Is Context All You Need? Non-Contextual vs Contextual Multiword Expressions Detection

Maciej Piasecki[0000−0003−1503−0993] and Kamil Kanclerz[0000−0002−7375−7544]

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

Abstract. Effective methods of the detection of multiword expressions are important for many technologies related to Natural Language Processing. Most contemporary methods are based on the sequence labeling scheme, while traditional methods use statistical measures. In our approach, we want to integrate the concepts of those two approaches. In this paper, we present a novel weakly supervised multiword expressions extraction method which focuses on their behaviour in various contexts. Our method uses a lexicon of Polish multiword units as the 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. Compared to the method based on non-contextual embeddings, we obtain gains of 15 percentage points of the macro F1-score for both classes and 30 percentage points of the F1-score for the incorrect multiword expressions. The proposed method can be quite easily applied to other languages.

Keywords: Natural Language Processing · Multiword Expressions · Detection of Multiword Expressions · Contextual Embeddings.

1 Introduction

Multiword expressions (henceforth MWEs) are defined in different ways in literature, e.g. see the overview in [29]. In this work, we consider MWEs from the lexicographic point of view as lexical units that "has to be listed in a lexicon" [12] and we focus on methods of automated extraction of MWEs from text corpora to be included in a large semantic lexicon as *multi-word lexical units*. Summarising a bit the definition of [29], 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. *skrzynia biegów* ('gearbox', lit. box of gears) , *pogoda ducha* (≈'optimism', 'good attitude', lit. 'weather of spirit'). A similar definition was adopted in the PARSEME Shared Task resource [30, 31]. As we target the construction of a general lexicon expressing good coverage for lexical units occurring frequently

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

enough in a very large corpus – and we will test our approach against such a resource, see Sec. 3 – we need also to take into account *multiword terms*, i.e. [29] "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 properties are postulated that can guide the extraction process, like arbitrariness, institutionalisation, limited semantic variability (especially non-compositionality and non-substitutability), domain-specificity and limited syntactic variability [29]. As we are interested in the lexicon elements, the frequency of potential MWEs should be taken into account and MWEs are indeed in some way specific with respect to the frequency of co-occurrence of its components. Extraction of MWEs and their description in a semantic lexicon (at least as a reference resource) is important for many NLP applications like semantic 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. Recent methods follow *sequence labelling* scheme and try to explore the specific behaviour of MWEs as language expressions in text. Due to our objective, we aim at combining the best of the two worlds. We propose a new weakly supervised method for MWE extraction from large text corpora that explores their peculiar properties as elements of language structures across various contexts. The proposed method combines neural language modelling with deep learning and a lexicon of MWEs as a knowledge base, i.e. the sole source of supervision. 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. We investigated and combined non-contextual representation of MWEs as lexical units and their contextual representation as elements of the sentence structures. For the latter purpose, we leverage deep neural contextual embeddings to describe the peculiarities of the semantic but also syntactic behaviour of MWEs in contrast to the behaviour of their components. What is more, the evidence for the whole MWEs and their components can be collected from different sentences in the corpus, not only those including whole MWEs. Our method can be quite easily adapted to any language, the only required elements are: a large corpus, an initial lexicon of MWEs, and a general contextual embeddings model. The proposed method, after training, may be applied to a list of collocations extracted from a corpus by association measures to distinguish MWEs from mere collocations.

2 Related Work

Initially, MWE recognition methods were based on statistical association measures based on co-occurrence statistics in text corpora [12] for weighting collocations as potential MWEs. Many association measures were examined and

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

combined into complex ones, e.g. by a neural network [27]. Syntactic information from parsing was used in counting statistics or post-filtering collocations [36]. Morpho-syntactic tagging and lexico-syntactic constraints were also used instead of parsing [7]. For Polish, association measures combined by a genetic algorithm and expanded with lexico-syntactic filtering were used to extract potential MWEs [28]. Several systems for MWE extraction were proposed, combining different techniques, e.g. *mwetoolkit* by Ramisch [29] 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 can be a feature source in ML-based MWE extraction [37]. Linguistic patterns were used to extract potential MWEs and post-filter out incorrect ones after association measures [2]. MWEs were also detected by tree substitution grammars [14] or finite state transducers [16].

Recently, attention was shifted to supervised ML and MWE extraction as a sequence labelling problem, e.g. [9], where corpora are annotated on the level of words, typically, BIO annotation format [32]: B – a word begins an MWE, I is inside, O – outside. Sequence labelling approaches can also be combined with heuristic rules [35] or supersenses of nouns or verbs [18]. Such heuristics are applied to extract linguistic features from texts for training a Bayesian network model [8]. Convolutional graph networks and self-attention mechanisms can be used to extract additional features [33]. 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 [4]. The model from [38] combines two learning tasks: MWE recognition and dependency parsing in parallel. The approach in [21] 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 [39].

MWEs can be also represented as subgraphs enriched with morphological features [6]. Graphs can be next combined with the *word2vec* [24] embeddings to represent word relations in the vector space and then used to predict MWEs on the basis of linguistic functions [3]. Morphological and syntactic information can be also delivered to a recurrent neural network [19]. Saied et al. [34] compared two approaches to MWE recognition within a transition system: 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 [30].

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 con-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

text. 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. On the basis of the assumption that the vast majority of MWEs are monosemous, e.g. the set of more than 50k MWEs in plWordNet [11], we performed an automated extraction of sentences containing the MWE occurrences and treated all sentences including a given MWE as representing the same multiword lexical unit.

3 Dataset

For evaluation, we used MWEs from plWordNet [11] marked as *multi-word lexical units* [23]. In addition, we utilised as negative data multiword lemmas removed from plWordNet as non-lexicalised over the years by the linguists. There is no information about all collocations considered for adding to plWordNet, but those that were once erroneously included must be more tricky ones. plWordNet contains 53,978 two-word MWEs and 6,369 longer than 2 words for Polish. English WordNet includes 59,079 two-word MWEs and 10,649 longer than 2 words. In the Polish part of the PARSEME corpus, there are also 3,427 two-word MWEs and 568 MWEs longer than 2 words (in the English part respectively, 457 two-word MWEs and 85 ones longer). Due to this high numerical prevalence of two-word MWEs, we concentrate on them in this paper. Two sample representations were compared: *non-contextual* and *contextual*. In the first case – a baseline – the representation is derived from word embeddings vectors. In the latter case of the contextual representation, we used the KGR10 Polish corpus [20], one of the largest Polish corpora (4,015,569,051 tokens, 18,084,712 unique ones) with a rich variety of text types.

3.1 plWordNet-based Non-Contextual Dataset

Context-free representation was built for both correct MWEs and incorrect 'MWEs' using the *fastText* skipgram model [5] (trained on the KGR10 corpus). It concatenates embeddings of the MWE components with vectors of differences between them. Fig. 1 and Eq. 1 show the generation of non-contextual MWE representation emb_{NC} from the vectors w_1 and w_2 of the component words. Such representation, including the difference of vectors, has been inspired by the sample representation used in the NLI domain and also in semantic relations extraction [13]. Moreover, the concatenation of the difference vector along with the word embeddings was also used to represent word relations in [22].

$$emb_{NC}(w_1, w_2) = \overrightarrow{w_1} \oplus \overrightarrow{w_2} \oplus (\overrightarrow{w_1} - \overrightarrow{w_2}) \quad (1)$$

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

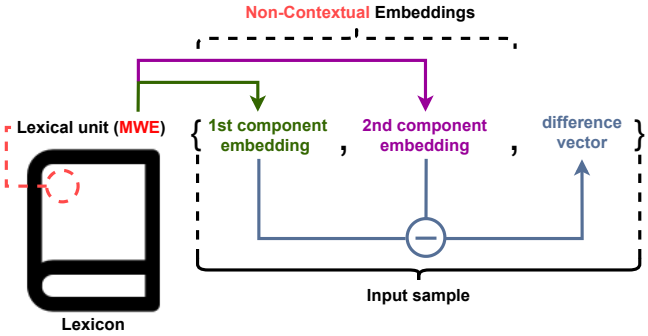

Fig. 1. Non-Contextual MWE representation generation

3.2 KGR10-based Contextual Dataset

For the contextual MWE representation, 687,900 sentences were extracted from the KGR10 corpus. Components of the correct MWEs were detected in 648,481 sentences and the incorrect in 39,419. We started by detecting the MWE component among the lemmas occurring in sentences. If lemmas of multiple MWEs were detected in a sentence, then it was associated with each of them as separate *training samples*, see Alg. 1. In order to test the performance of our method in detecting sentences containing MWE components, we prepared 4 randomly selected samples of 100 found sentences each. They were verified by linguists who found that 99% of all sentences contained correct MWE components.

Algorithm 1 Procedure of obtaining sentences (s) from the corpus (C), if they include MWEs or their components 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 ← [ ]
for s ∈ C do
    for l_i ∈ s do
        for m_j ∈ M do
            if l_i ∈ m_j then
                sentence_list.insert(s)
            end if
        end for
    end for
end for
return sentence_list
```

Eq. 2 describes the generation of contextual embeddings for MWEs as sample representations: an MWE embedding ($S_{m_{sent}}$) in the sentence context ($\overrightarrow{m_{sent}}$) is an average of the WordPiece subtoken vectors ($\overrightarrow{\nu_s}$) related to the MWE components. Next, we subsequently replaced the MWE occurrences in sentences with

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

each of their components and obtained 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. 3. The final contextual embedding (emb_C) of a training sample related to a sentence ($sent$) containing MWE (m) and one of its components (c) is described in Eq. 4. For each MWE occurrence, we generated the contextual embeddings corresponding to each of its components separately.

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

$$\overrightarrow{c_{sent}} = \frac{\sum_{s \in S_{c_{sent}}} \overrightarrow{\nu_s}}{|S_{c_{sent}}|} \tag{3}$$

$$emb_C(c, m, sent) = \overrightarrow{c_{sent}} \oplus \overrightarrow{m_{sent}} \oplus (\overrightarrow{m_{sent}} - \overrightarrow{c_{sent}}) \tag{4}$$

We aim at observing the difference between the contextual embedding of a whole MWE and each of its components across sentences. Thus, we calculated the difference vector between the representation of the complete expression and its component in the context of a sentence as is illustrated in Fig. 2.

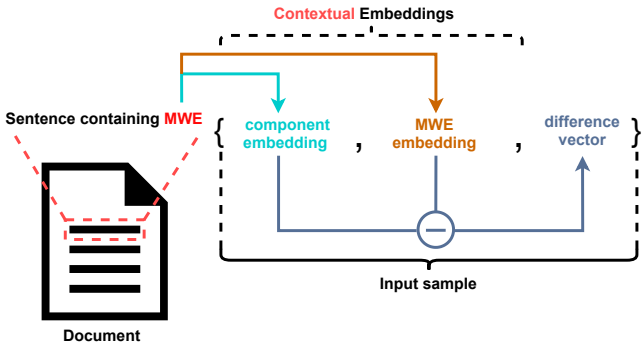

Fig. 2. MWE contextual representation generation.

4 Methods for Multiword Expression Detection

We assume that the context plays a significant role in the MWE detection. The first dataset from Sec. 3.1 contains training samples of non-contextual MWE representations (MWE vector, component vectors, and the difference vector). In this task, classifiers should focus on the semantic differences between the vector representations of the MWE components and the entire MWE. This is focused on non-compositional character of genuine MWEs. An incorrect 'MWE' example of

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

a *materiał opatrunkowy* (en. 'bandage cloth') is in fact compositional in contrast to a correct, genuine MWE: *głos serca* (en. lit. 'heart's voice'), whose semantics cannot be inferred from its component meanings.

In contrast to the first (baseline) non-contextual representation, the dataset from Sec. 3.2 includes samples of contextual MWE representations (contextual vectors of MWE components and the entire MWE, plus the difference vector). In this case, the task of classifiers is to decide on the correctness of an expression on the basis of knowledge extracted from the contexts of the expression occurrences and the interaction between the contexts and the semantic representation of the whole MWEs and their components. An example of an incorrect 'MWE' is *barwnik naturalny* (en. 'a natural pigment'), which is compositional in any context and an example of correct MWE is *ojciec chrzestny* (en. 'a godfather'), which is non-compositional and when occurs in different contexts, its components should receive significantly different contextual vectors from the MWE vector.

We prepared three different model architectures to measure the influence of context knowledge on the classification of collocations as MWEs:

- **Logistic Regression (LR)** – a statistical model, which utilizes the logistic function to model the probability of a discrete binary dependent variable,
- **Random Forest (RF)** – an ensemble learning method, aggregating multiple decision trees by calculating the mode of their predictions,
- **Convolutional Neural Network (CNN)** – a deep learning architecture, using convolution kernels, which move along the vector of the input data and provide translation outputs called feature maps.

Due to the nature of the MWE representation scheme shared between both representation types, we decided to use classifiers that work well with samples represented by concatenations of feature vectors. Contrary to the sequence labelling approaches, we decided to use logistic regression (LR), random forest (RF), and convolutional neural network (CNN). The RF model using an ensemble of decision trees focuses on the salient features of the vector representations. On the other hand, convolution operations allow the CNN model to derive additional knowledge from the data. We also used the LR model as a baseline to verify the quality of the RF and CNN classifiers knowledge extraction.

In the contextual representation, Sec 3.2, a single collocation or its component may occur in multiple sentences, so the same collocation may occur in several samples. As the vast majority of MWEs are monosemous, e.g. plWordNet [11], we leveraged this fact by preparing several voting strategies that aggregate the decisions of a selected model related to the same collocation:

- **Occurrence Classification (OC)** – each collocation occurrence is classified on its own, i.e. a separate decision, independent of the other occurrences, is made solely on the text of the given context,
- **Majority Voting (MV)** – predictions for all occurrences of a given collocation are collected and the final decision is made by majority voting,
- **Weighted Voting (WV)** – as the previous one, but the overall decision is made by weighted voting with confidence levels of a classifier as weights.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

5 Experiments

The task selected for all conducted experiments is a single-task binary classification, where each classifier had to predict the correct label out of the 2 available for the given expression as a potential MWE. We used the HerBERT model [25] to generate contextual embeddings as it is considered as one of the best transformer models trained and evaluated on texts in Polish. Implementations of the LR and RF classifiers come from the scikit-learn library [26], and CNN from the TensorFlow library [1]. The CNN architecture consists of three convolutional layers each followed by the pooling layer and the dropout layer, and is shown in Fig 3. To counter the impact of class imbalance in both datasets of samples (53,978 to 5,598 for the non-contextual one and 648,481 to 39,419 for the contextual one), we used the F1-macro measure to estimate the performance quality of classifiers. Moreover, we used the weighted loss function, depending on the number of instances of a given class in the training set. In addition, we applied 4 different variants of the SMOTE method (SMOTE [10], SVM-SMOTE [10], Borderline SMOTE [15], and ADASYN [17]) to generate additional synthetic training samples on the basis of the real sample embeddings. To avoid data leakage, we utilized the lexical split to counteract the risk of the same MWE appearing in both the training and test sets. We applied the 10-fold cross-validation in every experiment and used statistical tests to measure the significance of the differences between the models. We used the independent samples t-test with the Bonferroni correction if its assumptions were fulfilled. Otherwise the non-parametric Mann-Whitney U test was applied.

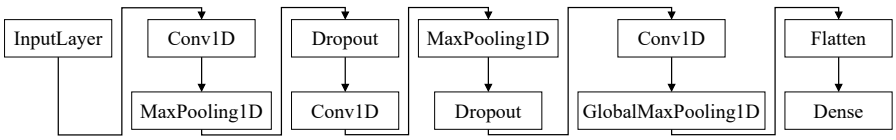

Fig. 3. Convolutional neural network classifier structure.

6 Results

Tab. 1 shows results averaged over ten folds for methods based on non-contextual and contextual representations. The expanded contextual knowledge resulted in significant improvements in the prediction quality of each classifier. The increase in the macro F1-score measure caused by the use of the contextual embeddings in comparison to non-contextual ones is presented in Fig. 4. The highest gain of 15% can be observed for the CNN model, as it was able to extract the most knowledge from the HerBERT embeddings due to its highest complexity.

Fig. 5 shows the performance improvement in the case of incorrect MWEs. In the case of the RF classifier, the use of contextual embeddings resulted in more

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

than sixfold and, in the case of CNN – over thirtyfold improvement in detection of incorrect expressions.

Model	Embedding	Inc F1	Cor F1	F1
LR	N-C	0.31	0.82	0.56
	C	**0.32**	**0.95**	**0.64**
RF	N-C	0.05	0.92	0.49
	C	**0.30**	**0.94**	**0.62**
CNN	N-C	0.01	0.96	0.48
	C	**0.31**	0.96	**0.63**

Table 1. F1-score values for incorrect MWEs (Inc F1), correct MWEs (Cor F1) and macro F1-score (F1) for non-contextual (N-C) and contextual (C) embeddings; models: LR, RF and CNN; values in **bold** are statistically significantly better in a given pair.

The performance of all voting strategies combined with each classifier is shown in Tab. 2. The use of weighted voting improved the value of the macro F1-score for the RF and CNN models by 2 and 4 percentage points, respectively, in relation to the results for occurrence classification. Moreover, the F1-score measure for incorrect MWEs increased by 6 percentage points for the CNN classifier. The improvement in evaluation performance may reflect the effect of using weighted voting to counteract the overfitting of more complex models, as this strategy benefits the most from the assumed monosemous nature of MWEs.

Model	Voting	Inc F1	Cor F1	F1
LR	OC	0.32	0.95	0.64
	MV	**0.33**	0.95	0.64
	WV	**0.33**	0.95	0.64
RF	OC	0.30	0.94	0.62
	MV	**0.33**	**0.95**	**0.64**
	WV	**0.33**	**0.95**	**0.64**
CNN	OC	0.31	0.96	0.63
	MV	0.36	0.96	0.66
	WV	**0.37**	**0.97**	**0.67**

Table 2. F1-score for the Contextual Dataset and incorrect MWEs (Inc F1), correct MWEs (Cor F1) and macro F1-score (F1) for LR, RF and CNN using three different voting strategies: occurrence classification (OC), majority voting (MV), and weighted voting (WV). Dataset. **Bold** values are statistically significantly better than others.

The evaluation results of different SMOTE methods used to counteract the class imbalance in the contextual representation samples, Sec. 3.2, are in Tab. 3. The use of SVM-SMOTE and Borderline SMOTE methods improved F1-score of the CNN model for incorrect MWEs by 14%. It also improved the overall F1-score by 5%. The CNN model was able to extract the most knowledge from

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

the synthetic samples generated by the SMOTE methods due to the fact that it has the most complex architecture among all used classifiers.

SMOTE Method	LR			RF			CNN		
	Inc F1	Cor F1	F1	Inc F1	Cor F1	F1	Inc F1	Cor F1	F1
None	0.30	0.94	0.62	0.30	0.94	0.62	0.17	**0.98**	0.58
SMOTE	0.31	**0.95**	0.63	0.30	0.94	0.62	0.27	0.93	0.60
SVM-SMOTE	**0.32**	**0.95**	**0.64**	0.30	**0.95**	0.62	**0.31**	0.95	**0.63**
Borderline SMOTE	0.31	**0.95**	0.63	0.30	0.94	0.62	**0.31**	0.96	**0.63**
ADASYN	0.31	**0.95**	0.63	0.30	0.94	0.62	0.30	0.94	0.62

Table 3. F1-score values for incorrect MWEs (Inc F1), correct MWEs (Cor F1), and macro F1-score (F1) LR, RF and CNN models trained on contextual embeddings based on the KGR10-based dataset with the use of four different SMOTE techniques: SMOTE, SVM-SMOTE, Borderline SMOTE, and ADASYN and no SMOTE (None). Values in **bold** are statistically significantly better than others.

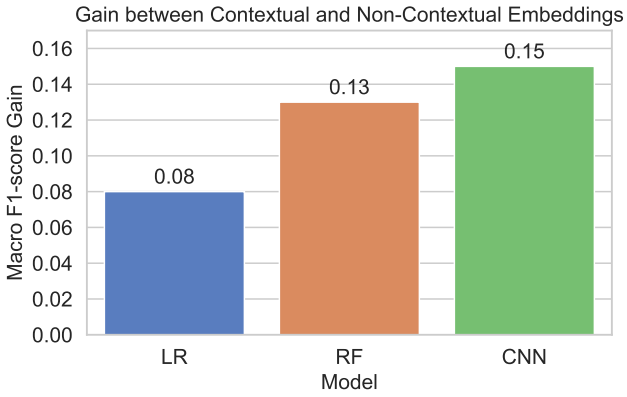

Fig. 4. F1-score improvement for the contextual vs non-contextual representations.

7 Discussion

One of the most important advantages of our method based on contextual representations is its ability to transform any text collection into a dataset, even if it has no annotations. We can leverage a MWE annotated corpus, but also any text collection, e.g. from web scraping. A seed MWE lexicon for a given language is enough. Time-consuming and expensive corpus annotation is avoided. Moreover,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

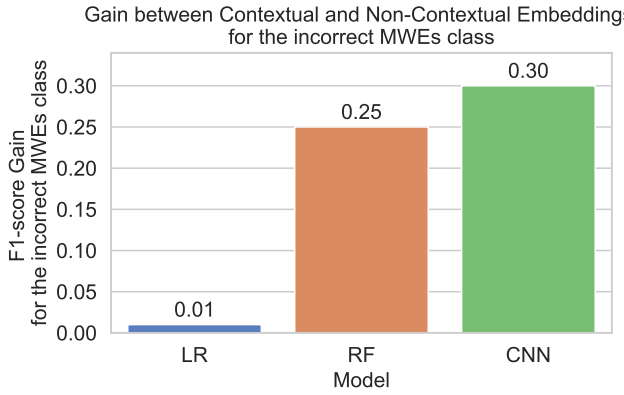

Fig. 5. F1-score increase for the incorrect MWEs class between the evaluation results for models trained contextual MWEs embeddings and the non-contextual ones.

it is easier to maintain high quality in a collection such as a lexicon, which can be annotated by several linguists, and metrics such as inter-annotator agreement can be easily calculated. Such a transformation of lexicon-based knowledge into a dataset enables the use of deep neural network models requiring a large number of training samples. Several linguistic resources can be also merged – both annotated texts, as well as lexicons. 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 relevant for low-resource languages. Another advantage of contextual representation is faster training and prediction compared to sequence labeling methods. In our case, the model gets the full sample representation only once before prediction. This shortens the inference time.

Non-contextual representations based only on word embeddings result in a smaller dataset with less noise and significantly reduce the training time. This approach also emphasizes the non-compositional nature of the MWEs, as the model focuses on the semantic differences between an MWE and its components.

Providing a full representation of a training sample to the model in one step enabled the use of SMOTE methods. Generating synthetic samples carries the risk of too much deviation from the actual data. This phenomenon has the greatest impact on sequence labeling methods that are vulnerable to outliers.

Our CNN method, pre-trained on contextual embeddings with weighted voting, applied to MWE recognition in the Polish part of the PARSEME corpus (mostly verbal) achieved significantly better results than the best results reported during Edition 1.2 of the PARSEME Shared Task [31] – our RF classifier using weighted voting scored 0.5244 on the macro F1-score measure, while the best result for PARSEME Edition 1.2 is 0.4344 macro F1-score, which indicates

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

a promising potential of our method. It is worth to emphasise that there is no overlap between the training set of our method and the PARSEME set of MWEs.

8 Conclusions and Future Work

Context plays a crucial role in MWE detection. Our three classifiers achieved significantly better results, with the CNN one on the top, with contextual embeddings than with the non-contextual ones. The context provided additional information on the MWE semantics, which improved the quality of the predictions. This is related to the non-compositional nature of the MWEs, the meaning of which cannot be inferred from the meanings of their components. The non-contextual representation forced the models to focus only on the nonstructural aspect meanings of the component meanings, but significantly reduced the training time. It may be more applicable in practice, when the training time and inference time are more important than the quality of prediction. This method is also faster to prepare as it requires no corpus data. 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.

The use of SMOTE methods was possible, because, in our setup, the model receives full data about the training sample in one step. The use of sequential methods with synthetic data generated by the SMOTE methods would carry too high a risk of overfitting the model due to the noise caused by synthetic fragments of the training sequences, potentially very different from the original data. In our approach, the generated synthetic data significantly improved the effectiveness of recognition of incorrect cases. In future work, we want to apply our methods in the multilingual MWEs detection, and to explore the transfer learning mechanism in a language-independent MWE detection.

Acknowledgements

This work was financed by the National Science Centre, Poland, project no. 2019/33/B/HS2/02814.

References

1. Abadi, M., Agarwal, A., Barham, P., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/
2. Agrawal, S., Sanyal, R., Sanyal, S.: Hybrid method for automatic extraction of multiword expressions. Int. Journal of Engineering & Technology **7**, 33 (2018)
3. Anke, L.E., Schockaert, S., Wanner, L.: Collocation classification with unsupervised relation vectors. In: Proc. of the 57th Annual Meeting of the ACL (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

4. Berk, G., Erden, B., Güngör, T.: 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). pp. 248–253. ACL (2018)
5. Bojanowski, P., Grave, E., Joulin, A., et al.: Enriching word vectors with subword information. Transactions of the ACL **5**, 135–146 (2017)
6. Boros, T., Burtica, R.: 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). pp. 254–260. ACL (2018)
7. Broda, B., Derwojedowa, M., Piasecki, M.: Recognition of structured collocations in an inflective language. Systems Science **34**(4), 27–36 (2008)
8. Buljan, M., Šnajder, J.: Combining linguistic features for the detection of Croatian multiword expressions. In: Proc. of the 13th Workshop on Multiword Expressions (MWE 2017). pp. 194–199. ACL (2017)
9. Chakraborty, S., Cougias, D., Piliero, S.: Identification of multiword expressions using transformers (2020)
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., et al.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
11. Dziob, A., Piasecki, M., Rudnicka, E.K.: plwordnet 4.1 - a linguistically motivated, corpus-based bilingual resource. In: Proc. of the Tenth Global Wordnet Conference: July 23-27, 2019, Wrocław (Poland). pp. 353–362 (2019)
12. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Univ. of Stuttgart (2004)
13. Fu, R., Guo, J., et al.: Learning semantic hierarchies via word embeddings. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 1199–1209. Baltimore, Maryland (2014)
14. Green, S., de Marneffe, M.C., Manning, C.D.: Parsing models for identifying multiword expressions. Computational Linguistics **39**(1), 195–227 (2013)
15. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. pp. 878–887 (2005)
16. Handler, A., Denny, M., Wallach, H., et al.: Bag of what? simple noun phrase extraction for text analysis. In: NLP+CSS@EMNLP (2016)
17. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks. pp. 1322–1328. IEEE (2008)
18. Hosseini, M.J., Smith, N.A., Lee, S.I.: 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). pp. 931–936. ACL (2016)
19. Klyueva, N., Doucet, A., Straka, M.: Neural networks for multi-word expression detection. In: Proc. of the 13th Workshop on Multiword Expressions (MWE 2017). pp. 60–65. ACL (2017)
20. Kocoń, J., Gawor, M.: Evaluating kgr10 polish word embeddings in the recognition of temporal expressions using bilstm-crf. Schedae Informaticae **27** (2018)
21. Kurfalı, M.: TRAVIS at PARSEME shared task 2020: How good is (m)BERT at seeing the unseen? In: Proc. of the Joint Workshop on Multiword Expressions and Electronic Lexicons. pp. 136–141. ACL (2020)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18

22. Levy, O., Remus, S., et al.: Do supervised distributional methods really learn lexical inference relations? In: Proc. of the 2015 Conference of the North American Chapter of ACL: Human Language Technologies. pp. 970–976 (2015)
23. Maziarz, M., Szpakowicz, S., Piasecki, M.: A procedural definition of multiword lexical units. In: Mitkov, R., Angelova, G., Boncheva, K. (eds.) Proc. of the International Conference Recent Advances in Natural Language Processing – RANLP'2015. pp. 427–435. INCOMA Ltd. (2015)
24. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013
25. Mroczkowski, R., Rybak, P., Wróblewska, A., et al.: HerBERT: Efficiently pretrained transformer-based language model for Polish. In: Proc. of the 8th Workshop on Balto-Slavic Natural Language Processing. pp. 1–10. ACL (2021)
26. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
27. Pečina, P.: Lexical association measures and collocation extraction. Language Resources and Evaluation **44**, 137–158 (2010)
28. Piasecki, M., Wendelberger, M., Maziarz, M.: Extraction of the multi-word lexical units in the perspective of the wordnet expansion. In: Proc. of the International Conference Recent Advances in Natural Language Processing (2015)
29. Ramisch, C.: Multiword Expressions Acquisition. A Generic and Open Framework. Springer (2015)
30. Ramisch, C., et al.: Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In: Proc. of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (2018)
31. Ramisch, C., Savary, A., Guillaume, B., et al.: Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In: Proc. of the Joint Workshop on Multiword Expressions and Electronic Lexicons (2020)
32. Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Third Workshop on Very Large Corpora (1995)
33. Rohanian, O., Taslimipoor, S., Kouchaki, S., et al.: 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). pp. 2692–2698. ACL (2019)
34. Saied, H.A., Candito, M., Constant, M.: Comparing linear and neural models for competitive MWE identification. In: Proc. of the 22nd Nordic Conference on Computational Linguistics. pp. 86–96. Linköping University Electronic Press (2019)
35. Scholivet, M., Ramisch, C.: Identification of Ambiguous Multiword Expressions Using Sequence Models and Lexical Resources. In: Proc. of the 13th Workshop on Multiword Expressions (MWE 2017). pp. 167 – 175 (2017)
36. Seretan, V.: Syntax-Based Collocation Extraction, Text, Speech and Language Technology, vol. 44. Springer Netherlands (2011)
37. Spasić, I., Owen, D., Knight, D., et al.: Unsupervised multi-word term recognition in Welsh. In: Proc. of the Celtic Language Technology Workshop. pp. 1–6 (2019)
38. Taslimipoor, S., Bahaadini, S., Kochmar, E.: MTLB-STRUCT @Parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In: Proc. of the Joint Workshop on Multiword Expressions and Electronic Lexicons. pp. 142–148. ACL (2020)
39. Yirmibeşoğlu, Z., Güngör, T.: ERMI at PARSEME shared task 2020: Embedding-rich multiword expression identification. In: Proc. of the Joint Workshop on Multiword Expressions and Electronic Lexicons. pp. 130–135. ACL (2020)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_18