A Supervised Model for Extraction of Multiword Expressions Based on Statistical Context Features

Meghdad Farahmand
The Computer Science Center
University of Geneva
Switzerland
meghdad.farahmand@unige.ch

Ronaldo Martins
UNDL Foundation
Geneva - Switzerland
r.martins@undl.ch

Abstract

We present a method for extracting Multiword Expressions (MWEs) based on the immediate context they occur in, using a supervised model. We show some of these contextual features can be very discriminant and combining them with MWE-specific features results in a relatively accurate extraction. We define context as a sequential structure and not a bag of words, consequently, it becomes much more informative about MWEs.

1 Introduction

Multiword Expressions (MWEs) are an important research topic in the area of Natural Language Processing (NLP). Efficient and effective extraction and interpretation of MWEs is crucial in most NLP tasks. They exist in many types of text and cause major problems in all kinds of natural language processing applications (Sag et al., 2002). However, identifying and lexicalizing these important but hard to identify structures need to be improved in most major computational lexicons (Calzolari et al., 2002). Jackendoff (1997) estimates that the number of MWEs is equal to the number of single words in a speaker's lexicon, while Sag et al. (2002) believe that the number is even greater than this. Moreover, as a language evolves, the number of MWEs consistently increases. MWEs are a powerful way of extending languages' lexicons. Their role in language evolution is so important that according to Baldwin and Kim (2010), "It is highly doubtful that any language would evolve without MWEs of some description".

The efficient identification and extraction of MWEs can positively influence many other NLP tasks, e.g., part of speech tagging, parsing, syntactic disambiguation, semantic tagging, machine translation, and natural language generation. MWEs also have important applications outside NLP. For instance in document indexing, information retrieval (Acosta et al., 2011), and cross lingual information retrieval (Hull and Grefenstette, 1996).

In this paper we present a method of extracting MWEs which is relatively different from most of the state of the art approaches. We characterize MWEs based on the statistical properties of the immediate context they occur in. For each possible MWE candidate we define a set of contextual features (e.g., prefixes, suffixes, etc.). The contextual feature vector is then enriched with a few MWE-specific features such as the frequency of its components, type frequency of the candidate MWE, and the association between these two (which is learned by a supervised model). Subsequently the MWEhood of the extracted candidates is predicted based on this feature representation, using a Support Vector Machine (SVM). The system reaches a relatively high accuracy of predicting MWEs on unseen data.

1.1 Previous Work

Attempts to extract MWEs are of different types. The most common techniques are primarily focused on collocations. Some of these techniques are rule-based and symbolic e.g., (Seretan, 2011; Goldman et al., 2001; Nerima et al., 2003; Baldwin, 2005; Piao et al., 2003; McCarthy et al., 2003; Jacquemin et al., 1997). Some rely on lexicons (Michiels and Dufour, 1998; Li et al., 2003) and (Pearce, 2001) that uses WordNet to evaluate the candidate MWE based on anti-collocations. Other approaches are hybrid in the sense that they benefit from both statistical and linguistic information. For instance (Seretan and Wehrli, 2006; Baldwin and Villavicencio, 2002; Piao and McEnery, 2001; Dias, 2003).

There are also fully statistical approaches. For instance (Pecina, 2010; Evert, 2005; Lapata and

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 10–16,
Gothenburg, Sweden, 26-27 April 2014. ©2014 Association for Computational Linguistics

Lascarides, 2003; Smadja et al., 1996), or the early work **Xtract** (Smadja, 1993).

Other approaches consider all types of MWEs (Zhang et al., 2006). Some of these approaches build upon generic properties of MWEs, for instance semantic non-compositionality (Van de Cruys and Moirón, 2007).

A different approach is presented in (Widdows and Dorow, 2005). The authors present a graph-based model to capture and assess fixed expressions in form of *Noun and/or Noun*.

There are also bilingual models which are mostly based on the assumption that a translation of the MWE in a source language exists in a target language. For instance (de Medeiros Caseli et al., 2010; Ren et al., 2009), and (Moirón and Tiedemann, 2006) which measures MWEs candidates' idiomaticity based on translational entropy. Another example is (Duan et al., 2009) which is a hybrid model that aims at extracting bilingual (English-Chinese) MWEs . It combines Multiple Sequence Alignment Model with some filtering based on hard rules to obtain an improved extraction.

A more generic model is presented in (Ramisch, 2012) where the author develops a flexible platform that can accept different types of criteria (from statistical to deep linguistic) in order to extract and filter MWEs. However, in this work, as the author claims, the quality of the extracted MWEs is highly dependent on the level of deep linguistic analysis, and thereby, the role of statistical criterion is less significant.

1.2 Motivation

We propose an original method to extract multiword expressions based on statistical contextual features, e.g., a set of immediate prefixes, suffixes, circumfixes, infixes to circumfixes, etc., (see Sec. 2). These features are used to form a feature representation, which together with a set of annotations train a supervised model in order to predict and extract MWEs from a large corpus.

We observed some discriminant behavior in contextual features (such as prefixes, suffixes, circumfixes, etc.) of a set of manually selected MWEs. A supervised model is then applied to learn MWEhood based on these features.

In general, modeling lexical and syntactic (and not semantic) characteristics of continuous MWEs is the focus of this paper. In order for the MWE decomposability condition to hold, we consider bigrams and above (up to size 4). Idiomaticity at some level is a necessary prerequisite of MWEs. Hereby, we consider idiomaticity at lexical, syntactic and statistical levels, and leave the semantic idiomaticity to the future work.

Relatively similar models have been previously applied to problems similar to MWEs, for instance named entity recognition (Nadeau and Sekine, 2007; Ratinov and Roth, 2009).

The focus on contextual features allows some degree of generalization, i.e., we can apply this model to a family of languages.[1] However, this work focuses only on English MWEs.

2 Proposed System

We prepared a corpus that comprises 100K Wikipedia documents for each of the mentioned languages.[1] After cleaning and segmenting the corpus, we extracted all possible n-grams (up to size 7) and their token and type frequencies. Then two basic statistical filters were applied in order to systematically decrease the size of our immense n-gram set: (i) *Frequency* filter, where we filter an n-gram if its frequency is less than the ratio between *tokens* and *types*, where for a given size of n-grams, the total number of n-grams and the number of distinct n-grams of that size, are considered *tokens* and *types*, respectively. (ii) *Redundancy* filter where we consider an n-gram to be redundant if it is subsumed by any other n'-gram, where $n' > n$. This gives us a pruned set of n-grams which we refer to as the *statistically significant* set. Table 1 presents a count-wise description of the filtering results on the English corpus.

	raw	frq flt	rdund flt
1-grams	1782993	64204	64204
2-grams	14573453	1117784	1085787
3-grams	38749315	3797456	3394414
4-grams	53023415	5409794	3850944
5-grams	53191941	2812650	2324912
6-grams	47249534	1384821	568645
7-grams	39991254	757606	757606

[1] We are adapting our model so that it can handle clusters of similar languages. So far we have processed the following 9 widely-spoken languages: English, German, Dutch, Spanish, French, Italian, Portuguese, Polish, and Russian. However, to study the efficiency of the presented model applied to languages other than English, remains a future work.

Table 1: Number of extracted n-grams for EN. First column indicates raw data, second and third columns indicate the number of n-grams after frequency and redundancy filters respectively.

For the set of significant n-grams a set of statistical features are extracted which will be described shortly. Fig. 1 illustrates the workflow of the system.

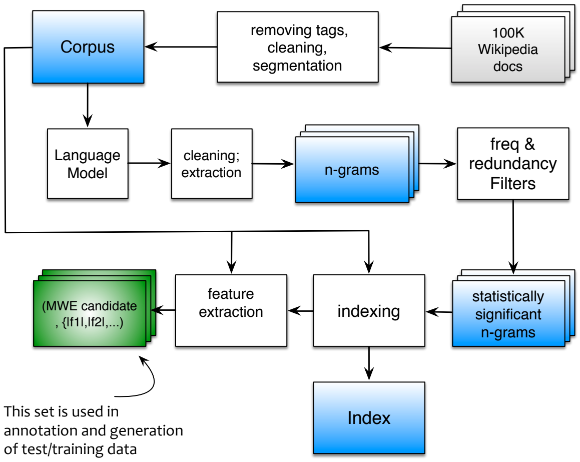
Corpus
removing tags, cleaning, segmentation
100K Wikipedia docs
Language Model
cleaning; extraction
n-grams
freq & redundancy Filters
(MWE candidate , {|f1|,|f2|,...)
feature extraction
indexing
statistically significant n-grams
Index
This set is used in annotation and generation of test/training data

Figure 1: Schematic of pre-processing, n-gram extraction and filtering. Blended and plain nodes represent resources, and operations respectively.

While studying the English corpus and different MWEs therein, it was observed that often, MWEs (as well as some other types of syntactic units) are followed, preceded or surrounded by a limited number of high frequency significant n-gram types. Moreover, our manual evaluation and constituency tests reveal that generally when a frequent significant prefix co-occurs with a frequent significant suffix, they form a circumfix whose significant infixes are (i) many, (ii) can mostly be considered syntactic unit, specifically when it comes to bi/trigrams. Table 2 illustrates a randomly selected sample of infixes of such circumfix (*the..of*). Remarkably, the majority of them are idiomatic at least at one level.

franz liszt academy	official list
most important albums	closest relatives
ministry of commerce	protestant church
executive vice president	peak period
famous italian architect	manhattan school
blessed virgin mary	rise and fall
world cup winner	former head

Table 2: Examples of bi/trigrams surrounded by the circumfix *the..of*

The immediate proximity of these particular context features to MWEs keeps emerging while evaluating similar circumfixes. We believe it suggests the presence of a discriminant attribute that we model with features 5-8 (see Table 3) and learn using a supervised model. Nevertheless, the fact that MWEs share these features with other types of syntactic units encourages introducing more MWE-specific features (namely, MWE's frequency, the frequency of its components, and their associations), then enforcing the learning model to recognize a MWE based on the combination of these two types of features. Note that the association between the type frequency of a MWE, and the frequency of its components is implicitly learned by the supervised model throughout the learning phase. A candidate MWE can be represented as:

$$\mathbf{y} = (x_1, ..., x_m, x_{m+1}, ..., x_n) \in \mathbb{N}_0 \qquad (1)$$

Where $x_1, ..., x_m$ are *contextual*, and $x_{m+1}, ..., x_n$ are *specific* features ($m = 8$, and $n = 11$). These features are described in Table 3.

	contextual features
x_1	# set of all possible prefixes of **y**
x_2	# set of distinct prefixes of **y**
x_3	# set of all possible suffixes of **y**
x_4	# set of distinct suffixes of **y**
x_5	# set of all possible circumfixes of **y**
x_6	# set of distinct circumfixes of **y** (**C**)
x_7	# set of all possible infixes to members of **C**
x_8	# set of distinct infixes to members of **C**
	specific features
x_9	the size of **y**
x_{10}	number of occurrences of **y** in the corpus
x_{11}	list of frequencies of the components of **y**

Table 3: Description of the extracted features

A prefix of $\mathbf{y}$ is the longest n-gram immediately before $\mathbf{y}$, if any or the boundary marker #, otherwise. A suffix of $\mathbf{y}$ is the longest n-gram immediately after $\mathbf{y}$, if any or the boundary marker #, otherwise. A circumfix ($c_i \in \mathbf{C}$) of $\mathbf{y}$ is the pair (p, s) where p and s are respectively the prefix and the suffix of a given occurrence of $\mathbf{y}$. An Infix of c_i is an n-gram that occurs between p and s.

Components to generate candidate MWEs, filter them and extract their relevant features were very memory and CPU intensive. To address the performance issues we implemented parallel programs and ran them on a high performance cluster.

3 Experimental Results

A set of $\approx$ 10K negative and positive English MWE examples were annotated. This set does not particularly belong in any specific genre, as the examples were chosen randomly from across a general-purpose corpus. This set comprises an equal number of positive and negative annotations. Part of it was annotated manually at UNDL foundation,[2] and part of it was acquired from the manually examined MWE lexicon presented in (Nerima et al., 2003). The set of positive and negative annotated n-grams is detailed in Table 4. The bias toward bigrams is due to the fact that the majority of manually verified MWEs that could be obtained are bigrams.

size	$+$ examples	$-$ examples
2-grams	$4,632$	$5,173$
3-grams	500	22
4-grams	68	15

Table 4: Annotations' statistics

This set was divided into $1/3$ test and $2/3$ training data, which were selected randomly but were evenly distributed with respect to positive and negative examples. The test set remains completely unseen to the model during the learning phase. We then train a linear SVM:

$$h(y) = \mathbf{w}^{\mathsf{T}}\,\mathbf{y} + b \qquad (2)$$

Where $h(y)$ is a discriminant hyperplane, $\mathbf{w}$ is the weight vector, and $\mathbf{y}$ is a set of MWE examples, where each example is defined as: $\mathbf{y}_j = x_1, ..., x_{11}$. Table 5 shows the results of the model's multiple runs on five different pairs of training and test sets.

	precision (%)	recall (%)	accuracy(%)
run 1	84.8	96.8	89.7
run 2	82.5	97.4	88.4
run 3	83.6	97.8	89.3
run 4	84.1	97.5	89.5
run 5	83.4	97.1	88.9

Table 5: Performance of the SVM which learns the MWEhood based on contextual and specific features ($x_1 - x_{11}$)

Table 6 illustrates the trained model's predictions on a set of randomly selected test examples. The overall performance of the model is shown in the form of a precision-recall curve in Fig. 2.

n-grams classified as MWE	
spend time	genetically modified
hijack a plane	fish tank
top dog	toy car
factory outlet	motorcycle racing
season nine	vintage car
video conference	chestnut tree
kill your	entry fee
safety precaution	quantum leap
version shown	make an appeal
flood damage	drug dealer
bargaining chip	lung transplant
grant her	tone like
postgraduate student	make a phone call
raise the price	ozone layer
n-grams classified as non-MWE	
score is	and dartmouth
the tabular	capped a
on sale	clarified his
liver was	the cancan
the regulating	an ending
the rabi	warns the
this manuscript	a few
an exponential	an institution
the petal	blades are
or ended	difficulties he
and workmen	the guidance
the eyelids	the examined
the vices	the episodes
they work	monument is

Table 6: Sample SVM's output on unseen data.

A t-test ranks the significance of the defined features in classifying n-grams into MWE, and non-MWE classes, as illustrated in Fig. 3. The most

[2] The Universal Networking Digital Language Foundation: http://www.undlfoundation.org/

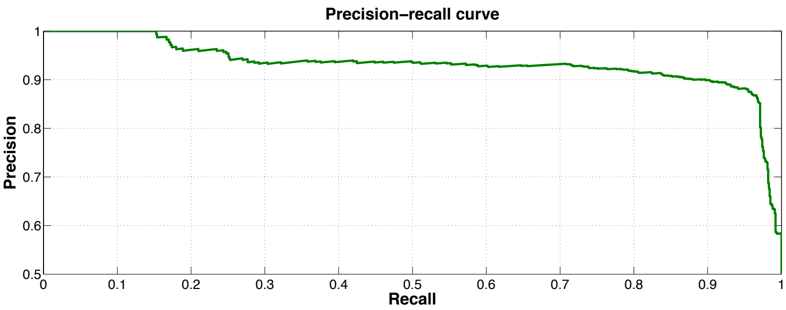

Figure 2: Precision-recall curve

important features are the size of examples (x_9), and the frequencies of their components (x_{11}). The significance of x_9 is due to the fact that in the training set majority of MWEs are bigrams. Therefore, by the SVM, being a bigram is considered as a substantial feature of MWEs. Nevertheless since the number of negative and positive examples which are bigrams are approximately the same, the bias toward x_9 in discriminating MWEs from non-MWE balances out. However its association with other features which is implicitly learned still has an impact on discriminating these two classes. x_7 and x_8 are the next two important features, as we expected. These two are the features whose magnitude suggests the presence or lack of contexts such as (*the..of*).

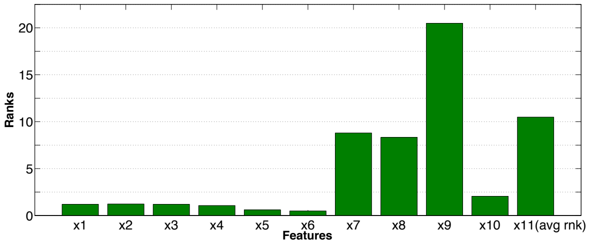

Figure 3: Ranks of the features that represent their discriminant impact.

The class separability of MWE (1), and non-MWE (-1) examples can be seen in Fig. 4, where the bidimentional projection of the examples of two classes is visualized. A star plot of a sample of 50 manually annotated examples is shown in Fig. 5. In many cases, but not always, non-MWEs can be discriminated from MWEs, in this eleven dimensional visualization. Same pattern was observed in the visualization of 500 examples (which would be hard to demonstrate in the present paper's scale).

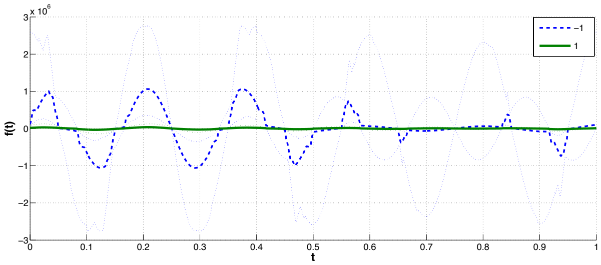

Figure 4: Andrews curve for the training examples. Bold line in the middle, and bold dotted line represent the median of MWE and non-MWE classes respectively.

4 Conclusions and Future Work

We presented a method to extract MWEs based on the immediate context they occur in, using a supervised model. Several contextual features were extracted from a large corpus. The size of the corpus had a profound effect on the effectiveness of these features. The presented MWE extraction model reaches a relatively high accuracy on an unseen test set. In future work, the efficiency of this approach on languages other than English will be studied. Furthermore, other features - specifically deep linguistic ones e.g., degree of constituency as described in (Ponvert et al., 2011) or POS tags, will be added to the feature representation of MWE candidates. Finally context-based probabilistic scores which are linguistically motivated can be investigated and compared with the supervised model. Another interesting work would be to introduce kernels so that we can go from statistics of contextual features to training the supervised model directly on the textual context.

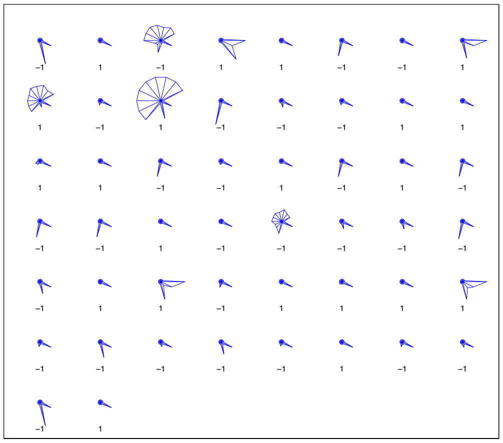

Figure 5: Star plot of 50 MWE (1), and non-MWE (−1) examples

References

Otavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. *Kordoni et al*, pages 101–109.

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

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

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

Nicoletta Calzolari, Charles J Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *LREC*.

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.

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.

Jianyong Duan, Mei Zhang, Lijing Tong, and Feng Guo. 2009. A hybrid approach to improve bilingual multiword expression extraction. In *Advances in Knowledge Discovery and Data Mining*, pages 541–547. Springer.

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

Jean-Philippe Goldman, Luka Nerima, and Eric Wehrli. 2001. Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocations*, pages 61–66.

David A Hull and Gregory Grefenstette. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57. ACM.

Ray Jackendoff. 1997. *The architecture of the language faculty*. Number 28. MIT Press.

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.

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.

Wei Li, Xiuhong Zhang, Cheng Niu, Yuankai Jiang, and Rohini Srihari. 2003. An expert lexicon approach to identifying english phrasal verbs. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 73–80. Association for Computational Linguistics.

Archibald Michiels and Nicolas Dufour. 1998. Defi, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In *Proceedings of the first international conference on language resources & evaluation*, pages 1179–1186.

Begona Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-wordexpressions in a multilingual context*, pages 33–40.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Luka Nerima, Violeta Seretan, and Eric Wehrli. 2003. Creating a multilingual collocation dictionary from large text corpora. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 131–134. Association for Computational Linguistics.

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.

Scott Songlin Piao and Tony McEnery. 2001. Multiword unit alignment in english-chinese parallel corpora. In *the Proceedings of the Corpus Linguistics 2001*, pages 466–475.

Scott SL Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 49–56. Association for Computational Linguistics.

Elias Ponvert, Jason Baldridge, and Katrin Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *ACL*, pages 1077–1086.

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.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

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, Kathleen R McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177.

Tim Van de Cruys and Begona Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 25–32. Association for Computational Linguistics.

Dominic Widdows and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 48–56. Association for Computational Linguistics.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44. Association for Computational Linguistics.