# Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking

Alfredo Maldonado, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel, Qun Liu

# Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking

**Alfredo Maldonado**[1], **Lifeng Han**[2], **Erwan Moreau**[1],
**Ashjan Alsulaimani**[1], **Koel Dutta Chowdhury**[2], **Carl Vogel**[1] **and Qun Liu**[2]

ADAPT Centre
[1]Trinity College Dublin, Ireland
[2]Dublin City University, Ireland
{firstname.lastname}@adaptcentre.ie

## Abstract

A description of a system for identifying Verbal Multi-Word Expressions (VMWEs) in running text is presented. The system mainly exploits universal syntactic dependency features through a Conditional Random Fields (CRF) sequence model. The system competed in the Closed Track at the PARSEME VMWE Shared Task 2017, ranking 2nd place in most languages on full VMWE-based evaluation and 1st in three languages on token-based evaluation. In addition, this paper presents an option to re-rank the 10 best CRF-predicted sequences via semantic vectors, boosting its scores above other systems in the competition. We also show that all systems in the competition would struggle to beat a simple lookup baseline system and argue for a more purpose-specific evaluation scheme.

## 1 Introduction

The automatic identification of Multi-Word Expressions (MWEs) or collocations has long been recognised as an important but challenging task in Natural Language Processing (NLP) (Sinclair, 1991; Sag et al., 2001). An effort in response to this challenge is the Shared Task on detecting multi-word, verbal constructions (Savary et al., 2017) organised by the PARSing and Multiword Expressions (PARSEME) European COST Action[1]. The Shared Task consisted of two tracks: a closed one, restricted to the data provided by the organisers, and an open track that permitted participants to employ additional external data.

The ADAPT team participated in the Closed Track with a system[2] that exploits syntactic dependency features in a Conditional Random Fields (CRF) sequence model (Lafferty et al., 2001), ranking 2nd place in the detection of full MWEs in most languages[3]. To the best of our knowledge, this is the first time that a CRF model is applied to the identification of verbal MWEs (VMWEs) in a large collection of distant languages.

In addition to our CRF-based solution officially submitted to the closed track, our team also explored an option to re-rank the top 10 sequences predicted by the CRF decoder using a regression model trained on word co-occurrence semantic vectors computed from Europarl. This semantic re-ranking step would qualify for the open track, however its results were not submitted to the official competition as we were unable to obtain its results in time for it.

This paper describes our official CRF-based solution (Sec. 3), as well as our unofficial Semantic Re-Ranker (Sec. 4). Since the Shared Task's main goal is to enable a discussion of the challenges of identifying VMWEs across languages, this paper also offers some observations (Sec. 5). In particular, we found that test files contain VMWEs that also occur in the training files, helping all systems in the competition, but also implying that a simple lookup system that only predicts MWEs it encountered in the training set will fare very well in the competition, and will in fact beat most systems. We also argue for a more purpose-based evaluation scheme. And we offer our conclusions and ideas for future work (Sec. 6).

## 2 Related Work

MWEs have long been discussed in NLP research and a myriad of identification techniques

[1]http://www.parseme.eu

[2]System details, feature templates, code and experiment instructions: https://github.com/alfredomg/ADAPT-MWE17

[3]Official results: http://bit.ly/2krOu05

have been developed, such as combining statistical and symbolic methods (Sag et al., 2001), single and multi-prototype word embeddings (Salehi et al., 2015), integrating MWE identification within larger NLP tasks such as parsing (Green et al., 2011; Green et al., 2013; Constant et al., 2012) and machine translation (Tsvetkov and Wintner, 2010; Salehi et al., ; Salehi et al., 2014).

More directly related to our closed-track approach are works such as that of Venkatapathy and Joshi (2006), who showed that information about the degree of compositionality of MWEs helps the word alignment of verbs, and of Boukobza and Rappoport (2009) who used sentence surface features based on the canonical form of VMWEs. In addition, Sun et al. (2013) applied a Hidden Semi-CRF model to capture latent semantics from Chinese microblogging posts; Hosseini et al. (2016) used double-chained CRF for minimal semantic units detection in SemEval task. And Bar et al. (2014) discussed that syntactic construction classes are helpful for verb-noun and verb-particle MWE identification. Schneider et al. (2014) also used a sequence tagger to annotate MWEs, including VMWEs, while Blunsom and Baldwin (2006) and Vincze et al. (2011) have used CRF taggers for identifying contiguous MWEs.

In relation to our open-track approach, Attia et al. (2010) exploited large corpora to identify Arabic MWEs, and Legrand and Collobert (2016) applied fixed-size continuous vector representations for various length of phrases and chunks in the MWE identification task. Constant et al. (2012) used a re-ranker for MWEs in an $n$-best parser.

## 3 Official Closed Track: CRF Labelling

We decided to model the problem of VMWE identification as a sequence labelling and classification problem. We operationalise our solution through CRFs (Lafferty et al., 2001), which encode relationships between observations in a sequence. We implemented our solution using the CRF++[4] system. CRFs have been successfully applied to such sequence-sensitive NLP tasks as segmentation, named-entity recognition (Han et al., 2013; Han et al., 2015) and part-of-speech tagging. Our team attempted 15 out of the 18 languages involved in the Shared Task. The data for the languages we did not attempt (Bulgarian, Hebrew and Lithuanian) lacked morpho-syntactic information, so we felt that we were unlikely to obtain good results with them. It should be noted that of these 15 languages, four (Czech, Farsi, Maltese and Romanian) were provided without syntactic dependency information, although morphological information (i.e. tokens' lemmas and parts of speech (POS)) was indeed supplied.

### 3.1 Features

We assume that features based on the relationships between the different types of morpho-syntactic information provided by the organisers will help identify VMWEs. Ideally, one feature set (or *feature template* in the terminology of CRF++) per language should be developed. Due to time constraints, we instead developed a feature set for a single language per broad language family (German, French and Polish), assuming that, for our purposes, morpho-syntactic relationships will behave similarly among closely related languages, but not among distant languages.

For each token in the corpora, the direct linguistic features available are its word surface (W), word lemma (L) and POS (P). In the languages where syntactic dependency information is provided, each token also has its head's word surface (HW), its head's word lemma (HL), its head's POS (HP) and the dependency relation between the token and its head (DR). It is possible to create CRF++ feature templates that combine these features in unigrams, bigrams, etc. In addition, it is also possible to combine the predicted output label of the previous token with the output label of the current token (B). We conducted preliminary 5-fold cross validation experiments on German, French and Polish training data independently, using feature templates based on different combinations of these features in unigram, bigram and trigram fashions. Templates exploiting token word surface features (W) performed unsurprisingly worse than those based on token lemmas and POS (L, P). Templates using head features (HL, HP, DR) in addition to token features (L, P) fared better than those relying on token features only. The three final templates developed can be summarised[5] as follows:

- FS3: B, L-2, L-1, L, L+1, L+2, L-2/L-1, L-1/L, L/L+1, L+1/L2, P, HL/DR, P/DR, HP/DR.
- FS4: FS3, P-2, P-1, P, P+1, P+2, P-1/P, P/P+1.
- FS5: FS4, L/HP.

Each template summary above consists of a name (FS3, FS4 or FS5) and a list of feature

[4] https://taku910.github.io/crfpp/

[5] Actual templates are on GitHub. See footnote 2.

abbreviations indicating a position relative to the current token and feature conditioning is indicated by a slash. After developing these templates through preliminary experimentation, a further 5-fold cross validation experiment on training data was conducted using each template against each of the 15 languages. For each language, the best performing template (regardless of the language family for which it was developed) was chosen for the final challenge, in which the CRF++ system was trained using that selected template on the full training data for the language, and the prediction output was generated from the blind test set provided. FS3 was chosen for Greek, Spanish, French, Slovenian and Turkish, whilst FS4 was chosen for Swedish and FS5 for the rest of the languages.

### 3.2 Offical Evaluation

Table 1 shows, under "crf", the F1 scores for each of the VMWE categories in the competition: ID (low-compositional verbal idiomatic expressions), IReflV (reflexive verbs), LVC (light verb constructions), VPC (verb-particle constructions) and OTH (a miscellaneous category for any other language-specific VMWE). The Overall score is also included. The column $n$ shows the count of MWEs in the test set for each category. Scores for which $n = 0$ are omitted as they are undefined. Sections 4 and 5 explain the "sem" and "PS" columns, respectively. On token-based evaluation, our system was ranked in first place in Polish, French and Swedish, second place in eight languages and third in three. For MWE-based scores, our system ranked second place on nine languages.

## 4 Unofficial Open Track: Semantic Re-Ranking

We implemented an optional post-processing stage intended to improve the performance of our CRF-based method using a distributional semantics approach (Schütze, 1998; Maldonado and Emms, 2011). Intuitively, the goal is to assess the likeliness of a given candidate MWE, and then, based on such features for all the candidate MWEs in a sentence, to select the most likely predicted sequence among a set of 10 potential sequences.

This part of the system receives the output produced by CRF++ in the form of the 10 most likely predictions for every sentence. For every such set of 10 predicted sequences, context vectors are computed for each candidate MWE, using a large third-party corpus. A set of features based on these context vectors is computed for each predicted sequence. These features are then fed to a supervised regression algorithm, which predicts a score for every predicted sequence; the one with the highest score among the set of 10 is the final answer.

### 4.1 Third-Party Corpus: Europarl

We use Europarl (Koehn, 2005) as third-party corpus, because it is large and contains most languages addressed in this Shared Task. It does not contain Farsi, Maltese and Turkish, which are therefore excluded from this part of the process. For each of the 12 remaining languages, we use only the monolingual Europarl corpus, and we tokenise it using the generic tokeniser provided by the organisers.[6]

### 4.2 Features

An instance is generated for every predicted sequence. For every candidate MWE in the sequence, we calculate context vectors (i.e. we count the words co-occurring with the MWE[7] in Europarl), and we compute three kinds of features: (1) Features comparing each pseudo-MWE consisting of a single word of the MWE against the full MWE; (2) Features comparing each pseudo-MWE consisting of the MWE minus one word against the full MWE; (3) Features comparing one of the other MWEs found in the 10 predicted sequences against the current MWE. For each category of features, the relative frequency and the similarity score obtained between the context vectors of the pseudo-MWEs and the full MWE are added as features, as well as the number of words (we implemented four kinds of similiarity measures: Jaccard index, Min/Max similarity, Cosine similarity with or without IDF weights).

The main difficulty in representing a predicted sequence as a fixed set of features is that each sentence can contain any number of MWEs, and each MWE can contain any number of words. We opted for "summarising" any non-fixed number of features with three statistics: minimum, mean and maximum. For instance, the similarity scores

[6] Discrepancies are to be expected between the tokenisation of the Shared Task corpus (language-specific) and the one performed on Europarl (generic).

[7] There are multiple ways to define the context window for a possibly discontinuous MWE. Here we simply aggregate the 4-words contexts (two words on the left, two on the right) of the words inside the MWE.

Table 1: F1 scores (per category and overall) on the test set for our official CRF-based ("crf") and our unofficial Semantic Re-Ranking ("sem") systems, with per category and overall MWE counts ("n") in the test set. PS refers to the MWEs in the test set that were *Previously Seen* in the training set: the % of Previously Seen MWEs and the F1 Score obtained by interpreting % as a Recall score and assuming a 100% Precision score.

| Lang | Eval | ID | | | IReflV | | | LVC | | | OTH | | | VPC | | | Overall | | | PS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *n* | crf | sem | *n* | crf | sem | *n* | crf | sem | *n* | crf | sem | *n* | crf | sem | *n* | crf | sem | % | F1 |
| **CS** | MWE | *192* | 5.48 | 5.65 | *1149* | 59.48 | 67.36 | *343* | 8.36 | 10.17 | *0* | | | *0* | | | *1683* | 57.72 | 65.20 | 92.26 | 95.97 |
| | Token | | 10.72 | 10.85 | | 74.49 | 75.76 | | 14.52 | 15.13 | | | | | | | | 72.86 | 74.55 | | |
| **DE** | MWE | *214* | 14.68 | 15.95 | *20* | 0.71 | 0.74 | *40* | 3.30 | 4.14 | *0* | | | *226* | 18.81 | 23.95 | *500* | 22.80 | 26.93 | 39.96 | 57.10 |
| | Token | | 28.92 | 26.61 | | 4.81 | 4.50 | | 8.48 | 8.73 | | | | | 33.61 | 35.37 | | 40.48 | 40.41 | | |
| **EL** | MWE | *127* | 12.45 | 13.62 | *0* | | | *336* | 27.28 | 32.86 | *21* | 0.91 | 0.88 | *16* | 2.30 | 2.24 | *500* | 31.34 | 36.73 | 34.20 | 50.97 |
| | Token | | 19.11 | 19.57 | | | | | 38.18 | 40.15 | | 3.97 | 3.67 | | 3.30 | 2.82 | | 43.14 | 45.33 | | |
| **ES** | MWE | *166* | 13.75 | 14.60 | *223* | 42.13 | 45.09 | *109* | 18.27 | 17.89 | *11* | 0.00 | 1.18 | *3* | 0.00 | 0.00 | *500* | 44.33 | 48.61 | 52.20 | 68.59 |
| | Token | | 21.99 | 22.45 | | 43.44 | 46.06 | | 24.04 | 22.20 | | 0.00 | 1.16 | | 0.00 | 0.00 | | 49.17 | 52.64 | | |
| **FA** | MWE | *0* | | | *0* | | | *0* | | | *500* | 80.08 | | *0* | | | *500* | 80.08 | | 98.80 | 99.40 |
| | Token | | | | | | | | | | | 85.36 | | | | | | 85.36 | | | |
| **FR** | MWE | *119* | 35.59 | 35.39 | *105* | 37.12 | 40.00 | *271* | 15.38 | 20.93 | *5* | 0.00 | 0.00 | *0* | | | *500* | 50.88 | 56.24 | 28.00 | 43.75 |
| | Token | *316* | 44.78 | 42.79 | *210* | 40.90 | 40.56 | *577* | 23.07 | 25.19 | *5* | 0.00 | 0.00 | *0* | | | *1108* | 61.52 | 62.68 | | |
| **HU** | MWE | *0* | | | *0* | | | *146* | 15.16 | 15.88 | *0* | | | *354* | 68.89 | 69.29 | *499* | 66.89 | 67.92 | 79.76 | 88.74 |
| | Token | | | | | | | | 24.84 | 26.23 | | | | | 65.69 | 66.45 | | 66.10 | 67.85 | | |
| **IT** | MWE | *250* | 19.18 | 19.77 | *150* | 17.36 | 13.11 | *87* | 9.90 | 8.84 | *2* | 0.00 | 0.00 | *11* | 4.76 | 3.81 | *500* | 23.09 | 20.20 | 37.00 | 54.01 |
| | Token | | 22.33 | 22.40 | | 16.12 | 12.34 | | 11.39 | 9.39 | | 0.00 | 0.00 | | 3.97 | 3.15 | | 25.11 | 21.93 | | |
| **MT** | MWE | *185* | 8.63 | | *0* | | | *259* | 3.98 | | *56* | 0.00 | | *0* | | | *500* | 6.41 | | 47.20 | 64.13 |
| | Token | | 10.76 | | | | | | 5.57 | | | 1.57 | | | | | | 8.87 | | | |
| **PL** | MWE | *66* | 8.41 | 8.24 | *265* | 64.21 | 67.88 | *169* | 26.31 | 28.72 | *0* | | | *0* | | | *500* | 67.95 | 72.40 | 66.80 | 80.10 |
| | Token | | 13.17 | 12.73 | | 67.90 | 68.63 | | 30.27 | 30.80 | | | | | | | | 72.74 | 74.34 | | |
| **PT** | MWE | *90* | 19.41 | 20.04 | *81* | 18.15 | 19.60 | *329* | 46.24 | 52.67 | *0* | | | *0* | | | *500* | 58.14 | 64.64 | 59.40 | 74.53 |
| | Token | | 28.52 | 27.80 | | 19.68 | 19.76 | | 57.08 | 56.83 | | | | | | | | 70.18 | 71.01 | | |
| **RO** | MWE | *75* | 17.15 | 18.05 | *290* | 51.11 | 57.74 | *135* | 37.83 | 37.79 | *0* | | | *0* | | | *500* | 73.38 | 79.26 | 87.80 | 93.50 |
| | Token | | 23.51 | 23.57 | | 57.96 | 59.90 | | 41.02 | 39.46 | | | | | | | | 81.90 | 83.41 | | |
| **SL** | MWE | *92* | 2.67 | 3.65 | *253* | 40.00 | 44.77 | *45* | 1.22 | 1.19 | *2* | 0.00 | 0.00 | *108* | 15.90 | 16.50 | *500* | 37.08 | 41.41 | 41.60 | 58.76 |
| | Token | | 5.94 | 7.77 | | 49.90 | 49.62 | | 4.30 | 3.97 | | 0.39 | 0.36 | | 21.31 | 20.20 | | 45.06 | 46.35 | | |
| **SV** | MWE | *51* | 6.33 | 6.33 | *14* | 1.65 | 1.65 | *14* | 6.61 | 6.61 | *2* | 0.00 | 0.00 | *155* | 32.06 | 32.82 | *236* | 30.32 | 30.90 | 5.51 | 10.44 |
| | Token | | 8.00 | 8.00 | | 3.27 | 3.27 | | 6.48 | 6.48 | | 0.00 | 0.00 | | 33.40 | 34.16 | | 31.49 | 32.04 | | |
| **TR** | MWE | *249* | 25.86 | | *0* | | | *199* | 27.55 | | *53* | 9.60 | | *0* | | | *501* | 42.83 | | 58.88 | 74.12 |
| | Token | | 33.18 | | | | | | 35.31 | | | 12.00 | | | | | | 52.85 | | | |

between each individual word and the MWE ($n$ scores) are represented with these three statistics computed over this set of scores. Finally, the probability of the predicted sequence (given by CRF++) is included as a feature. In training mode, the instance is assigned score 1 if it corresponds exactly to the sequence in the gold standard, or 0 otherwise. It might happen that none of the 10 sequences corresponds to the gold sequence: in such cases all the instances are left as negative cases.

### 4.3 Regression and Sequence Selection

We use the Weka (Hall et al., 2009) implementation of Decision Trees regression (Quinlan, 1992) to train a model which assigns a score in $[0, 1]$ to every instance. Among each group of 10, the predicted sequence with the highest score is selected. We use regression rather than classification because a categorical answer would cause problems in cases where there is either no positive or multiple positive answers for a set of predicted sequences.

### 4.4 Evaluation

F1 scores on the test set for the Semantic Re-Ranking of CRF outputs can be seen in Table 1 under the "sem" heading. As can be seen, in nearly every language the Semantic Re-Ranking improves the CRF best prediction considerably. These promising results are obtained with the first "proof of concept" version of the Semantic Re-Ranking component, that we plan to develop further in future work.

## 5 Discussion

The "%" column under "PS" (henceforth PS%), in Table 1, shows the proportion of MWE instances found in the test set that occurred at least once in the training set, i.e. they are "Previously Seen" MWEs. It is reasonable to expect that most systems would benefit from having a large number of previously seen MWEs in the test set. Our systems tend to perform well when PS% is high (e.g. Farsi, Romanian) and poorly when PS% is low (e.g. Swedish), although not in all cases. In fact, this is a trend observed in the other competing systems: the Pearson correlation coefficient be-

Table 2: Number of languages each system ranked at. Systems in grey italics are open systems, the rest are closed. PS and sem are unofficial systems.

| Rnk | PS | TRA | *sem* | MUM | SZE | crf | RAC | *LAT* | LIF |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 13 | 1 | *1* | | | | | | |
| **2** | | 11 | *3* | | 1 | | | | |
| **3** | | 2 | *5* | | 1 | 5 | 2 | | |
| **4** | | | *3* | 1 | 2 | 5 | 3 | | |
| **5** | | 1 | | 5 | 1 | 3 | 3 | *1* | |
| **6** | 2 | | | 3 | 3 | 2 | 2 | | |
| **7** | | | | 2 | | | 2 | | 1 |

tween PS% and all official systems' scores is 0.63. It would indeed be interesting to re-run the competition using a test set that featured MWEs not present in the training set.

PS could be potentially regarded as a baseline system that simply attempts to find matches of training MWEs in the test set. Such a simple lookup system, which could compete in the Closed Track, would achieve very high scores in several languages. In fact, it would beat all other systems in the competition in most languages. PS% can be interpreted as its Recall score. Since such a lookup system is incapable of "predicting" MWEs it has not seen, we assume it would always achieve a 100% Precision score, allowing us to compute an F1 score, presented in the "F1" column in Table 1, for the baseline PS system. Table 2 shows the number of languages in which each system would rank at each position if we include PS and our unofficial Semantic Re-Ranker scores. Only the 15 languages we attempted are counted. PS would always rank first except only in French and Swedish, the two languages with the lowest proportion of previously seen MWEs. One might contest PS's 100% Precision assumption as it depends on the accuracy of the actual VMWE matching method used. However, under this assumption PSF1 measures the best performing lookup method possible. This reasoning feeds into the simple matching method used: VMWEs are extracted from training and test set files according to their gold standard. PS% is their intersection divided by the total number of test set VMWEs. A VMWE is deemed to be present in both portions if its extracted dependency structure (if provided), lemmas and POS tags are identical in both files. For languages without dependencies, MWEs are matched based on lemmas and POS linear sequences only.

Interesting questions about the Shared Task's F1-based evaluation can also be raised. F1 considers Precision and Recall to be equally important, when in reality their relative importance depends on the purpose of an actual VMWE identification exercise. In a human-mediated lexicographic exercise, for example, where coverage is more important than avoiding false positives, Recall will take precedence. Conversely, in a computer-assisted language learning application concerned with obtaining a small but illustrative list of VMWE examples, Precision will take priority. We suggest that for future iterations of the Shared Task, a few candidate applications be identified and subtasks be organised around them. The identification task's purpose will also inform on the appropriateness of including previously seen MWEs in the test set. In a lexicographic or terminological task, there is usually an interest in identifying *new*, *unseen* MWEs as opposed to *known* ones, whereas in Machine Translation, the impact of known MWEs in new, unseen sentences is of interest.

## 6 Conclusions and Future Work

In this paper, we described our VMWE identification systems based on CRF and Semantic Re-Ranking, achieving competitive results. We analysed the role of previously seen MWEs and showed that they help all systems in the competition, including a hypothetical, simple lookup system that would beat all systems in most languages. We also argued for a more purpose-based evaluation scheme. Our future work will focus on language-specific features, rather than on language families. We also intend to explore tree-based CRF methods to better exploit syntactic dependency tree structures. The promising first results obtained with the Semantic Re-Ranker deserve to be explored further. Aspects such as parameter tuning, feature selection and other semantic vector types, like word embeddings (Legrand and Collobert, 2016), might help improve the performance. Finally, we want to explore alternative evaluation methods based on lexicographic and terminological tasks (Maldonado and Lewis, 2016) on the one hand and Machine Translation tasks (Xiong et al., 2016) on the other.

## Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

## References


Mohammed Attia, Lamia Tounsi, Pavel Pecina, Josef van Genabith, and Antonio Toral. 2010. Automatic extraction of Arabic multiword expressions. In *Proceedings of the COLING workshop on multiword expressions: from theory to applications (MWE 2010), Beijing*.

Kfir Bar, Mona Diab, and Abdelati Hawwari, 2014. *Arabic Multiword Expressions*, pages 64–81. Springer, Berlin.

Phil Blunsom and Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 164–171, Sydney.

Ram Boukobza and Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 468–477, Singapore.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 204–212, Jeju.

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 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations*, 11(1):10–18.

Aaron L-F Han, Derek F Wong, and Lidia S Chao. 2013. Chinese named entity recognition with conditional random fields in the light of chinese characteristics. In *Language Processing and Intelligent Information Systems*, pages 57–68. Springer.

Aaron Li-Feng Han, Xiaodong Zeng, Derek F Wong, and Lidia S Chao. 2015. Chinese named entity recognition with graph-based semi-supervised learning model. *Eighth SIGHAN Workshop on Chinese Language Processing*, page 15.

Mohammad Javad Hosseini, Noah A. Smith, and Su-In Lee. 2016. UW-CSE at semeval-2016 task 10: Detecting multiword expressions and supersenses using double-chained conditional random fields. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 931–936.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Joël Legrand and Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, Berlin.

Alfredo Maldonado and Martin Emms. 2011. Measuring the compositionality of collocations via word co-occurrence vectors: Shared task system description. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 48–53, Portland, OR.

Alfredo Maldonado and David Lewis. 2016. Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In *Proceedings of The 12th International Conference on Terminology and Knowledge Engineering*, pages 91–100, Copenhagen.

J.R. Quinlan. 1992. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, pages 343–348.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

Bahar Salehi, Paul Cook, and Timothy Baldwin. Detecting Non-compositional MWE Components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pages 977–983.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, Valencia.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Xiao Sun, Chengcheng Li, Chenyi Tang, and Fuji Ren. 2013. Mining Semantic Orientation of Multiword Expression from Chinese Microblogging with Discriminative Latent Model. In *Proceedings of 2013 International Conference on Asian Language Processing*, pages 117–120, Urumqi.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1256–1264, Beijing.

Sriram Venkatapathy and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27, Sydney.

Veronika Vincze, István Nagy, and Gábor Berend. 2011. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 289–295, Hissar.

Deyi Xiong, Fandong Meng, and Qun Liu. 2016. Topic-based term translation models for statistical machine translation. *Artificial Intelligence*, 232:54–75.