The TALP-UPC Machine Translation Systems for WMT18 News Shared Translation Task

Noe Casas, Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa
contact@noecasas.com, carlos.escolano@tsc.upc.edu
{marta.ruiz,jose.fonollosa}@upc.edu
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona

Abstract

In this article we describe the TALP-UPC research group participation in the WMT18 news shared translation task for Finnish-English and Estonian-English within the multi-lingual subtrack. All of our primary submissions implement an attention-based Neural Machine Translation architecture. Given that Finnish and Estonian belong to the same language family and are similar, we use as training data the combination of the datasets of both language pairs to paliate the data scarceness of each individual pair. We also report the translation quality of systems trained on individual language pair data to serve as baseline and comparison reference.

1 Introduction

Neural Machine Translation (NMT) has consistently maintained state of the art results in the last years. However, due to its need for large amounts of training data, low resource language pairs need to resort to extra techniques to achieve acceptable translation quality.

In the WMT18 news shared translation task, two of the languages to translate are Finnish and Estonian (that are to be translated to and from English). Both can be considered low-resource languages in general, and also in particular for this shared task, based on the volume of data made available for training, especially Estonian.

In this report we describe the participation of the TALP research group from *Universitat Politècnica de Catalunya* (UPC) at the aforementioned WMT18 news shared translation task, specifically in the multi-lingual subtrack, as our systems make use of the data from both Finnish and Estonian language to improve the translation quality.

2 Linguistic Background

Finnish and Estonian are respectively the official languages of Finland and Estonia, having 5.4 and 1.1 million native speakers (Lewis, 2009). They are **Finnic Languages**, a branch within the Uralic Language family.

Estonian and Finnish make use of the Latin alphabet with some additional letters, each one incorporating extra letters (e.g. ä, ö, ü, õ, š, ž).

Finnish and Estonian are morphologically-rich **agglutinative languages**. Estonian presents fourteen grammatical cases while Finnish presents fifteen. Verb conjungations are very regular in both languages. Neither of them has grammatical gender nor definite or indefinite articles. Both have flexible word order, but the basic order is subject-verb-object.

Like other Finnic languages, both Finnish and Estonian present consonant gradation (consonants are classified in grades according to phonologic criteria, and such grades condition the combined appearance of the consonants in a derived word), but the gradation patterns each one follows are different.

While Finnish has kept most of its late Proto-Finnic linguistic traits, Estonian has lost some of its former characteristics, like vowel harmony (vowels in a word cannot appear freely but their allowance is constrained by rules), which in Finnish affects case and derivational endings. Also, Estonian mostly lost the word-final sound, making its inflectional morphology more fusional for nouns and adjectives (Fortescue et al., 2017). German language influence also led Estonian to use more postpositions where Finnish uses cases. Geographical location has also led to differences in the loanwords borrowed by each language.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 359–364

https://doi.org/10.18653/v1/W18-6406

3 Attention-based NMT

The first competitive NMT systems were based on the sequence-to-sequence architecture (Cho et al., 2014; Sutskever et al., 2014), especially with the addition of attention mechanisms (Bahdanau et al., 2014; Luong et al., 2015), either using Gated Recurrent Units (GRU) (Cho et al., 2014) or Long-Short Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997).

Sequence-to-sequence with attention was the state of the art NMT model until the Transformer architecture (Vaswani et al., 2017) was proposed. This model does not rely on recurrent units or convolutional networks, but only on attention layers, combining them with several other architectural elements: positional embeddings (Gehring et al., 2017), layer normalization (Ba et al., 2016), residual connections (He et al., 2016) and dropout (Srivastava et al., 2014).

The type of attention mechanism used by the Transformer model is a multi-headed version of the dot-product attention, applied both as self-attention to source and target (prefix) sentences and as encoder-decoder attention mechanism.

4 Low resource NMT

The application of NMT to low resource language pairs needs extra techniques to achieve good translation quality. These are some of the frequently used approaches:

Back-translation (Sennrich et al., 2015a) consists in training an auxiliary translation system from target language to source language and use it to translate a large target language monolingual corpus into the source language, and then use such synthetic source-target sentence pairs to augment the originally available parallel corpus and train on it a new source language to target language translation system.

Pivoting approaches use a third resource-rich language as *pivot* and train translation systems from source language to pivot and from pivot to target language. These auxiliary systems can either be used in *cascade* to obtain source-to-target translations, or be used to build syntethic parallel source-target corpora (i.e. *pseudocorpus approach*). A recent application of pivoting techniques to NMT can be found in (Costa-jussà et al., 2018).

Adversarial learning (Lample et al., 2018; Artetxe et al., 2018) in a multi-task learning setup, with an auxiliary text (denoising) auto-encoding loss, whose internal sentence representation is aligned with the ones from the translation task by means of a discriminator in feature space.

Pre-trained cross-lingual embeddings (Artetxe et al., 2016, 2017) can be used complementarily to further reduce the need for parallel data.

Finding parallel data from a similar source language and the same target language (or vice versa) and adding it to the original parallel corpus. With such a composite training data set, a wordpiece-level vocabulary can leverage the common word stems between the similar languages and profit from the combined amount of data. This approach is used in the present work, as described in sections 5 and 6.1.

Multilingual zero-shot translation (Johnson et al., 2017) also uses parallel corpora from different source and target language pairs, but mixes together every available language pair, regardless of how linguistically close they are. This way, there is a single shared word-piece vocabulary for all languages, and the system is trained on a corpus that combines data from several different language pairs. In order to convey the association between a source sentence and its translation to a specific target language, the source sentence is prefixed with a token that specifies which language the target sentence belongs to. This approach aims at implicitly learning language-independent internal representations, enabling the translation of low resource language pairs (and even language pairs where there is zero parallel data available) to profit from the combined language pair training data.

5 Corpora and Data preparation

All proposed systems in this work are constrained using exclusively parallel data provided by the organization. For the English - Finnish language pair the data employed is the Europarl corpus version 7 and 8, Paracrawl corpus, Rapid corpus of EU press releases and Wiki Headlines corpus. For the English - Estonian data the Europarl v8 corpus, Paracrawl and Rapid corpus of EU press releases corpus were employed.

All language pairs have been preprocessed following the proposed scripts by the organization of the conference. The pipeline consisted in normalizing punctuation, tokenization and truecasing using the standard *Moses* (Koehn et al., 2007)

scripts. With the addition that, for tokenization, no escaping of special characters was performed.

For the language pair of English - Estonian we found that from Paracrawl corpus a considerable number of sentences were not suitable sentences in the intended languages, but apparently random sequences of upper case characters. In order to remove them, an additional step of language detection was performed using library `langdetect` (Danilák, 2017), which is a port to Python of library `language-detection` (Shuyo, 2010). The criteria for removing noisy sentences from the dataset was that either one of the languages of the pair could not be identified as a language.

The sizes of the different data sets compiled for each language pair and once cleaned as described earlier in this section are presented in Table 1.

Table 1: Corpus statistics in number of sentences and words for both parallel corpora, English - Estonian and English - Finnish.

corpus	lang	set	sentences	words
En-Et	*En*	train	998547	23056922
		test	2000	44305
	Et	train	998547	17376004
		test	2000	34733
En-Fi	*En*	train	3064124	62208347
		dev	3000	64611
		test	3002	63417
	Fi	train	3064124	45692989
		dev	3000	48839
		test	3002	46572

As described in sections 2 and 4, as Finnish and Estonian belong to the Finnic language family and are similar to each other, we aimed at combining the individual parallel corpora (*En - Fi* and *En - Es*) into a single larger corpus. For the translation directions where English is the target language (i.e. *Fi* → *En* and *Et* → *En*) we prepared a combined *Fi* + *Et* → *En* corpus by simply concatenating the original ones. This approach was not applicable to the reverse directions, as we needed some way to convey the information about whether to generate either Finnish or Estonian as part of the input to the neural network. Following the approach in (Johnson et al., 2017), we modify the individual parallel corpora to add a prefix to the English sentences to mark whether the associated target sentence was Finnish or Estonian, and then proceed to concatenate both corpora into the final combined one *En* → *Fi* + *Et*. The prefixes used were respectively `<fi>` and `<et>`. This prefix needs to be added likewise to the test English sentences when decoding them into Finnish or Estonian.

As the combined corpora are concatenations of the individual ones, their sizes can be computed from the figures in Table 1 by mere addition of the individual sizes of each language pair.

6 System Description

In this section we present the translation systems used for our submissions, both in terms of vocabulary extraction strategies followed (section 6.1), of neural architecture used (section 6.2) and of needed post-processing (section 6.3).

6.1 Vocabulary Extraction

The NMT models used for all of our submissions, which are described in section 6.2 make use of predefined sets of discrete tokens that comprise the *vocabulary*.

The vocabulary of each of our translation systems (both the final submissions and the systems trained for reference described in section 7) was based on wordpiece extraction (Wu et al., 2016). For each system, the source and target vocabularies were extracted separately, aiming at a vocabulary size of 32K tokens. Vocabularies are not shared between source and target languages in any case.

Word-piece vocabularies (or the very similar Byte-Pair Encoding (BPE) vocabularies (Sennrich et al., 2015b)) are usually applied to extract vocabularies from corpora that contain data from similar languages in order to try to find common stems and derivational suffixes so that the language commonalities can be leveraged by the neural network training.

6.2 NMT Models

All the submissions presented to the task make use of the Transformer NMT architecture, which is described in section 3. We used the implementation released by the authors of (Vaswani et al., 2017) [1]

The complete hyperparameter configuration used for all the attention-based neural machine

[1]The authors of (Vaswani et al., 2017) made the source code available at `https://github.com/tensorflow/tensor2tensor`. For this work, version `1.2.9` was used.

translation models in our submissions (which consisted in the `transformer_base` parameter set in `tensor2tensor`) is shown in Table 2.

Table 2: Hyperparameters of the neural model.

hyperparameter	value
attention layers	6
attention heads per layer	8
hidden size (embedding)	512
batch size (in tokens)	4096 (4 GPU)
training steps	800000
tokenization strategy	wordpiece
vocabulary size	32K
optimization algorithm	Adam
learning rate	warmup + decay

After the training, the weights of the last 5 checkpoints (having checkpoints stored every 2000 optimization steps) are averaged to obtain the final model.

6.3 Post-processing

Following the inverse steps of the processing described in section 5, the decoded outputs of NMT model need to be de-truecased and de-tokenized by means of the appropriate *Moses* scripts.

7 Experiments

The hypothesis on which we base this work is that, given the similarity between Estonian and Finnish, a system trained with the combination of the data from both languages would outperform systems trained on the individual language datasets.

In order to validate this hypothesis, we conducted direct experiments, training systems on the individual language datasets and also on the combined datasets (as described in section 5), and comparing their translation quality. The datasets used for testing the performance were `newsdev2018` for Estonian - English and `newstest2017` for Finnish - English. The results of the experiments are shown in Table 3, were all figures represent case-insensitive BLEU score over the aforementioned reference test corpora.

While the results for Finnish are not very different between the individual and combined data trainings [2], the results for Estonian show an important improvement of the training on the combined data over the individual data. This correlates with the fact that the Estonian - English training set is less than one third the size of the Finnish - English, therefore the size increase in the Finnish - English combined training corpus is much smaller than the increase for Estonian - English, as shown in Table 1.

[2]Improvements of less than 1 BLEU point are normally considered neglectable.

Table 3: Comparison between translation quality (case-insensitive BLEU) of systems trained on the individual language data vs. systems trained on the combined data .

direction	individual	combined	ΔBLEU
En → *Fi*	24.36	25.21	+0.85
Fi → *En*	29.39	30.00	+0.61
En → *Et*	15.97	18.92	+2.95
Et → *En*	21.66	25.66	+4.00

8 Conclusions

In this article we described the TALP-UPC submissions to the the multi-lingual subtrack of the WMT18 news shared translation task for Finnish - English and Estonian - English language pairs.

Our experiments suggest that for low resource languages, enlarging the training data with translations from a similar language can lead to important improvements in the translation quality when using subword-level vocabulary extraction strategies. In this line, further research should be conducted to understand how subwords have captured the differences between Estonian and Finnish cognates and to leverage such an insight to devise more effective vocabulary extraction strategies.

Acknowledgements

We would like to thank Magdalena Biesialska for her help in the news task human evaluation.

This work is partially supported by Lucy Software / United Language Group (ULG) and the Catalan Agency for Management of University and Research Grants (AGAUR) through an Industrial PhD Grant. This work is also supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, contract TEC2015-69266-P (MINECO/FEDER,EU) and contract PCIN-2017-079 (AEI/MINECO).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Marta R. Costa-jussà, Noe Casas, and Maite Melero. 2018. English-catalan neural machine translation in the biomedical domain through the cascade approach. In *Proceedings of the 11th Language Resources and Evaluation Conference of the European Language Resources Association*.

Michal Danilák. 2017. Langdetect. https://github.com/Mimino666/langdetect.

Michael Fortescue, Marianne Mithun, and Nicholas Evans. 2017. *The Oxford Handbook of Polysynthesis*. Oxford Handbooks. Oxford University Press.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Nakatani Shuyo. 2010. Language detection library for java.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.