Robust Automated Natural Language Processing with Multiword Expressions and Collocations

Valia Kordoni and Markus Egg
Humboldt-Universität zu Berlin (Germany)
kordonie@anglistik.hu-berlin.de,
markus.egg@anglistik.hu-berlin.de

1 Introduction

This tutorial aims to provide attendees with a clear notion of the linguistic and distributional characteristics of *multiword expressions* (MWEs), their relevance for robust automated natural language processing and language technology, what methods and resources are available to support their use, and what more could be done in the future. Our target audience are researchers and practitioners in language technology, not necessarily experts in MWEs, who are interested in tasks that involve or could benefit from considering MWEs as a pervasive phenomenon in human language and communication.

2 Topic Overview

Multiword expressions (MWEs) like *break down*, *bus stop* and *make ends meet*, are expressions consisting of two or more lexical units that correspond to some conventional way of saying things (Sag et al., 2001). They range over linguistic constructions such as fixed phrases (*per se, by and large*), noun compounds (*telephone booth, cable car*), compound verbs (*give a presentation*), idioms (*a frog in the throat, kill some time*), etc. They are also widely known as collocations, for the frequent co-occurrence of their components (Manning and Schütze, 2001).

From a natural language processing perspective, the interest in MWEs comes from the very important role they play forming a large part of human language, which involves the use of linguistic routines or prefabricated sequences in any kind of text or speech, from the terminology of a specific domain (*parietal cortex, substantia nigra, splice up*) to the more colloquial vocabulary (*freak out, make out, mess up*) and the language of the social media (*hash tag, fail whale, blackbird pie*). New MWEs are constantly being introduced in the language (*cloud services, social networking site, security apps*), and knowing how they are used reflects the ability to successfully understand and generate language.

While easily mastered by native speakers, their treatment and interpretation involves considerable effort for computational systems (and non-native speakers), due to their idiosyncratic, flexible and heterogeneous nature (Rayson et al., 2010; Ramisch et al., to appear). First of all, there is the task of identifying whether a given sequence of words is an MWE or not (e.g. *give a gift* vs. *a presentation*) (Pecina, 2008; Green et al., 2013; Seretan, 2012). For a given MWE, there is also the problem of determining whether it forms a compositional (*take away the dishes*), semi-idiomatic (*boil up the beans*) or idiomatic combination (*roll up your sleeves*) (Kim and Nakov, 2011; Shutova et al., 2013). Furthermore, MWEs may also be polysemous: *bring up* as carrying (*bring up the bags*), raising (*bring up the children*) and mentioning (*bring up the subject*). Unfortunately, solutions that are successfully employed for treating similar problems in the context of simplex works may not be adequate for MWEs, given the complex interactions between their component words (e.g. the idiomatic use of *spill* in *spilling beans* as revealing secrets vs. its literal usage in *spilling lentils*).

3 Content Overview

This tutorial consists of four parts. Part I starts with a thorough introduction to different types of MWEs and collocations, their linguistic dimensions (idiomaticity, syntactic and semantic fixedness, specificity, etc.), as well as their statistical characteristics (variability, recurrence, association, etc.). This part concludes with an overview of linguistic and psycholinguistic theories of MWEs to date.

For MWEs to be useful for language technology, they must be recognisable automatically.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 7–8,
Sofia, Bulgaria, August 4-9 2013.

Hence, Part II surveys computational approaches for MWEs recognition, both manually-authored approaches and using machine learning techniques, and for modeling syntactic and semantic variability. We will also review token identification and disambiguation of MWEs in context (e.g. *bus stop* in *Does the bus stop here?* vs. *The bus stop is here*) and methods for the automatic detection of the degree of compositionality of MWEs and their interpretation. Part II finishes with a discussion of evaluation for MWE tasks.

Part III of the tutorial describes resources made available for a wide range of languages as well as MWE-related multi-level annotation platforms and examples of where MWEs treatment can contribute to language technology tasks and applications such as parsing, word sense disambiguation, machine translation, information extraction and information retrieval. Part IV concludes with a list of future possibilities and open challenges in the computational treatment of MWEs in current NLP models and techniques.

4 Tutorial Outline

1. PART I – General overview:
   (a) Introduction
   (b) Types and examples of MWEs and collocations
   (c) Linguistic dimensions of MWEs: idiomaticity, syntactic and semantic fixedness, specificity, etc.
   (d) Statistical dimensions of MWEs: variability, recurrence, association, etc.
   (e) Linguistic and psycholinguistic theories of MWEs
2. PART II – Computational methods
   (a) Recognising the elements of MWEs: type identification
   (b) Recognising how elements of MWEs are combined: syntactic and semantic variability
   (c) Token identification and disambiguation of MWEs
   (d) Compositionality and Interpretation of MWEs
   (e) Evaluation of MWE tasks
3. PART III – Resources, tasks and applications:
   (a) MWEs in resources: corpora, lexica and ontologies (e.g. Wordnet and Genia)
   (b) Tools for MWE identification and annotation (e.g. NSP, mwetoolkit, UCS and jMWE)
   (c) MWEs and Collocations in NLP tasks: Parsing, POS-tagging, Word Sense Disambiguation (WSD)
   (d) MWEs and Collocations in Language Technology applications: Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT)
4. PART IV – Future challenges and open problems

References

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Su Nam Kim and Preslav Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In *EMNLP*, pages 648–658.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 636–644, Los Angeles, California, June. Association for Computational Linguistics.

Christopher D. Manning and Hinrich Schütze. 2001. *Foundations of statistical natural language processing*. MIT Press.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors, *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors, *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53.

Carlos Ramisch, Aline Villavicencio, and Valia Kordoni. to appear. *Special Issue on Multiword Expressions*. ACM TSLP.

Paul Rayson, Scott Songlin Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón. 2010. Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, 44(1-2):1–5.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15.

Violeta Seretan. 2012. *Syntax-Based Collocation Extraction*, volume 44, Text, Speech and Language Technology. Springer.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Comput. Linguist.*, 39(2):301–353, June.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043, Prague, Czech Republic, June. Association for Computational Linguistics.