# Neural Networks for Multi-Word Expression Detection

**Natalia Klyueva[1], Antoine Doucet[2] and Milan Straka[1]**

[1]Charles University, Prague, Czech Republic
[2]University of La Rochelle, France
{klyueva,straka}@ufal.mff.cuni.cz
{antoine.doucet}@univ-lr.fr

## Abstract

In this paper we describe the MUMULS system that participated to the 2017 shared task on automatic identification of verbal multiword expressions (VMWEs). The MUMULS system was implemented using a supervised approach based on recurrent neural networks using the open source library TensorFlow. The model was trained on a data set containing annotated VMWEs as well as morphological and syntactic information. The MUMULS system performed the identification of VMWEs in 15 languages, it was one of few systems that could categorize VMWEs type in nearly all languages.

## 1 Introduction

Multiword expressions (MWEs) present groups of words in which the meaning of the whole is not derived from the meaning of its parts. The task of processing multiword expressions is crucial in many NLP areas, such as machine translation, terminology extraction etc.

This paper describes the MUMULS system[1] which was evaluated through its participation to the PARSEME shared task on automatic identification of verbal MWEs[2] (VMWEs).

The experimental data set of the shared task is the result of a massive collaborative effort that produced training and evaluation data sets, available in 18 languages. The subsequent corpus was built by experts of each of the languages who manually annotated all VMWEs. The training and test sets respectively consist of a total of about 4.5 and 0.9 million tokens, containing 52,724 and 9,494 annotated VMWEs.

For most languages, a `.conllu` file provided morphological and syntactic information for each token. In addition, the training data set was indicating for each token, whether it belonged to an MWE, which one, and the type of that MWE. The MWE types are IReflV(inherently reflexive verb), LVC (light verb construction), VPC (Verb-particle construction), ID (idiomatic expression) and OTH - other types.

The goal of systems is to identify the VMWEs from text and to recognize to what type they belong. The data set and full evaluation procedure is more extensively described in the overview paper of the PARSEME shared task(Savary et al., 2017).

Since MUMULS did not make use of any other resources than those provided by the shared task organisers, the system participated in the "closed track" (as opposed to the open track, in which participants could make use of any external resources).

The rest of the paper is organised as follows. Section 2 describes the MUMULS system. We then present the results (Section 3) which are analysed in Section 4, before we conclude and suggest future works.

## 2 System description

For the task of automatic detection of multiword expression researchers use language-independent approaches that combine association measures like mutual information or dice

[1]MUltilingual MULtiword Sequences
[2]`http://multiword.sourceforge.net/sharedtask2017`

coefficient with machine learning approaches (Tsvetkov and Wintner, 2011), (Pecina, 2008). Neural networks were exploited in a number of papers for the task very related to ours, e.g. (Martínez-Santiago et al., 2002). Our system does not directly use the techniques presented in the mentioned papers, but some ideas behind are very similar to ours. Now that the annotated data described above are available for multiple languages, the natural thing is to exploit is a supervised approach, for which we have chosen deep artificial neural networks.

Deep learning algorithms have recently been applied to a vast majority of NLP tasks. Several frameworks to train deep models were introduced that simplify a lot the deploying process, like Theano, Torch, CNTK and recently an open source framework from Google TensorFlow,[3] which we used for training our MWE tagger, called *mwe_tagger*.[4]

Generally the task at hand resembles POS tagging, with inputs as various columns from them the CoNLL-U files, and outputs as the respective mwe tags from parsemetsv files.

Our model is based on a bi-directional recurrent neural network (Graves and Schmidhuber, 2005) with gated-recurrent units – GRUs (Cho et al., 2014). In (Chung et al., 2014) the GRUs performance is empirically evaluated and demonstrates sufficient results for long distance dependencies, which is especially important for processing discontinuous MWEs.

The linguistic attributes (features) used to predict the output tag and the output tag itself is extracted from the training data files `train.conllu` and `train.parsemetsv` combined and transformed into the following form (example for French):

```
Steffi  Steffi  PROPN  _
rend    rendre  VERB   LVC
visite  visite  NOUN   CONT
à       à       ADP    _
Monica  Monica  PROPN  _
```

Our model cannot take into account the numbering of MWEs in case more of them are present in one sentence, and we delete the numbers leaving only the name of MWE tags and substituting the continuation of the MWE with the symbol CONT.[5] For Romanian, the extended POS tag with more morphological features was used instead of UPOS tag. If the CoNLL-U file was not provided for a language, the lemma/POS attributes were substituted by underscores.

In the neural network, every input word is represented as a concatenation of embeddings of its form, lemma and POS tag. We use randomly initialized embeddings with dimension 100 for those three attributes.

We then process the words using a bi-directional recurrent neural network with single-layer GRUs of 100 cells. Finally we map the results for each word to an output layer with softmax activation function returning the distribution of possible output tags.

The network is trained using Adam optimizer (Kingma and Ba, 2014) to minimize the cross-entropy loss, using fixed learning rate of 0.001 and default hyperparameters.

The model was trained using batches of 64 sentences, for 14 epochs. Increasing the number of epochs or batch size did not lead to any improvement in the accuracy.

We trained the model on a cluster with multi-core CPU machines with 8 parallel threads.

The converted data were split into training, development and test sets to set the initial model, taking the first 10 % of the corpus as a development set, consequent 80 % as a training set and the last 10 % as a test set. We did not perform any cross-validation using different parts for train, test and dev while training which may result in poor score for some languages when the blind test data might be very different from the training. The final model that was used to tag the blind test data was trained on the joined train and test sets from the initial experiments, with the development set staying the same.

The final evaluation of the system was made by the script provided by the organizers which measures precision, recall and F-score for token-based and MWE-based predictions.

[3] `www.tensorflow.org`

[4] The scripts are available at `https://github.com/natalink/mwe_sharedtask`

[5] Our architecture unfortunately does not allow to handle properly neither embedding nor overlapping of MWEs.

## 3 Experiment Results

Table 1 presents the results of the MUMULS system for all the languages for which it produced non-zero results. Out of 18 available languages, MUMULS was experimented over 17. We found the bug that was introduced during data pre-processing for Czech language that caused recall issues, the re-trained model with very same setup as for other languages had higher score, which we additionally included in the result table. We did not include the languages for which we were not able to produce any predictions.

Table 2 provides the accuracy in terms of f-measure for the individual types of VMWEs. It can be seen that the system scored better in more 'syntactic' MWEs like IReflV, LVC or VPC, and generally (with the exception of French) the score for those categories is higher than for idioms.

## 4 Linguistic evaluation

We provide a short errors analysis for a couple of languages looking for possible reasons for the errors in tagging. Just to note, we do not do any statistical analysis, rather just observations on the test data.

Those observations should be taken with caution because slightly changing parameters of the algorithm may lead to different annotations (tags), making the provided observations inappropriate.

### 4.1 MWEs not seen in the training data

We did not use cross-validation, and one of the natural questions is how much the model overfit the training data and fail to generalize. Next are the examples of MWEs which are not present in the training data, but a construction was tagged in the test:

- Czech LVC: *přicházet s náměty* – 'come with proposals'. In training data a very similar construction with a synonymous predicative noun *přicházet s návrhy* – 'come with suggestions' is annotated, whereas in gold test the first one is not
- Bulgarian IReflV: The verb *се консултира* – 'consulting' is not in train.parsemetsv, but yet marked by the mwe_tagger.

Thus, we can say that the mwe_tagger can make generalizations to some extent.

### 4.2 Analysis of distinct types of MWEs

We observe the following errors for several MWE types and for several languages:

- not all the tokens of an MWE are marked. This entails the difference between MWE-based and token-based scores from the Table 2. Examples:
  – In Czech the verb is marked as reflexive, but the particle is not tagged as the continuation of the MWE
  – Some of the LVC part is not tagged, generally it is a predicative noun. E.g. in Polish *mieć problem* – 'have problem' the word *problem* was not tagged.
  – The particular case is analytical tense formation, like e.g. future tense in Czech. In the MWE *se bude hodit* – 'will be useful' mwe_tagger marked only the reflexive particle and the verb, but not the auxiliary verb *bude* – 'will be' which has to be annotated according to the annotation guidelines, so it was also penalized by the evaluation script.

- a token is marked as MWE, while it should not.
  – Often the reason is that some similar construction is tagged in the training data, e.g. in French *Comment Angiox agit -il* – 'How does Angiox work' learned from numerous examples of an idiom *il s'agit* – 'it's about'.
  – Sometimes more tokens around LVC are marked without any logical explanation. In Polish, *po zgaszeniu-LVC zadawał-LVC pytanie-LVC* – 'after switching_off (he) put question' the word totally unrelated to LVC was marked, while it did not occur at all in the training data.

In addition to the above, we present observations on individual MWE types and the issues our tagger had with them.

| Lang | P-MWE | R-MWE | F-MWE | P-token | R-token | F-token | Rank-MWE | Rank-token |
|---|---|---|---|---|---|---|---|---|
| DE | 0.3277 | 0.1560 | 0.2114 | 0.6988 | 0.2286 | 0.3445 | 3 | 3 |
| BG | 0.3581 | 0.3362 | 0.3468 | 0.7686 | 0.4809 | 0.5916 | 2 | 2 |
| CS | 0.4413 | 0.1028 | 0.1667 | 0.7747 | 0.1387 | 0.2352 | 4 | 4 |
| CS-fixed | 0.6241 | 0.6875 | 0.6548 | 0.7629 | 0.7784 | 0.7705 | 1 | 1 |
| PL | 0.6562 | 0.5460 | 0.5961 | 0.8310 | 0.6013 | 0.6977 | 3 | 3 |
| SL | 0.3557 | 0.2760 | 0.3108 | 0.6142 | 0.3628 | 0.4562 | 3 | 2 |
| ES | 0.3673 | 0.3100 | 0.3362 | 0.6252 | 0.3995 | 0.4875 | 3 | 3 |
| FR | 0.1466 | 0.0680 | 0.0929 | 0.5089 | 0.2067 | 0.2940 | 5 | 4 |
| PT | 0.5358 | 0.3740 | 0.4405 | 0.8247 | 0.4717 | 0.6001 | 3 | 3 |
| RO | 0.7683 | 0.7760 | 0.7721 | 0.8620 | 0.8112 | 0.8358 | 2 | 1 |
| EL | 0.2087 | 0.2580 | 0.2308 | 0.4294 | 0.4143 | 0.4217 | 4 | 3 |
| HU | 0.6291 | 0.6152 | 0.6221 | 0.7132 | 0.6657 | 0.6886 | 4 | 1 |
| TR | 0.4557 | 0.2774 | 0.3449 | 0.6452 | 0.3502 | 0.4540 | 4 | 4 |

Table 1: Results of MUMULS, organized by language groups, separated by horizontal lines (Germanic, Slavic, Romance, others).

| | LVC | | IReflV | | VPC | | ID | | OTH | |
|---|---|---|---|---|---|---|---|---|---|---|
| Lang | mwe | token | mwe | token | mwe | token | mwe | token | mwe | token |
| DE | 0.00 | 0.06 | 0.01 | 0.02 | 0.14 | 0.24 | 0.04 | 0.22 | 0.00 | 0.00 |
| BG | 0.00 | 0.08 | 0.41 | 0.64 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 |
| CS | 0.12 | 0.19 | 0.65 | **0.73** | 0.00 | 0.00 | 0.05 | 0.11 | 0.00 | 0.00 |
| PL | 0.18 | 0.28 | 0.53 | 0.61 | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 |
| SL | 0.00 | 0.01 | 0.33 | 0.47 | 0.10 | 0.20 | 0.02 | 0.09 | 0.00 | 0.00 |
| ES | 0.10 | 0.18 | 0.31 | 0.42 | 0.00 | 0.00 | 0.06 | 0.18 | 0.00 | 0.00 |
| FR | 0.02 | 0.09 | 0.02 | 0.11 | 0.00 | 0.00 | 0.12 | **0.33** | 0.00 | 0.00 |
| PT | 0.37 | **0.49** | 0.12 | 0.18 | 0.00 | 0.00 | 0.06 | 0.18 | 0.00 | 0.00 |
| RO | 0.26 | 0.35 | 0.55 | 0.62 | 0.00 | 0.00 | 0.03 | 0.13 | 0.00 | 0.00 |
| EL | 0.16 | 0.30 | 0.00 | 0.00 | 0.03 | 0.04 | 0.03 | 0.15 | 0.00 | 0.03 |
| HE | 0.07 | 0.16 | 0.00 | 0.00 | 0.06 | 0.13 | 0.00 | 0.03 | 0.03 | 0.12 |
| HU | 0.12 | 0.29 | 0.00 | 0.00 | 0.37 | **0.51** | 0.00 | 0.00 | 0.00 | 0.00 |
| TR | 0.30 | 0.38 | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.24 | 0.09 | 0.14 |

Table 2: F-score for the distinct MWE categories

#### 4.2.1 IReflV

IREflV is the most frequent MWE tag, and it is relatively easy to identify reflexives in the text with the help of some rules. However, the mwe_tagger encountered several problems that we will demonstrate for a few languages:

- it is hard for an algorithm to distinguish between inherently reflexive verbs and other very structurally similar "deagentive", passive or reciprocal constructions, more see (Kettnerová and Lopatková, 2014), (Bejček et al., 2017) or the guidelines manual[6]. E.g. in Bulgarian, *ce убедят* – '(they will) be convinced' was tagged by the mwe_tagger, but it was just passivisation from *убедят* – 'convince', not the true reflexive verb. In Polish *oblizując się* – 'licking (lips)' was also tagged, whereas it should not according to the guidelines definition.
- For French, there are two forms that clitic takes - full and contracted (in case it comes before a vowel) This might lead to some bias and thus influence the prediction results.
- For Portuguese, the system was supposedly confused by the clitic being either 1) separated by a hyphen within one token or 2) with a hyphen ending the verb and clitic on the next line: e.g. MWEs *refiro-me* – 'refer', *corresponder- se(next token)* – 'correspond' were not marked as such by mwetagger. The verb-clitic IReflV as two separate tokens without a hyphen were generally tagged by the system properly.

Overall, it seems like inherently reflexive verbs are more probable to be detected correctly for Slavic languages with the exception of Romanian. We can suggest that for Slavic languages the role of clitics is different than

[6] `http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/?page=060_Specific_tests_-_categorize_VMWEs/040_Inherently_reflexive_verbs`

that in Romance languages, but that claim will need more thorough analysis of the annotated data.

#### 4.2.2 LVC

The second most frequent MWE tag was LVC - light verb construction - an MWE generally formed by a verb and a noun where the verb looses its initial meaning and the whole construction takes the semantics of the noun. There are no consistent criteria on which expressions should be considered as LVC, and for this shared task the special tests were created on how to distinguish LVC from non-LVC.

Below are some examples of how the tagger tackled LVCs for different languages.

- Some LVC tokens might be marked as idioms (ID). In Czech,e.g. *dali pokoj* – 'lit. give piece – let alone' was predicted as LVC, whereas it is marked as idiom in the gold test file.
- In some cases the LVC are not marked, even though they are present in the training data, like LVC in Romanian *face referire* – 'referred to' was not tagged, though was quite frequent in the training data
- Discontinuous LVCs where the components are separated by a number of other tokens, are often not detected. E.g. in Romanian *pune astfel accent* – 'put such emphasis' only one word in between the LVC components led to the predicative noun not to be tagged

In general, the score for LVC predictions is lower than that for IReflV.

#### 4.2.3 Idioms

ID - idiom - was a tag which was very hard to detect. The F-measure for this tag never got more than 0.3 (for French), it was 0.1 in average. We have studied a Czech output file and all the idioms were coming from the training data.

The generalizations like in the case of IReflV or LVC constructions will not work and are not desirable in this case as this can lead to improper tagging, like in the following example in Czech. *nestál na vrcholu* – '(did not) stand on the top' was detected as an idiom(ID), though the meaning was literal in this case (*stand on the mountain top*). probably from one single example from the training data: *dosahnout vrcholu* – 'reach the top'. For French language, the detection of idioms worked better than that for other categories. This may be, above all, attributed to the fact that idioms annotated in French were quite frequent in the training data, e.g. *il faut* – 'it is necessary' or *pris en compte* – 'take into account'.

For proper handling of the idioms, using special lexical resources will be the most efficient measure.

## 5 Conclusion

We have presented the system MUMULS that participated in the shared task of identification of MWEs. MUMULS was a neural network deployed within the framework TensorFlow that learned to detect MWEs based on manually annotated corpora. Overall, the systems participating in the closed track for some languages have approximately the same F-score while for others it may vary. The results of the shared task might as well depend on the consistency and quality of the annotations of the training data.

We are waiting for further details on other approaches so as to be able to better understand why our system outperformed other systems for some languages, and why it underperformed for some others.

## Acknowledgments

The research was held during the PARSEME Short Term Scientific Mission funded by the grant IC1207-070117-081755. This work has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071)

## References

Eduard Bejček, Jan Hajič, Pavel Straňák, and Zdeňka Urešová. 2017. Extracting verbal multiword data from rich treebank annotation. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT 15)*, pages 13–24. Indiana University, Bloomington, Indiana University, Bloomington.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio, 2014. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, chapter On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, pages 103–111. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Václava Kettnerová and Markéta Lopatková. 2014. Reflexive verbs in a valency lexicon: The case of czech reflexive morphemes. In Andrea Abel, Chiara Vettori, and Natascia Ralli, editors, *Proceedings of the XVI EURALEX International Congress: The User in Focus*, pages 1007–1023, Bolzano/Bozen, Italy. EURAC research, EURAC research.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Fernando Martínez-Santiago, Manuel Carlos Díaz-Galiano, Maite Teresa Martín-Valdivia, Víctor Manuel Rivas-Santos, and Luis Alfonso Ureña-López. 2002. Using neural networks for multiword recognition in ir. In *Proceedings of Conference of International Society of Knowledge Organization (ISKO-02), Granada, Spain*, pages 559–564.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions*, pages 54–57, Marrakech, Morocco. ELRA.

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

Yulia Tsvetkov and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 836–845. Association for Computational Linguistics.