A fresh look at modelling and evaluation of multiword expression tokens

Shiva Taslimipoor
University of Wolverhampton

Omid Rohanian
University of Wolverhampton

Ruslan Mitkov
University of Wolverhampton

Afsaneh Fazly
Thomson Reuters

Automatic identification of Multiword Expressions (MWEs) in running text has recently received much attention among researchers in computational linguistics. The wide range of reported results for the task in the literature has prompted us to take a closer look at the algorithms and evaluation methods. For supervised classification of Verb+Noun expressions, we propose a context-based methodology in which we find word embeddings to be appropriate features. We discuss the importance of train and test splitting in validating the results and present type-aware train and test splitting. Given our specialised data, we also discuss the benefits of framing the task as classification rather than tagging.

1 Introduction

Ambiguity is a pervasive phenomenon in natural language. It starts from single words, and can propagate through larger linguistic constructs. Multiword Expressions (MWEs) which are idiosyncratic combinations of two or more words, behave differently in their separate usages in running text. In natural language processing (NLP) tasks such as part-of-speech tagging, parsing and machine translation, these expressions should be treated either before the task (Nivre & Nilsson

\lsCollectionPaperCitationText.

2004) or combined with the process (Constant & Tellier 2012; Kordoni, Ramisch & Villavicencio 2011; Nasr et al. 2015).

Examples of such expressions are: *take action*, *make sense* and *set fire*[1]. MWEs are a recurring theme in any language with some sources estimating their number to be in the same range as single words (Jackendoff 1997) or even beyond (Sag et al. 2002). Besides, new expressions come to languages on a regular basis. It is therefore not practically feasible to simply list MWEs in dictionaries or thesauri.

More importantly, most idiomatic expressions can also have literal meaning depending on context. For instance consider the expression *play games*. It is opaque with regards to its status as an MWE and depending on context could mean different things. For example in *He went to play games online* it has a literal sense but is idiomatic in *Don't play games with me as I want an honest answer*. Resolving these cases is critically important in many NLP applications (Katz 2006). Katz (2006) has framed the task as sense disambiguation. Tagging corpora for MWEs or token-based identification of MWEs is an example of a task where it is necessary to differentiate between idiomatic and literal usages of each expression type.

Studies on MWEs can be divided into two main categories. One includes works regarding the canonical forms of expressions, their lexical properties and their potential to be considered as MWEs, namely type-based extraction of MWEs or MWE discovery; the other regards studies on tagging texts for the idiomatic usages of expressions, namely MWE tagging or token-based identification of MWEs. The former is a traditional approach which is of use to lexicographers (as pointed in Ramisch (2014)); the latter though, is more practical for NLP applications (Schneider et al. 2016).

Although discovering canonical forms of multiword expressions is still an active research area (Salehi & Cook 2013; Farahmand & Martins 2014), recently there is a significant move towards automatic tagging of corpora for MWEs (Schneider et al. 2014; Constant & Tellier 2012).

The focus of our study is token-based identification of MWEs, and we model it as a classification, rather than a sequence labelling problem. To determine the idiomaticity of each Verb+Noun occurrence, we experiment with using solely context features without any sophisticated linguistic information. We do not exploit parsing, tagging or external lexicon-based information.

For token-based identification of MWEs, there is a wide range of results in the literature reported as the state-of-the art: from F-score of 64% with the DiMSUM

[1] MWEs combine words from many different parts of speech. The pattern in our datasets is Verb+Noun, so all the examples in this chapter are of this kind.

dataset (Schneider et al. 2016) to 90% (Al Saied, Constant & Candito 2017) for a dataset in the last Parseme shared task (Savary et al. 2017). We find that in order for the performance results not to be misleadingly high, the distribution of the tokens between train and test should be controlled. Failure to do so will result in a kind of overfitting which could be overlooked during evaluation. For instance, an expression like *take advantage* is idiomatic consistently in all its usages in text. When different occurrences of this expression exist in both train and test, the model memorises it from training data and predicts it very well in the test. Since such expressions are highly frequent, this memorisation helps the model to achieve erroneously high performance scores.

In the process of supervised identification of MWEs, we make observations with regards to the following: (1) the effect of train and test splitting of the tokens on generalisability of a model; (2) comparison between sequence labelling (tagging) and sequence classification.

1.1 Literature review

Identification of MWEs has been shown to be effective in different NLP tasks, such as machine translation (Pal, Chakraborty & Bandyopadhyay 2011) and automatic parsing (Constant, Sigogne & Watrin 2012). There exists a considerable body of work in the literature attempting to investigate lexical and syntactic properties of expressions to account for their potential for being MWEs (Ramisch 2014; Baldwin & Kim 2010). However, recently there has been a great attention given to identifying where exactly this potential takes effect by tagging a running text for each individual occurrence (token) of an expression (Schneider et al. 2014; Constant & Tellier 2012; Gharbieh, Bhavsar & Cook 2017). Token-based identification of MWEs is effective in disambiguating between different behaviours of expressions in their individual usages. Evaluating all occurrences of expressions in the whole corpus of big size is not feasible. For this reason we have gathered a specialised dataset of concordances of particular expressions.

To the best of our knowledge, there are very few comprehensive tagged corpora for MWEs available, among which DiMSUM by Schneider et al. (2016) is very recent and well-cited. This corpus was used in the SemEval (2016) shared task in Detecting Minimal Semantic Units and their Meanings (DiMSUM). It is not particularly clear if the current methodologies applied to this corpus are capable of disambiguating between different usages of one specific canonical form.

Cook, Fazly & Stevenson (2008) have prepared a dataset of English Verb+Noun constructions, categorising expressions based on their idiomaticity and how consistent they behave in their different usages. Fazly, Cook & Stevenson (2009) have used that dataset for classifying Verb+Noun tokens into idiomatic or literal categories.

Katz (2006) has used context features for identifying the idiomaticity/non-compositionality of MWEs in a different way. They represent different occurrences of an expression using LSA vectors and show that the vectors of the expressions in their idiomatic sense are very different from those of the same expressions in literal sense. Based on this observation they classify a test expression token depending on whether it is more similar to the idiomatic sense of the expression in training data or to the literal sense.

Scholivet & Ramisch (2017) recently have tried to disambiguate a number of opaque French expressions using their contexts. They have proposed a tagging approach using unigram and bigram features of the word forms and their POS. Qu et al. (2015) have found word embedding representation of the words in context very useful for tagging a text with MWEs. We have also used word vector representations of the verb and noun components of the expression and the words in a window size of two on the right of the expression as features for classifying expressions as MWE or not.

While most of the previous work on token-based identification of MWEs have applied sequence tagging approaches using some kind of IOB labeling, Legrand & Collobert (2016) have looked at the problem as classification. They have proposed a neural network based approach that learns fixed-size representations for arbitrary sized chunks which is able to classify these representations as MWE or not. They have shown better performance in MWE identification over the CRF-based approach in Constant, Candito & Seddah (2013).

1.2 Summary of contributions

In almost all of the previous work on supervised modelling of MWE tokens, data is randomly split into train and test sets. However, we have observed that by doing so the test data tends to overlap with the training data. In a random splitting, it is possible for occurrences of the same expression type to occur in both train and test sets. There are many instances where the expression almost always behaves idiomatically (e.g. *take part*, *make progress*) or literally (e.g. *eat food*, *give money*). In such cases a model learns every feature related to the POS and lemma form of the expression, and naturally can predict the correct tag for the expression perfectly in the test set (regardless of the expression being idiomatic or literal).

Having observed this issue, for evaluation we propose and perform type-aware train and test splitting. To this end we divide expression types into train and test folds and gather all occurrences of each type into the same fold. This makes the predication rigorous, since the model performs cross-type learning. One interesting study that considered cross-type learning of MWEs is the one by Fothergill & Baldwin (2012). However, they have not clearly explained the general advantages and effects of cross-type classification in evaluation. They have used the approach in order to learn better features from specialised MWE resources.

We propose type-aware splitting of the data as a supplementary benchmark for evaluating MWE identification. We design experiments to show the effectiveness of this kind of evaluation in assessing the generalisability of models.

The direction of all recent studies on token-based identification of MWEs is towards using structured sequence tagging models. The choice of the model based on the data is an important issue. Our data includes occurrences of specific Verb+Noun expressions with the context around them. This makes it possible to have a huge data annotated for a specific type of MWE in order to have a thorough evaluation. We observe that our data cannot benefit from sequence tagging and a regular classification approach can more reasonably model the data. We show better results from classification over a tagging model. Other than traditional machine learning classification approaches, we also propose a neural-network model by combining convolutional neural network and long short term memory models for identifying MWEs. Although some deep learning models have already been investigated for tagging MWEs by Gharbieh et al. (2017), to the best of our knowledge this is the first time this approach has been applied for classifying MWE instances.

We extensively discuss the following: 1) the division of data into train and test sets for evaluation and 2) the choice of model (classification versus tagging) based on the data.

2 Context-based identification of MWEs

In this study we use context features in a supervised environment to identify the idiomaticity of Verb+Noun expression tokens. In order to construct context features, for our first set of experiments (4.1), given each occurrence of a Verb+Noun combination, we concatenate four different word vectors corresponding to the verb, noun, and their two following adjacent words while preserving the original order (following the previous work by Taslimipoor et al. 2017). Concatenated word vectors are fed into different classification models to be evaluated in terms of their performance.

The classification algorithms that have been used are Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Multi Layer Perceptron (MLP) and Support Vector Machine (SVM). We have also experimented with different neural network-based classification models. The best result is achieved with a combination of bidirectional Long Short-Term Memory network with a convolutional layer as a front-end (ConvNet+LSTM).

For the second set of experiments (4.2), in which we compare Conditional Random Field (CRF) as a tagger with a simple Naive Bayes Classifier (NBC), we consider simple word forms of the verb, the noun, and the two words after as lexical context features.

We have conducted our extensive experiments with Italian. The experiments are augmented by applying the approach also for smaller data in Spanish and English.

3 Experiments

3.1 Data

We first experiment with two similarly formatted datasets in Italian and Spanish and later also on a standard available dataset for English.

For Italian, our data includes a huge set of concordances of Verb+Noun expressions. [2] Each item in the dataset is one concordance of a Verb+Noun expression and the whole item is annotated with 1 if the Verb+Noun inside is an MWE and with 0 otherwise. The data as explained in Taslimipoor et al. (2016) has been annotated by two native speakers with Kappa agreement measure of 0.65. We have resolved the disagreements by employing a third annotator who has decided on most (but not all) cases of disagreements. This results in $20,030$ concordances of $1,564$ expression types. The Italian data is very imbalanced and almost 70% of the data is marked as MWE. To resolve this issue, we ignore the 15 most frequent expression types which are exclusively marked as MWE and also the expressions with frequency lower than 3. As a result we run the experiments on $18,540$ concordances of 940 expression types.

For Spanish, we extracted concordances of Verb+Noun expressions in the same way using SketchEngine (Kilgarriff et al. 2004).[3] After ignoring the concordances for five most frequent expressions, $3,918$ usages have been marked by two native

[2] The data as described in Taslimipoor et al. (2016) has been gathered for four light verbs *fare*, *dare*, *trovare* and *prendre*. For some examples of expression instances refer to the same work.

[3] For Spanish, we focused on four light verbs *tener*, *hacer*, *dar* and *tomar*, similar to the ones we use for Italian.

speakers. The Kappa inter-annotator agreement was 0.55. Having seen the observed agreement of 0.79, we ignored all cases of disagreements and considered only the concordances on which both annotators agreed. This results in $3,090$ concordances of 747 expression types.

For English, we employ a standard dataset called VNC-tokens prepared by Cook, Fazly & Stevenson (2008).[4] The dataset is a benchmark for English verb-noun idiomatic expressions and has been used for identifying MWE tokens in a number of previous studies such as Fazly, Cook & Stevenson (2009) and Salton, Ross & Kelleher (2016). The dataset includes sentences from the BNC corpus including occurrences of Verb+Noun expressions and is suitable for our task since it contains expressions with both skewed and balanced behaviour in being literal or idiomatic. Rather than concordances, it includes sentences from BNC containing occurrences of Verb+Noun expressions. Two English native speakers have selected the expression types based on whether they have the potential for occurring in both idiomatic or literal senses. Although this dataset is slightly different from our Italian and Spanish data (which are extracted randomly), it has the same favourable pattern of different occurrences of same expression types that can be split into train and test. We find it interesting to investigate our observations on a differently gathered but standard dataset. The Verb+Nouns in this dataset are not necessarily continuous. We ignore the cases where the Verb+Noun occurs in passive form and the ones that the annotators were unsure of and this results in $2,499$ sentences consisting of Verb+Noun expressions. The statistics of the data for all three languages are reported in Table 1.

Table 1: Distribution of the data

	Italian	Spanish	English
Expression types	940	747	53
Expression tokens	18,540	3,090	2,499
MWE tokens	10,804 (58.27%)	2,094 (66.57%)	1,981 (79.27%)

For all the three datasets, we consider the same context words as features for classification: we extract the vectors of the verb, noun and the two words after the noun.

[4] The dataset is available in https://sourceforge.net/projects/multiword/files/MWE_resources/20110627/

3.2 Evaluation

In all cases classifier performance has been measured using 10-fold cross-validation.

3.2.1 Standard splitting of data into train and test

In the standard method of performing cross-validation, the whole data is randomly divided into k folds and then the model is repeatedly trained on the data of $k-1$ folds and tested on the data of the remaining fold. The result is averaged among k different iterations. In our task, we find the result of this evaluation misleading, since the repetition of the same expression in both train and test partitions helps the model to predict those specific types of expressions well, while the model might not work for new unseen expressions in test. Even stratified cross-validation suffers from the same kind of overfitting. In standard stratified cross-validation, imbalance is coped with by controlling the distribution of labels alone, so that all folds have the same distribution of labels. Similar to standard cross-validation, this method is not informed about the idiosyncratic distribution of types and tokens.

Therefore, these methods of evaluation cannot precisely reflect the effectiveness of the model or features and show better results for models that are more prone to overfitting. It is not particularly clear from this kind of evaluation if a good performing model could be generalised to unseen expressions and also to ambiguous expressions that have balanced distribution of occurrences as literal or idiomatic. We show the performance computed using this type of evaluation for different classifiers in Table 2.

3.2.2 Type-aware splitting and evaluation

We perform a custom cross-validation by splitting the expression occurrences into different folds considering their types/canonical forms. We split the expression types into k groups and all the occurrences of the expressions in the k^{th} group goes into the k^{th} fold. This method ensures that the model performs cross-type learning and generalises to tokens from unseen types in the test fold. In other words, the model is learning the features and general patterns and does not overfit on highly recurrent token occurrences. The results for all classifiers evaluated using this approach is reported in Table 3.

4 Results

In this section, first a comparison of several classifiers using different train and test splitting methods is reported; then we present experiments using sequence tagging for identifying MWEs; and finally, the effectiveness of neural network-based word embeddings compared with count-based representations has been analysed using one of the best classifiers.

4.1 Regular and type-aware evaluation

Evaluation performance for all classifiers using two different kinds of train and test splitting, namely regular (random) and our proposed type-aware, are reported in Tables 2 and 3. The columns of the tables represent the results for Italian (it), Spanish (es) and English (en). All traditional classifiers in this experiment use the same vectorised context features. The word vectors used in this study are available online.[5] The generated Italian and Spanish word embeddings have applied Gensim's skipgram word2vec model with the window size of 10 to extract vectors of size 300. For English we use word embeddings of the same dimension trained using Glove (Pennington, Socher & Manning 2014) algorithm available via spaCy. [6]

We also report the results from a more sophisticated neural network based architecture comprising of a BiLSTM with an additional convolutional layer as a front-end (ConvNet+LSTM). For this architecture the context window size is 2 (two words before and two words after the Verb+Noun expression). [7] Implementation details of these experiments can be found at https://github.com/shivaat/VN-tokens-clf.

Different classifiers show high performance with not much difference using regular cross-validation in which tokens are distributed into separate folds regardless of their types (Table 2). ConvNet+LSTM, in particular, performs the best, which we believe is the result of overfitting arising from this method of train and test splitting. However, we can see notable differences between classifiers in Table 3 where we cross validate in a way that no same expression type occurs in both train and test partitions.

In the case of cross-type learning (Table3), the SVM classifier has shown the

[5] http://hlt.isti.cnr.it/wordembeddings/ for Italian and https://github.com/Kyubyong/wordvectors for Spanish

[6] https://spacy.io/docs/usage/word-vectors-similarities

[7] The difference in results were negligible when considering only the two context words on the right.

Table 2: Regular evaluation results: accuracy (standard deviation)

Classifiers	it	es	en
Majority Baseline	0.5827	0.6657	0.7927
LR	0.8869 (0.007)	0.9129 (0.011)	0.8651 (0.020)
DT	0.8905 (0.008)	0.9065 (0.017)	0.8799 (0.018)
RF	0.9218 (0.005)	0.9337 (0.019)	0.9024 (0.017)
MLP	0.9069 (0.006)	0.933 (0.009)	0.9056 (0.016)
SVM	0.9116 (0.005)	0.9207 (0.009)	0.7927 (0.021)
ConvNet+LSTM	**0.9220 (0.007)**	**0.9668 (0.01)**	**0.8860 (0.024)**

Table 3: Type-aware evaluation results: accuracy (standard deviation)

Classifiers	it	es	en
Majority Baseline	0.5827	0.6657	0.7927
LR	0.6909 (0.06)	0.8178 (0.074)	0.8092 (0.149)
DT	0.6048 (0.03)	0.7483 (0.078)	0.6327 (0.128)
RF	0.6337 (0.08)	0.7604 (0.097)	0.7321 (0.19)
MLP	0.7053 (0.06)	0.8319 (0.086)	0.7294 (0.169)
SVM	**0.7369 (0.07)**	0.8460 (0.093)	0.8062 (0.152)
ConvNet+LSTM	0.6601 (0.053)	**0.8681 (0.072)**	**0.8112 (0.106)**

best results in identifying MWEs using vectorised context features for Italian, and close to the second best for Spanish and English data for which ConvNet+LSTM is the best. The performance of this classifier is followed by that of MLP and LR for both Italian and Spanish. For English the results of SVM and LR are comparable. Computed performance for other classifiers like DT and RF have dropped sharply when we use our type-aware cross-validation. This is also the case for ConvNet+LSTM for Italian data. This experiment determines how well a classifier can generalise among different expression types. SVM and LR in particular are shown to be fairly suitable for cross-type identification of MWEs. MLP also performs relatively well overall.

As for the English data it is worth noting that the VNC data is very imbalanced with the majority baseline of 0.7927 which is difficult to beat by classifiers.

4.2 Sequence classification versus sequence tagging

The experimental data in this study can be perfectly processed with standard classification approach, since the goal is to predict idiomaticity of an expression in a given context. However, Scholivet & Ramisch (2017) have modelled such a data with sequence tagging. We believe that since not all the words in a sequences are going to be tagged, MWE identification using such a data cannot benefit from sequence labelling. We have applied sequence tagging on the data to properly investigate the effects. Specifically, simple Naive Bayes Classifier (NBC) has been considered as a simple sequence classification methodology and Conditional Random Field (CRF) has been used as the sequence tagging approach. Both of the models use simple nominal features: the verb, the noun and the two words after the noun. The results are reported in Table 4 in terms of accuracy.

Table 4: Performance of sequence classification versus sequence tagging

	regular cross-validation			type-aware cross-validation		
	it	es	en	it	es	en
NBC	0.9504	0.9601	0.8560	0.7291	0.7298	0.6013
CRF	0.9165	0.9142	0.8176	0.6447	0.7199	0.6848

As can be seen in Table 4, CRF cannot even beat the simple naive bayes classifier except in the case of English data (when we apply cross-type learning). This is because our data is naturally suited for sequence classification and cannot benefit from sequence labelling models.

4.3 Effectiveness of word embedding representation

To specifically show the effectiveness of neural network-based embeddings for the classifiers to identify Verb+Noun MWEs, we have performed an experiment using sparse bag-of-words count vectors with tf-idf weighting. Similar to previous experiments, we feed the vectors to a Multi Layer Perceptron (MLP) which works reasonably well compared to other classifiers based on the previous experiment. Note that the execution time for the best performing model, SVM, is almost 5 times that of MLP which makes it inefficient in comparison. The results of this comparison can be seen at Table 5.

The results in Table 5 show the improvement in performance when using word embeddings rather than the vanilla count-based vectors for all three languages

Table 5: The accuracy of MLP in identifying Verb+Noun MWEs using word2vec and count-based embedding

	Accuracy (std.)		
	it	es	en
MLP with count based embedding	0.6504 (0.0354)	0.7851 (0.042)	0.7002 (0.099)
MLP with word2vec	0.7053 (0.06)	0.8319 (0.086)	0.7294 (0.169)

(although less significant for English).

5 Discussion

In order to understand the argument behind type-aware evaluation and decide its applicability, we have to look at the distribution of data points. In the Italian data, for instance, the majority of data points belong to MWE types whose tokens occur invariably as idiomatic or literal only. In other words, if we plot the distribution of tokens with regards to the degree of idiomaticity of their corresponding types, we would see a skewed distribution (even after ignoring the 15th most frequent expressions), where only a smaller portion of tokens belong to MWE types whose usages can be fluid between literal and idiomatic. In such a scenario, a model easily overfits on the majority of the data, where labels have been assigned invariably. However, this skewedness is not necessarily reflected in the distribution of MWE labels, as we might see a relatively balanced distribution of literal and idiomatic labels. This means there might be no severe class imbalance in the dataset, but within-class imbalance (Ali, Shamsuddin & Ralescu 2015).

To illustrate the point, we operationalise two categories for MWE types, namely Consistent (C) and Fluid (F). Those types whose tokens occur more than 70% or less than 30% of the time as only literal or idiomatic are tagged as C, and the rest are considered F. Accordingly, Figure 1 shows the distribution of the expression types with regards to the behavior of their corresponding tokens. As can be seen, the majority of expressions with higher token frequencies are from the sub-class C. For this reason, evaluation using a vanilla cross validation or even stratified cross validation would not provide us with reliable results, since splitting of train and test disregards the within-class imbalance inherent in the data.

Since this is the case with data in real world, we propose type-aware train and test splitting as a supplementary approach for modelling the data and eval-

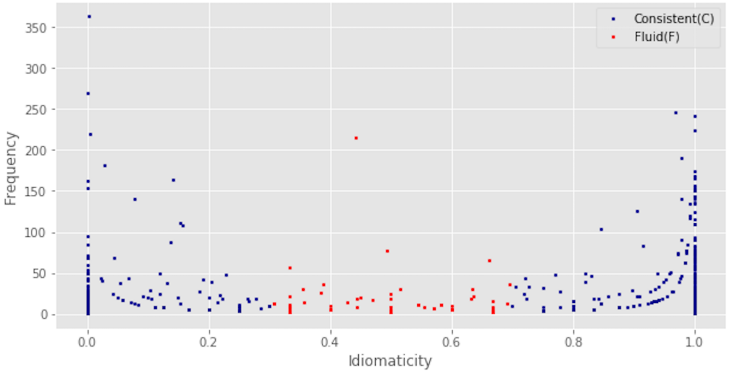

Figure 1: Distribution of expression types.

uating the results. This way, we make sure that a model has the best ability for generalisation, learns general properties for MWEs and is not merely based on memorising the words that construct MWEs.

It is worth noting that we have not used any linguistic or lexical features and we expect vector representation of context to be generalisable enough. Even with these generalisable features we observe substantial differences between regular and type-aware cross-validation. A proper method for train and test splitting is even more essential to validate the evaluation when a model trains on more exact features such as lexical ones.

With regards to previous data and models for MWEs, DiMSUM is one of the most noteworthy shared tasks. DiMSUM includes a recent tagged corpus for MWEs with a fairly small size of $4,799$ sentences in train and $1,000$ in test, including all types of MWEs. With such limited data, we have observed only a few number of expressions of the form Verb+Noun occurring in both train and test. To give an example, with a selection of 6 most frequent light verbs, all their combinations with nouns are only 13 occurrences in the test data, out of which only 3 are MWEs. There are no repeated occurrences of these cases in both train and test data. Therefore, we believe that this data inherently does not lead to misleading results. In other words, a model that works well on this data could be fairly generalised.

Gharbieh, Bhavsar & Cook (2017) have shown better performance when using deep neural network models compared with traditional machine learning on DiMSUM. However, in our experiment of type-aware classification, SVM has performed the best, even outperforming LSTM and ConvNet and their combinations for Italian and Spanish. Since neither of DiMSUM or our data is big enough for a proper analysis with deep learning, more studies are required to find the most effective model to identify MWEs.

Another data for token-based identification of MWEs in English that we have also used in this study is VNC-tokens (Cook, Fazly & Stevenson 2008). One advantage of this corpus is that the data is particularly gathered for the task of disambiguation between idiomatic and literal usages of expressions. Before the annotation, they have selected only the expressions that have the potential for occurring in both idiomatic and literal senses. Although for this study we have not followed the initial development splitting of the data (i.e. we followed our proposed way of splitting the data into train and test), the development and test splitting of this data is type-aware. Therefore, an experiment with this data, is able to truly measure generalisation.

In Parseme shared task (Savary et al. 2017), which features the most recent multi-lingual data for MWEs, Maldonado et al. (2017) have presented statistics on the percentage of previously seen data in test sets of all languages (i.e. proportion of MWE instances in the test set that have been seen also in the training set). The correlation between these percentages and the results stress the need for proper train and test splitting. The experiments with the data for the Parseme shared task would definitely benefit from such type-aware train and test splitting.

6 Conclusions

In this study, we have explored a context-based classification method for identification of Verb+Noun expressions. We have employed word embedding to represent context features for MWEs. We have evaluated the methodology using type-aware cross-validation and discussed its effectiveness compared with standard evaluation. We argue that only this proposed method properly accounts for the generalisability of a model. We have also shown that our data (and similar ones) for this task cannot benefit from structured sequence tagging models.

The effectiveness of word embeddings as context features for identifying MWEs should be examined in more detail with datasets of larger size and with more sophisticated embeddings that consider linguistic features. We would also like to analyse the effect of our proposed approach on unseen and less frequent data.

Acknowledgements

The authors would like to thank Manuela Cherchi, Anna Desantis, Martina Cotella, Andrea Silvestre and Alondra Nava Zea for their help in annotations.

Abbreviations

Full form	Abbreviation
Bi-directional Long Short-term Memory	BiLSTM
Conditional Random Field	CRF
Convolutional Neural Network	ConvNet
Decision Tree	DT
Inside-outside-beginning	IOB
Latent Semantic Analysis	LSA
Logistic Regression	LR
Long Short-term Memory	LSTM
Multi Layer Perceptron	MLP
Multiword Expressions	MWEs
Naive Bayes Classifier	NBC
Natural Language Processing	NLP
Part of Speech	POS
Random Forest	RF
Support Vector Machine	SVM

References

Al Saied, Hazem, Matthieu Constant & Marie Candito. 2017. The ATILF-LLF system for Parseme Shared Task: A transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 127–132. Association for Computational Linguistics. http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1717.pdf. DOI:10.18653/v1/W17-1717

Ali, Aida, Siti Mariyam Shamsuddin & Anca L. Ralescu. 2015. Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and its Application* 7(3). 687–719.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing, Second edition*, 267–292. Boca Raton: CRC Press.

Constant, Matthieu, Marie Candito & Djamé Seddah. 2013. The LIGM-Alpage Architecture for the SPMRL 2013 Shared Task: Multiword expression analysis and Dependency Parsing. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages*, 46–52.

Constant, Matthieu, Anthony Sigogne & 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*, 204–212. Association for Computational Linguistics.

Constant, Matthieu & Isabelle Tellier. 2012. Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), 646–650. European Language Resources Association (ELRA).

Cook, Paul, Afsaneh Fazly & Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions* (MWE '08), 19–22. Association for Computational Linguistics.

Farahmand, Meghdad & 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 '14), 10–16. Association for Computational Linguistics.

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103. http://aclweb.org/anthology/J09-1005.

Fothergill, Richard & Timothy Baldwin. 2012. Combining resources for MWE-token classification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation* (SemEval '12), 100–104. Association for Computational Linguistics.

Gharbieh, Waseem, Virendra Bhavsar & Paul Cook. 2017. Deep learning models for multiword expression identification. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, 54–64. Association for Computational Linguistics.

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge, MA.: MIT Press.

Katz, Graham. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (MWE '06), 12–19. Association for Computational Linguistics.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. In Geoffrey Williams & Sandra Vessier (eds.), *Proceedings of the 11th EURALEX International Congress*, 105–116. Lorient, France: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Kordoni, Valia, Carlos Ramisch & Aline Villavicencio. 2011. *Proceedings of the ACL Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE '11). Association for Computational Linguistics.

Legrand, Joël & Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions* (MWE '16), 67–71. Association for Computational Linguistics. http://anthology.aclweb.org/W16-1810.

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 114–120. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1715. DOI:10.18653/v1/W17-1715

Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1116–1126. Association for Computational Linguistics. http://www.aclweb.org/anthology/P15-1108.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications* (MEMURA 2004), 39–46. http://stp.lingfil.uu.se/~nivre/docs/mwu.pdf.

Pal, Santanu, Tanmoy Chakraborty & Sivaji Bandyopadhyay. 2011. Handling multiword expressions in phrase-based Statistical Machine Translation. In *Proceedings of the 13th Machine Translation Summit* (MT Summit 2011), 215–224. September 19-23, 2011.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: global vectors for word representation. In *Proceedings of the Conference on*

Empirical Methods in Natural Language Processing (EMNLP 2014), 1532–1543. October 25–29, 2014. http://www.aclweb.org/anthology/D14-1162.

Qu, Lizhen, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, Nathan Schneider & Timothy Baldwin. 2015. Big Data Small Data, In Domain Out-of Domain, Known Word Unknown Word: the impact of word representations on sequence labelling tasks. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning* (CoNLL 2015), 83–93.

Ramisch, Carlos. 2014. *Multiword expressions acquisition: A generic and open framework*. Vol. XIV (Theory and Applications of Natural Language Processing). Springer. http://link.springer.com/book/10.1007%2F978-3-319-09207-2.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. London, UK: Springer-Verlag.

Salehi, Bahar & Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics*, vol. 1 (* SEM 2013), 266–275. June 13-14, 2013.

Salton, Giancarlo, Robert Ross & John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 194–204. Berlin, Germany: Association for Computational Linguistics.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1704.pdf. DOI:10.18653/v1/W17-1704

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* 2(1). 193–206. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/281.

Schneider, Nathan, Dirk Hovy, Anders Johannsen & Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), 546–559. Association for Computational Linguistics. http://www.aclweb.org/anthology/S16-1084.

Scholivet, Manon & Carlos Ramisch. 2017. Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 167–175. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1723.

Taslimipoor, Shiva, Anna Desantis, Manuela Cherchi, Ruslan Mitkov & Johanna Monti. 2016. Language resources for Italian: Towards the development of a corpus of annotated Italian multiword expressions. In *Proceedings of 3rd Italian Conference on Computational Linguistics* (CLiC-it 2016) *& 5fth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (EVALITA 2016) (Collana dell'Associazione Italiana di Linguistica Computazionale), 5–6 December 2016. online. Torino: Accademia University Press.

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2017. Investigating the opacity of verb-noun multiword expression usages in context. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 133–138. Association for Computational Linguistics.