COST Action: IC1207
STSM title: Enchancing the Maximal Frequent Sequences method with morphosyntactic relations for verbal MWEs identification
Reference:
ECOST-STSM-IC1207-070117-081755
STSM dates: from 07-01-2017 to 17-01-2017
Location: University of La Rochelle, La Rochelle, France

STSM grantee: Natalia Klyueva
Charles University in Prague
E-mail: klyueva@ufal.mff.cuni.cz

Host: Prof Antoine Doucet
University of La Rochelle
E-mail: antoine.doucet@univ-lr.fr

The main goal of the Short Term Scientific Mission in the Laboratory L3i in University of La Rochelle was to create a system of automatic detection of verbal Multiword Expressions, participate in the shared task on automatic identification of verbal multiword expressions[1] and to write a system description paper. The system MUMULS [2] participated in the 'closed track' as we did not use any additional resources other than those provided within the shared task.

According to the plan, we exploited MFS algorithm, and run it on the data from the shared task. The input was presented as a text with words substituted by lemmas+ underscore + their morphological tags.

At the same time, the participant started to work with Neural Networks[3] setting the baseline for the shared task. It soon became apparent that much more effort and time was required to integrate MFS with TensorFlow, so we proceeded working

[1] http://multiword.sourceforge.net/sharedtask2017

[2] MUltilingual MULtiword Sequences

[3] Colleagues from L3I gave the participant some valuable advice on training NN within TensorFlow

solely on Neural Networks. The detailed description of the system and the results are presented below, the text largely intersects with the paper submitted to the MWE workshop. The system is open source, the scripts are available at `https://github.com/natalink/mwe_sharedtask/`.

1 System description

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

The underlying concept of the task ensembles the pos-tagging, with inputs as attributes from the conllu files, and outputs as the respective mwe tags from parsemetsv files. As this model can not take into account the numbering of MWEs in case more of them are present in one sentence, we deleted the numbers leaving only the name of MWE tags (IREflV - Inherently reflexive verbs, ID - idioms etc.) and substituting the continuation of the MWE with the symbol CONT. The input factors included word, lemma and pos tag as those attributes were present in almost all languages. For Romanian, the extended pos tag with more morphological features was used as the one extracted from the respective position in the conllu file. If the conllu file was not provided for a language, the lemma/pos positions were substituted by the underscore. Example of the training file for French:

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

The nature of the experiment unfortunately does not allow to handle properly neither embedding nor overlapping of MWEs.

The converted data were split into training, development and test sets to set the initial model. The final model that was used to tag the blind test data was trained on all the training data.

We represent each input word as a concatenation of embeddings of its form, lemma and POS tag. We used randomly initialized word embeddings for those three attributes exploiting the tensorflow implementation of word embedding algorithm[5], without pre-trained models.

We then process the words using a bi-directional RNN with GRU units and map the results for each word to an output layer with softmax activation function returning the distribution of possible output tags.

[4] `www.tensorflow.org`

[5] https://www.tensorflow.org/tutorials/word2vec/

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
LT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	n/a	n/a
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
HE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	n/a	n/a
HU	0.6291	0.6152	0.6221	0.7132	0.6657	0.6886	4	1
MT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	n/a	n/a
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, balto-slavic, romance, others).

The loss was defined as softmax loss cross entropy function, and we used Adam optimizer to minimize the loss.

As for the hardware, we exploited the multi-core CPU cluster machines with 8 parallel threads .

2 Experiment Results

Table 1 presents the results of the MUMULS system for all the languages for which it produced results. Out of 18 available languages, MUMULS was experimented over 17. Only after the announcing the results 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.

In the paper submitted to the workshop we provided detailed linguistic analysis of errors that our system made.

3 Conclusion and results

The main output of the STSM:

- the participant got hands-on experience on training Neural Networks using TensorFlow sharpening her programming skills

- the system participated in the shared task and for Romanian language it scored the first
- we wrote the system description paper which was accepted to EACL workshop

4 Future collaboration

The participant established close connections in L3i Laboratory, the future collaboration might include further work on Neural Networks, better pre-processing of the data and some fine-tuning like dropout and other measures to increase the score.

5 Confirmation of the host

This is to confirm that Natalia Klyueva from Charles University in Prague (Czech Republic) visited the University of La Rochelle in France in January 2017 for an STSM as planned. Her research visit allowed for great collaboration and networking. Its direct outcome was the design and implementation of a system detecting verbal MWEs that participated in the Parseme shared task in most languages. It obtained the best overall results in Romanian and is one of the few that are able to predict VMWE category. A paper to describe the system was recently accepted for publication at the MWE workshop of EACL 2017.