiword expressions. In *Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing*, 2010.

[4] Baldwin, Timothy. Deep lexical acquisition of verb–particle constructions. *Computer Speech & Language*, 19(4):398 – 414, 2005. Special issue on Multiword Expression.

[5] Baldwin, Timothy and Aline Villavicencio. Extracting the unextractable: A case study on verb-particles. In *Proceedings of CoNLL-2002*, pages 98–104. Taipei, Taiwan, 2002.

[6] Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, 2003.

[7] Bannard, Colin. Statistical techniques for automatically inferring the semantics of verb-particle constructions. Lingo working paper no. 2002-06, University of Edinburgh, 2002.

[8] Bekavac, B. and M. Tadic. A generic method for multi word extraction from wikipedia. In *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on*, pages 663–668, June 2008.

[9] Bekavac, Marko and Jan Šnajder. Gpkex: Genetically programmed keyphrase extraction from croatian texts. *ACL 2013*, page 43, 2013.

[10] Bharati, Akshar, Samar Husain, Bharat Ambati, Sambhav Jain, Dipti Sharma, and Rajeev Sangal. Two semantic features make all the difference in parsing accuracy. *Proc. of ICON*, 8, 2008.

[11] Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum. Identifying bilingual multi-word expressions for statistical machine translation. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[12] Bu, Fan, Xiaoyan Zhu, and Ming Li. Measuring the non-compositionality of multiword expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 116–124, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[13] Carpuat, Marine and Mona Diab. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proceedings of HLT: The 2010 Annual Conference of the North American Chapter of the ACL (NAACL 2003)*, pages 242–245, Los Angeles, California, June 2010. ACL.

[14] Constant, Matthieu and Isabelle Tellier. Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[15] Constant, Matthieu, Anthony Sigogne, and Patrick Watrin. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 204–212, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[16] da Silva, Joaquim Ferreira, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA 1999, pages 113–132, London, UK, 1999. Springer.

[17] Daille, Beatrice. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Ucrel technical papers 5, Dept. of Linguistics, University of Lancaster, UK, 1995.

[18] de Medeiros Caseli, Helena, Carlos Ramisch, Maria das Gracas Volpe Nune, and Aline Villavicencio. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77, April 2010.

[19] Evert, Stefan and Brigitte Krenn. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language Special issue on MWEs*, 19(4):450–466, 2005.

[20] Farahmand, Meghdad and Ronaldo Martins. A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 10–16, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[21] Fazly, Afsaneh and Suzanne Stevenson. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[22] Finlayson, Mark and Nidhi Kulkarni. Detecting multiword expressions improves word sense disambiguation. In Kordoni, Valia, Carlos Ramisch, and Aline Villavicencio, editors, *Proceedings of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, pages 20–24, Portland, OR, USA, June 2011. ACL.

[23] Fujita, Sanae and Akinori Fujino. Word sense disambiguation by combining labeled data expansion and semi-supervised learning method. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12 (2):7, 2013.

[24] Girju, Roxana, Dan Moldovan, Marta Tatu, and Daniel Antohe. On the semantics of noun compounds. *Computer Speech & Language*, 19(4):479 – 496, 2005. Special issue on Multiword Expression.

[25] Green, Spence, Marie-Catherine de Marneffe, and Christopher D. Manning. Parsing models for identifying multiword expressions. *Comput. Linguist.*, 39(1): 195–227, March 2013.

[26] Justeson, John S. and Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 3 1995.

[27] Katz, Graham and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06, pages 12–19, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[28] Korkontzelos, Ioannis. *Unsupervised Learning of Multiword Expressions*. PhD thesis, University of York, York, UK, 2011.

[29] Lin, Dekang. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 317–324, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[30] Nazar, Rogelio. A statistical approach to term extraction. *International Journal of English Studies*, 11(2), 2011.

[31] Pearce, Darren. Synonymy in collocation extraction. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the*

Association for Computational Linguistics, pages 41–46, 2001.

[32] Pecina, Pavel. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions*, pages 54–57, Marrakech, Morocco, 2008. European Language Resources Association.

[33] Pecina, Pavel and Pavel Schlesinger. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 651–658, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[34] Ramisch, Carlos, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and MariaJosé Finatto. A hybrid approach for multiword expression identification. In Pardo, ThiagoAlexandreSalgueiro, António Branco, Aldebaro Klautau, Renata Vieira, and VeraLúciaStrube de Lima, editors, *Computational Processing of the Portuguese Language*, volume 6001 of *Lecture Notes in Computer Science*, pages 65–74. Springer Berlin Heidelberg, 2010.

[35] Schneider, Nathan, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics (TACL)*, 2014.

[36] Seretan, V. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer, 2011.

[37] Sinha, R. Mahesh K. Mining complex predicates in hindi using a parallel hindi-english corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 40–46, Singapore, August 2009. Association for Computational Linguistics.

[38] Smadja, Frank. Retrieving collocations from text: Xtract. *Comput. Linguist.*, 19(1):143–177, March 1993.

[39] Šnajder, Jan, Bojana Dalbelo Bašić, Saša Petrović, and Ivan Sikirić. Evolving new lexical association measures using genetic programming. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 181–184. Association for Computational Linguistics, 2008.

[40] Tsvetkov, Yulia and Shuly Wintner. Extraction of multiword expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1256–1264, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[41] Tsvetkov, Yulia and Shuly Wintner. Extraction of multiword expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573, 2012.

[42] Tsvetkov, Yulia and Shuly Wintner. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2): 449–468, 2014.

[43] Tufiş, Dan, AnaMaria Barbu, and Radu Ion. Extracting multilingual lexicons from parallel corpora. *Computers and the Humanities*, 38(2):163–189, 2004.

[44] Van de Cruys, Tim. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 16–20, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[45] Van de Cruys, Tim and Begoña Villada Moirón. Semantics-based multiword expression extraction. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, 2007.

[46] Vechtomova, Olga and Stephen Robertson. Integration of collocation statistics into the probabilistic retrieval model. In *the 22nd BCS IRSG conference*, pages 165–177, Cambridge, UK, 2000.

[47] Vincze, Veronika, István Nagy T., and Richárd Farkas. Identifying english and hungarian light verb constructions: A contrastive approach. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[48] Wehrli, Eric, Violeta Seretan, and Luka Nerima. Sentence analysis and collocation identification. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 28–36, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[49] Wermter, Joachim and Udo Hahn. Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[50] Wiechmann, Daniel. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4(2):253–290, 2008.

[51] Xu, Ying, Randy Goebel, Christoph Ringlstetter, and Grzegorz Kondrak. Application of the tightness continuum measure to Chinese information retrieval. In Laporte, Éric, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio, editors, *Proceedings of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 54–62, Beijing, China, August 2010. ACL.

[52] Zarrieß, Sina and Jonas Kuhn. Exploiting translational correspondences for pattern-independent mwe identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 23–30, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[53] Zilio, Leonardo, Luiz Svoboda, Luiz Henrique Longhi Rossi, and Rafael Martins Feitosa. Automatic extraction and evaluation of MWE. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 2011.