The Universitat d'Alacant submissions to the English-to-Kazakh news translation task at WMT 2019

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, Spain
{vmsanchez,japerez,fsanchez}@dlsi.ua.es

Abstract

This paper describes the two submissions of Universitat d'Alacant to the English-to-Kazakh news translation task at WMT 2019. Our submissions take advantage of monolingual data and parallel data from other language pairs by means of iterative backtranslation, pivot backtranslation and transfer learning. They also use linguistic information in two ways: morphological segmentation of Kazakh text, and integration of the output of a rule-based machine translation system. Our systems were ranked 2nd in terms of chrF++ despite being built from an ensemble of only 2 independent training runs.

1 Introduction

This paper describes the Universitat d'Alacant submissions to the WMT 2019 news translation task. Our two submissions address the low-resource English-to-Kazakh language pair, for which only a few thousand in-domain parallel sentences are available.

In order to build competitive neural machine translation (NMT) systems, we generated synthetic training data. We took advantage of the available English–Russian (`en-ru`) and Kazakh–Russian (`kk-ru`) parallel data by means of pivot backtranslation and transfer learning, and integrated monolingual data by means of iterative backtranslation.

In addition, we used linguistic information in two different ways: we morphologically segmented the Kazakh text to make the system generalize better from the training data; and we built a hybrid system combining NMT and the Apertium English-to-Kazakh rule-based machine translation (RBMT) system (Forcada et al., 2011; Sundetova et al., 2015).

The rest of the paper is organized as follows. Section 2 describes how corpora were filtered and preprocessed, and the steps followed to train NMT systems from them. Section 3 outlines the process followed to obtain synthetic training data. Sections 4 and 5 describe respectively morphological segmentation and hybridization with Apertium. The model ensembles we submitted are then presented in Section 6. The paper ends with some concluding remarks.

corpus	pair	raw	cleaned
News Commentary	`en-kk`	7.7k	7.4k
Wikititles	`en-kk`	117k	113k
web crawled	`en-kk`	97.6k	27.2k
web crawled	`kk-ru`	4.5M	4.4M
concatenation of WMT19 data	`en-ru`	31.7M	31.1M

Table 1: Number of segments in the parallel corpora used for training.

2 Data preparation and training details

In our submissions, we only used the corpora allowed in the constrained task. Parallel corpora were cleaned with the script `clean-corpus-n.perl` shipped with Moses (Koehn et al., 2007), that removes unbalanced sentence pairs and those with at least one side longer than 80 tokens. Additional filtering steps, described below, were applied to the web crawled corpora. Tables 1 and 2 depict the number of segments in the parallel and monolingual corpora used, and their sizes after cleaning.

The English–Kazakh web crawled corpus allowed in the constrained task presented a high proportion of parallel segments that were not translation of each other. We filtered it with Bicleaner (Sánchez-Cartagena et al., 2018). We applied the *hardrules* and the detection of misaligned sentences described by Sánchez-Cartagena et al. (2018), but not the fluency filtering.[1]

[1] We extracted probabilistic bilingual dictionaries from the

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 356–363
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

corpus	lang.	raw	cleaned
News Crawl	kk	783k	783k
Wiki dumps	kk	1.7M	1.7M
Common Crawl	kk	10.9M	5.4M
News Crawl	en	200M	200M

Table 2: Number of segments in the monolingual corpora used for training.

The Kazakh–Russian crawled corpus was cleaned in a shallower way: we just removed those sentence pairs that contained less than 50% of alphabetic characters in either side, as we did not consider them fluent enough to be useful for NMT training. The same filtering was applied to the monolingual Kazakh Common Crawl corpus. In addition, inspired by Iranzo-Sánchez et al. (2018), we ranked its sentences by perplexity computed by a character-based 7-gram language model and discarded the half of the corpus with the highest perplexity. The language model was trained[2] on the high-quality Kazakh monolingual News Commentary corpus.

Training corpora were tokenized and truecased with the Moses scripts. Truecaser models were learned independently for each trained system from the very same training parallel corpus. Unless otherwise specified, for each trained system, words were split with 50 000 byte pair encoding (BPE; Sennrich et al., 2016c) operations learned from the concatenation of the source-language (SL) and target-language (TL) training corpora.

As described in Section 6, our submissions were ensembles of Transformer (Vaswani et al., 2017) and recurrent neural network (RNN; Bahdanau et al., 2015) NMT models trained with the Marian toolkit (Junczys-Dowmunt et al., 2018). We used the Transformer hyperparameters[3] described by Sennrich et al. (2017) and the RNN hyperparameters[4] described by Sennrich et al. (2016a). Early stopping was based on perplexity and patience was set to 5. We selected the checkpoint that obtained the highest BLEU (Papineni et al., 2002) score on the development set.

Since the only evaluation corpus made available was `newsdev2019`, we split it in two halves, and we respectively used them as development and test set in all the training runs previous to the submission (those reported in all sections but Section 6). Throughout the paper, we report BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) scores.[5] The latter is known to correlate better than BLEU with human judgements when the TL is highly inflected (Bojar et al., 2017), as is the case. Where reported, we assess whether differences between systems' outputs are statistically significant for $p < 0.05$ with 1 000 iterations of paired bootstrap resampling (Koehn, 2004).

3 Data augmentation

This section describes the process followed to select the best strategy to take advantage of parallel corpora from other language pairs (Section 3.1) and monolingual corpora (Section 3.2).

3.1 Data from other language pairs

In order to take advantage of the parallel corpora listed in Table 1 for other language pairs, we applied the transfer learning approach proposed by Kocmi and Bojar (2018). We experimented with the *parent* models listed next (models trained on other high-resource language pairs) and used the concatenation of the genuine English–Kazakh parallel data as the *child* corpus (corpus of a low-resource language pair used to continue training a parent model):[6]

- A Russian-to-Kazakh model trained on the crawled parallel corpus depicted in Table 1.
- An English-to-Russian model trained on all the available parallel data for the English–Russian language pair in this year's news translation task (depicted in Table 1).
- A multilingual system (Johnson et al., 2017) trained on the concatenation of the corpora of the two previous models. This strategy aims at making the most of the data available for related language pairs.

We also explored pivot backtranslation (Huck and Ney, 2012): we translated the Russian side of the crawled Kazakh–Russian parallel corpus with

Wikititles parallel corpus and extracted the positive and negative training examples from News Commentary. We kept those sentences with a classifier score above 0.6.

[2] The language model was trained with KenLM (Heafield, 2011) with modified Kneser-Ney smoothing (Ney et al., 1994).

[3] `https://github.com/marian-nmt/marian-examples/tree/master/wmt2017-transformer`

[4] `https://github.com/marian-nmt/marian-examples/tree/master/training-basics`

[5] Following Popović (2017), we set β to 2.

[6] In the 3 set-ups evaluated, BPE models were trained from the concatenation of the parent and the child corpora.

a Russian-to-English NMT system to produce a synthetic English–Kazakh parallel corpus. The NMT system was a Transformer trained on the English–Russian parallel data depicted in Table 1. We concatenated the pivot-backtranslated corpus to the genuine English–Kazakh parallel data and fine-tuned the resulting system only on the latter.

The results of the evaluation of these strategies, reported in the upper part of Table 3, show that the multilingual/transfer learning strategy outperforms the pure transfer learning approaches, probably because it takes advantage of more resources. Moreover, it performs similarly to pivot backtranslation, which we chose for our submission. All the strategies evaluated clearly outperformed the system trained only on the genuine parallel data.

As a Kazakh-to-English MT system is needed to backtranslate the Kazakh monolingual data (see Section 3.2), we also explored the best strategy for taking advantage of data from other language pairs for that direction. We experimented only with transfer learning and discarded pivot backtranslation since we wanted to avoid training a system on a parallel corpus with a synthetic TL side.

We evaluated the same parent-child configurations as in the English-to-Kazakh experiments, but we inverted their direction to ensure that either the SL of the parent corpora is Kazakh or the TL is English. Results are reported in the lower part of Table 3 and show that, as in the opposite direction, transfer learning brings a clear improvement over training only on the genuine parallel data, and the best parent model is the multilingual one.

3.2 Monolingual data: iterative backtranslation

Backtranslation (Sennrich et al., 2016b) is a widespread method for integrating TL monolingual corpora into NMT systems. In order to integrate the available Kazakh monolingual data into our submission, we need a Kazakh-to-English MT system as competitive as possible, since the quality of a system trained on backtranslated data is usually correlated with the quality of the system that perform the backtranslation (Hoang et al., 2018, Sec. 3). We followed the iterative backtranslation algorithm (Hoang et al., 2018) outlined below with the aim of obtaining strong English-to-Kazakh and Kazakh-to-English systems using monolingual English and monolingual Kazakh corpora:

1. The best strategies from Section 3.1 were applied to build systems in both directions without backtranslated monolingual data.
2. English and Kazakh monolingual data were backtranslated with the previous systems.
3. Systems in both directions were trained on the combination of the backtranslated data and the parallel data.
4. Steps 2–3 were re-executed 2 more times. Backtranslation in step 2 was always carried out with the systems built in the most recent execution of step 3.

The Kazakh monolingual corpus used was the concatenation of the corpora listed in Table 2, while the English monolingual corpus was a subset of the News Crawl corpus in the same table. The size of the subset was duplicated after each backtranslation and started at 5 million sentences in the first one. The objective of the first 2 executions of steps 2–3 (from now on, *iterations*) was building a strong Kazakh-to-English system. The remainder of this section explains how MT systems were trained in these 2 iterations. The objective of the 3rd iteration, in which only English-to-Kazakh systems were trained, was building the submissions, and the corresponding details are described in Section 6.

We explored different ways of training NMT systems with backtranslated data. First, we carried out transfer learning from the multilingual models described in Section 3.1. In this case, the child model was trained on a parallel corpus built from the concatenation of the genuine parallel data and the backtranslated data. The genuine parallel data was oversampled to match the size of the backtranslated data (Chu et al., 2017).

As an alternative to transfer learning, we experimented with corpus concatenation and fine-tuning. For the English-to-Kazakh direction, we concatenated the backtranslated data to the pivot-backtranslated corpus and the genuine parallel corpora, trained a model from scratch, and fine-tuned it only on the genuine parallel data. For the opposite direction, we trained a system only on the concatenation of the backtranslated and the genuine parallel data, and fine-tuned it on the latter (note that in this set-up we dispensed with parallel data from other language pairs).

Table 4 shows the automatic evaluation scores obtained in the 1st iteration by the strategies being evaluated. Only the best performing strategies in the 1st iteration were used in the subsequent ones; the scores obtained on the 2nd iteration are also depicted. The results show the positive impact of the introduction of backtranslated data in both directions. Concatenation plus fine-tuning outperformed

strategy	BLEU	chrF++
en→kk		
only parallel en→kk	4.36	27.80
transfer from ru→kk	10.22	39.93
transfer from en→ru	9.66	39.67
transfer from en→ru,ru→kk	11.81	42.87
pivot backtranslation	11.80	42.86
kk→en		
only parallel kk→en	8.15	30.43
transfer from kk→ru	17.03	42.90
transfer from ru→en	15.77	41.33
transfer from ru→en,kk→ru	20.58	46.24

Table 3: Results obtained by the different strategies evaluated for combining the available parallel corpora.

strategy	it.	BLEU	chrF++
en→kk			
transfer learning	0	11.80	42.86
transfer learning	1	12.63	44.46
concatenate + fine-tune	1	13.46	44.99
concatenate + fine-tune	2	13.79	45.24
kk→en			
transfer learning	0	20.58	46.24
transfer learning	1	21.58	47.65
concatenate + fine-tune	1	22.66	48.91
concatenate + fine-tune	2	23.28	49.45

Table 4: Results obtained by the different strategies evaluated for combining parallel corpora and the back-translated data.

transfer learning in both directions. This result is surprising for Kazakh-to-English, where the transfer learning strategy makes use of more resources. One possible explanation could be that, with concatenation plus fine-tuning, the system is trained mostly on data from the news domain, as the English monolingual data is extracted only from News Crawl. Finally, the repetition of steps 2–3 helped to further improve translation quality.

4 Morphological segmentation

Morphological segmentation is a strategy for segmeting words into sub-word units that consists in splitting them into a *stem*, that carries out the meaning of the word, and a *suffix* or sequence of suffixes that contain morphological and syntatic information. When that strategy has been followed to segment the training corpus for an NMT system, it has been reported to outperform BPE for highly inflected languages such as Finnish (Sánchez-Cartagena and Toral, 2016), German (Huck et al., 2017) or Basque (Sánchez-Cartagena, 2018).

In our submissions, we morphologically segmented the Kazakh text with the Apertium Kazakh morphological analyzer.[7] For each word, the analyzer provides a set of candidate analyses made of a lemma and morphological information. Those analyses in which the lemma is a prefix of the word are considered valid analyses for segmentation and involve that the word can be morphologically segmented into the lemma and the remainder of the word.[8] When there are multiple valid analyses for a word, they are disambiguated as explained below. When a word has no valid analyses for segmentation, we generate as many segmentation candidates as known suffixes match the word (plus the empty suffix, since a possible option could be no segmenting at all). Known suffixes are extracted in advance from those words with a single valid analysis.

Multiple segmentation candidates (either coming from multiple valid analyses or from suffix matching) are disambiguated by means of the strategy described by Sánchez-Cartagena (2018), which relies on the semi-supervised morphology learning method Morfessor (Virpioja et al., 2013). We trained the Morfessor model on all the available Kazakh corpora listed in Tables 1 and 2. Finally, as suggested by Huck et al. (2017), we applied BPE splitting with a model learned on the concatenation of all training corpora after performing the morphological segmentation.

Table 5 depicts some examples of Kazakh words, their analyses and their morphological segmentation. The first word is the genitive form of университет (*university*). The morphological segmentation allows the NMT system to generalize to other inflected forms of the same word, while BPE does not split it because it is a rather frequent term in the corpus. The second word is an inflected form of the verb жаса (*to do*), although it is also analyzed as a inflected form of жасал due to an error in the analyzer. The Morfessor model preferred the wrong analysis, but the plain BPE segmentation made translation even more difficult for the MT system by choosing the prefix жас, which means *young*. BPE introduced more ambiguity, as the token жас can encode both the verb *to do* and the adjective *young*.

[7] https://github.com/apertium/apertium-eng-kaz

[8] We can safely apply this strategy because in Kazakh the stem usually corresponds to the lemma.

word	analyses	morph. seg.	plain BPE
университетінің	университет-n.px3sp.gen	университет@@ інің	университетінің
жасалмайды	жаса-v.tp.n.p3 жасал-v.i.n.p3*	жасал@@ майды	жас@@ алмайды

Table 5: Examples of Kazakh words, their morphological analyses, and their segmentation.

system	**BLEU**	**chrF++**
RNN	10.13	40.54
hybrid RNN	10.53	**41.03**
Transformer	11.71	42.65
hybrid Transformer	11.20	42.23
Apertium	1.59	26.60

Table 6: Results obtained by the different strategies evaluated for integrating the Apertium English-to-Kazakh rule-based machine translation system into an NMT system. Scores of hybrid systems are shown in bold if they outperform the corresponding pure NMT system by a statistically significant margin.

5 Hybridization with rule-based machine translation

The Apertium platform contains an English-to-Kazakh RBMT system (Sundetova et al., 2015) that may encode knowledge that is not present in the corpora available in the constrained task. In order to take advantage of that knowledge, we built a hybrid system by means of multi-source machine translation (Zoph and Knight, 2016). Our hybrid system is a multi-source NMT system with two inputs: the English sentence to be translated, and its translation into Kazakh provided by Apertium. This very same set-up has been successfully followed in the WMT automated post-editing task (Junczys-Dowmunt and Grundkiewicz, 2018).

In order to assess the viability of this approach, we trained and automatically evaluated multi-source and single-source English-to-Kazakh systems on the concatenation of the genuine English–Kazakh parallel corpora and the backtranslation of the Kazakh monolingual corpora News Crawl and Wiki dumps.[9]

Results, depicted in Table 6, show that the multi-source system is able to outperform the single-source one only with the RNN architecture (the difference is statistically significant for chrF++). Apertium output seems to be of very low quality according to the scores reported in the table.[10] Despite that, the multi-source RNN is able to extract useful information from it. The poor performance of the multi-source Transformer architecture could be related to the low quality of the Apertium output. In order to prevent that the errors in the Apertium translation are propagated to the output, the decoder should focus mostly on the SL input. However, according to the analysis of attention carried out by Libovický et al. (2018), in the serial multi-source architecture of Marian the output seems to be built with information from all inputs. We plan to explore more multi-source architectures in the future. Due to the poor performance of the Transformer multi-source architecture, we used only the multi-source RNN in our submission, as explained in the next section.

6 Final submissions

We submitted a constrained and an unconstrained ensemble for the English-to-Kazakh direction. This section describes how the individual models of the ensembles were trained and selected, and presents the results of an automatic evaluation.

Training details. All the ensembled models were trained on the genuine parallel corpora, the pivot-backtranslated corpus, and the backtranslated corpus obtained in the 3rd iteration, in a similar way to what has been described in Section 3.2. Preprocessing steps and training parameters were those described in Section 2, with the following exceptions: we applied morphological segmentation to the Kazakh text as described in Section 4, we used the full `newsdev2019` as the development corpus, and we oversampled the News Commentary parallel corpus for fine-tuning to match the size of the concatenation of all the other genuine English–Kazakh parallel corpora.

Ensemble building. Our constrained submission was an ensemble of 2 transformer models and 2 RNN models. For each architecture, the 2 models

[9] We backtranslated with the best system from Section 3.1.

[10] Sundetova et al. (2015) state that the system is only able to translate simple sentences and questions.

were checkpoints from the same training run, thus our submission only contained models from 2 independent training runs. In both cases, the first model in the ensemble was the last saved checkpoint of the main training run (that was carried out on the concatenation of all the corpora), after being fine-tuned on the genuine parallel corpora. The second model in the ensemble was the checkpoint of the main training run which, after being fine-tuned on the genuine parallel corpora and ensembled with the first model, maximized chrF++ on the development set. We gave the Transformer and RNN models different weights on the final ensemble, which were also optimized on the development set. Our unconstrained submission was created in a similar way, but the two RNN models were multi-source models such as those described in Section 5. Additionally, we built an ensemble of 5 independently trained Transformer models that could not be submitted due to time constraints.

Automatic evaluation. Table 7 shows the values of the BLEU and chrF++ automatic evaluation metrics obtained by our systems on the `newstest2019` test set. In order to assess the impact of the enhancements applied, we also show scores for single models, and for alternatives without morphological segmentation and without the additional RBMT input. We can observe that morphological segmentation slightly improves the results. In line with the results in Section 5, adding the additional Apertium input to a single model also brings an improvement according to both evaluation metrics. However, that gain vanishes when we compare the ensembles, probably because the scores obtained by the RNN models are far below those obtained by the Transformer models. Moreover, the ensemble of 5 independently trained Transformers outperforms our submitted systems, which were ensembles of only 2 independent training runs.

Comparison with other teams. Table 7 also depicts the scores obtained by the top 3 constrained systems submitted by other teams with the highest chrF++. In comparison with them, our constrained submission is ranked in 2$^{\text{nd}}$ position in terms of chrF++ and 3$^{\text{rd}}$ in terms of BLEU. Our ensemble of 5 Transformer models, built after the submission deadline, reaches the 1$^{\text{st}}$ position in terms of chrF++. There are no statistically significant differences for any of the evaluation metrics between our 5-Transformer ensemble and the best performing contestant.

system	**BLEU**	**chrF++**
single Transformer	9.25	39.48
+ morph. seg.	9.57	39.76
single RNN + morph. seg.	8.43	37.24
+ Apertium	8.68	37.99
constrained submission	9.97	40.28
unconstrained submission	9.90	40.31
ensemble 5 Transformer	10.65	41.00
`NEU`	11.11	40.78
`CUNI-T2T-transfer-enkk`	8.70	39.30
`rug_enkk_bpe`	10.30	37.65

Table 7: Results obtained by our submissions, single-model alternatives, and systems submitted by other teams, computed on `newstest2019`. There are no statistically significant differences for any of the evaluation metrics between our 5-Transformer ensemble and the `NEU` submission.

7 Concluding remarks

We have presented the Universitat d'Alacant submissions to the WMT 2019 news translation shared task for the English-to-Kazakh language pair. As it is a low-resource pair, we took advantage of parallel corpora from other language pairs via pivot backtranslation and transfer learning. We also iteratively backtranslated monolingual data and made the most of the noisy, crawled corpora after filtering it with automatic classifiers and language models. We morphologically segmented Kazakh text to improve the generalization capacity of the NMT system and successfully used multi-source machine translation to build a hybrid system that integrates the Apertium RBMT English-Kazakh RBMT engine. Our constrained submission was ranked 2$^{\text{nd}}$ in terms of chrF++.

We plan to continue exploring the hybridization of NMT and RBMT. More multi-source Transformer architectures need to be evaluated to better fit the nature of the RBMT input. Another research line involves using RBMT to generate synthetic training data.

Acknowledgments

We would like to thank Mikel L. Forcada for its contribution to the morphological segmentation algorithm, and Barry Haddow for the advice on iterative backtranslation. This work has been funded by European Union's Horizon 2020 research and innovation programme under grant agreement No 825299 (GoURMET project).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR 2015*, San Diego, CA, USA.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Matthias Huck and Hermann Ney. 2012. Pivot lightly-supervised training for statistical machine translation. In *Proc. 10th Conf. of the Association for Machine Translation in the Americas*, pages 50–57.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67.

Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan. 2018. The MLLP-UPV German-English machine translation system for WMT18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 422–428, Belgium, Brussels. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 244–252, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

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 association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1 – 38.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Víctor M Sánchez-Cartagena. 2018. Prompsit's Submission to the IWSLT 2018 Low Resource Machine Translation Task. In *Proceedings of the 15th International Workshop on Spoken Language Translation*.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Víctor M Sánchez-Cartagena and Antonio Toral. 2016. Abu-matran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 362–370.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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)*, volume 1, pages 1715–1725.

Aida Sundetova, Mikel Forcada, and Francis Tyers. 2015. A free/open-source machine translation system for English to Kazakh. In *Proceedings of the International Conference Turkic Languages Processing (Turk-Lang 2015)*, pages 78–90, Kazan, Tatarstan, Russia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. D4 julkaistu kehittämis- tai tutkimusraportti tai -selvitys.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34.