Modeling the Statistical Idiosyncrasy of Multiword Expressions

Meghdad Farahmand
University of Geneva
Geneva, Switzerland
meghdad.farahmand@unige.ch

Joakim Nivre
Uppsala University
Uppsala, Sweden
joakim.nivre@lingfil.uu.se

Abstract

The focus of this work is statistical idiosyncrasy (or collocational weight) as a discriminant property of multiword expressions. We formalize and model this property, compile a 2-class data set of MWE and non-MWE examples, and evaluate our models on this data set. We present a possible empirical implementation of collocational weight and study its effects on identification and extraction of MWEs. Our models prove to be more effective than baselines in identifying noun-noun MWEs.

1 Introduction

Multiword Expressions (MWEs) are sequences of words that show some level of idiosyncrasy. For instance they can be semantically idiosyncratic (i.e., their meaning cannot be readily inferred from the meaning of their components, e.g., f*lea market*), syntactically idiosyncratic (their syntax cannot be extracted from the syntax of their components, e.g., *at large*), statistically idiosyncratic (their components tend to co-occur more often than expected by chance, e.g., *drug dealer*), or have other forms of idiosyncrasy. MWEs comprise several types and sub-types. Although it is not always clear where to draw the line between various types of MWEs, the two broadest categories are lexicalized MWEs and institutionalized MWEs (Sag et al., 2002). The main property of lexicalized MWEs is syntactic or semantic idiosyncrasy and the main property of institutionalized MWEs is statistical idiosyncrasy. Semantic idiosyncrasy is closely related to the concept of non-compositionality. It is important to note that a MWE is often idiosyncratic in more than one way (Baldwin and Kim, 2010). This means lexicalized MWEs can be statistically idiosyncratic, and institutionalized MWEs can be semantically idiosyncratic. Institutionalized MWEs are closely related to collocations.[1] They can be compositional (*seat belt*) or non-compositional (*hard drive*), but statistically they co-occur more often than expected by chance.

Efficient extraction and identification of MWEs can positively influence some important Natural Language Processing (NLP) tasks such as parsing (Nivre and Nilsson, 2004) and Statistical Machine Translation (Ren et al., 2009). Identification and extraction of MWEs are therefore important research questions in the area of NLP.

In this work we refer to statistical idiosyncrasy as collocational weight and present a method of modeling this property for noun-noun compounds. Comparative evaluation reveals better performance of proposed models compared to that of the baselines.

In previous work, it has often been suggested that collocations can be identified by their non-substitutability. This means we cannot replace a collocation's components with their near synonyms (Manning and Schütze, 1999). For instance we cannot say *brief film* instead of *short film*. Pearce (2001) defines collocations as pairs of words where "one of the words significantly prefers a particular lexical re-

[1] Although the major property of collocations is known to be statistical idiosyncrasy, in many works, semantically idiosyncratic multiword expressions have also been regarded as collocation.

Proceedings of NAACL-HLT 2015, pages 34–38,
Denver, Colorado, May 31 – June 5, 2015.

alization of the concept the other represents." To the best of our knowledge, however, non-substitutability (with near synonyms) or in other words collocational weight has never been explicitly and empirically tested. In this work, we present two models that partially, and fully, model collocational weight, and investigate its effects on extraction of MWEs.

2 Related work

Extraction of MWEs has been widely researched from different perspectives. Various models from rule-based to statistical have been employed to address this problem.

Examples of rule-based models are Seretan (2011) and Jacquemin et al. (1997) who base their extraction on linguistic rules and formalism in order to identify and filter MWE candidates, and Baldwin (2005) who aims at extracting verb particle constructions based on their linguistic properties using a chunker and dependency grammar.

Examples of statistical models are Pecina (2010), Evert (2005), Lapata and Lascarides (2003), and the early work **Xtract** (Smadja, 1993). Farahmand and Martins (2014) present a method of extracting MWEs based on their statistical contextual properties and Hermann et al. (2012) employ distributional semantics to model non-compositionality and use it as a way of identifying lexicalized compounds.

There are also hybrid models in the sense that they benefit from both statistical and linguistic information (Seretan and Wehrli, 2006; Dias, 2003). Ramisch (2012) implements a flexible platform that accepts both statistical and deep linguistic criteria in order to extract and filter MWEs.

There are also bilingual models which are mostly based on the assumption that a translation of a source language MWE exists in a target language (Smith, 2014; Caseli et al., 2010; Ren et al., 2009).

A similar work to ours is Pearce (2001) who uses WordNet in order to produce anti-collocations from synonyms of the components of a MWE candidate, and decides about "MWEhood" based on these anti-collocations. Another similar work is Ramisch et al. (2008) who use WordNet Synsets as one of their resources in order to calculate the entropy between the components of verb particle constructions.

3 Method

Following previous work by Manning and Schütze (1999), and Pearce (2001), we define collocational weight -a discriminant property of mainly institutionalized but also lexical MWEs, for noun-noun pairs according to the following hypotheses:

Simplified Hypothesis *For a given two-word compound, the head word is more likely to co-occur with the modifier than with synonyms of the modifier.*

Main Hypothesis *For a given two-word compound, the head word is more likely to co-occur with the modifier than with synonyms of the modifier, and the modifier is more likely to co-occur with the head than with synonyms of the head.*

We formalize these hypotheses in the form of M_1 and M_2 models which implement the simplified and main hypotheses and are described by equations (1) and (2), respectively.

$$M_1 : P(w_2|w_1) > \alpha P(w_2|Syns(w_1)) \quad (1)$$

where:

$$P(w_2|w_1) = \frac{\#(w_1 w_2)}{\#(w_1)}$$

and

$$P(w_2|Syns(w_1)) = \frac{\sum_{w_1' \in Syns(w_1)} \#(w_1' w_2)}{\sum_{w_1' \in Syns(w1)} \#(w_1' + \mathcal{L})}$$

$w_1 w_2$ represents a compound. $Syns(w)$ represents a set of synonyms of w, and in order to obtain such a set we use WordNet's $synset()$ function. $\mathcal{L}$ is the smoothing factor, which is set to 0.1, and α is a parameter that we altered between $[1 - 30]$. $\mathcal{L}$ and α are also present in M_2 and are assigned the same values as in M_1.

$$M_2 : P(w_2|w_1) > \alpha P(w_2|Syns(w_1)) \quad (2)$$

$$\&\& \; P(w_1|w_2) > \alpha P(w_1|Syns(w_2))$$

where:

$$P(w_2|w_1) = \frac{\#(w_1 w_2)}{\#(w_1)}$$

$$P(w_1|w_2) = \frac{\#(w_1 w_2)}{\#(w_2)}$$

and

$$P(w_2|Syns(w_1)) = \frac{\sum\limits_{w_1' \in Syns(w_1)} \#(w_1' w_2)}{\sum\limits_{w_1' \in Syns(w1)} \#(w_1') + \mathcal{L}}$$

$$P(w_1|Syns(w_2)) = \frac{\sum\limits_{w_2' \in Syns(w_2)} \#(w_1 w_2')}{\sum\limits_{w_2' \in Syns(w2)} \#(w_2') + \mathcal{L}}$$

4 Experiments

In order to test our hypotheses, we implement the two models described above and two baselines, and run a comparative evaluation. We divide our data into two subsets: development and test sets. The evaluation is carried out in two phases. In the first phase we perform model selection and find the optimal parameters for various models on the development set. In the second phase we evaluate the selected models with optimal parameters on the test set, which remains unseen by the models up to this phase.

4.1 Data

Although there exist a few data sets for English compounds (Baldwin and Kim, 2010; Reddy et al., 2011), to the best of our knowledge there is no data set with annotations for both MWE and non-MWE classes. We required this for the evaluation of our models therefore we compiled our own data set. We randomly extracted a set of 3000 noun-noun pairs that had the frequency of greater than 10 from across POS-tagged English Wikipedia. We kept only the pairs whose both head and modifier had more than one synonym according to WordNet. In cases were a given compound had different POS tags, we selected the most frequent tags. We asked two computational linguists with background in MWE research to annotate the pairs as MWE and non-MWE. Pairs which were either semantically or statistically idiosyncratic, or both were annotated as MWE. Pairs which were neither semantically nor syntactically nor statistically idiosyncratic were annotated as non-MWE. To asses the inter annotator agreement we calculated Cohen's kappa (κ) and to measure the pairwise correlation among the annotators we calculated Spearman's rank correlation coefficient (ρ). The Spearman ρ was equal to 0.66. The Cohen's kappa was equal to 0.64 (with the error of 0.02) which can be interpreted as "substantial agreement" according to Landis and Koch (1977). In the final data set, the instances which were judged as MWE by both annotators were regarded as MWE and the instances which were judged as non-MWE by both annotators were regarded as non-MWE. This resulted in a set of 262 instances of MWE and 560 instances of non-MWE classes. To avoid the possible bias of the results towards non-MWE class, we reduced the size of non-MWE class to 262 by randomly removing 298 instances. Afterward we divided the data into development $(2/3)$ and test $(1/3)$ sets, which contain the same proportion of MWE and non-MWE instances. An overview of the data set is presented in Table 1.

Set	MWE	non-MWE
original set	262	262
dev. set	174	174
test set	88	88
examples	gold rush, role model, family tree, city center, bow saw, life cycle	chess talent, bus types, attack damage, player skill, oil storage, lobby area

Table 1: Dataset statistics.

4.2 Evaluation

We implement the following two baselines: (1) Multinomial likelihood (Evert, 2005), which calculates the probability of the observed contingency table for a given pair under the null hypothesis of independence. (2) Mutual information (Church and Hanks, 1990), which calculates the mutual depen-

dency of words of a co-occurrence, and has been proved efficient in identification and extraction of MWEs (Pecina, 2010; Evert, 2005). With respect to the range of scores, we set and alter a threshold for multinomial likelihood ($M.N.L$ hereafter) and mutual information ($M.I.$ hereafter). Pairs that obtain a score above the threshold are considered MWE, and pairs that obtain a score below the threshold are considered non-MWE. Figure 1 illustrates the precision-recall curve for our models and the baselines on the development set.

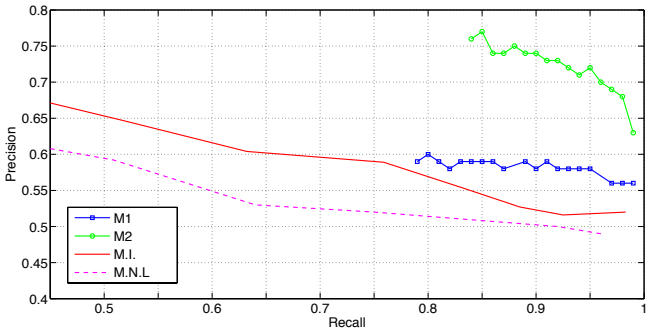

Figure 1: Precision-recall curve for various models.

The two baseline models i.e., $M.N.L.$ and $M.I.$ reach a high precision only at the cost of a dramatic loss in recall. They behave similarly, however, $M.I.$ in general performs better. M_2 clearly performs better compare to all other models. It reaches a high precision and recall, however, its precision declines rather quickly when recall increases. M_1 shows a more steady behaviour in the sense that reaching a higher recall doesn't significantly impact its precision. Figure 2 shows how F_1 score changes for various models when changing parameters in order to go from high precision to high recall. M_1 and M_2 constantly have a higher F_1 score, where $M.I.$ and $M.N.L.$ start off with a low score and reach a score which is comparable with that of the other models.

Out of the four tested models, with respect to F_1 scores, we select M_1, M_2, and $M.I.$ for further experiments. We set the relevant parameters to optimal values[2] (obtained by looking at the highest F_1 scores) and run the next experiments on the test set, which has remained unseen by the models up to this

[2]Optimal values of the parameters are as follows: α in M_1 : 15, α in M_2 : 20 and threshold for $M.I.$: 0.2

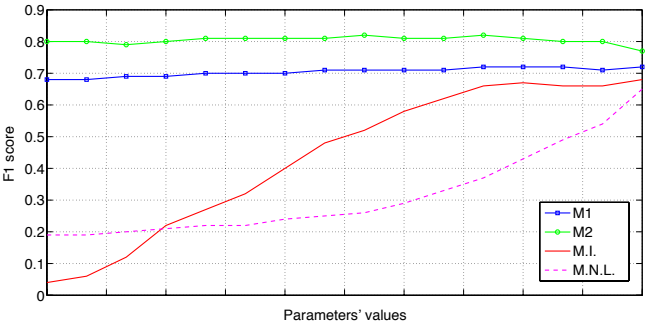

Figure 2: F_1 score for various models.

point. Table 2 shows the result of these experiments. The performance of all three models on the test set is consistent with their performance on the development set. M_2 reaches the highest precision and F_1 score. $M.I.$ has the highest recall but a low precision, and M_1 has a high recall and a reasonable but not very high precision.

model	precision	recall	F_1
M_1	0.57	0.88	0.69
M_2	0.75	0.86	0.80
$M.I.$	0.51	0.95	0.66

Table 2: Evaluation results in terms of precision, recall and F_1 score for the three selected models.

5 Conclusions

We showed that statistical idiosyncrasy can play a significant role in identification and extraction of MWEs. We showed that this property can be used efficiently to extract idiosyncratic noun compounds which constitute the largest subset of English MWEs. We referred to statistical idiosyncrasy as collocational weight and formalized this property and implemented two corresponding models. We empirically tested the performance of these models against two baselines and showed that one of our models constantly outperforms the baselines and reaches an F_1 score of 0.80 on the test set.

Acknowledgments

We would like to thank James Henderson and Aaron Smith for discussions of various points and their help in carrying out this work.

References

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

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

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, 44(1-2):59–77.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Gaël Dias. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 41–48. Association for Computational Linguistics.

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

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.

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.

Christian Jacquemin, Judith L Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 24–31. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Mirella Lapata and Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 235–242. Association for Computational Linguistics.

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

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *In Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*.

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.

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

Carlos Ramisch, Aline Villavicencio, Leonardo Moura, and Marco Idiart. 2008. Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 49–56. Association for Computational Linguistics.

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.

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

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.

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.

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.

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

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

Aaron Smith. 2014. Breaking bad: Extraction of verb-particle constructions from a parallel subtitles corpus. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 1–9. Association for Computational Linguistics.