Modeling the Non-Substitutability of Multiword Expressions with Distributional Semantics and a Log-Linear Model

Meghdad Farahmand
Department of Computer Science
University of Geneva
meghdad.farahmand@unige.ch

James Henderson
Xerox Research Centre Europe
james.henderson@xrce.xerox.com

Abstract

Non-substitutability is a property of Multiword Expressions (MWEs) that often causes lexical rigidity and is relevant for most types of MWEs. Efficient identification of this property can result in the efficient identification of MWEs. In this work we propose using distributional semantics, in the form of word embeddings, to identify candidate substitutions for a candidate MWE and model its substitutability. We use our models to rank MWEs based on their lexical rigidity and study their performance in comparison with association measures. We also study the interaction between our models and association measures. We show that one of our models can significantly improve over the association measure baselines, identifying collocations.

1 Introduction

Multiword expressions (MWEs), commonly referred to as collocations,[1] are idiosyncratic sequences of words whose idiosyncrasy can be broadly classified into semantic, statistical, and syntactic classes. Semantic idiosyncrasy (also referred to as non-compositionality) means that the meaning of an MWE cannot be inferred from the meaning of its components, as in *loan shark*. Syntactic idiosyncrasy refers to the situation where the syntax of an MWE does not follow syntactic rules, as in *in short*. Statistical idiosyncrasy means that components of a statistically idiosyncratic MWE co-occur more than expected by chance, as in *swimming pool*. The range of types of idiosyncrasy included in MWEs has been characterized in several other ways (Baldwin and Kim, 2010; Sag et al., 2002). To avoid getting mired down in this uncertainty, which mainly emerges while dealing with borderline MWEs, between completely idiosyncratic and fully compositional, we subscribe to the viewpoint of McCarthy et al. (2007) and treat idiosyncrasy as a spectrum and focus only on the (very) idiosyncratic end of this spectrum. MWEs have application in different areas in NLP and linguistics, for instance statistical machine translation (Ren et al., 2009; Carpuat and Diab, 2010); shallow parsing (Korkontzelos and Manandhar, 2010); language generation (Hogan et al., 2007); opinion mining (Berend, 2011); corpus linguistics and language acquisition (Ellis, 2008). In general, as Green et al. (2011) point out, "MWE knowledge is useful, but MWEs are hard to identify."

In this work, we propose a method of identifying MWEs based on their non-substitutability. Non-substitutability means that the components of an MWE cannot be replaced with their synonyms (Manning and Schütze, 1999; Pearce, 2001). It implies statistical idiosyncrasy, which is relevant for all kinds of MWEs, and identifying non-substitutability in text results in the identification of a wide range of MWEs. In MWE research, non-substitutability has been widely considered but never thoroughly studied, except for a few work that present low coverage and limited models of this concept.

We develop a model that takes into account the semantics of words for identifying statistical idiosyncrasy, but is highly generalizable and does not require supervision or labor-intensive resources. The proposed model uses distributional semantics, in the form of word embeddings, and

[1] In older work, the term collocation refers to all kinds of MWEs. In more recent work, however, it mainly refers to statistically idiosyncratic MWEs. In any case, statistical idiosyncrasy can be considered as a general property of all kinds of MWEs, regardless of other forms of idiosyncrasy they may have.

Proceedings of the 12th Workshop on Multiword Expressions, pages 61–66,
Berlin, Germany, August 7-12, 2016.

uses them to identify semantically similar words for the components of a candidate MWE. Non-substitutability is then measured for the candidate MWE using log-linear model(s), also computed using word embeddings. Our proposed models result in an improvement over the state-of-the-art.

1.1 Syntactic Categories of MWEs

From a syntactic point of view, MWEs are very heterogeneous, including light verb constructions, phrasal verbs, noun compounds, verb-object combinations and others. In this work, however, we focus only on noun compounds for the following reasons: (i) They are the most productive and frequent category of MWEs.() (ii) There are more datasets of compounds available for evaluation. (iii) Focusing on one controlled category allows us to focus on modeling and detecting idiosyncrasy in isolation, avoiding complexities such as gappy MWEs. We also focus only on two-word noun compounds, because higher order ones are relatively rare.

2 Related Work

Identification of statistical idiosyncrasy of MWEs seems to have been first formally discussed in Choueka et al. (1983) by proposing a statistical index to identify collocates and further developed into more efficient measures of collocation extraction such as Pointwise Mutual Information (Church and Hanks, 1990), t-score (Church et al., 1991; Manning and Schütze, 1999), and Likelihood Ratio (Dunning, 1993). Smadja (1993) proposes a set of statistical scores that can be used to extract collocations. Evert (2005) and Pecina (2010) study a wide range of association measures that can be employed to rank and classify collocations, respectively.

Farahmand and Nivre (2015) assume that a word pair is a true MWE if the conditional probability of one word given the other is greater than the conditional probability of that word given synonyms of the other word, and Riedl and Biemann (2015), and Farahmand and Martins (2014) use contextual features to identify MWEs.

The above-mentioned methods target statistical idiosyncrasy of MWEs. There are however many other approaches to extraction of MWEs which do not explicitly focus on statistical idiosyncrasy. For instance, some identify MWEs based on their semantic idiosyncrasy (Yazdani et al., 2015; Im Walde et al., 2013; Hermann et al., 2012; Reddy et al., 2011; Baldwin et al., 2003; McCarthy et al., 2003), some approaches are rule-based (Seretan, 2011; Baldwin, 2005), and some are both rule-based and statistical (Ramisch, 2012; Seretan and Wehrli, 2006).

3 Modeling Non-Substitutability

As discussed earlier, we model statistical idiosyncrasy based on an assumption inspired by *non-substitutability*, which means that the components of an MWE cannot be replaced with their near synonyms. Let w_1w_2 represent a word pair. We make the same assumption as Farahmand and Nivre (2015) that w_1w_2 is statistically idiosyncratic if:

$$P(w_2|w_1) > P(w_2|sim(w_1)) \quad (1)$$

where $sim(w_i)$ (defined below in Section 3.1) represents the words that are similar to w_i. With respect to noun noun compounds, this inequality roughly means that for an idiosyncratic compound, the probability of the headword (w_2) co-occurring with the modifier (w_1) is greater than the probability of the headword co-occurring with "synonyms" of the modifier (e.g. *climate change* is more probable than *weather change*). This, however, is not the case for non or less idiosyncratic compounds (e.g. *film director* which is substitutable with *movie director*).

Farahmand and Nivre (2015) estimate a similar probability, in both directions, with the help of WordNet *synset*s. They show that the model that considers the probabilities in both directions outperforms the model that considers only one direction (head conditioned on modifier).

To study and model the effects of *direction* we also consider the following inequality:

$$P(w_1|w_2) > P(w_1|sim(w_2)) \quad (2)$$

Intuitively, inequality 1 plays a more important role in lexical rigidity than inequality 2, but this is something we study in section 4.

In related work, (Pearce, 2001) extracts the synonyms of the constituents of a compound, creates new phrases called *anti-collocations*, and based on the number of *anti-collocations* of the candidate MWE decides whether it is a true MWE.

3.1 Modeling Semantically Similar Words

In previous work, WordNet synsets were employed to model the $sim()$ function. The obvious

limitation of such an approach is coverage. Other limitations include costliness and labor intensiveness of updating and expanding such a knowledge base. In this work, we use cosine similarity between word embeddings to represent semantically similar words (that include but are not limited to synonyms). This may result in a drop in precision, but the coverage will be immensely improved. Moreover, similarity in the word embedding space is shown to provide a relatively good approximation of synonymy (Chen et al., 2013).

3.2 Ranking with Log-Linear Models

We estimate the probabilities presented in (1) and (2) using a log linear model. Let $\phi(w_i)$ represent the word embedding of w_i where $\phi \in \mathbb{R}^{50}$.

$$P(w_2|w_1) = \frac{exp(v_{w_2} \cdot \phi(w_1))}{\sum\limits_{w_2'} exp(v_{w_2'} \cdot \phi(w_1))} \quad (3)$$

where v_{w_i} is a parameter vector and v is the model's parameter matrix. The analogous equation is used to define $P(w_1|w_2)$.

Let S_{w_i} represent the set of top-n $\phi(w_j)$ that are most similar to $\phi(w_i)$, $S_{w_i} = \{w_j|w_j \in nGreatest(w_i.w_j)\}$. $P(w_2|sim(w_1))$ can then be estimated as:

$$P(w_2|sim(w_1)) = \frac{1}{|S_{w_1}|} \sum_{w_j \in S_{w_1}} P(w_2|w_j)$$

where $P(w_2|w_j)$ is defined in (3).

And again, the analogous equation defines $P(w_1|sim(w_2))$.

Combining these gives us the following version of (1), and an analogous version of (2).

$$\frac{exp(v_{w_2} \cdot \phi(w_1))}{\sum\limits_{w_2'} exp(v_{w_2'} \cdot \phi(w_1))} > \frac{1}{|S_{w_1}|} \sum_{w_j \in S_{w_1}} \frac{exp(v_{w_2} \cdot \phi(w_j))}{\sum\limits_{w_2'} exp(v_{w_2'} \cdot \phi(w_j))} \quad (4)$$

Given that MWEs lie on a continuum of idiosyncrasy, it is natural to treat identification of MWEs as a ranking problem. We therefore define an unsupervised ranking function as follows:

$$\delta_{21} = \frac{exp(v_{w_2} \cdot \phi(w_1))}{\sum\limits_{w_2'} exp(v_{w_2'} \cdot \phi(w_1))} - \frac{1}{|S_{w_1}|} \sum_{w_j \in S_{w_1}} \frac{exp(v_{w_2} \cdot \phi(w_j))}{\sum\limits_{w_2'} exp(v_{w_2'} \cdot \phi(w_j))} \quad (5)$$

And an analogous function δ_{12}.

4 Evaluation

As our evaluation set we used the dataset of Farahmand et al. (2015) who annotate 1042 English noun compounds for statistical and semantic idiosyncrasy. Each compound is annotated by four judges with two binary votes, one for their semantic and one for their statistical idiosyncrasy.

As our baselines we use three measures that have been widely used as a means of identifying collocations: Pointwise Mutual Information (PMI) (Church and Hanks, 1990; Evert, 2005; Bouma, 2009; Pecina, 2010), t-score (Manning and Schütze, 1999; Church et al., 1991; Evert, 2005; Pecina, 2010), and Log-likelihood Ratio (LL_r) (Dunning, 1993; Evert, 2005).

Since we are concerned with the idiosyncratic end of the spectrum of MWEs, we look at the identification of MWEs as a ranking problem. To evaluate this ranking, we use precision at k ($p@k$) as the evaluation metric, considering different values of k.

4.1 Individual Models

To train the log-linear model, we first extracted all noun-noun compounds from a POS-tagged Wikipedia dump (only articles) with a frequency of at least 5. This resulted in a list of $\approx 560,000$ compounds. We created word embeddings of size 50 for words of Wikipedia that had the frequency of at least 5 using *word2vec*[2]. These word embeddings were used both to determine the set of similar words for each word of a compound and to train the log-linear model by stochastic minimization of the cross entropy. We discarded 30 instances of the evaluation set because (having type frequency of below 5) word embeddings were not available for at least one of their components.

To measure precision, we assume those evaluation set instances that were annotated as statistically or semantically idiosyncratic by three or more judges (out of four) are MWE and other instances are not. This results in the total of 369 positive instances. Figure 1 shows the performance of the different models.

At the top of the ranked list, δ_{21} outperforms one of the baselines (t-score) but performs similarly to the other two baselines, PMI and $\mathrm{LL_r}$. It, however, shows a more steady performance up

[2] https://code.google.com/archive/p/word2vec/

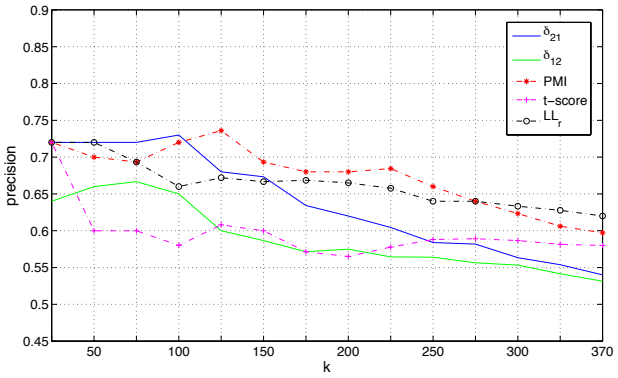

Figure 1: $p@\,k$ for our models and the baselines.

until $p@\,100$. As it moves further from the idiosyncratic end of the spectrum its precision drops further. δ_{12}, on the other hand, shows a weaker performance. It, however, outperforms t-score for the most part. The best baseline is PMI, the worst is t-score. Again, considering lexicalization, the main process that MWEs should undergo to become useful for other NLP applications, a high precision at a small (proportional) k is what we should be really concerned about: lexicons cannot grow too large so every multi-word entry should be sufficiently idiosyncratic and lexically rigid. On the other hand, we do not want to limit a model's ability to generalize by lexicalizing every word sequence that appears slightly idiosyncratic. Looking back at the models, we know that δ_{21}, PMI, and LL_r independently perform well at the top of their ranked list. On the other hand, we know that in theory δ_{21} bases its ranking on relatively different criteria from PMI and $\mathrm{LL_r}$. The question we seek to answer in the next section is whether merging these criteria (semantic non-substitutability and statistical association) can improve on the best performance.

4.2 Combining Non-Substitutability and Association

Our first combined model of non-substitutability integrates both directions (head to modifier and modifier to head). To emphasize precision, we propose a combination function H_1 that requires both δ_{21} and δ_{12} to be high.

$$H_1 = min(\delta_{21}, \delta_{12})$$

By ranking according to the minimum of the scores δ_{21} and δ_{12}, each highly-ranked data point must be highly ranked by both individual models.[3]

To combine an association measure with our non-substitutability models we chose PMI because its performance at the top of the ranked list is better than other baselines. The values of PMI and the δs have different scales. We measured the linear correlation in terms of Pearson r between PMI and δs in order to see whether we can scale up the δs' by a linear factor. The correlation was very small and almost negligible, so instead of using $min()$ we combined the two rankings as:

$$H_2 = H_1 \circ \mathrm{PMI}$$

where $\circ$ denotes the element-wise product.

We perform the same experiments as in Section 4.1 with the combined models[4] and compare their performance with the best models from the previous experiments. The results can be seen in Figure 2.

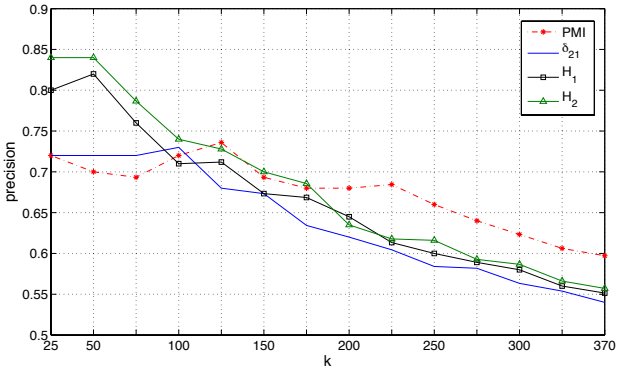

Figure 2: $p@\,k$ for H_1, H_2 and previous best models.

H_2 clearly outperforms other models at the top of the ranked list. It reaches a significantly higher precision than other models. This confirms our assumption that in practice association measures and substitutability based models that are semantically motivated[5] base their decisions on different pieces of information that are complementary. Also, the results for H_1 show that combining both δ_{21} and δ_{12} gives us an improvement for high precision and performs similarly to the best one (δ_{21}) at lower k.

[3] We also tried element-wise multiplication in order to combine these models. The performance of $min()$, however, was slightly better.

[4] We also combined different association measures which resulted in models with performances that were mainly similar to the performance of their sub-models.

[5] Assuming that word embeddings represent semantics in a slightly more meaningful way than first order statistical association.

5 Conclusions

We presented a method for identifying MWEs based on their semantic non-substitutability. We assumed that non-substitutability implies statistical idiosyncrasy and modeled this property with word embedding representations and a log-linear model. We looked at MWE identification as a ranking problem due the nature of idiosyncrasy, which is better defined as a continuum than as a binary phenomenon. We showed our best model can reach the same performance as the best baseline. We showed that joining our models lead to a better performance compared to that of the baselines and individual models. We also showed that joining our models -that are aware of semantic non-substitutability, and association measures (baselines) can result in a model with a performance that is significantly higher than the performance of the baselines.

References

[Baldwin and Kim2010] Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool*.

[Baldwin et al.2003] Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics.

[Baldwin2005] Timothy Baldwin. 2005. Deep lexical acquisition of verb–particle constructions. *Computer Speech & Language*, 19(4):398–414.

[Berend2011] Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *IJCNLP*, pages 1162–1170. Citeseer.

[Bouma2009] G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen.

[Carpuat and Diab2010] Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245. Association for Computational Linguistics.

[Chen et al.2013] Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226*.

[Choueka et al.1983] Yaacov Choueka, Shmuel T Klein, and E Neuwitz. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Association for Literary and Linguistic Computing Journal*, 4(1):34–38.

[Church and Hanks1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

[Church et al.1991] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum.

[Dunning1993] Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

[Ellis2008] Nick C Ellis. 2008. The periphery and the heart of language. *Phraseology: An interdisciplinary perspective*, pages 1–13.

[Evert2005] Stefan Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, Dissertation, Stuttgart University.

[Farahmand and Martins2014] Meghdad Farahmand and Ronaldo Martins. 2014. 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. Association for Computational Linguistics.

[Farahmand and Nivre2015] Meghdad Farahmand and Joakim Nivre. 2015. Modeling the statistical idiosyncrasy of multiword expressions. In *Proceedings of NAACL-HLT*, pages 34–38.

[Farahmand et al.2015] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions (MWE-NAACL 2015)*. Association for Computational Linguistics.

[Green et al.2011] Spence Green, Marie-Catherine De Marneffe, John Bauer, and Christopher D Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 725–735. Association for Computational Linguistics.

[Hermann et al.2012] Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 132–141. Association for Computational Linguistics.

[Hogan et al.2007] Deirdre Hogan, Conor Cafferkey, Aoife Cahill, and Josef Van Genabith. 2007. Exploiting multi-word units in history-based probabilistic generation. Association for Computational Linguistics.

[Im Walde et al.2013] Sabine Schulte Im Walde, Stefan Müller, and Stephen Roller. 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*, pages 255–265.

[Korkontzelos and Manandhar2010] 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. Association for Computational Linguistics.

[Manning and Schütze1999] Christopher D Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

[McCarthy et al.2003] Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acqusation and Treatment*, pages 73–80.

[McCarthy et al.2007] Diana McCarthy, Sriram Venkatapathy, and Aravind K Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *EMNLP-CoNLL*, pages 369–379.

[Pearce2001] Darren Pearce. 2001. Synonymy in collocation extraction. In *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 41–46. Citeseer.

[Pecina2010] Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158.

[Ramisch2012] Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, pages 61–66. Association for Computational Linguistics.

[Reddy et al.2011] Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *IJCNLP*, pages 210–218.

[Ren et al.2009] Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54. Association for Computational Linguistics.

[Riedl and Biemann2015] Martin Riedl and Chris Biemann. 2015. A single word is not enough: Ranking multiword expressions using distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, Lisboa, Portugal.

[Sag et al.2002] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

[Seretan and Wehrli2006] Violeta Seretan and Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 953–960. Association for Computational Linguistics.

[Seretan2011] Violeta Seretan. 2011. *Syntax-based collocation extraction*, volume 44. Springer.

[Smadja1993] Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

[Yazdani et al.2015] Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal, September. Association for Computational Linguistics.