STSM Report

Handling MWEs in neural machine translation systems

Participant: Matīss Rikters
University of Latvia
E-Mail: matiss@lielakeda.lv

Host: Ondrej Bojar
Institute of formal and Applied Linguistics
Charles University in Prague
E-Mail: bojar@ufal.mff.cuni.cz

1. Purpose of the STSM

The goal of the visit was to experiment with the way multi-word expressions (MWEs) are handled in neural machine translation (MT) systems. It is well known that neural MT has defined the new state of the art in the last few years, but the many specific aspects of neural MT outputs are not yet explored.

Depending on properties like morphological richness of the languages in question, MWEs may be harder to memorize for neural MT, because it represents the whole sentence in a high-dimensional vector. Standard phrase-based MT will be able to copy MWEs verbatim but will probably suffer in grammaticality.

During the visit, together with scientists from the institute, we aimed to compare how neural MT pays attention to MWEs during translation, using a test set particularly targeted at handling of MWEs, and if that can be improved by populating the training data for the NMT systems with parallel corpora of MWEs.

Depending on the outcomes of these experiments, we would like to propose new ways of handling MWEs in neural MT.

2. Description of the work carried out during the STSM

The final target objective of this mission was to obtain a comparison of how NMT with regular training data and NMT with synthetic MWE data pays attention to MWEs during the translation process as well as to improve the final NMT output. To achieve this objective, it needed to be broken down into smaller sub-objectives:

- train baseline NMT systems
- extract parallel MWE corpora from the training data
- train the NMT systems with synthetic MWE data
- develop a tool for inspecting alignments produced by the NMT

2.1. Baseline NMT

The experiments were conducted on English – Czech and English – Latvian language pairs. To be able to compare the results with other MT systems, training and development corpora were used from the WMT[1] shared tasks – data from the News task for English – Latvian and data from the Neural MT Training task for English –

[1] SECOND CONFERENCE ON MACHINE TRANSLATION (WMT17) - http://www.statmt.org/wmt17/

Czech. The English – Czech data consists of about 49 million parallel sentence pairs and the English – Latvian – about 4.5 million.

For training the NMT systems Neural Monkey – an open-source NMT system developed in the institute - was used with configurations from the WMT17 Neural MT Training task 8GB setup for English – Czech and 4GB setup for English – Latvian.

2.2. Parallel MWE Extraction

To extract MWEs, the corpora were first tagged with morphological taggers (UDPipe[2] (Straka et al., 2016) for English and Czech, LV Tagger[3] (Paikens et al., 2013) for Latvian), then processed with the Multiword Expressions toolkit[4] (Ramisch, 2012) and finally aligned with the MPAligner[5] (Pinnis, 2013) (intermittently pre-processing and post-processing with a set of custom tools[6]). To extract MWEs from the corpora with the MWE Toolkit, patterns were required for each of the involved languages. Patterns from Skadiņa (2016) were used for Latvian (210 patterns) and English (57 patterns) languages and patterns from Majchráková et al. (2012) and Pecina (2008) for Czech (23 patterns). This workflow allowed to extract a parallel corpus of about 400 000 multiword expressions for English – Czech and about 60 000 for English – Latvian.

2.3. NMT with synthetic MWE data

The final training datasets for training NMT systems with synthetic data were composed of parts from the baseline sets mixed with the extracted parallel MWE corpora. This was done due to the fact that for a neural network to learn something better, it is beneficial to show the specific examples again and again.

Figures 1 and 2 illustrate how the training data was divided into portions, where 1xMWE corresponds to a full set of extracted MWEs (400K for En-Cs, 60K for En-Lv) and 2xMWE – a double set of MWEs (800K for En-Cs, 120K for En-Lv). For En-Lv the full corpus was used, but for En-Cs – only the first 15 million sentences, due to it being too large to train multiple epochs on the available hardware effectively. The MWEs get repeated five times in both language pairs. By doing this the En-Cs data set was reduced from 49M to 17M and the En-Lv data set increased to 4.8M parallel sentences.

Figure 1: Portions of the final training data set for English-Czech

[2] UDPipe - https://ufal.mff.cuni.cz/udpipe
[3] Latvian morphological tagger - https://github.com/PeterisP/LVTagger
[4] The Multiword Expressions toolkit - http://mwetoolkit.sourceforge.net
[5] MPAligner - https://github.com/pmarcis/mp-aligner
[6] MWE-Tools - https://github.com/M4t1ss/MWE-Tools

Figure 2: Portions of the final training data set for English-Latvian

2.4. Alignment Inspection

For inspecting the alignments, a tool was developed that takes data that is produced by Neural Monkey - a 3d array (tensor) filled with the alignment probabilities, source and target byte pair encodings (BPEs) – and produces a soft alignment matrix (Figure 3) of the BPEs that highlights all BPEs that get attention when translating a specific BPE. The tool also allows to output the soft alignments in a different perspective of connections between BPEs as visible in Figures 5 and 6.

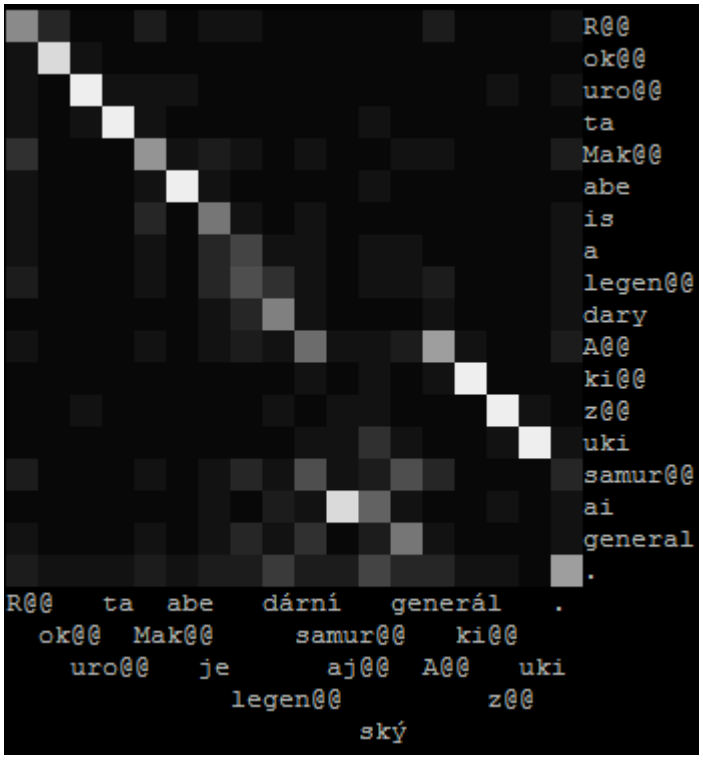

Figure 3: Soft alignment matrix

Source	It should be noted that this is not the first time that Facebook has been actively involved in determining what network users see in their news feeds .
Reference	Jāteic , ka šī nav pirmā reize , kad Facebook aktīvi iesaistās , nosakot , ko tīkla lietotāji redz savās jaunumu plūsmās .
MWEs	Ir jāatzīmē , ka šis ir pirmā reize , kad Facebook ir aktīvi iesaistījusies , nosakot to , ko tīkla lietotāji dara viņu ziņu formātā .
Baseline	Jāpiebilst , ka šis nav pirmais laiks , kas Facebook , ir aktīvi iesaistīts , lai noteiktu , kas ir pārziņa , kas ir viņu ziņas par viņu ziņas .

Figure 4: Differences of the example sentence

While the models are still being trained, a snapshot was copied to examine translations produced on a test set of sentences, of which each includes at least one MWE. One sentence was chosen for a closer check-up. It contains one MWE that was identified by the MWE Toolkit - “network users”. The differences in translations are highlighted in Figure 4. The alignment inspection tool allows to see that the baseline

NMT in Figure 5 has multiple faded alignment lines for both words “network” and “users”, which outlines that the neural network is unsure and looking all around for traces to the correct translation. However, in Figure 6 it is visible that both these words have strong alignment lines to the words “tīkla lietotāji”, that were also identified by the MWE Toolkit as an MWE.

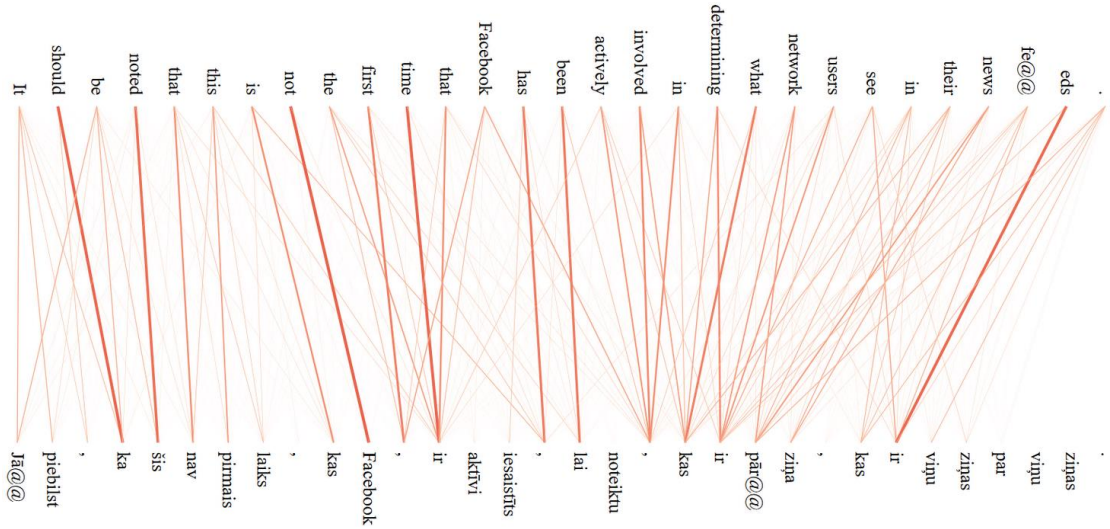

Figure 5: Soft alignments of the example sentence from the baseline NMT

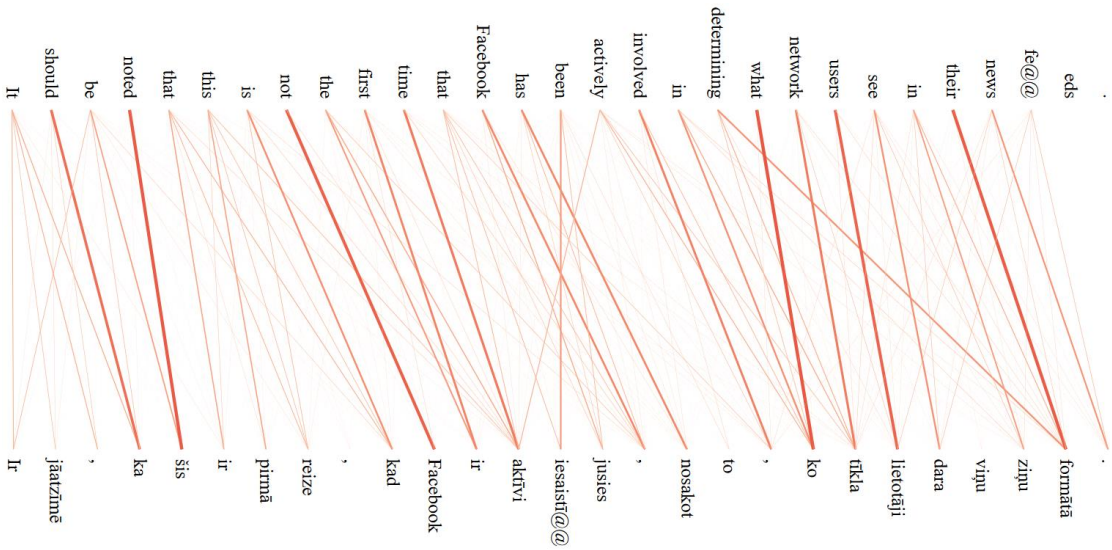

Figure 6: Soft alignments of the example sentence from NMT with synthetic MWE data

3. Description of the main results obtained

The main results consist of three parts:

- A polished workflow for obtaining parallel corpora of MWEs from parallel sentences. Available as a part of the MWE tools on GitHub - https://github.com/M4t1ss/MWE-Tools (MWE translation workflow).
- A tool for obtaining and inspecting visualizations of NMT soft alignments. Available on GitHub - https://github.com/M4t1ss/SoftAlignments.

- The ongoing NMT training experiments and possible submissions to the WMT17 Neural MT Training task, and a paper for NMT[7] or MWE[8] workshops.

4. Future collaboration with host institution

New and useful contacts have been established with the scientists of the institute. Currently, future collaboration is planned until the end of the WMT17 shared tasks. In addition to the NMT systems already training, at least one more hybrid MT system is planned to train for the English – Latvian News Translation Task of WMT17. That may involve a step for processing MWEs in which case the further collaboration would be required. After that, depending on the results, we might meet up in the WMT17 conference and discuss further intents of collaboration.

5. Foreseen publications to result from the STSM

If the Czech – English system produces better results than the baseline, it will be submitted to the WMT17 Neural MT Training task.

Also, when the NMT systems finish training, the final results and analysis will be summarized in a paper and submitted to the 1st Workshop on Neural Machine Translation[7] and/or the 3rd Workshop on Multi-Word Units in Machine Translation and Translation Technology [8] conference.

6. Confirmation by the host institution of the successful execution of the STSM

Matiss Rikters spent three weeks (Feb 13th till Mar 3rd) at our institute. The collaboration was very fruitful. Matiss learned quickly how to run Neural Monkey and started experimenting very early. This was particularly good because neural MT systems can take weeks to train and without an early start, no preliminary results would be available by this time at all. The outcomes of the stay are useful already now, for instance we will make use of the visualization tool. The experiments with using MWEs in neural MT are still rather intriguing than promising, but Matiss has made a great start and we will continue with the collaboration also remotely after Matiss' return home.

References

Majchráková, D., Dušek, O., Hajič, J., Karčová, A., & Garabík, R. (2012). Semi-automatic Detection of Multiword Expressions in the Slovak Dependency Treebank.

Paikens, P., Rituma, L., & Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*; May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16 (No. 085, pp. 267-277). Linköping University Electronic Press.

[7] 1st Workshop on Neural Machine Translation - https://sites.google.com/site/acl17nmt/
[8] The 3rd Workshop on Multi-Word Units in Machine Translation and Translation Technology - http://rgcl.wlv.ac.uk/europhras2017/mumttt-2017/

Pecina, P. (2008). Reference data for Czech collocation extraction. In *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008) (pp. 11-14)*.

Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *RANLP (pp. 562-570)*.

Ramisch, C. (2012). A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop (pp. 61-66)*. Association for Computational Linguistics.

Skadiņa, I. (2016). Multi-Word Expressions in English-Latvian SMT: Problems and Solutions. *Human Language Technologies–The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016 (Vol. 289)*. IOS Press.

Straka, M., Hajic, J., & Straková, J. (2016). UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016).