Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources

Yulia Tsvetkov Shuly Wintner

Abstract

Multi-word Expressions (MWEs) are lexical items that consist of multiple orthographic words (e.g., *ad hoc, by and large, New York, kick the bucket*). MWEs are numerous and constitute a significant portion of the lexicon of any natural language (Jackendoff, 1997; Erman and Warren, 2000; Sag et al., 2002). They are a heterogeneous class of constructions with diverse sets of characteristics, distinguished by their idiosyncratic behavior. Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or preventing) the inflection of other constituents. In some cases MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation. Syntactically, some MWEs behave like words while other are phrases; some occur in one rigid pattern (and a fixed order), while others permit various syntactic transformations. Semantically, the compositionality of MWEs is gradual, ranging from fully compositional to idiomatic.

Because of their prevalence and irregularity, MWEs must be stored in lexicons of natural language processing applications. Handling MWEs correctly is beneficial for a variety of applications, including information retrieval (Doucet and Ahonen-Myka, 2004), building ontologies (Venkatsubramanyan and Perez-Carballo, 2004), text alignment (Venkatapathy and Joshi, 2006), and machine translation (MT) (Baldwin and Tanaka, 2004; Uchiyama et al., 2005).

We propose an architecture for expressing various linguistically-motivated features that help identify multi-word expressions in natural language texts. The architecture combines various linguistically-motivated classification features in a Bayesian Network, a classification device that is optimal for this task. Our methodology is almost entirely unsupervised and completely language-independent; it relies only on few language resources and is thus suitable for a large number of languages. Furthermore, unlike much recent work, our approach can identify expressions of various lengths, types and syntactic constructions. We demonstrate a significant improvement in identification accuracy, compared with less sophisticated baselines.

References

Timothy Baldwin and Takaaki Tanaka. Translation by machine of complex nominals: Getting it right. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Antoine Doucet and Helana Ahonen-Myka. Non-contiguous word sequences for information retrieval. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 88–95, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Britt Erman and Beatrice Warren. The idiom principle and the open choice principle. *Text*, 20(1):29–62, 2000.

Ray Jackendoff. *The Architecture of the Language Faculty.* MIT Press, Cambridge, USA, 1997.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 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)*, pages 1–15, Mexico City, Mexico, 2002.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. Disambiguating Japanese compound verbs. *Computer Speech & Language*, 19(4):497–512, October 2005.

Sriram Venkatapathy and Aravind Joshi. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, July 2006.

Shailaja Venkatsubramanyan and Jose Perez-Carballo. Multiword expression filtering for building knowledge. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 40–47, Barcelona, Spain, July 2004. Association for Computational Linguistics.