The TALP-UPC Machine Translation Systems for WMT19 News Translation Task: Pivoting Techniques for Low Resource MT

Noe Casas, José A. R. Fonollosa, Carlos Escolano, Christine Basta, Marta R. Costa-jussà
{noe.casas,jose.fonollosa,carlos.escolano}@upc.edu,
{christine.raouf.saad.basta,marta.ruiz}@upc.edu
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona

Abstract

In this article, we describe the TALP-UPC research group participation in the WMT19 news translation shared task for Kazakh-English. Given the low amount of parallel training data, we resort to using Russian as pivot language, training subword-based statistical translation systems for Russian-Kazakh and Russian-English that were then used to create two synthetic pseudo-parallel corpora for Kazakh-English and English-Kazakh respectively. Finally, a self-attention model based on the decoder part of the Transformer architecture was trained on the two pseudo-parallel corpora.

1 Introduction

Attention-based models like the Transformer architecture (Vaswani et al., 2017) or the Dynamic Convolution architecture (Wu et al., 2019) are currently the dominant approaches for Machine Translation (MT). Nevertheless, these architectures offer best results when trained on large training corpora. When faced with a low-resource scenario, other supporting techniques are needed in order to obtain good translation results. In the WMT19 news translation shared task, two low-resourced language pairs where proposed, namely Gujarati-English and Kazakh-English.

In this report, we describe the participation of the TALP Research Group at Universitat Politècnica de Catalunya (UPC) at the WMT19 news translation shared task (Barrault et al., 2019) in Kazakh→English and English→Kazakh translation directions.

The amount of available parallel Kazakh-English data is very low. In order to overcome this problem in the frame of the shared task, we made use of Russian as an pivot language. This way, we used English-Russian and Kazakh-Russian data to train intermediate translation systems that we then used to create synthetic pseudo-parallel Kazakh-English data. This data enabled us to train the final Kazakh-English translation systems.

This work is organized as follows: in section 2 we describe some techniques normally used in low-resource scenarios, to frame our proposal; in section 3 we provide an overview of other works addressing Kazakh-English as language pair for translation; in section 4 we study the available data sets, both in terms of amount and quality of the data, and describe the processing performed over it; in section 5 we describe the proposed system, together with the details about, including the data augmentation techniques used and the final NMT model trained; in section 6 we describe the experiments carried out to evaluate the translation quality prior to submitting and the obtain results; finally, in section 7 we describe the conclusions drawn from this work.

The source code used for the data download, data preparation and training of the pivot and final systems is available at https://github.com/noe/wmt19-news-lowres.

2 Low-resource NMT

There are several different approaches that can improve translation quality in under-resourced scenarios. In this section, we provide an overview of some of the dominant techniques and justify their application in the frame of this shared task.

While for low resource languages there is limited parallel data, monolingual data is often available in greater quantities. A common strategy to integrate this monolingual data into the NMT system is back-translation (Sennrich et al., 2016a), which consists in generating synthetic data by translating monolingual data of the target language into the source language that would be then fed to the system to further train it.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 155–162
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

Another common scenario is that few or no parallel data is available between the source and target languages but there is a third language or pivot. for which there is parallel data to both source and target. In this case, two systems can be trained, one from the source to the pivot language and another from the pivot to the target language. Inference will be performed as a cascade using the source to pivot system output as synthetic data to input to the pivot to the target system, obtaining a source to target translation.

An alternative to this approach could be the generation of a synthetic pseudo-parallel corpus of translated data between the source and target language through the pivot, and train a system as done in the back translation approach.

Finally, multilingual systems are recently showing nice improvements. Among the different types of multilingual systems there are the many-to-one approaches and the many-to-many approaches. The former is aiming to translate to one single language and can simply concatenate source languages (Zoph and Knight, 2016; Tubay and Costa-jussà, 2018). However, the latter either needs to use independent encoders and decoders (Schwenk and Douze, 2017; Firat et al., 2016; Escolano et al., 2019) or when using universal encoder and decoders (Johnson et al., 2017) needs to add a tag in the source input to let the system know to which language it is translating. This many-to-many systems are an alternative to pivot systems. However, most these multilingual systems are not able to achieve the level of performance of pivot systems yet.

In the frame of the WMT19 news translation shared task several of the aforementioned techniques are applicable.

An English+Russian→Kakakh multilingual system could be trained, but the amount of Kazakh-Russian data is much larger than Kazakh-English, which would bias the encoder toward Russian; as Russian is not similar to English this would decrease the effectiveness of the approach, as opposed to what happens for similar languages (Casas et al., 2018b).

Back-translation could also be applied in this context, but the amount of Kazakh monolingual data is not very large and it is crawled data, with presumably low quality. It could have been used additionally to other techniques, though.

Finally, pivoting approaches are also applicable to this scenario. The cascade approach, however, would not allow to profit from the existing parallel English-Kazakh data, making the pseudo-parallel corpus approach the most sensible option.

3 Related Work

In this section we provide an overview of the different approaches proposed in the literature for Kazakh-English machine translation.

The Apertium Rule-based Machine Translation (RBMT) system (Forcada et al., 2011) offers a generic platform to implement transfer-based rule systems for translation. This platform was used by Assem and Aida (2013) and Sundetova et al. (2014) to implement transfer rules for English→Kazakh and Kazakh→English respectively.

Assylbekov and Nurkas (2014) and Bekbulatov and Kartbayev (2014) studied the effectiveness of Statistical Machine Translation (SMT) of Kazakh to English with different segmentation strategies, trying to cope with the large amount of surface forms of Kazakh in relation to the low amount of available training data. Kartbayev (2015) studied the influence of different alignment models in SMT for Kazakh to English SMT.

Finally, Tukeyev et al. (2019) study the application of NMT to Kazakh to English translation by augmenting the training data with synthetically sentences generated with a rule-based procedure that computes variations of surface forms over simple sentence templates.

4 Corpora and Data Preparation

In order to train our MT systems, we used the data made available by the shared task organizers, including the not only Kazakh-English data but also the English-Russian and Kazakh-Russian data to train pivot translation systems. In this section we describe the data used for each language pair and the processing applied to each of them in order to compile appropriate training datasets.

4.1 Kazakh-English

The available parallel Kazakh-English corpora for the shared task included News Commentary v14, Wiki Titles v1 and a crawled corpus prepared by Bagdat Myrzakhmetov of Nazarbayev University.

Wiki Titles accounts for half of the available parallel segments, but its sentences are around 2 tokens long in average. Therefore, we decided not to include it in the training data, to avoid biasing the trained systems toward short translations.

After concatenating the training corpora, we used the standard Moses scripts to preprocess them, including tokenization, truecasing and cleaning. The statistics of the resulting training data are shown in table 1.

Table 1: Summary statistics of the Kazakh-English training data.

Lang.	Sents.	Words	Vocab.	L_{max}	L_{mean}
Kazakh	99.6K	1.2M	139.6K	85	11.7
English		1.5M	85.3K	102	14.9

The WMT organization split a part of News Commentary to use as development[1]. From this data, we left 500 parallel sentences as hold-out to assess final system translation quality and left the remaining 1566 segments as development data.

4.2 English-Russian

The available parallel English-Russian corpora for the shared task included News Commentary v14, Wiki Titles v1, Common Crawl corpus, ParaCrawl v3, Yandex Corpus and the United Nations Parallel Corpus v1.0 (Ziemski et al., 2016).

Following the rationale exposed for the English-Kazakh Wiki Titles data, we also dropped the English-Russian Wiki Titles data.

Among the other corpora, some are of very large size. In order to assemble a manageable final training dataset and taking into account the high presence of garbage in the crawled datasets, before combining the individual corpora, we filtered each corpus and selected from each a random sample of segments.

For the filtering, we applied heuristic criteria based on our visual inspection of the data, including elimination of lines with repeated separation characters (like ++++ or ----), elimination of fixed expressions (like `The time is now`, which appeared several times in some corpora) and eliminating lines with high ratio of numbers and punctuation characters.

For the random sample, from UN Corpus we took 2M segments out of 23M, from Common Crawl we took 200K out of 900K, from ParaCrawl we took 4M out of 12M and from the Yandex Corpus we took all the 1M segments. These samples were then combined and went through standard processing with Moses scripts, including tokenization, truecasing and cleaning. After combining them, we applied Moses corpus cleaning with more aggressive settings (sentences between 5 and 80 words and a maximum length ratio of 3.0 between source and target). From the combined corpus, we extracted 4000 random lines as development data and 1000 segments as hold out test set, leaving the rest for training. The statistics of the resulting training data are shown in table 2.

Table 2: Summary statistics of the English-Russian training data.

Lang.	Sents.	Words	Vocab.	L_{max}	L_{mean}
Russian	6.1M	125.6M	3.2M	80	20.7
English		144.9M	2.0M	80	23.9

4.3 Kazakh-Russian

The available parallel Kazakh-Russian corpora for the shared task included News Commentary v14 and a crawled Russian-Kazakh corpus prepared by Bagdat Myrzakhmetov of Nazarbayev University.

After concatenating the training corpora, we used the Moses scripts for preprocessing, including tokenization, truecasing and cleaning, using the same settings as for the aggressive English-Russian data cleaning described before. From the combined corpus, we extracted 4000 lines as development data and 1000 segments as hold out test set, leaving the rest for training. The statistics of the resulting training corpus are shown in table 3.

Table 3: Summary statistics of the Russian-Kazakh training data.

Lang.	Sents.	Words	Vocab.	L_{max}	L_{mean}
Russian	4.2M	78.8M	1.4M	96	18.9
Kazakh		75.3M	1.6M	70	18.0

5 System Description

The amount of available parallel training data for English-Kazakh is scarce. When an NMT system is directly trained on this data, the resulting translation quality is very low, as shown in section 6.

Given the amount of available English-Russian and Kazakh-Russian parallel training data, we decided to use Russian as pivot language. Taking into account the availability of some parallel Kazakh-English data, the pivoting approach that best suits this case is to prepare pseudo-parallel English-Kazakh and Kazakh-English cor-

[1] The part of News Commentary provided as development data was excluded from the training set.

pora based on the Russian data and then combine it with the parallel English-Kazakh data. Further justification of the technique used can be found in section 2.

In pivoting approaches, the final translation quality does not get influenced significantly if synthetic data is used for the source language side; on the other hand, using synthetic data for the target language side results in degraded translation quality in the final system (Casas et al., 2018a; Costa-Jussà et al., 2019). Therefore, we will create two different pseudo-parallel corpora for English→Kazakh and Kazakh→English.

In order to create the English→Kazakh synthetic data, we translated the Russian side of the Russian-Kazakh corpus into English. To perform this translation, we need an intermediate Russian→English system. We made use of the Russian-English corpus to train this pivot system.

In order to create the Kazakh→English synthetic data, we translated the Russian side of the Russian-English corpus into Kazakh. To perform this translation, we need an intermediate Russian→Kazakh system. We made use of the Russian-Kazakh corpus to train this pivot system.

The preparation and training of the two pivot translation systems is further described in section 5.1

Once the synthetic data was prepared by means of the pivot translation systems, we combined each synthetic corpus with the parallel data, obtaining the respective training datasets for the two translation directions. This is further described in section 5.2.

Finally, we trained the English→Kazakh and Kazakh→English translation systems on the previously described mix of parallel and synthetic corpora. The NMT model used is presented in section 5.3.

5.1 Pivot SMT Systems

For the Russian→English and Russian→Kazakh pivot translation systems we decided to use Moses (Koehn et al., 2007), a popular phrase-based Statistical Machine Translation (SMT) software package. The use of pivot approaches for SMT has been studied previously, like the works by De Gispert and Marino (2006), Wu and Wang (2007) or Utiyama and Isahara (2007).

Another option would have been to use a Neural Machine Translation (NMT) approach, but this would have required large amounts of GPU time to translate the pseudo-parallel corpora.

While the English language presents simple morphology, Russian is morphologically rich and Kazakh is agglutinative. Therefore, the amount of surface forms in a word-level vocabulary of the two latter languages is very high. This way, we decided to apply subword-level tokenization before training the SMT systems. For this, we used Byte-Pair Encoding (BPE) (Sennrich et al., 2016b) to extract a vocabulary of subword parts based on frequency statistics. We prepared separate BPE vocabularies for each language, with 32K merge operations each. Although not frequent, there are some precedents for subword tokenization in SMT, like the work by Kunchukuttan and Bhattacharyya (2016, 2017).

The use of subword tokenization leads to longer token sequence lengths compared to the usual word-based vocabularies of SMT systems. In order to cope with this fact, we configured the subword-based SMT systems to have longer n-gram order for their Language Models (LM) and phrase tables: the typical n-gram order used is 3 and we used 6. All other Moses configuration settings are the standard ones, using KenLM as language model (Heafield, 2011; Heafield et al., 2013) and MGIZA++ (Gao and Vogel, 2008) for alignment.

The data used to create the respective target-side LMs consisted of the target side of the parallel data used for training. Some improvement could have been gained by using the available extra monolingual English and Kazakh data for the LMs.

5.2 Combination of Parallel and Synthetic Data

The process followed to combine the parallel data with the synthetic data was the same for English-Kazakh and for Kazakh-English: we oversampled at 300% the parallel data and concatenated it with the synthetic data, obtaining the final training datasets on which the translation systems for the submissions were trained.

5.3 Joint Source-Target Self-Attention NMT

The translation system trained on the augmented Kazakh-English data and used for the final WMT submissions is based on the architecture proposed by (He et al., 2018; Fonollosa et al., 2019). This approach is based on the self-attention blocks from (Vaswani et al., 2017), but breaks from the

Table 4: BLEU scores (cased) of the Rule-based baseline (**RBMT**), the Moses system trained on the parallel Kazakh-English data with word-level tokenization (**SMT(w)**), the Moses system trained on the parallel Kazakh-English data with subword-level tokenization (**SMT(sw)**), the **NMT** system trained on the parallel Kazakh-English data, and the final systems trained on the augmented pseudo-parallel corpus data (**NMT pseudo-p.**)

Direction	**RBMT**	**SMT (w)**	**SMT (sw)**	**NMT**	**NMT pseudo-p.**
Kazakh→English	1.51	6.34	7.48	2.32	21.00
English→Kazakh	1.46	3.53	3.82	1.42	15.47

encoder-decoder structure and has only a single decoder block that is fed both the source and target sentences, therefore learning joint source-target representations from the initial layers. This model resembles how a language modeling architecture is trained and used for inference.

The positional encodings are applied separately to source and target. An extra embedded vector representation is added to the combination of token and position in order to distinguish source and target parts.

The attention weights can be masked to control the receptive fields (Fonollosa et al., 2019). Both source-source and target-target receptive fields are constrained to a local window around each token, while target-source receptive fields are unconstrained.

The hyperparameter configuration used was the same as the one originally used by the authors for WMT'14 English-German (14 layers, 1024 as embedding dimensionality, feedforward expansion of dimensionality 4096 and 16 attention heads).

For Kazakh-English we used separate BPE vocabularies with 32K merge operations, while for English-Kazakh we used a joint BPE vocabulary with 32K merge operations, together with shared source-target embeddings.

6 Experiments and Results

In order to assess the translation quality of the systems, we computed the BLEU score (Papineni et al., 2002) over the respective held out test sets.

As there is not much literature of current NMT approaches being applied to English-Kazakh, we prepared different baselines to gauge the range of BLEU values to expect:

- Rule-based machine translation system (RBMT): we used the Apertium system (Forcada et al., 2011; Sundetova et al., 2014; Assem and Aida, 2013), which is based on transfer rules distilled from linguistic knowledge. Using the BLEU score to compare an RBMT system with data-driven systems is not fair (see (Koehn, 2010) §8.2.7) but we included it to have a broader picture.

- Statistical Machine Translation with word-level tokenization (SMT(w)): we trained a Moses system on the parallel Kazakh-English data, using normal word-level tokenization

- Statistical Machine Translation with subword-level tokenization (SMT(sw)): we trained a Moses system on the parallel Kazakh-English data, using BPE tokenization with 10K merge operations[2]. Moses default values were used for the rest of configuration settings .

- Neural Machine Translation (NMT): we trained a Transformer model on the parallel Kazakh-English data, using BPE tokenization with 10K merge operations, separately for source and target. We used the fairseq (Ott et al., 2019) implementation with the same hyperparameters as the IWSLT model, namely an embedding dimensionality of 512, 6 layers of attention, 4 attention heads and 1024 for the feedwordward expansion dimensionality.

The translation quality BLEU scores of the aforedescribed baselines were very low, as shown in table 4.

In order to evaluate the pivot translation systems described in section 5.1, we also measured the BLEU scores in the respective held out test sets, obtaining 36.05 BLEU for the Russian→English system and 21.06 for the Russian→Kazakh system. With these pivot systems, we created two pseudo-parallel synthetic corpora, merged them with the parallel data and trained a self-attention NMT model that obtained BLEU scores one order of magnitude above the chosen baselines, as shown in table 4.

[2] The low number of BPE merge operations is justified with the low amount of training data

When we tested the final Kazakh→English system on the shared task test set, we identified several sentences that remained completely in Cyrillic script. In order to mitigate this problem, we trained a SMT system on the augmented Kazakh-English data and used it for the sentences that had a large percentage of Cyrillic characters. This lead to a mere 0.1 increase in the case-insensitive BLEU score and no change for the uncased one.

7 Conclusion

In this article we described the TALP-UPC submissions to the WMT19 news translation shared task for Kazakh-English. Our experiments showcase the effectiveness of pivoting approaches for low resourced scenarios, making use of SMT to support the data augmentation process, while using the more effective attention-based NMT approaches for the final translation systems.

Acknowledgments

We would like to thank Lluis Guardia for his 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 supported in part by a Google Faculty Research Award

This work is supported in part by the Catalan Agency for Management of University and Research Grants (AGAUR) through the FI PhD Scholarship.

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

S. Assem and S. Aida. 2013. Machine translation of different systemic languages using a apertium platform (with an example of english and kazakh languages). In *2013 International Conference on Computer Applications Technology (ICCAT)*, pages 1–4.

Zhenisbek Assylbekov and Assulan Nurkas. 2014. Initial explorations in kazakh to english statistical machine translation. In *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 12.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Eldar Bekbulatov and Amandyk Kartbayev. 2014. A study of certain morphological structures of kazakh and their impact on the machine translation quality. In *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE.

Noe Casas, Marta R. Costa-jussà, and José A. R. Fonollosa. 2018a. 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.*

Noe Casas, Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2018b. The TALP-UPC machine translation systems for WMT18 news shared translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 355–360, Belgium, Brussels. Association for Computational Linguistics.

Marta R. Costa-Jussà, Noé Casas, Carlos Escolano, and José A. R. Fonollosa. 2019. Chinese-catalan: A neural machine translation approach based on pivoting and attention mechanisms. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):43.

Adrià De Gispert and Jose B Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68. Citeseer.

Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1905.06831.*

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

José A. R. Fonollosa, Noe Casas, and Marta R. Costa-jussà. 2019. Joint source-target self attention with locality constraints. *arXiv preprint arXiv:1905.06596.*

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.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7955–7965. Curran Associates, Inc.

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.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

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.

Amandyk Kartbayev. 2015. Learning word alignment models for kazakh-english machine translation. In *Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 326–335, Cham. Springer International Publishing.

Philipp Koehn. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, New York, NY, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*, ACL '07, pages 177–180. Association for Computational Linguistics.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Orthographic syllable as basic unit for SMT between related languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1912–1917, Austin, Texas. Association for Computational Linguistics.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2017. Learning variable length units for SMT between related languages via byte pair encoding. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 14–24, Copenhagen, Denmark. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Aida Sundetova, Aidana Karibayeva, and Ualsher Tukeyev. 2014. Structural transfer rules for kazakh-to-english machine translation in the free/open-source platform apertium. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 7(2):48–53.

Brian Tubay and Marta R. Costa-jussà. 2018. Neural machine translation with the transformer and multi-source romance languages for the biomedical WMT 2018 task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 667–670, Belgium, Brussels. Association for Computational Linguistics.

Ualsher Tukeyev, Aidana Karibayeva, and Balzhan Abduali. 2019. Neural machine translation system for the kazakh language based on synthetic corpora. In *MATEC Web of Conferences*, volume 252, page 03006. EDP Sciences.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491.

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 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.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

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, San Diego, California. Association for Computational Linguistics.