Broad-Coverage Analysis of English MWEs: From Annotation to Sequence Tagging

Nathan Schneider, University of Edinburgh
nschneid@inf.ed.ac.uk

1 Introduction

Multiword expressions (MWEs) are both *numerous*, occurring frequently in text, and *diverse*—they are not restricted to particular syntactic constructions or semantic domains (Baldwin and Kim, 2010). I will present a comprehensive and broad-coverage framework for manually **annotating** diverse MWEs in corpora, without requiring a lexicon, and then automatically **identifying** MWE instances in context with a statistical sequence tagger. Key contributions of the framework that will be highlighted below include:

- a formal **representation** of shallow token groupings into "strong" MWEs (including noncompositional expressions and proper names) and "weak" collocations
- **annotation guidelines** for applying such a representation to free text;
- a **corpus** of 55k words of informal English text comprehensively annotated for MWEs (Schneider et al., 2014b);
- a supervised **sequence model** that identifies gappy as well as contiguous MWEs (Schneider et al., 2014a); and
- an **evaluation scheme** appropriate to our representation which allows us to quantify the benefit of the statistical model over a lookup-based heuristic baseline.

Data and tools developed in this framework are available at www.ark.cs.cmu.edu/LexSem/.

2 Representation

The proposed approach to MWEs in context is *comprehensive*, meaning that it is not restricted to a particular lexical or even syntactic inventory of candidates. Included are the full spectrum of MWE classes—ranging from the most fixed (proper names, nominal compounds, connectives like *as well as*, idioms like *by and large*) to the most flexible (especially verb phrase expressions subject to internal modification or other syntactic processes affecting word order and/or contiguity). For example, the expression whose citation form is *pay attention to* could be instantiated as ***paid*** *no* ***attention to*** or ***attention*** *was* ***paid to***, both of which contain **gaps** between the lexicalized parts of the expression. Further, the object of the preposition is not part of the MWE, so the MWE is not a complete constituent by a standard syntactic analysis.

The approach taken here is to bypass the difficult issue of syntactic representation altogether: the (very shallow) MWE representation simply assigns tokens to groups, where each group reflects the lexicalized part of an MWE. Tokens within a group are not required to be contiguous. Two kinds of groups are allowed: **weak** groups for statistically idiomatic collocations, such as *highly recommended*, and **strong** groups for all MWEs involving an element of noncompositionality. A strong group may include one or more weak groups, but otherwise there is no nesting of groups. See Schneider et al. (2014b) for details.

To facilitate automatic sequence tagging, the group annotations are mapped to an encoding similar to the traditional BIO scheme for chunking (Ramshaw and Marcus, 1995): namely, 8 tags—`O`, `o`, `B`, `b`, `Ī`, `ī`, `Ĩ`, `ĩ`—allow for the distinctions of tokens:

- positioned in the gap of some MWE (lowercase tags) vs. not (uppercase tags), and
- not belonging to an MWE (`O`/`o`), beginning an MWE (`B`/`b`), continuing a strong MWE (`Ī`/`ī`), or continuing a weak MWE (`Ĩ`/`ĩ`).

This encoding allows for MWEs with multiple gaps (e.g., ***putting*** *me* ***at*** *my* ***ease***). It prohibits any MWE with a gap from occurring in the gap of another MWE, and also excludes weak MWEs consisting of a gappy strong MWE and one or more tokens inside the gap. That these constraints are linguistically reasonable is empirically supported by the annotated corpus (counterexamples exist but are extremely rare).

3 Annotation

A corpus of 723 online reviews from the English Web Treebank (Bies et al., 2012) has been annotated in this framework. The text in this corpus is written in an informal style and colloquial idioms are frequent. The comprehensive annotations cover 3,800 sentences (55k words); 3,024 strong and 459 weak MWEs are annotated. Each sentence was independently annotated by at least two annotators, who then negotiated a consensus for any disagreements. All annotators hold bachelor's degrees in linguistics. Inter-annotator agreement estimates and other details are given in (Schneider et al., 2014b). Because the corpus is from a treebank, the shallow MWE annotations could be aligned post hoc to syntactic parses.

4 Statistical Model and Evaluation

The shallow MWE annotations described above can, with the 8-tag encoding, be used to train and evaluate a statistical sequence tagger, similar to other shallow MWE identification systems (Constant and Sigogne, 2011; Constant et al., 2012; Vincze et al., 2013, *inter alia*). The strong vs. weak distinction and the way in which gaps are allowed are novel (Bar et al.'s (2014) shallow Arabic MWE tagger takes an alternate approach to gaps). Details appear in Schneider et al. (2014a). In brief: Schneider et al. adapt the feature representation of Constant et al. (2012), incorporating several MWE lexicons and training a discriminative first-order Markov model with the structured perceptron (Collins, 2002).

POS pattern	**#**	**examples (lowercased lemmas)**
Noun Noun	53	*customer service*, *oil change*
Verb Prep	36	*work with*, *deal with*, *yell at*
PropN PropN	29	*eagle transmission*, *comfort zone*
Adj Noun	21	*major award*, *top notch*
Verb Part	20	*move out*, *end up*, *pick up*, *pass up*
Verb Adv	17	*come back*, *come in*, *come by*
Prep Noun	12	*on time*, *in fact*, *in cash*, *for instance*
Verb Noun	10	*take care*, *make money*, *give crap*
Verb Pron	10	*thank you*, *get it*
Prep Prep	8	*out of*, *due to*, *out ta*, *in between*
Adv Adv	6	*no matter*, *up front*, *at all*, *early on*
Det Noun	6	*a lot*, *a little*, *a bit*, *a deal*
Verb Det Noun	6	*answer the phone*, *take a chance*
Noun Prep	5	*kind of*, *care for*, *tip on*, *answer to*

Table 1: Top predicted POS patterns and counts.

Experiments on held-out data show that the statistical model is vastly superior to a baseline involving heuristic matching against MWE lexicons. However, utilizing those lexicons in a soft way, through features, is beneficial. Schneider et al. (2014a) quantify these comparisons with a new evaluation measure for automatic shallow MWE analyses. The main idea is that partial credit is given for partial overlap between a gold MWE instance and a predicted MWE instance by computing precision and recall not over the full MWE, but over *links* between consecutive tokens belonging to each MWE. The best result (without gold POS tags) is 64% precision, 56% recall, and 59% F_1 on the test set. Table 1 shows a sample of the system's output.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA.

Kfir Bar, Mona Diab, and Abdelati Hawwari. 2014. Arabic multiword expressions. In Nachum Dershowitz and Ephraim Nissan, editors, *Language, Culture, Computation. Computational Linguistics and Linguistics*, number 8003 in Lecture Notes in Computer Science, pages 64–81. Springer, Berlin.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.

Michael Collins. 2002. Discriminative training methods for Hidden Markov Models: theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, pages 1–8. Philadelphia, Pennsylvania, USA.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56. Portland, Oregon, USA.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proc. of ACL*, pages 204–212. Jeju Island, Korea.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proc. of the Third ACL Workshop on Very Large Corpora*, pages 82–94. Cambridge, Massachusetts, USA.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 455–461. Reykjavík, Iceland.

Veronika Vincze, István Nagy T., and János Zsibrita. 2013. Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing*, 10(2):6:1–6:25.