Deep Neural Representations for Multiword Expressions Detection

Kamil Kanclerz and **Maciej Piasecki**
Department of Artificial Intelligence,
Wrocław University of Science and Technology, Wrocław, Poland
{kamil.kanclerz,maciej.piasecki}@pwr.edu.pl

Abstract

Effective methods for multiword expressions detection are important for many technologies related to Natural Language Processing. Most contemporary methods are based on the sequence labeling scheme applied to an annotated corpus, while traditional methods use statistical measures. In our approach, we want to integrate the concepts of those two approaches. We present a novel weakly supervised multiword expressions extraction method which focuses on their behaviour in various contexts. Our method uses a lexicon of English multiword lexical units acquired from The Oxford Dictionary of English as a reference knowledge base and leverages neural language modelling with deep learning architectures. In our approach, we do not need a corpus annotated specifically for the task. The only required components are: a lexicon of multiword units, a large corpus, and a general contextual embeddings model. We propose a method for building a silver dataset by spotting multiword expression occurrences and acquiring statistical collocations as negative samples. Sample representation has been inspired by representations used in Natural Language Inference and relation recognition. Very good results (F1=0.8) were obtained with CNN network applied to individual occurrences followed by weighted voting used to combine results from the whole corpus. The proposed method can be quite easily applied to other languages.

1 Introduction

Multiword expressions (henceforth MWEs) have been studied for decades, defined in different ways in literature with different denotations of this term, e.g. see the overview in (Ramisch, 2015). Probably, the most genuine, but the least operational, definition is multiword lexemes stored as single lexical units in the mental lexicon ready to be retrieved. In the spirit of this fundamental property, we consider MWEs from the lexicographic point of view as lexical units that "has to be listed in a lexicon" (Evert, 2004) and we seek for methods of automated extraction of MWEs from text corpora to expand a large semantic lexicon with *multi-word lexical units*. Summarising a longer definition given in (Ramisch, 2015), MWEs are "lexical items decomposable into multiple lexemes", "present idiomatic behaviour at some level of linguistic analysis" and "must be treated as a unit" and, thus, should be described in a semantic lexicon, e.g. from (Stevenson, 2010) *air corridor* (an agreement between two countries), *slow food* ("traditional food and ways of producing, cooking and eating it"), *fast food*, *fire door*, *first lady* etc. A similar definition was adopted in the PARSEME Shared Task resource (Ramisch et al., 2018, 2020a). As we target the construction of a general lexicon expressing good coverage for lexical units occurring frequently enough in a very large corpus, we need also to take into account *multiword terms*, i.e. (Ramisch, 2015) "specialised lexical units composed of two or more lexemes, and whose properties cannot be directly inferred by a non-expert from its parts because they depend on the specialised domain".

Several MWE characteristics or identifying properties have been postulated, e.g.: arbitrariness, institutionalisation, limited semantic variability (especially non-compositionality and non-substitutability), domain specificity, and limited syntactic variability (Ramisch, 2015). Among them, semantic non-compositionality seems to be one of the strongest identifying factors. However, the challenge is to trace them using some corpus-based evidence and guide the extraction process. In addition, MWEs should be some how correlated with higher or more prominent frequency in language use in order to be worth inclusion in a lexicon.

Extraction of MWEs and their description in a semantic lexicon (e.g. as a reference resource) is important for many NLP applications like semantic

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 444 - 453
May 22-27, 2022 ©2022 Association for Computational Linguistics

indexing, knowledge graph extraction, vector models, topic modelling etc. Due to the specific properties of MWEs as whole units, their automated description by the distributional semantics method, e.g. embeddings, is not guaranteed, especially in the case of MWEs of lower frequency.

Traditionally, MWEs extraction is preceded by finding collocations (frequent word combinations) by statistical or heuristic *association measures* and filtering them by syntactic patterns. However, in this way mainly the frequency-related aspect is covered. The peculiar behaviour of MWEs as a language unit may be observed in linguistic contexts, and methods based on the well-known *sequence labelling* scheme try to do that. They explore MWE specific behaviour of as a language expressions across text contexts, where the contexts are represented by contextual embeddings (neural language models). However, such approaches require a lot of hard manual work on text annotation. In addition, due to the corpus size limitation, most potential MWEs are observed only in a few, if not singular uses, while a lexicon element by a definition is a ready-to-use unit to be included in different contexts and, as such, should be studied.

Thus, we want to fully explore the expected MWE characteristic aspects, including frequency, and to reduce the amount of manual work required. MWE annotated corpora are very rare and small, e.g. PARSEME (Ramisch et al., 2020b), but MWEs are listed in dictionaries and lexical resources. We propose a weakly supervised approach in which a lexicon of MWEs is used to build a kind of silver data on the basis of general text corpus. Concerning negative examples, i.e. language expressions rejected to be MWEs, that are hardly listed in any lexical resources, we use association measures (frequency aspect) to find collocations very likely not being MWEs. Next we feed a system combining contextual embeddings, deep neural learning and weighted voting scheme across individual MWE occurrences with the silver data. As a result, the system can be next used to filter potential MWEs extracted from a corpus with association measures (the frequency aspect in a positive role). In contrast to many methods from literature, we neither need a corpus laboriously annotated with MWE occurrences, nor language models specially trained for this task. In addition we aim at jointly encompass most of the MWE characteristic aspects with the majority of them recognised in a kind of overlap of MWE contextual embeddings across their different occurrences. The proposed approach is illustrated with good results achieved on English MWEs coming from several dictionaries and the British National Corpus. However, our method can be quite easily adapted to any language, the only required elements are: a corpus and an initial lexicon of MWEs, and a general contextual embeddings model.

2 Related Work

Initially statistical association measures calculated on the basis of word co-occurrence statistics in corpora were used for discovering and ranking collocations as potential MWEs (Evert, 2004). Single measures can be also combined into complex ones, e.g. by a neural network (Pečina, 2010). Syntactic information from parsing (Seretan, 2011) or from lexico-syntactic constraints based on morpho-syntactic tagging (Broda et al., 2008) were used in counting statistics and post-filtering collocations. Several systems for MWE extraction were proposed, combining different techniques, e.g. *mwe-toolkit* by Ramisch (Ramisch, 2015) combines statistical extraction and morpho-syntactic filtering, but also describes collocations with feature vectors to train Machine Learning (ML) classifiers. Lexico-syntactic patterns, measures, length and frequency are used as features in ML-based MWE extraction (Spasić et al., 2019). Linguistic patterns were used to extract MWEs and post-filter the outcome of association measures (Agrawal et al., 2018). MWEs were also detected by tree substitution grammars (Green et al., 2013) or finite state transducers (Handler et al., 2016).

Recently, attention was shifted to MWE extraction perceived as a sequence labelling problem, e.g. (Chakraborty et al., 2020), where corpora are annotated on the level of words, typically, BIO annotation format (Ramshaw and Marcus, 1995): B – a word begins an MWE, I is inside, O – outside. Sequence labelling approaches can also be combined with heuristic rules (Scholivet and Ramisch, 2017) or supersenses of nouns or verbs (Hosseini et al., 2016). Such heuristics are applied to extract linguistic features from texts for training a Bayesian network model (Buljan and Šnajder, 2017). Convolutional graph networks and self-attention mechanisms can be used to extract additional features (Rohanian et al., 2019). There are many challenges related to the nature of the MWEs, e.g.: discontinuity – another token occurs between the MWE components or overlapping – another MWE occurs between the components of the given MWEs. To counteract this, a model based on LSTM, the long short-term memory networks and CRF is proposed (Berk et al., 2018). The model from (Taslimipoor et al., 2020) combines two learning tasks: MWE recognition and dependency parsing in parallel. The approach in (Kurfalı, 2020) leverages feature-independent models with standard BERT embeddings. mBERT was also tested, but with lower results. An LSTM-CRF architecture combined with a rich set of features: word embedding, its POS tag, dependency relation, and its head word is proposed in (Yirmibeşoğlu and Güngör, 2020).

MWEs can be also represented as subgraphs enriched with morphological features (Boros and Burtica, 2018). Graphs can be next combined with the *word2vec* (Mikolov et al., 2013) embeddings to represent word relations in the vector space and then used to predict MWEs on the basis of linguistic functions (Anke et al., 2019). Morphological and syntactic information can be also delivered to a recurrent neural network (Klyueva et al., 2017). Two approaches to MWE recognition within a transition system were compared in (Saied et al., 2019): one based on a multilayer perceptron and the second on a linear SVM. Both utilise only lemmas and morphosyntactic annotations from the corpus and were trained and tested on PARSEME Shared Task 1.1 data (Ramisch et al., 2018).

However, such sequence labeling approaches focus on word positions and orders in sentences, and seem to pay less attention to the semantic incompatibility of MWEs or semantic relations between their components. Furthermore, sequence labeling methods do not emphasize the semantic diversity of MWE occurrence contexts. Thus, they overlook one of the most characteristic MWE factors: components of a potential MWE co-occur together regardless of the context. It allows us to distinguish a lexicalised MWE from a mere collocation or even a term strictly related to one domain. To the best of our knowledge, the concept of using deep neural contextual embeddings to describe the semantics of the MWEs components and the semantic relations between them in a detection task has not been sufficiently studied, yet. Moreover, due to the sparsity of the MWEs occurrences in the corpus, the corpus annotation process is very time consuming and can lead to many errors and low inter-annotator agreement. For this reason, we propose a lexicon-based corpus annotation method. We assume that the vast majority of MWEs are monosemous, automatically extract the sentences containing the MWE occurrences, and treat all sentences including a given MWE (as a word sequence) as representing the same multiword lexical unit.

3 Datasets

The conducted analysis of the existing resources has shown that it is difficult to find a large annotated dataset for the multiword expressions detection task. PARSEME shared task and multilingual corpus (Ramisch et al., 2020b) is a very valuable initiative, but focused mainly on verbal MWEs and quite small, especially its English part. Moreover, dictionaries containing MWEs follow different definitions and lexicographic practices, which makes it difficult to unambiguously determine whether a given multiword entity is a valid MWE. Therefore, in order to obtain a large dataset, we followed our idea of silver dataset and selected The Oxford Dictionary of English (ODE) (Stevenson, 2010) as a reference point to obtain the list of correct MWEs. The proposed method is in some way parameterised by a selected reference dictionary.

Concerning language expressions that are not MWEs, i.e. negative samples from the ML perspective, they are not listed or mentioned in the dictionaries. Having a corpus annotated with MWE occurrences we could extract expressions that are not as negative samples. However genuine MWEs are more frequent or statistically specific. Thus, 'normal' language expressions would be too obviously different. Instead, we noticed that statistical association measures produce very long ranking lists of collocations. Further down the ranking, MWE occurrences are quickly dwindling away. In addition, we are interested only in specific structural types of collocations that match structural types of MWEs acquired from a dictionary.

To generate the list of incorrect MWEs, we selected three popular association measures[1]: (1) the Pointwise Mutual Information (PMI) (Church and Hanks, 1990), (2) the Sørensen–Dice coefficient (Dice) (Dice, 1945), and (3) Pearson's chi-square (Chi2) (Manning and Schutze, 1999) and used them

[1] A combined association measure could produced a better ranking, but only moderately better and would require optimisation on the given dictionary and corpus. Moreover, our dictionary seems to be too small, with too small coverage for the optimisation.

to extract collocation ranking list from the British National Corpus (BNC) (Burnard, 1995). In order to find relevant examples of multiword units, we decided to select those collocations that were in the third quartile of the list sorted in descending order based on the value of the selected measure. We quickly skimmed the list in order to ensure that it is hard to spot anything looking as a MWE (but we do not exclude the possibility that some MWEs may occur, perfect precision does not seem to be necessary). We combined the list of the correct MWEs (from the dictionary) separately with the lists of collocations obtained via each of the three selected measures. In all experiments we concentrated on two word MWEs and collocations, as the statistical association measures we applied are naturally defined for two word combinations. However, as it will be presented later, some of the MWE representation we propose can be easily expanded to k-word cases. Moreover, two word MWEs form the vast majority of all in the dictionaries. Collocations extracted from the corpus were restricted only to those that represent structural types of MWEs from the dictionary.

We then used the three resulting lists to search for sentences including collocation or MWE occurrences in the BNC corpus. The searched expressions were simply recognised by comparing lemma sequences. Some recognition error may appear, but the potential error ration seems to be very small (single percents). If multiple MWE/collocation lemma sequences were detected among the sentence lemmas, then their occurrences were considered as separate *training samples* (positive or negative), see Alg. 1. In order to evaluate our method of detecting sentences containing MWEs, we extracted 4 randomly selected samples containing 100 found sentences each. A linguist conducted the analysis and found that 99% of the sentences contained correct MWE occurrences. The analysis was performed only on sentences corresponding to positive samples – MWEs from the dictionary, but similar results can be expected for collocations from the lists. Our work resulted in the creation of three datasets of MWE and collocation occurrences, named on the basis of the sources of knowledge:

- **ODE–PMI dataset** – dataset containing occurrences of correct MWEs from the ODE dictionary and the incorrect ones obtained via the PMI measure,
- **ODE–Dice dataset** – dataset containing occurrences of correct MWEs from the ODE dictionary and the incorrect ones obtained via the Dice measure,
- **ODE–Chi2 dataset** – dataset containing occurrences of correct MWEs from the ODE dictionary and the incorrect ones obtained via the Chi2 measure.

Algorithm 1 Procedure of obtaining sentences (s) containing MWEs from the corpus (C) by comparing sentence word lemmas ($l_i \in [l_0, l_1, \ldots, l_n]$) to the list ($M$) of lemmatised MWEs ($m_j \in [m_0, m_1, \ldots, m_k]$)

```
$sentence\_list \leftarrow [\,]$
for $s \in C$ do
    for $l_i \in s$ do
        for $m_j \in M$ do
            if $l_i \in m_j$ then
                $sentence\_list$.insert($s$)
            end if
        end for
    end for
end for
return $sentence\_list$
```

4 Deep Neural Representations for MWE Detection

4.1 Baseline

As our *baseline*, we decided to use a concatenation of vectors consisting of:

1. a component embedding ($\overrightarrow{c_{sent}}$),
2. an MWE embedding ($\overrightarrow{m_{sent}}$) in the context of the sentence ($sent$),
3. the absolute difference between the MWE embedding and the component embedding ($|\overrightarrow{m_{sent}} - \overrightarrow{c_{sent}}|$),
4. and the Hadamard product between the MWE embedding and the component embedding ($\overrightarrow{m_{sent}} \odot \overrightarrow{c_{sent}}$).

The proposed representation has been inspired by the ones often used in the Natural Language Inference domain and also in the task of semantic relations extraction (Fu et al., 2014; Levy et al., 2015). Our idea is to represent syntactic and semantic relations between the whole MWE and its components. We want to analyse the relation between the picture of the whole MWE used in a context and one of its components used in the same context, but separately, i.e. we exchange the whole MWE with one of its components and vice versa to see their contextual picture and interactions alone. The obvious target is the potential compositionality of an expressions: MWE or non-lexicalised collocation. In the case of compositional expressions we expect to see some kind of inclusion relation. However, we assumed that contextual embeddings allow us to go beyond focusing only on semantic compositionality, e.g. some syntactic idiosyncrasy should be also visible in relation between contextual embeddings of the whole expression and its component. Moreover, in order to minimise the effect of accidental properties of some specific context we try to collect representations of the same expressions (MWEs and collocations) across as many contexts as possible.

The obtaining of contextual MWE embeddings is described in Eq. 1. An MWE embedding ($\overrightarrow{m_{sent}}$) in the sentence context ($sent$) is an average of the WordPiece subtoken ($s \in S_{m_{sent}}$) vectors ($\overrightarrow{v_s}$) related to the MWE components.

$$\overrightarrow{m_{sent}} = \frac{\sum_{s \in S_{m_{sent}}} \overrightarrow{\nu_s}}{|S_{m_{sent}}|} \tag{1}$$

In the next step, the MWE occurrence was replaced subsequently with each of its components in order to obtain their contextual embeddings ($\overrightarrow{c_{sent}}$) by averaging the corresponding subtoken vectors representations ($\overrightarrow{\nu_s}$) related to the substituted components ($S_{c_{sent}}$), see Eq. 2.

$$\overrightarrow{c_{sent}} = \frac{\sum_{s \in S_{c_{sent}}} \overrightarrow{\nu_s}}{|S_{c_{sent}}|} \tag{2}$$

The final baseline embedding ($\overrightarrow{B}$) of a training sample related to a sentence ($sent$) containing MWE (m) and one of its components (c) is described in Eq. 3.

$$\begin{aligned}\overrightarrow{B_{c,m,sent}} =& \overrightarrow{c_{sent}} \oplus \overrightarrow{m_{sent}} \oplus (\overrightarrow{m_{sent}} - \overrightarrow{c_{sent}}) \\ & \oplus (\overrightarrow{m_{sent}} \odot \overrightarrow{c_{sent}})\end{aligned} \tag{3}$$

4.2 Difference vector based representation Diff-Emb

Our element-wise difference vector based representation *Diff-Emb* ($\overrightarrow{D}$), described in Eq. 5 leverages the absolute difference between non-contextual component embeddings ($\overrightarrow{w_1} - \overrightarrow{w_2}$) obtained via the skipgram model from the *fastText* library (Bojanowski et al., 2017) and the averaged element-wise difference between the component embeddings and MWE embedding ($avg\_diff_{m,sent}$) in the context of the sentence ($sent$). Eq. 4 describes the averaged difference vector for the MWE (m) containing components ($c \in m$). The non-contextual, static word embeddings were introduced into the representation in order to take into account semantic characteristics of expression components collected from a large corpus. In this way we want to take a yet another perspective on relation between the components.

$$\overrightarrow{avg\_diff_{m,sent}} = \frac{\sum_{c \in m}(\overrightarrow{m_{sent}} - \overrightarrow{c_{sent}})}{|m|} \tag{4}$$

$$\overrightarrow{D_{m,sent}} = |\overrightarrow{w_1} - \overrightarrow{w_2}| \oplus \overrightarrow{avg\_diff_{m,sent}} \tag{5}$$

4.3 Product based representation

We also decided to consider the relevance of Hadamard product vectors, which we included in our *Prod-Emb* representation ($\overrightarrow{P}$), explained in Eq. 7. It consists of the Hadamard product of non-contextual *fastText* component embeddings ($\overrightarrow{w_1} \odot \overrightarrow{w_2}$) and the averaged vector of Hadamard products between the component ($c \in m$) embeddings and MWE (m) embedding ($avg\_prod_{m,sent}$) in the context of the sentence ($sent$) described in Eq. 6

$$\overrightarrow{avg\_prod_{m,sent}} = \frac{\sum_{c \in m}(\overrightarrow{m_{sent}} \odot \overrightarrow{c_{sent}})}{|m|} \tag{6}$$

$$\overrightarrow{P_{m,sent}} = (\overrightarrow{w_1} \odot \overrightarrow{w_2}) \oplus \overrightarrow{avg\_prod_{m,sent}} \tag{7}$$

4.4 Combined representation: differences and products

In order to combine the difference-based and product-based approaches we developed the *Mean-Emb* representation ($\overrightarrow{M}$), explained in Eq. 8. It consists of the averaged difference vector ($\overrightarrow{avg\_diff_{m,sent}}$) and the averaged Hadamard product vector ($\overrightarrow{avg\_prod_{m,sent}}$) described in Eq. 4 and 6 respectively.

$$\overrightarrow{M_{m,sent}} = \overrightarrow{avg\_diff_{m,sent}} \oplus \overrightarrow{avg\_prod_{m,sent}} \tag{8}$$

5 Experimental Setup

For all conducted experiments we selected a single-task binary classification, where the classifier aims to predict the correct label out of 2 possible ones (lexicalised vs non-lexicalised) for the expression represented by one of the vector representations: baseline, Diff-Emb, Prod-Emb or Mean-Emb. In the process of generating the contextual embeddings we used the XLM-RoBERTa (Conneau et al., 2020) language model as it is considered as one of the best transformer models for English. We decided to use the convolutional neural network (CNN) architecture as the classifier to better extract the knowledge from our vector representations. We used the TensorFlow library (Abadi et al., 2015) to implement the CNN model. Our convolutional neural network contains three convolutional layers, each followed by the pooling layer and the dropout layer and is shown in Fig. 1. We used the F1-macro metric to measure the performance of the classifier on each of the representations. To prevent data leakage, we applied the *lexical split* to avoid the risk of testing on the same multiword unit, which was used in the training procedure (even if the sentence samples are obviously not overlapping). We leveraged the 10-fold cross-validation and used statistical tests to measure the significance of the difference between different experiment configurations. We checked the assumptions and then applied the independent samples *t*-test with the Bonferroni correction if they were met. Otherwise we used the Mann-Whitney *U*-test.

6 Results

Tab. 1 shows the evaluation results for each representation on the ODE–PMI dataset. Each value is averaged over ten folds. The Mean-Emb representation combining both the knowledge based on the difference vector and the Hadamard product vector achieved the best results.

The performance of the CNN model trained on all representations and evaluated on the ODE–Dice dataset is shown in Tab. 2. The best performance can be observed for the Mean-Emb model. Each of the developed representations achieved better results than the baseline vector representation.

The evaluation results for the classifier trained on each representation and evaluated on the ODE–Chi2 dataset are shown in Tab. 3. The Mean-Emb model achieved the best results among other representations. The worst performance can be observed for the baseline vector representation.

Representation	Cor F1	Inc F1	F1
baseline	0.77	0.77	0.77
Diff-Emb	0.77	0.78	0.78
Prod-Emb	0.78	0.78	0.78
Mean-Emb	**0.79**	**0.79**	**0.79**

Table 1: The results of the CNN model trained on various representations on the ODE–PMI dataset. Measures: Cor F1 – F1 score for lexicalised MWEs; Inc F1 – F1 score for non-lexicalised MWEs; F1 – macro average of the F1 scores for lexicalised and non-lexicalised MWEs. Values in **bold** are significantly better than others.

Representation	Cor F1	Inc F1	F1
baseline	0.75	0.75	0.75
Diff-Emb	0.76	0.76	0.76
Prod-Emb	0.76	0.76	0.76
Mean-Emb	**0.77**	**0.77**	**0.77**

Table 2: Evaluation results on the ODE–Dice dataset. Measures: Cor F1 – F1 score for lexicalised MWEs; Inc F1 – F1 score for non-lexicalised MWEs; F1 – macro average of the F1 scores for lexicalised and non-lexicalised MWEs. Values in **bold** are significantly better than others.

Representation	Cor F1	Inc F1	F1
baseline	0.77	0.77	0.77
Diff-Emb	0.77	0.78	0.78
Prod-Emb	0.77	0.78	0.78
Mean-Emb	**0.79**	**0.80**	**0.80**

Table 3: Evaluation results on the ODE–Chi2 dataset. Measures: Cor F1 – F1 score for lexicalised MWEs; Inc F1 – F1 score for non-lexicalised MWEs; F1 – macro average of the F1 scores for lexicalised and non-lexicalised MWEs. Values in **bold** are significantly better than others.

7 Discussion

The idea of silver dataset enables transformation of any corpus into a dataset for MWE extraction, only if a limited lexicon of MWE examples is provided as a starting point – a kind of seed lexicon to be expanded. We can leverage a MWE annotated corpus, too, in the same way as a lexicon to extract the initial list of MWEs, but a large non-annotated corpus stays the basis. Several linguistic resources can be also merged, any MWE annotated text, as well as lexicons. Time-consuming and expensive corpus annotation is avoided. Moreover, it seems to be

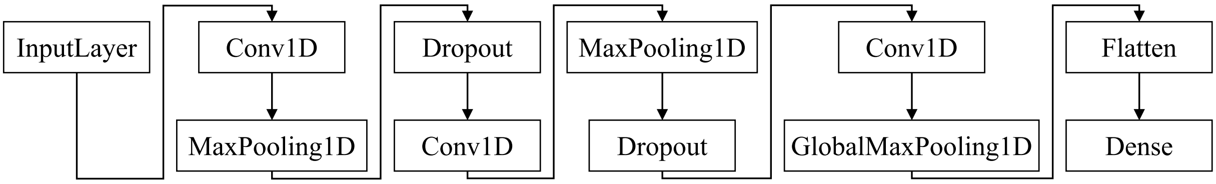

Figure 1: Convolutional neural network classifier structure.

easier to maintain high quality lexicon than corpus annotation, e.g. due to potential errors and discrepancies between single annotations. A lexicon can be edited by several linguists, and metrics such as inter-annotator agreement can be easily calculated.

What is more, such a transformation of lexicon-based knowledge into a dataset enables the use of deep neural network models that require large number of training samples. This is one of the reasons why our CNN method, pre-trained on contextual embeddings with weighted voting, applied to MWE recognition achieved several times better results than methods based on contextual embeddings and recurrent neural in the PARSEME shared task in general (Ramisch et al., 2020a), not mentioning the English part alone that is very small.

Our approach may be applied to texts in different languages, both to obtain multilingual collections and to apply transfer learning to facilitate the knowledge about MWEs in one language to MWE recognition in another language. This may be particularly relevant for low-resource languages, and it definitely a direction for further research.

Another advantage of the proposed method is faster training and prediction in comparison to sequence labeling methods. In our case, the model gets the full sample representation only once before prediction. This shortens the inference time.

Our vector representations support MWEs longer than two words. In the case of multiword units containing three and more words, the difference and product vectors calculated between two MWE components can be replaced with the vector obtained via the same operation, but averaged over all MWE component pairs.

The obtained results show that non-lexicalised representations, i.e. those that do not include vectors for components and the whole expression[2] perform better independently of the kind of a measure used to extract collocations. All representations except the baseline are built from differences and products of vectors, not the vectors itself. Thus they are more focused on representing relations between a potential MWE and its components. It is worth to be emphasised that lexical split was also implemented in order to prevent the models to remember concrete words instead of learning patterns for behaviour of proper MWEs. There are no large differences between results for different measure, but, with some caution, we can observe that results obtained with PMI are slightly better, while in the case of PMI the measure is naturally is filtered by 0 threshold and produces potentially more interesting collocations, thus harder to be distinguished from the proper MWEs.

8 Conclusions and Future Work

Our three representations allowed classifier to achieve significantly better results in comparison to the baseline approach focused on the component and MWE embedding.

The context provided additional information on the MWE semantics, which improved the model performance. This is related to the non-compositional nature of the MWEs, which meaning cannot be inferred from their component meanings.

Our approach based on difference and product vectors forced the models significantly reduced the training time. It may be more important in practice, when the training time and inference time are more important than the quality of prediction. On the other hand, the method based on contextual embeddings allows transforming any set of texts with the use of dictionary knowledge into an annotated corpus containing occurrences of the MWEs and their components. The model, by examining the semantic differences between the component and the entire expression, takes into account the variability of the context, which should allow for the extraction of the MWE meaning following the assumption of its monosemous character.

In future work, we want to use our methods to generate corpora in other languages, which will be later used to train models in the multilingual MWEs

[2] A contextual vector of the whole expression somehow includes a picture of the particular expression and its lexemes.

detection task and to explore the transfer learning mechanism in a language-independent MWE detection.

Acknowledgements

This work was financed by (1) the National Science Centre, Poland, project no. 2019/33/B/HS2/02814; (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19; (4) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology.

References

Martín Abadi, Ashish Agarwal, Paul Barham, and et al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.

Shaishav Agrawal, Ratna Sanyal, and Sudip Sanyal. 2018. Hybrid method for automatic extraction of multiword expressions. *Int. Journal of Engineering & Technology*, 7:33.

Luis Espinosa Anke, Steven Schockaert, and Leo Wanner. 2019. Collocation classification with unsupervised relation vectors. In *Proc. of the 57th Annual Meeting of the ACL*, pages 5765–5772, Florence, Italy. Association for Computational Linguistics.

Gözde Berk, Berna Erden, and Tunga Güngör. 2018. Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. In *Proc. of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 248–253, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and et al. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Tiberiu Boros and Ruxandra Burtica. 2018. GBD-NER at PARSEME shared task 2018: Multi-word expression detection using bidirectional long-short-term memory networks and graph-based decoding. In *Proc. of the Joint Work. on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 254–260, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

B. Broda, M. Derwojedowa, and M. Piasecki. 2008. Recognition of structured collocations in an inflective language. *Systems Science*, 34(4):27–36.

Maja Buljan and Jan Šnajder. 2017. Combining linguistic features for the detection of Croatian multiword expressions. In *Proc. of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 194–199, Valencia, Spain. Association for Computational Linguistics.

Lou Burnard. 1995. *British National Corpus: Users Reference Guide British National Corpus Version 1.0.* Oxford Univ. Computing Service.

Sritanu Chakraborty, Dorian Cougias, and Steven Piliero. 2020. Identification of multiword expressions using transformers.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Univ. of Stuttgart.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209, Baltimore, Maryland. Association for Computational Linguistics.

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

Abram Handler, M. Denny, H. Wallach, and et al. 2016. Bag of what? simple noun phrase extraction for text analysis. In *NLP+CSS@EMNLP*, pages 114–124, Austin, Texas. Association for Computational Linguistics.

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 *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 931–936, San Diego, California. Association for Computational Linguistics.

Natalia Klyueva, Antoine Doucet, and Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proc. of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 60–65, Valencia, Spain. Association for Computational Linguistics.

Murathan Kurfalı. 2020. TRAVIS at PARSEME shared task 2020: How good is (m)BERT at seeing the unseen? In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 136–141, online. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

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

Tomás Mikolov, Kai Chen, Greg Corrado, and et al. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proc.*

P. Pečina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition. A Generic and Open Framework*. Springer.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020a. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020b. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, and et al. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. In *Proc. of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2692–2698, Minneapolis, Minnesota. Association for Computational Linguistics.

Hazem Al Saied, Marie Candito, and Mathieu Constant. 2019. Comparing linear and neural models for competitive MWE identification. In *Proc. of the 22nd Nordic Conference on Computational Linguistics*, pages 86–96, Turku, Finland. Linköping University Electronic Press.

Manon Scholivet and Carlos Ramisch. 2017. Identification of Ambiguous Multiword Expressions Using Sequence Models and Lexical Resources. In *Proc. of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 167–175, Valencia, Spain. Association for Computational Linguistics.

V. Seretan. 2011. *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer Netherlands.

Irena Spasić, David Owen, Dawn Knight, and et al. 2019. Unsupervised multi-word term recognition in Welsh. In *Proc. of the Celtic Language Technology Workshop*, pages 1–6, Dublin, Ireland. European Association for Machine Translation.

Angus Stevenson. 2010. *Oxford Dictionary of English*. Oxford University Press.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Zeynep Yirmibeşoğlu and Tunga Güngör. 2020. ERMI at PARSEME shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 130–135, online. Association for Computational Linguistics.