*Human Language Technologies – The Baltic Perspective*
*I. Skadiņa and R. Rozis (Eds.)*

*doi:10.3233/978-1-61499-701-6-84*

# Towards Hybrid Neural Machine Translation for English-Latvian

Mārcis PINNIS [1]
*Tilde, Latvia*

**Abstract.** This paper investigates a hybrid method for translation from English into Latvian by chaining an NMT system with an SMT system in order to cover out-of-vocabulary word translation. Different from other works, the primary translation is handled by the NMT system, and the SMT system acts as a secondary system. Automatic evaluation results have shown that the hybrid method allows improving NMT translation quality by up to three BLEU points.

**Keywords.** hybrid system, neural machine translation, English, Latvian

## 1. Introduction

Neural network ability to store deep sentence representations with long distance dependencies has motivated active research of neural network technologies in machine translation (MT) system development. Researchers have analysed different methods for neural network integration in existing statistical MT (SMT) models. For instance, neural network technologies can be used to improve concrete existing SMT solutions by a) replacing specific SMT models, for instance, language models [1], b) introducing new features in SMT models based on NN calculations [2], c) performing re-scoring of SMT translation hypotheses [3][4], and d) performing out-of-vocabulary word translation [5].

Recently, researchers have developed the first promising end-to-end neural machine translation (NMT) methods [6][7][8] in which a single neural network allows performing the whole translation (without involving other previous rule-based or statistical MT technologies). However, research has mainly focussed on large languages (e.g. English, Spanish and German), and there are minimal or no efforts carried out for smaller morphologically rich and highly inflected languages, such as Latvian, for which MT solution development is more complex due to (relatively) free word order and richness of surface forms (or word forms). Furthermore, the current word-level NMT methods are limited in terms of vocabulary size, which is an issue for Latvian as a morphologically rich language.

This paper investigates a hybrid method for translation from English into Latvian by chaining an NMT system together with an SMT system in order to cover out-of-vocabulary word translation. Different from other works, the primary translation will be handled by the NMT system, and the SMT system will act as a secondary system when the NMT system's translation produces unknown word tokens.

[1] Corresponding author, SIA Tilde, Vienibas gatve 75A, Riga, Latvia, LV-1004; E-mail: marcis.pinnis@tilde.lv

The paper is further structured as follows: Section 2 describes the data used in MT system training and evaluation, Section 3 describes NMT and SMT system training, Section 4 provides details about the hybridisation method, Section 5 describes a method for word alignment extraction from probabilistic word alignment weights acquired from the NMT system, section 6 provides evaluation results, and section 7 concludes the paper.

## 2. Data

In this paper, we present experiments with two different data sets: 1) smaller publicly available parallel corpora in the legislation domain and 2) relatively large broad domain parallel corpora consisting of publicly available and proprietary data. For training of the small NMT and SMT systems, the DGT-TM parallel corpora [9] (releases from 2007 to 2015) were used. Prior to training, the data was de-duplicated, cleaned, tokenised, and truecased on the LetsMT platform [10]. Then, non-translatable tokens (e.g. file paths, web site and e-mail addresses, XML tags, etc.) were identified and substituted with class labels. The statistics of both training data sets before and after filtering are given in Table 1.

**Table 1.** Training data statistics.

| | Sentences before filtering | Sentences after filtering |
|---|---|---|
| ***Data for small systems*** | | |
| Parallel corpus | 4,325,065 | 2,695,355 |
| Monolingual corpus | 4,325,065 | 2,310,979 |
| ***Data for large systems*** | | |
| Parallel corpus | 22,058,109 | 7,300,666 |
| Monolingual corpus | 246,044,055 | 74,741,452 |

For tuning of the small systems, 2,000 randomly selected sentences from the training data were used. For tuning of the large systems, a balanced tuning corpus[2] of 1,000 sentences was used. For evaluation of the small systems, we use 1,000 randomly selected sentences from the training data and a balanced (broad domain) evaluation corpus of 512 sentences. The broad domain corpus is also used for the evaluation of the large systems.

## 3. MT System Training

For NMT system development, we use the NMT toolkit *DL4MT-tutorial*[3] developed by Kyunghyun Cho and Orhan Firat, which allows training attention based encoder-decoder models with gated recurrent units. The small NMT system was trained using a vocabulary of the 40,000 most frequent tokens and a batch size of 12 sentences. The training of the NMT model (without data preparation) on an Nvidia GeForce 960 graphics card took approximately eight days using the Adam stochastic gradient-based

[2] The balanced tuning and evaluation corpora were created within the ACCURAT project (http://accurat-project.eu/).

[3] The DL4MT-Tutorial is an NMT toolkit that can be found online at: https://github.com/nyu-dl/dl4mt-tutorial. In the paper, we used the implementation of the commit 4377b76.

optimisation algorithm [11]. The large NMT system was trained using a vocabulary of the 100,000 most frequent tokens and a batch size of 32. The training on an Nvidia GTX Titan X graphics card took approximately five days using the Adam algorithm and four days using the Adadelta algorithm [12]. The parameters were selected to efficiently utilise the available resources of the graphics cards. Other parameters were not changed.

For SMT system development, we use the LestMT[4] platform that is based on the Moses [13] SMT system. Comparing to NMT systems, the complete training (with data preparation) of the small and large SMT systems took approximately 8.5 and 16.5 hours respectively.

## 4. NMT and SMT Hybrid System

Word level NMT systems can be limited by a vocabulary size that is significantly smaller than for SMT systems. For instance, due to restrictions of computation resources, NMT systems typically handle a vocabulary of up to 100,000 tokens [8]. In comparison, the small SMT system's translation model alone handles a vocabulary of 287,373 English and 449,767 Latvian tokens. The large SMT system's translation model handles 551,894 English and 899,086 Latvian tokens. Although there have recently been promising attempts to address this issue by segmenting words using byte-pair encoding [14] and by performing importance sampling [7], in this paper we will focus on pure NMT systems.

When an NMT system produces unknown word tokens in the output, the context around the tokens can still be translated correctly. In such situations (see Table 2 for an example), it would be natural to add a post-processing step that could fill the gaps (or in other words, to translate the unknown word tokens) by taking into account the already translated content. We see that this task could be performed by an SMT system, thereby creating a hybrid neural machine translation system.

**Table 2.** An example of unknown word tokens in NMT output.

| | |
|---|---|
| Source sentence: | model with pink and beige lines , four pages . |
| NMT output: | paraugs ar UNK un UNK līnijām , četras lappuses . |
| Hybrid translation: | paraugs ar rozā un smilškrāsas līnijām , četras lappuses . |

The hybrid neural machine translation method works as follows (see Table 3 for an example):

1) A sentence is pre-processed with LetsMT pre-processing workflows. I.e. the sentence is normalised (e.g. quotation marks and apostrophes are replaced with alternatives according to standard writing styles), tokenised, truecased and non-translatable tokens are replaced with class labels.
2) The pre-processed sentence is then translated with the NMT system.
3) The word alignment information between the source sentence and the translation provided by the NMT system is extracted from the probabilistic word alignment matrix. The example in Table 3 shows that functional word and auxiliary verb alignment is ambiguous in some cases. Because such words may transform into specific morphological characteristics (i.e. inflected forms

[4] The LetsMT platform can be found online at: www.letsmt.eu.

of nouns, verbs, etc.) in morphologically rich languages, these words may disappear in translations. Our experiments show that it is possible to handle such cases by: a) ignoring non-aligned words and b) identifying the closest content words to which these words can be assigned. The word alignment extraction method is described in more detail in Section 5.

4) The source sentence is transformed into Moses XML[5], where known words are marked with translations from the NMT system and unknown words are kept without translation hypotheses. For example, in the example provided in Table 3, the English phrase "*the workshop*" corresponds to the unknown word token "*UNK*" in the NMT system's output, and, therefore, it is not marked with Moses XML tags and will be translated with the SMT system. The phrases are ordered in the sequence as they appear in the translation of the NMT system.
5) The sentence is translated with the SMT system. When translating, phrase reordering is disabled because the NMT system already provides target language phrases in a possibly correct sequence.

**Table 3.** An example of a sentence translated with the hybrid NMT method; colours in the example indicate correct (green with a solid underline) and incorrect (red with a dotted underline) translations; orange (with a curly underline) depicts alternative (not correct, but in other contexts valid) translations.

| | |
|---|---|
| Source sentence | Accepted papers will be published in **the workshop** proceedings. |
| Pre-processed sentence (input for NMT) – normalised, tokenised, truecased and with identified non-translatable tokens | accepted papers will be published in **the workshop** proceedings . |
| NMT output | pieņemtie dokumenti tiks publicēti **UNK** gaitā . |
| Moses XML (input for SMT) – with phrases ordered identically to the NMT output. The unknown word is marked in bold. | <nmt translation="pieņemtie"> accepted </nmt> <nmt translation="dokumenti"> papers </nmt> <nmt translation="tiks"> be </nmt> <nmt translation="publicēti"> published </nmt> **the workshop** <nmt translation="gaitā"> in proceedings </nmt> <nmt translation="."> . will </nmt> |
| SMT output (the hybrid system) | Pieņemtie dokumenti tiks publicēti semināra norises gaitā. |
| SMT-only output | Pieņemt dokumenti tiks publicēti semināra norisi. |
| Reference sentence | Akceptētie darbi tiks publicēti semināra rakstu krājumā. |

## 5. Word Alignment Extraction

In order to chain the NMT and SMT systems together, it is important to acquire word alignment information that can be used to identify which unknown word tokens in the NMT system's output are linked to which tokens in the source sentence. The alignment information can be acquired from alignment weights (i.e. from the attention mechanism of the neural network) that are calculated by the NMT model [7]. This alignment information is different from the usual word alignment information that is provided by SMT systems in the sense that the alignment is probabilistic. I.e. a source word is linked to a target word with a weight from 0 to 1 (see Figure 1 for an example). This makes it difficult to extract the correct word alignment sequence.

[5] More details about the Moses XML mark-up can be found online at: http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc7.

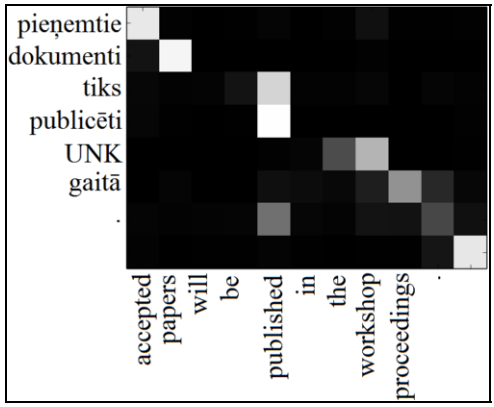


**Figure 1.** Example of probabilistic word alignment from an English-Latvian NMT system where unknown words are replaced with the token "UNK" (black corresponds to 0 and white to 1).

To acquire the word alignments from the probabilistic alignments, we developed a heuristic algorithm that operates as follows:

1) First, we identify cells in the alignment matrix that correspond to maximum source-to-target and target-to-source alignments (i.e. we find those source and target token pairs for which the alignment is unambiguous). In the example given in Figure 1, this would create the following alignments: "accepted→*pieņemtie*", "*papers→dokumenti*", "*published→*publicēti", "*workshop→UNK*" and "*proceedings→gaitā*".
2) Then, we perform a leftward and rightward search around each identified unambiguous target token to identify additional source tokens for which the maximum alignment weight is assigned to the target token (i.e. we expand the previously identified source and target token pairs with unambiguous many-to-one alignments). In the example given in Figure 1, this would create the alignment "the→*UNK*". In this step, we implemented two different sub-scenarios. It is possible that a target token is aligned to two or more source words between which there are other tokens. The first sub-scenario ("*with gaps*") does not allow pairing the split source tokens. The second scenario ("*without gaps*") allows pairing such source tokens. The two sub-scenarios are visually depicted in Figure 2.

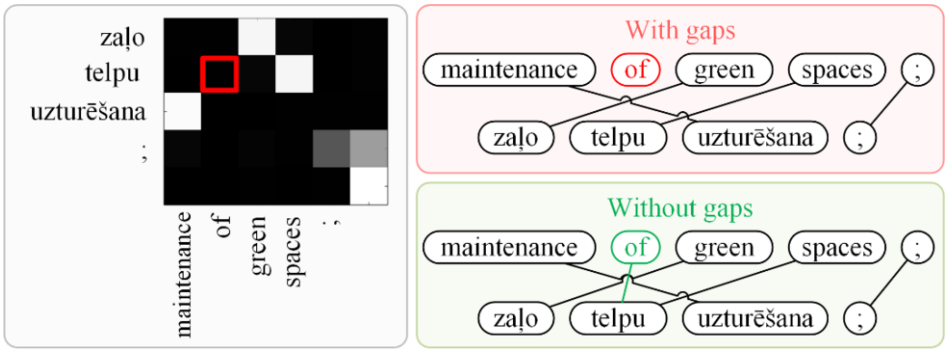


**Figure 2.** An example of extracted word alignments with and without gaps.

3) Next, among the remaining unpaired target and source tokens, we identify the respective source and target tokens with the highest alignment weights. This is the first ambiguous step as this alignment will not be reciprocal (i.e. for a yet unpaired target token, if we identify that a particular unpaired source token has the highest alignment weight, then the maximum alignment weight for the source token is with a different token). In the example given in Figure 1, this would create the alignments "*be→tiks*", "*in→gaitā*" and ".→.".
4) Next, the second step is performed for the target tokens that were paired in the third step. This allows capturing possibly missing many-to-one alignments. In the example given in Figure 1, this would create the alignment "*will→.*". This is obviously a false alignment. However, it will not affect the hybrid system's translation result as the target token is known. Only those source tokens that are paired with unknown target tokens will be able to affect the output of the hybrid system.
5) Finally, unpaired source tokens are assigned to the target tokens with the highest alignment weights. As the remaining source tokens often have low weights and could potentially be assigned to the wrong target tokens, this step is optional. Examples of word alignments where this step is enabled ("*with low confidence tokens*") and where it is not enabled are given in Figure 3.

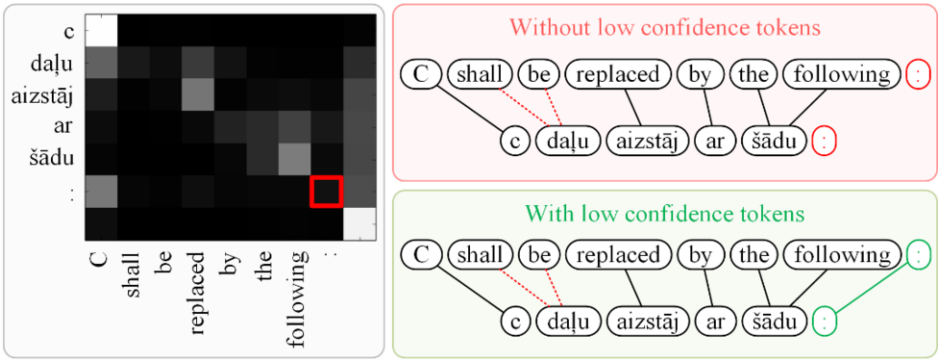


**Figure 3.** An example of extracted word alignments with and without low confidence tokens.

## 6. Evaluation

The automatic evaluation results for the small systems are given in Table 4. The results show that the NMT system produces better translations for out-of-domain sentences according to BLEU (although the confidence intervals do overlap, the results are significant with $p$ equal to 0.01). However, in terms of in-domain translation quality, the baseline SMT system achieves better results. The mixed domain evaluation set is comprised of 512 sentences from both the in-domain and out-of-domain evaluation sets.

The NMT system's translations contained 1,185 (5.34% of all) and 1,660 (16.35% of all) unknown word tokens in in-domain and out-of-domain data sets respectively. It is evident that the out-of-vocabulary rate for out-of-domain data is significantly (three times) higher. The results show that the hybrid system translations allow achieving an approximately 3 BLEU point increase and 2 BLEU point increase for in-domain and out-of-domain data sets respectively. This shows the usefulness of the hybrid method

for both in-domain and out-of-domain translation scenarios. The differences between the different word hybrid system word alignment extraction scenarios are insignificant.

**Table 4.** Automatic evaluation results for small systems using three evaluation data sets.

| System | ACCURAT (512) out-of-domain | DGT (1000) in-domain | COMBINED (1024) mixed domain |
|---|---|---|---|
| Baseline (SMT) | 17.65 (15.81-19.64) | **46.11 (44.22-47.75)** | **33.82 (31.81-35.77)** |
| NMT 40K | 19.60 (17.79-21.54) | 35.30 (33.70-36.78) | 29.01 (27.38-30.58) |
| NMT 40K Without UNK | 19.58 (17.75-21.49) | 35.38 (33.80-36.82) | 29.05 (27.38-30.68) |
| Hybrid (alignment with gaps and without low conf. source words) | 21.59 (19.63-23.46) | 38.83 (37.11-40.53) | 31.70 (30.16-33.46) |
| Hybrid (alignment with gaps and with low conf. source words) | 21.58 (19.62-23.53) | 38.83 (37.16-40.40) | 31.70 (30.15-33.25) |
| Hybrid (alignment without gaps and without low conf. source words) | **21.59 (19.68-23.54)** | 38.84 (37.20-40.40) | 31.70 (29.97-33.34) |
| Hybrid (alignment without gaps and with low conf. source words) | 21.58 (19.71-23.47) | 38.84 (37.08-40.49) | 31.70 (30.10-33.33) |

The automatic evaluation results for the large systems are given in Table 5. The results show that none of the NMT systems (including the hybrid systems) produced better translations than the baseline system according to BLEU points. In fact, the baseline system achieved more than 11.7 BLEU points in comparison to the best hybrid system. In terms of optimisation algorithms, the Adadelta algorithm showed that it allows training of a better NMT system. However, the difference is not statistically significant.

The translations of the NMT systems, which were trained using the Adadelta and Adam algorithms, contained 634 (5.94% of all) and 713 (6.60% of all) unknown word tokens respectively. Due to significantly lower out-of-vocabulary token rates compared to the small systems, the hybrid system translations allow achieving a slightly less than 1 BLEU point improvement. However, the hybrid method still allows improving translation quality.

**Table 5.** Automatic evaluation results for big systems using the ACCURAT balanced evaluation data set.

| System | Adadelta (stopped after 360,000 iterations) | Adam (stopped after 570,000 iterations) |
|---|---|---|
| Baseline (SMT) | **39.78 (37.36-42.20)** | **39.78 (37.36-42.20)** |
| NMT 100K | 26.49 (24.61-28.47) | 26.08 (24.04-28.21) |
| NMT 100K Without UNK | 27.17 (25.10-29.20) | 27.06 (25.13-28.94) |
| Hybrid (alignment with gaps and without low conf. source words) | 28.11 (26.24-30.01) | 27.80 (25.69-29.81) |
| Hybrid (alignment with gaps and with low conf. source words) | 28.08 (26.23-30.04) | 27.75 (25.63-29.91) |
| Hybrid (alignment without gaps and without low conf. source words) | 28.04 (26.10-29.91) | 27.68 (25.49-29.66) |
| Hybrid (alignment without gaps and with low conf. source words) | 28.04 (26.24-29.95) | 27.68 (25.45-29.68) |

Although the difference between the baseline system and the best hybrid system is 11.7 BLEU points, we performed a small-scale human comparative evaluation experiment to identify whether the NMT systems really do produce lower quality translations. Four evaluators analysed a total of 140 sentences (a random subset of the evaluation data set). The results showed that the baseline SMT system's translations were preferred 52 times, and the hybrid NMT system's translations were preferred 62 times. The remaining 26 sentences were translated equally good or bad by both systems.

The main result of this small-scale evaluation is that it cannot be confirmed with a statistically significant confidence that one of the systems produces better translations than the other system. This means that the large BLEU difference does not correctly represent the quality difference between the SMT and NMT systems. The result also shows that human comparative evaluation is crucial when comparing MT systems from fundamentally different approaches.

## 7. Conclusion

The authors presented a hybridisation method for chaining NMT and SMT systems to address the out-of-vocabulary word issue of NMT systems. Evaluation results from two experiments with small and large MT systems showed that the method allows improving the translation quality by up to 3 BLEU points (depending on the proportion of unknown word tokens in the NMT output and the domain of the source text).

A small-scale human comparative evaluation showed that it cannot be statistically proved that, in terms of human perceived quality, the large SMT and NMT system quality differs (despite the 11.7 BLEU point difference). However, a comprehensive SMT and NMT qualitative analysis is necessary to validate the results of the small-scale experiment.

## References

[1] Mikolov, T., Karafiát, M., Burget, L., Cernock\`y, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In INTERSPEECH (Vol. 2, p. 3).

[2] Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R. M., & Makhoul, J. (2014). Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In ACL (1) (pp. 1370–1380).

[3] Schwenk, H. (2010). Continuous-space language models for statistical machine translation. The Prague Bulletin of Mathematical Linguistics, 93, 137–146.

[4] Zoph, B., Vaswani, A., May, J., & Knight, K. (2016). Simple, Fast Noise-Contrastive Estimation for Large RNN Vocabularies.

[5] Sennrich, R., Haddow, B., & Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. arXiv Preprint arXiv:1508.07909.

[6] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. CoRR, abs/1409.0. Retrieved from http://arxiv.org/abs/1409.0473

[7] Jean, S., Firat, O., Cho, K., Memisevic, R., & Bengio, Y. (2015). Montreal Neural Machine Translation Systems for WMT15. In Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 134–140).

[8] Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1412–1421). Lisbon, Portugal: Association for Computational Linguistics.

[9] Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlter, P. (2012). DGT-TM: A Freely Available Translation Memory in 22 Languages. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (pp. 454–459).

[10] Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: a Cloud-Based Platform for Do-It-Yourself Machine Translation. In Proceedings of the ACL 2012 System Demonstrations (pp. 43–48).

[11] Kingma, D. P. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[12] Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

[13] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 177–180).

[14] Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh Neural Machine Translation Systems for WMT 16. In arXiv preprint arXiv:1606.02891.