First Experiments and Results in English–Hungarian Neural Machine Translation

László Tihanyi, Csaba Oravecz

I.R.I.S. / European Commission, Directorate-General for Translation
e-mail:{laszlo.tihanyi,csaba.oravecz}@ext.ec.europa.edu

Abstract. Neural machine translation (NMT) has emerged recently as a promising alternative of standard rule-based or phrase-based statistical approaches especially for languages which have so far been considered challenging for the two paradigms. Since Hungarian has long been one of these challenging languages, it is a natural candidate for neural machine translation to explore whether this approach can bring some improvement in a task which translation models have so far been unable to cope with. The paper presents our first results of applying neural models to English to Hungarian translation and shows that with the right configuration and data preparation, publicly available NMT implementations can significantly outperform the previous state-of-the-art systems on standard benchmarks.

Keywords: neural machine translation, attention based model, source side linguistic analysis, morphology-aware subword units, alignment-based unknown word replacement

1 Introduction

Neural machine translation is a relatively young but already very successful approach in one of the most difficult areas of natural language processing. Recent results, in particular those at WMT'16 [1] have made the attention of the machine translation community shift considerably towards neural networks, and it seems more and more plausible that the field is experiencing a paradigm shift. After the early rule-based approaches phrase-based SMT has been dominant for many years but now the new models promise progress even for languages deemed difficult for standard translation systems. Although some of the main drawbacks of current neural systems are already widely known, such as the occasional lack of adequacy, NMT has been reported to work well with rich morphology and significant word reordering, producing more fluent output [2,3]. This motivates the present work in an attempt to create an English to Hungarian end-to-end neural translation system.

The aim of this paper is not so much to give a detailed account of all the system components and provide an extensive description of the many experiments but rather to present a general overview with focus on the best setups and most

important findings. We compare the performance of a Moses based [4] in house translation architecture (MT@EC [5]) with neural systems based on two publicly available implementations and report results that even in this early stage convincingly show that NMT is not only comparable but promisingly better on the EU domain than earlier phrase-based statistical or rule-based approaches [6] for the En-Hu direction.

2 Neural Machine Translation

Standard phrase-based statistical machine translation is built around a number of components, while a(n attention based) NMT system is in principle an end-to-end encoder-decoder framework to model the entire translation process [7]. The role of the encoder, which is often implemented as a bidirectional recurrent network, is to summarise the source sentence into a set of context vectors, and the decoder acts as a recurrent language model to generate a target sentence word after word by leveraging information from the weighted sum of the context vectors at each step, where the weights are computed by the attention mechanism [8]. In learning, the whole model is jointly trained to maximise the conditional probability of a gold standard translation given a source segment from a training corpus of parallel sentences. Learning is regulated by various optimisation algorithms with backpropagation.

3 Previous Work

The evolution of neural translation systems has been rapid. In the beginning, neural networks were only used as a component in a classical SMT system [9,10]. But soon after the success of the end-to-end neural network based translation system from Montreal University at WMT15 [11], more and more work has been invested to develop direct neural translation models for several language pairs, predominantly for Indo-European languages and Chinese [12,13] but some early results have already been published on for example Arabic as well [14].

At the latest WMT conference, the Edinburgh Neural Machine Translation System was clearly dominant [15], and since then NMT has already found its way even into production systems [16,17]. The field is still evolving swiftly and many techniques have been and are being developed to further improve performance, some of them alrady being close to standard, such as the attention mechanism or using subword units to overcome the unknown/rare word problem [18]. Further attempts have been made to change the basic input (output) unit from the wordform to the character [19], to better model translation coverage [20], or to use a more suitable optimisation technique, which instead of maximising the likelihood of the training data, can take evaluation metrics as loss functions and aims to minimise expected loss on the training data [21]. Many more new directions are investigated and there is constant progress with the promise of offering new solutions to old problems of MT.

4 System Architecture

Our pilot En-Hu system consists of five basic building blocks (steps). The core component is the translation module (see Section 4.4), which can be of different types and used more or less interchangeably within a pre- and post-processing pipeline. This, however, must eventually be tailored a bit to each of the core translation modules. In our full fledged neural architecture the components are the following: i) pre-processing pipeline, ii) configuration frontend to the NMT toolkits, iii) core NMT module iv) post-processing module, v) evaluation and visualisation module.

4.1 Experiment management

Finding the optimal settings for all (hyper-)parameters for an NMT system requires a large number of experiments with a wide scale of different values. In the NMT toolkit implementations this is not (yet) conveniently supported. Therefore we have developed a simple Python frontend which supports the setting of all relevant parameter values in one unified configuration file, which is used by the core neural training, the pre- and post-processing and the test process at the same time. This not only makes running experiments with a range of parameter values simple but also keeps a record of all the settings that define a specific translation model.

The whole system is managed by two wrapper shell scripts that take the configuration file and run the processing steps with the parameter settings as defined therein. One script is responsible for pre-processing and training while the other for translation (including pre-processing), post-processing and evaluation.

4.2 Solving the rare/unknown word problem

For neural translation models two types of approaches have been proposed to alleviate the problem that only a limited (50-100k word) vocabulary is allowed for the model to remain computationally manageable. The first branch is based on extending the base encoder-decoder model to incorporate external information, such as alignment information and dictionary look-up for out-of-dictionary words [22], the second tries to transform the input data into units smaller than a full word form, thereby constraining the number of possible forms [18]. Our solution attempts to utilise the advantages of both approaches.

On the Hungarian target side we apply morphological analysis [23] to facilitate morphology-aware tokenisation and split the target word forms into linguistically motivated stems and suffixes, [1] and so reduce the number of possible units. Further reduction is achieved by an extensive placeholder mechanism tailored for each language pair. This mechanism is different from the one used in the in house MT@EC translation system. In MT@EC, placeholder replacement

[1] It is possible to further process this format with the algorithm of [18], but we do not yet have results for this type of setup.

is based upon hard alignment information, which is extractable from a standard phrase-based (Moses) system. In NMT, however, this information is not available, and so placeholder replacement uses segment level unique identifiers and language specific mapping tables to find the appropriate target forms.

For the remaining unknown words external information is utilised in the form of an alignment dictionary generated by a pre-processing step on the parallel data using the tool `fast_align` [24]. The source side equivalents of target side unknown tokens are identified from the "soft" alignment[2] based on the attention weights, potentially constrained by the guided alignment strategy of [25].

4.3 Pre- and Post-processing

Most of the pre-processing steps contain standard transformations on the data, such as tokenisation with the default Moses tokeniser, some normalisation, segment filtering by length, truecasing and placeholder replacement. It is possible to filter out training segments according to the ratio of unknown words but we have not experienced significant differences in performance by adding this filter to the pre-processing pipeline. For the unknown word replacement, the alignment dictionary is generated after the target side morphological split to find better translation equivalents between English and Hungarian input units. To support the processing of full documents, sentence segmentation is also implemented in the workflow.

In neural models it is very easy to support the incorporation of source side linguistic analysis into the model [26] by simply concatenating the additional information with the input token representation, creating a "neural equivalent" of a standard phrase-based factored model. We use the Wapiti toolkit [27] to provide part-of-speech and chunk labels on the English source side for some of the experiments (Figure 1). Although performance improvements have been reported using the output of higher level of linguistic analysis (full parsing) [26], in a future production environment processing time is critical and full parsing being prohibitively slow we try to use tools which are sufficiently fast not to delay the translation time significantly.

raw	OJ L 302, 19.10.1989, p. 1. Directive as last amended by Commission Decision 2006/60/EC (OJ L 31, 3.2.2006, p. 24).
pp	`_oj_num-1|NN ,|PUNCT _num_datum-1|NN ,|PUNCT p.|NN _num-1|CD .|PUNCT Directive|NNP as|IN last|JJ amended|VBN by|IN Commission|NNP Decision|NNP _num_inst-1|NN (|-LRB- _ojpage-1|CD )|-RRB- .|PUNCT`

Fig. 1. Raw and pre-processed source segment.

The translation output is post-processed to replace unknown tokens using the soft alignment information and the dictionary, and to insert the target surface forms for the the placeholders. If target side splitting is switched on then

[2] Which is in effect a probability distribution over possible tokens.

target surface forms are generated from stems and suffixes. In the end, the standard steps of segment recasing and detokenisation restore the final output of the system, which can be evaluated with BLEU and sentence level METEOR scores. For detailed investigation of the translation process, the visualisation of soft alignments is also implemented (Figure 2). In document translation mode translation memories in TMX format are generated as output.

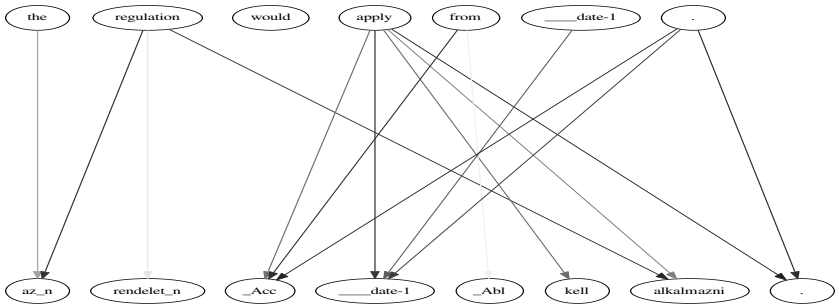

Fig. 2. Soft alignment between source and morpheme split target.

4.4 MT Systems

The baseline system is a Moses-based framework that is currently used in the MT@EC service at the European Commission. It does not use the most recent Moses version but in its current setup newer Moses versions give no increase in translation quality.[3] There have been many in house experiments to improve the En-Hu system in recent years but no significant progress has been made so we use a fairly basic phrase-based architecture in lack of any better alternative as our baseline system.

The neural systems are built around two publicly available implementations. The first one is called Nematus[4] [15], originally forked from Kyunghyun Cho's DL4MT tutorial repository[5], and is a Theano based toolkit using gated recurrent units (GRU) [28], which supports the output of alignment information, right-to-left rescoring, subword units and source side linguistic factors. The other is a Torch implementation[6] of an LSTM [29] based network architecture, maintained by the HarvardNLP group and SYSTRAN [16] and so is already used

[3] The novelty of recent versions is mostly in the support of advanced, more complex models, which on the one hand cannot be efficiently used in a production environment and on the other hand do not work better for the En-Hu language pair.

[4] `https://github.com/rsennrich/nematus`

[5] `https://github.com/nyu-dl/dl4mt-tutorial`

[6] `https://github.com/harvardnlp/seq2seq-attn`

in a production environment. This toolkit offers a wide range of parameter settings and configurations including the use of character level models or external embeddings, supports linguistic input features as well and comes with a built-in mechanism for alignment-based unknown word replacement using an external dictionary.

5 Experiments

5.1 Datasets

To find the best settings and values for the set of parameters that most influence translation quality it is necessary to run a large numbers of experiments, each taking significant time to run through. For this and other circumstantial reasons the initial experiments are run on the DGT-TM public dataset [30], originally with 2 million segments used for training, 3000 for validation and 1000 for testing. It has turned out, however, that the datasets contains a lot of duplicate segments and filtering them out not only shortens experiments, makes the model more compact but most importantly has no negative impact on translation qualittranslation quality. In fact, using only the filtered 1.1 million training set improved the Nematus-based system considerably (see Section 6).

5.2 Hardware environment

Training neural models requires substantial resources with dedicated hardware. Our models have been trained on two different architectures, one with NVIDIA GeForce GTX 980 GPUs with 4GB RAM, which allowed for one single training per GPU with some of the settings limited to not necessarily the optimal values, such as minibatch size or vocabulary, although our initial experiments have so far shown that for this magnitude of training data a larger vocabulary (100k word) does not lead to performance improvement. The other environment is provided in the IBM cloud by the CEF eTranslation project and contains 4 powerful servers with dual core Tesla K80 (2x12GB RAM) GPUs, which allow for at least two trainings and further two test runs in parallel. Training time is not significantly different on the two architectures, it takes around 2 days to reach optimal performance on these medium size datasets. Decoding time is also manageable, with less than one second per segment using the GPU.

5.3 Training Procedure

During training the optimisation algorithm and related hyper-parameters are pivotal to reach the best performance of the model. Similarly to the Google NMT research group [17] in our best systems (so far) we use a combination of the Adam gradient descent optimisation algorithm [31] and Stochastic gradient descent (SGD) but the exact values always depend on the particular dataset and language pair, hence there are no globally valid settings for all scenarios.

On average, Adam quickly converges at around 150k updates when we switch to SGD starting with a learning rate of 0.1, which we anneal gradually by a magnitude after a certain number of steps, depending on the actual dataset. On average, in our case this process could lead to 1 point increase in the BLEU score.

We have not experimented with dropout[7] to avoid overfitting. Dropout is normally used for limited datasets and we expect it will not be necessary once full size parallel data can be fed to the systems, which in general contain well more than 10M segments for most of the EU language pairs.

Ensembles of neural models have been shown to perform better than a single model[8], however, we have not extensively tested them so far. Running an ensemble always requires more resources, which can be a problem in our production environment so if ensembling is not absolutely necessary a single model can still be a good compromise.

Table 1. BLEU scores for various En-Hu MT systems.

	Model	Abbreviation	BLEU
0	Baseline phrase-based	SMT	23.37
1	First basic neural model	NMT-BASE	22.05
2	Neural model with target side split	NMT-MORPH	21.94
3	Increased segment length (75)	NMT-LENGTH	23.02
4	Hyphenated compound split	NMT-SPLIT	24.58
5	Unique training segments	NMT-UNIQE	26.11
6	Subword units	NMT-BPE	25.95
7	Source side factors	NMT-FACTOR	26.13
8	HarvardNLP toolkit base model	HNLP	24.90

6 Results

We have run a large number of experiments with many settings and are still in search of the best models hoping to increase the translation quality further with better pre-processing or more optimal parameter values. In Table 1 we summarise the best results of selected models while Figure 3 illustrates how model performance improves during training. In most of the models, we use a 50k word vocabulary, and a minibatch size of 32.

We note that initially we experimented with a maximal segment length of 50-60, inherited from the phrase-based system. Using a higher value (75-100) significantly increased the performance of the NMT systems (Model 3) but interestingly had only negligible effect on SMT. The same tendency seems to be

[7] Randomly dropping units from the network during training.
[8] See eg. [15] but this technique is ubiquitous in all NMT systems.

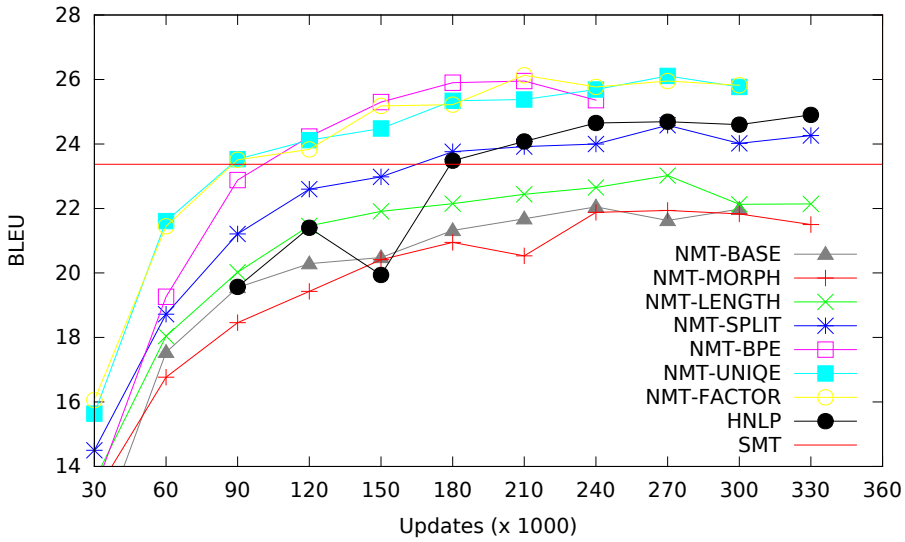

Fig. 3. BLEU score vs. number of updates during training.

true for the use of source side factors (Model 7), splitting the hyphenated compunds on both source and target side (Model 4) or duplicate segment filtering (Model 5): they are useful for NMT but almost totally useless for the Moses-based system (therefore we do not separately report these scores for the SMT system). Morphology aware tokenization in itself does not give improvement (Model 1 vs. Model 2) but our current implementation is extremely rudimentary without any disambiguation and with many errors in generation. We expect much better results from the new processing tools for Hungarian that we are now building into our system.

Our best model achieves a BLEU score of 26.13 which is about 3 points higher than the phrase-based baseline. In Figure 4 we illustrate this difference with sample translations from selected engines together with the reference translation and the segment level METEOR scores. On a casual inspection the good quality (and fluency) of the neural translation is striking, and (informal) manual evaluation tends to prefer the neural translation even in cases where the automatic score favours the phrase-based alternative. In this stage of the project, there has not yet been rigorouos qualitative evaluation but some early feedback from professional translators seems to confirm much of the findings of [3]: the NMT output has better morphology, word order and agreement, and post editing effort is estimated much lower. It is true, however, that without a large scale manual evaluation campaign of full size engines it is too early to jump to final conclusions.

SRC	Ms Peeters, the UWV, the Netherlands, Czech, Danish and German Governments and the European Commission submitted written observations to the Court.
REF	M. A. Peeters, az UWV, a holland, a cseh, a dán, és a német kormány, valamint az Európai Bizottság írásbeli észrevételeket terjesztett a Bíróság elé.
SMT	A ms peeters UWV, Hollandia, a cseh, a dán és a német kormány
0.554	és az Európai Bizottság írásbeli észrevételeket a Bíróság elé.
NMT	Peeters úr, az UWV, Hollandia, cseh, dán és német kormány és
0.230	az európai Bizottság írásos észrevételt nyújtott be a bírósághoz.
SRC	This does not mean that by regulating access to public infrastructure, a resource has been transferred (or indeed forgone).
REF	Ez nem jelenti azt, hogy a nyilvános infrastruktúrához való hozzáférés szabályozásával sor kerül forrásátruházásra (vagy valójában -kiesésre).
SMT	Ez nem jelenti azt, hogy szabályozása által való hozzáférés,
0.329	infrastruktúra állami forrásnak át (vagy akár negatív).
NMT	E nem jelenti az, hogy az állami infrastruktúrához való hozzáférés
0.329	szabályozása révén a forrásokat átruházták (vagy).
SRC	On 24 October 2013, the Council adopted Decision 2013/527/CFSP [2] amending and extending the mandate of the EUSR for the Horn of Africa until 31 October 2014.
REF	A Tanács 2013. október 24-én elfogadta az Európai Unió Afrika szarváért felelős EUKK megbízatásának módosításáról és 2014. október 31-ig történő meghosszabbításáról szóló 2013/527/KKBP határozatot [2].
SMT	A Tanács 2013. október 24-én elfogadta a 2013/527/KKBP határozat [2],
0.421	amely az EUKK megbízatását az Afrika szarva térségre vonatkozóan 2014. október 31-ig.
NMT	A tanács 2013. október 24-én elfogadta az Afrika szarváért felelős
0.917	EUKK megbízatásának módosításáról és 2014. október 31-ig történő meghosszabbításáról szóló 2013/527/KKBP határozatot [2].

Fig. 4. Sample translations by the different systems.

7 Conclusions and future work

We have presented our first promising results of machine translating English to Hungarian with neural models. Due to the novelty of this line of work and technology in our working environment there are limitations in the current setups which will be soon overcome and more extensive tests on much larger datasets can be carried out. This will hopefully lead to a machine translation system that can significantly help human translators at the European Commission, even in languages where the current MT@EC system cannot offer sufficient support and is therefore rarely used by the translators.

Neural translation models are robust, flexible and seem to respond more favourably than phrase-based systems to the changes in parameters of the experiments in many respects; they can utilise linguistic information without introducing much complexity in the model, and seem to capture the properties of the training data better. We hope that they will lead to a breakthrough in the translation of difficult language pairs and soon be mature enough to be used even in our production environments. They are easier to train, customise and

manage than a Moses based system and can benefit from advances on neural models from other fields as well. In the multilingual environment of the EC the possibility of multilingual translation with a single model [32] is a promise which needs serious consideration.

In the near future we plan to further improve the precision of language dependent pre-processing tools, decrease the vocabulary by targeting named entities with special pre-processing, further narrow the distance between source and target by for example adding morphological split for the source side, and set up a large scale manual evaluation campaign of full size NMT engines.

Acknowledgements

This work is carried out in the framework of the CEF eTranslation project. The authors would like to thank the Research Group for Mathematical Linguistics and the Research Group for Language Technology at the Research Institute for Linguistics for giving access to their hardware infrastructure for some of the experiments, Attila Novák for developing and providing the core components of the Hungarian morphological analyser and generator, and the reviewers for their comments and suggestions to improve the paper.

References

1. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., Zampieri, M.: Findings of the 2016 conference on machine translation. In: Proceedings of the First Conference on Machine Translation, Berlin, Germany, Association for Computational Linguistics (2016) 131–198
2. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal (2015) 1412–1421
3. Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M.: Neural versus phrase-based machine translation quality: a case study. In: Proceedings of EMNLP 2016, Austin, USA (2016)
4. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. ACL '07, Stroudsburg, PA, USA, Association for Computational Linguistics (2007) 177–180
5. MT@EC: Secure machine translation for the European Union, European Commission, Directorate-General for Translation (2014)
6. Novák, A., Tihanyi, L., Prószéky, G.: The metamorpho translation system. In: Proceedings of the Third Workshop on Statistical Machine Translation. StatMT '08, Stroudsburg, PA, USA, Association for Computational Linguistics (2008) 111–114

7. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. (2014) 3104–3112
8. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR. (2015)
9. Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., Makhoul, J.: Fast and robust neural network joint models for statistical machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, Association for Computational Linguistics (2014) 1370–1380
10. Cho, K., Van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Association for Computational Linguistics (2014) 1724–1734
11. Jean, S., Firat, O., Cho, K., Memisevic, R., Bengio, Y.: Montreal neural machine translation systems for WMT'15. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, Association for Computational Linguistics (2015) 134–140
12. Lu, Z., Li, H., Liu, Q.: Memory-enhanced decoder for neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas (2016) 278–286
13. Zhang, B., Xiong, D., su, j., Duan, H., Zhang, M.: Variational neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Association for Computational Linguistics (2016) 521–530
14. Almahairi, A., Cho, K., Habash, N., Courville, A.C.: First result on Arabic neural machine translation. CoRR **abs/1606.02680** (2016)
15. Sennrich, R., Haddow, B., Birch, A.: Edinburgh neural machine translation systems for wmt 16. In: Proceedings of the First Conference on Machine Translation, Berlin, Germany, Association for Computational Linguistics (2016) 371–376
16. Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., Enoue, S., Geiss, C., Johanson, J., Khalsa, A., Khiari, R., Ko, B., Kobus, C., Lorieux, J., Martins, L., Nguyen, D.C., Priori, A., Riccardi, T., Segal, N., Servan, C., Tiquet, C., Wang, B., Yang, J., Zhang, D., Zhou, J., Zoldan, P.: SYSTRAN's pure neural machine translation systems. CoRR **abs/1610.05540** (2016)
17. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016)
18. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 1715–1725
19. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation. In: Proceedings of the 54th Annual Meeting

of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 1693–1703
20. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 76–85
21. Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Minimum risk training for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 1683–1692
22. Luong, T., Sutskever, I., Le, Q., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics (2015) 11–19
23. Novák, A., Siklósi, B., Oravecz, Cs.: A new integrated open-source morphological analyzer for Hungarian. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
24. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of IBM Model 2. In: Proceedings of NAACL-HLT, Atlanta, Georgia (2013) 644–648
25. Chen, W., Matusov, E., Khadivi, S., Peter, J.T.: Guided alignment training for topic-aware neural machine translation. In Green, S., Schwartz, L., eds.: Proceedings of The Twelfth Conference of The Association for Machine Translation in the Americas. Volume 1., Austin, Texas (2016) 121–134
26. Sennrich, R., Haddow, B.: Linguistic input features improve neural machine translation. In: Proceedings of the First Conference on Machine Translation, Berlin, Germany, Association for Computational Linguistics (2016) 83–91
27. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics (2010) 504–513
28. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv e-prints **abs/1412.3555** (2014) Presented at the Deep Learning workshop at NIPS2014.
29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9** (1997) 1735–1780
30. Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P.: DGT-TM: A freely available translation memory in 22 languages. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul (2012) 454–459
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. (2015)
32. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation (2016) arXiv:1611.04558.