Investigating the Opacity of Verb-Noun Multiword Expression Usages in Context

Shiva Taslimipoor[†], Omid Rohanian[†], Ruslan Mitkov[†] and Afsaneh Fazly[‡]

[†]Research Group in Computational Linguistics, University of Wolverhampton, UK
[‡]VerticalScope Inc., Toronto, Canada
{shiva.taslimi, m.rohanian, r.mitkov}@wlv.ac.uk
afsaneh.fazly@gmail.com

Abstract

This study investigates the supervised token-based identification of Multiword Expressions (MWEs). This is an ongoing research to exploit the information contained in the contexts in which different instances of an expression could occur. This information is used to investigate the question of whether an expression is literal or MWE. Lexical and syntactic context features derived from vector representations are shown to be more effective over traditional statistical measures to identify tokens of MWEs.

1 Introduction

Multiword expressions (MWEs) belong to a class of phraseological phenomena that is ubiquitous in the study of language (Baldwin and Kim, 2010). Scholarly research in MWEs immensely benefit both NLP applications and end users (Granger and Meunier, 2008). Context of an expression has been shown to be discriminative in determining whether a particular token is idiomatic or literal (Fazly et al., 2009; Tu and Roth, 2011). However, in-context investigation of MWEs is an under-explored area.

The most common approach to treat MWEs computationally in any language is by examining corpora using statistical measures (Evert and Krenn, 2005; Ramisch et al., 2010; Villavicencio, 2005). These measures are broadly applied to identifying the types [1] of MWEs. While there is ongoing research to improve the type-based investigation of MWEs (Rondon et al., 2015; Farahmand and Martins, 2014; Salehi and Cook, 2013), the challenge of token-based identification of MWEs (as in tagging corpora for these expressions) requires more attention (Schneider et al., 2014; Brooke et al., 2014; Monti et al., 2015).

In this study, we focus on a specific variety of MWEs, namely Verb + Noun combinations. This type of MWEs doesn't always correspond to fixed expressions and this leads to computational challenges that make identification difficult (e.g. while *take place* is a fixed expression, *makes sense* is not and can be altered to *makes perfect sense*). The word components in such cases may or may not be inflected and the meaning of the components may or may not be exposed to the meaning of the whole expression. This paper outlines investigation of MWEs of the class Verb + Noun in Italian. Examples of these cases in Italian are *fare uso* 'to make use', *dare vita* 'to create' or *fare paura* 'to frighten'.

We propose a supervised approach that utilises the context of the occurrences of expressions in order to determine whether they are MWEs. Having the whole corpus tagged for our purpose of training a classifier would be a labour-intensive task. A more feasible approach would be to use a special-purpose data, labeled with concordances containing Verb + Noun combinations. We report the preliminary results on the effectiveness of context features extracted from this special-purpose language resource for identification of MWEs.

We differentiate between expressions whose instances occur with a single fixed idiomatic or literal behaviour and the ones that show degrees of ambiguity with regards to potential usages. We partition the dataset in a way to account for both of these groups and the experiments are run separately for each.

To extract context features, we use a word embedding approach (word2vec) (Mikolov et al., 2013) as the state of the art in the study of dis-

[1] Type refers to the canonical form of an expression, while token refers to each instance (usage) of the expression in any morphological form in text.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 133–138,
Valencia, Spain, April 4.

tributional similarity. We extract features from the raw corpus without any pre-processing. While we report the results for Italian, the approach is language-independent and can be used for any resource-poor language.

2 Motivation

It is important to consider expressions at the token level when deciding if they are MWEs. The reason being, there are expressions that in some cases occur with an idiomatic sense whereas with a literal sense in others. This could be determined by the context in which they appear. For example take 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*. A traditional classification model that is blind to linguistic context proves to be insufficient in such cases. The following is an example of the same phenomenon in Italian which is the language of interest in this study:

1) Per migliorare il sistema dei trasporti, si dovrebbero **creare ponti** anche verso e da le isole minori.

   'In order to improve the transportation system, the government should **build bridges** both to and from the smaller islands.'

2) Affinch possiamo migliorare la convivenza fra popoli diversi, bisognerebbe **creare ponti**, non sollevare nuovi muri!

   'In order to improve coexistence among different people, we should **build bridges** not raise new walls!'

3 Related Work

With regards to context-based identification of idiomatic expressions, Birke and Sakar (2006) use a slightly modified version of an existing word sense disambiguation algorithm for supervised token-based identification of MWEs. Katz and Giesbrecht (2006) rely primarily on the local context of a token without considering linguistic properties of expressions. Fazly et al. (2009) take into account both linguistic properties and local context in their analysis of MWE tokens. They have employed and evaluated an unsupervised approach on a small sample of human annotated expressions. Their method uses grammatical knowledge about the canonical form of expressions.

There is some recent interest in segmenting texts (Brooke et al., 2014; Schneider et al., 2014) based on MWEs. Brook et al. (2014) propose an unsupervised approach for identifying the types of MWEs and tagging all the token occurrences of identified expressions as MWEs. This methodology might be more useful in the case of longer idiomatic expressions that is the focus of that study. Nevertheless for expressions with fewer words, the aforementioned challenges regarding opacity of tokens limit the efficacy of such techniques. The supervised approach posited by Schneider et al. (2014) results in a corpus of automatically annotated MWEs. However, the literal/idiomatic usages of expressions have not been dealt with in particular in their work.

The idea behind our work is to use concordances of all the occurrences of a Verb + Noun expression in order to decide the degree of idiomaticity of a specific Verb + Noun expression. Our work is very related to the work of Tu and Roth (2011), in that they have also particularly considered the problem of in-context analysis of light verb construction (as a specific type of MWEs) using both statistical and contextual features. Their approach is also supervised, but it requires parsed data from English. Their contextual features include POS tags of the words in context as well as information from Levin's classes of verb components. Our approach requires little pre-processing and is best suited for languages that lack ample tagged resources. The present study is in the same vein as the approach taken by Gharibeh et al. (2016). Here, we have specifically analysed expressions that have more ambiguous usages, running separate experiments on partitions of the dataset.

4 Methodology

Our goal is to classify tokens of Verb + Noun expressions into literal and idiomatic categories. To this end, we exploit the information contained in the concordance of each occurrence of an expression. Given each concordance, we extract vector representations for several of its words to act as syntactic and lexical features. Compared to literal Verb + Noun combinations, idiomatic combinations are expected to appear in more restricted lexical and syntactic forms (Fazly et al., 2009). One

traditional approach in quantifying lexical restrictions is to use statistical measures. (Ramisch et al., 2010).

We target syntactic features by extracting vectors for the verb and the noun contained in the expression. Here we extract the vectors of the verb and the noun components in their raw form hoping to indirectly learn lexical and syntactic features for each occurrence of an expression. We believe that the structure of the verb component is important in extracting fixedness information for an expression. Also, the distributional representation of the noun component is informative since Verb + Noun expressions are known to have some degrees of semi-productivity (Stevenson et al., 2004).

Additionally, we extract vectors for co-occurring words around a target expression. Specifically, we focus on the two words immediately following the Verb + Noun expression. We expect the arguments of the verb and the noun components that occur following the expression to play a distinguishing role in these kinds of so-called complex predicates[2] (Samek-Lodovici, 2003).

The word vectors in this study come from the Italian word2vec embedding which is available online[3]. The generated word embedding approach has applied Gensim's skipgram word2vec model with the window size of 10 to extract vectors of size 300 for Italian words from Wikipedia corpus.

In order to construct our context features, 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. In other words, given each expression, the context feature consists of a combined vector with the dimension of 4 * 300 = 1200.

Concatenated feature vectors are fed into a logistic regression classifier. The details with regards to training the classifier are explained in Section 6.

5 Experiments

5.1 Experimental Data

The data used in this study is taken from an Italian language resource for Verb + Noun expressions (Taslimipoor et al., 2016). The resource focuses on four most frequent Italian verbs: *fare*, *dare*, *prendere* and *trovare*. It includes all the concordances of these verbs when followed by any noun, taken from the itWaC corpus (Baroni and Kilgarriff, 2006) using SketchEngine (Kilgarriff et al., 2004).

The concordances include windows of ten words before and after an expression; hence, there are contexts around each Verb + Noun expression to be used for the classification task[4]. $30,094$ concordances are annotated by two native speakers and can be used as the gold-standard for this research. The Kappa measure of inter-annotator agreement between the two annotators on the whole list of concordances is 0.65 with the observed agreement of 0.85 (Taslimipoor et al., 2016). Since the agreement is substantial, we continue with the first annotator's annotated data for evaluation.

5.2 Partitioning the Dataset

The idea is to evaluate the effect of context features to identify the literal/idiomatic usages of expressions, particularly for the type of expressions that are likely to occur in both senses. In our specialised data, around 32% of expression types have been annotated in both idiomatic and literal form in different contexts. For this purpose, we divide the data into two groups:

(1) Expressions with a skewed division of the two senses (e.g., with more than 70% of instances having either a literal or idiomatic sense).[5]

(2) Expressions with a more balanced division of instances (e.g., with less than or equal to 70% of instances having either a literal or idiomatic sense).

We develop different baselines to evaluate our approach on these two groups as explained in the following section.

5.3 Baseline

5.3.1 Majority baseline

We devise a very informed and supervised baseline based on the idiomatic/literal usages of ex-

[2]Most of the Verb + Noun expressions that we investigate belong to the category of complex predicates which is the focus of Samek-Lodovici (Samek-Lodovici, 2003)

[3]http://hlt.isti.cnr.it/wordembeddings/

[4]Cases where components of a potential MWE occur with in-between gaps (intervening words) are not considered.

[5]Expressions such as *dare inizio* 'to start' and *trovare cose* 'to find things' which most of the times occur as MWE and non-MWE respectively.

pressions in the gold-standard data. According to this baseline a target instance vn_{ins}, of a test expression type vn, gets the label that it has received in the majority of vn occurrences in the gold-standard set. The baseline approach labels all instances of an expression with a fixed label (1 for MWE and 0 for non-MWE). This is a high precision model when working with Group 1, due to the more consistent behaviour of instances there. However, its results are suitable for evaluating the results of our developed model over expressions of Group 2.

5.3.2 Association measures as a baseline

The data in Group 1 include the expressions that mostly occur in either idiomatic or literal forms. These expressions are commonly categorised as being MWE or non-MWE using association measures. Association measures are computed by statistical analysis through the whole corpus, hence the values are the same for all instances of an expression. In other words, these methods are blind to the contexts in which different instances of an expression could occur.

To evaluate our model over data in Group 1, these association measures are used as features to develop a baseline. We focus on two widely used association measures, log-likelihood and Salience as defined in SketchEngine. We also use frequency of occurrence as a statistical measure to rank MWEs. The statistical measures are computed using SketchEngine on the whole of itWac. The statistical measures are then given to an SVM classifier to identify MWEs.

6 Evaluation

6.1 Evaluation Setup

There are $1,480$ types of expressions with $28,483$ occurrences in Group 1 and 169 types of expressions with $1,611$ occurrences in Group 2. For each group, we extract context features to train logistic regression classifiers.

Our proposed context features are vector representations of the raw form of the verb component, the raw form of the noun component and a window of two words after the target expression. We refer to the combination of these vectors as the `Context` feature. We apply a 5-fold cross validation approach to compute accuracies for each classifier. We split the dataset into five separate folds so that no instance of the same expression could occur in more than one fold. This is to make sure that the test data is blind enough to the training data. The classifiers are compared against the baselines using different features. The results are reported in Tables 1 and 2.

6.2 Results and Analyses

Table 2 shows the results of our model over data in Group 2 compared to the majority baseline. Recall that the data instances in Group 2 are highly unpredictable in their occurrence as MWE or non-MWE. We expect that our supervised model using context features (`Context`) be able to disambiguate between different instances of an expression. Here, our model performs slightly better than the informed majority baseline.

Table 1: Classification accuracies (%) using different features over Group 1 and the whole data.

Features	all data	Group 1
Freq	70.77	69.20
Likelihood	72.11	70.64
Salience	73.83	72.81
Likelihood+Salience+Freq	73.90	73.29
Context (word2vec)	**75.42**	74.13
Salience + Context	**78.40**	**80.13**
Likelihood+Salience+Freq+Context	**76.95**	**80.07**

Table 2: Classification accuracies (%) over data in Group 2 compared to the majority baseline.

Model	Group 2
Majority Baseline	59.52
Logistic regression with Context features	**63.21**
Logistic regression with Context+Salience	54.37

Statistical measures are expected to be promising features when identifying MWEs among expressions with consistent behaviour. However, the results in Table 1 show that our `Context` features are more effective in MWE classification even when applied over Group 1 and also over the whole data.

The good performance when using word context features leads us to think that their usefulness can be attributed to the information obtained from external arguments of the verb and the noun constituents of expressions. More experiments need to be done to confirm this and also to find the best

suitable window size for the word context around a target expression[6].

We have also trained the logistic regression model with the combination of the `Context` features and the association measures in Table 1. According to these results, the combination of features improves the accuracies of our model in identifying idiomatic expressions specially when applied over the consistent data in Group 1. The results lead us to believe that context features are even more useful in cases where we expect the best result from statistical measures due to the more consistent behaviour of the data. The better performance when using `Context` and statistical measures together, compared with when we use `Context` features alone is also a remarkable observation visible at Table 1.

Our experiment using the combination of Context and Salience (as the best statistical measure) for training over Group 2 expressions (Table 2), shows that the statistical measure is not helpful for the class of ambiguous expressions.

7 Conclusions and Future Work

We investigate the inclusion of concordance as part of the feature set used in supervised classification of MWEs. We have shown that context features have discriminative power in detecting literal and idiomatic usages of expressions both for the group of expressions with high potential of occurring in both literal/idiomatic senses or otherwise. Our results suggest that, when used in combination with traditional features, context can improve the overall performance of a supervised classification model in identifying MWEs.

In future, we intend to consider incorporating linguistically motivated features into our model. We will also experiment with constructing features that would consider long-distance dependencies in cases of MWEs with gaps in between their components.

Acknowledgments

The authors would like to thank Manuela Cherchi and Anna Desantis for their annotations and input on Italian examples.

[6]We have realised through trial-and-error that a window size of two after a target expression leads to better results compared with no context or contexts of bigger size.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, second edition.*, pages 267–292. CRC Press.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations*, EACL '06, pages 87–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of EACL-06*, pages 329–336.

Julian Brooke, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2014. Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 753–761.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19(4):450–466.

Meghdad Farahmand and Ronaldo Martins. 2014. A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 10–16, Gothenburg, Sweden, April. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Waseem Gharbieh, Virendra Bhavsar, and Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions, MWE@ACL 2016, Berlin, Germany, August 11, 2016.*

Sylviane Granger and Fanny Meunier. 2008. *Phraseology: an interdisciplinary perspective*. John Benjamins Publishing Company.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Kilgarriff, Pavel Rychl, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *EURALEX 2004*, pages 105–116, Lorient, France.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Johanna Monti, Federico Sangati, and Mihael Arcan. 2015. TED-MWE: a bilingual parallel corpus with mwe annotation. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, May. European Language Resources Association.

Alexandre Rondon, Helena Caseli, and Carlos Ramisch. 2015. Never-ending multiword expressions learning. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 45–53, Denver, Colorado, June. Association for Computational Linguistics.

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

Vieri Samek-Lodovici. 2003. The internal structure of arguments and its role in complex predicate formation. *Natural Language & Linguistic Theory*, 21(4):835–881.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *TACL*, 2:193–206.

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shiva Taslimipoor, Anna Desantis, Manuela Cherchi, Ruslan Mitkov, and Johanna Monti. 2016. Language resources for italian: towards the development of a corpus of annotated italian multiword expressions. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016), Napoli, Italy*.

Yuancheng Tu and Dan Roth. 2011. Learning english light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39, Portland, Oregon, USA, June. Association for Computational Linguistics.

Aline Villavicencio. 2005. The availability of verb–particle constructions in lexical resources: How much is enough? *Computer Speech & Language*, 19(4):415–432.