The AFRL-MITLL WMT17 Systems: Old, New, Borrowed, BLEU

Jeremy Gwinnup, Timothy Anderson, Grant Erdmann, Katherine Young
Air Force Research Laboratory
{jeremy.gwinnup.1,timothy.anderson.20,
grant.erdmann,katherine.young.1.ctr}
@us.af.mil

Michaeel Kazi, Elizabeth Salesky, Brian Thompson, Jonathan Taylor
MIT Lincoln Laboratory
{michaeel.kazi,elizabeth.salesky,
brian.thompson,jonathan.taylor}
@ll.mit.edu

Abstract

This paper describes the AFRL-MITLL machine translation systems and the improvements that were developed during the WMT17 evaluation campaign. This year, we explore the continuing proliferation of Neural Machine Translation toolkits, revisit our previous data-selection efforts for use in training systems with these new toolkits and expand our participation to the Russian–English, Turkish–English and Chinese–English translation pairs.

1 Introduction

As part of the 2017 Conference on Machine Translation (WMT, 2017) news-translation shared task, the MITLL and AFRL human language technology teams participated in the Russian–English, English–Russian, Turkish–English and Chinese–English tasks.

Our machine translation systems this year are a departure from our previous Moses (Koehn et al., 2007) based systems from WMT16 (Gwinnup et al., 2016). We employ systems built with the Nematus (Sennrich et al., 2017) toolkit as in our IWSLT2016 (Kazi et al., 2016) systems, the Nematus-compatible Marian training toolkit and AmuNMT decoder (Junczys-Dowmunt et al., 2016) and the OpenNMT (Klein et al., 2017) toolkit.

For the Russian–English and Turkish–English language pairs, we submitted an entry comprising the best systems combined using the Jane system combination method (Freitag et al., 2014) and the best-scoring single system for that language pair. For the Chinese–English and English-Russian language pairs, we only submitted our single-best system.

Portions of this work are sponsored by the Air Force Research Laboratory under Air Force contracts FA-8721-05-C-0002 and FA-8650-09-D-6939-029.

2 Data and Preparation

2.1 Data Used

We utilized all available data sources provided for the language pairs we participated in, including the Commoncrawl (Smith et al., 2013), Yandex[1], UN v1.0 (Ziemski et al., 2016), SETimes (Tyers and Alperen, 2010) corpora.

2.2 Data Preparation

The Russian/English files were cleaned to remove blank lines, replace carriage returns with line feed characters, remove wrong-language text, and correct mixed alphabet spellings, following techniques outlined in (Young et al., 2016) and (Schwartz et al., 2014).

The number of non-parallel blank lines in the Russian/English news commentary files indicated some sentence alignment errors, so these files were re-processed using the NLTK Punkt (Kiss and Strunk, 2006) sentence segmenter and the Champollion sentence aligner (Ma, 2006) before cleaning. Altogether, 9537 of the original 236,314 newscommentary lines were removed during the clean-up process.

The Chinese files were word-segmented with Jieba[2] and the Stanford Chinese segmenter (Chang et al., 2008). The Chinese–English parallel data was cleaned to replace carriage returns and to remove wrong-language text. Lines with URLs in *http* were also removed, because of a difference in the Chinese and English tokenization. Altogether, the clean-up process removed 310,121 of the 21,248,495 lines in the combined file.

[1] https://translate.yandex.ru/corpus?lang=en
[2] https://github.com/fxsjy/jieba

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 303–309
Copenhagen, Denmark, September 711, 2017.

2.3 Subselection

We use our corpus subselection algorithm, defined in (Gwinnup et al., 2016). We use a vocabulary of up to 4-grams for subselection, after using byte-pair encoding (see Section 3) to produce sub-word units. We believe that selecting from subwords is especially beneficial in morphologically-complex languages like Turkish and Russian.

For Russian we conducted monolingual selection from provided Common Crawl, to match test sets from 2012-2016 (15K lines total). This corpus was broken into 571 chunks of one million lines each, and five thousand lines were selected from each (2.9M lines total). 3-gram and 4-gram subword subselection vocabulary was used.

For Turkish we conducted monolingual selection from Common Crawl, to match SE Times and dev/test 2016 corpora (212K lines total). This corpus was broken into 502 chunks of one million lines each, and 25 thousand lines were selected from each (12.5M lines total). 4-gram subword subselection vocabulary was used.

After this subselection process completed for various languages we then sampled the first 3000 (e.g. top-scoring) English sentences from each selected chunk. For Russian and Turkish, we utilized the entire subselected chunk. Final line-counts for these selected data sets are listed in Table 1.

Language	Final Lines
English	8,921,942
Russian	2,856,141
Turkish	5,011,001

Table 1: Final count of subselected lines per language used in training AFRL's backtranslation systems.

3 MT System Descriptions

This year we participated in the Russian–English, English–Russian, Turkish–English and Chinese–English translation pairs using a variety of toolkits and techniques. Of particular note, we employed byte-pair encoding (Sennrich et al., 2016b) (BPE) of the source and target training data to address the out-of-vocabulary(OOV) problem.

3.1 Russian–English

The Russian–English language pair has been our largest focus since our participation in WMT14. We spent significant effort building a variety of systems described as described below.

3.1.1 AFRL Nematus/Marian Systems

Our Nematus/Marian systems follow the general approach of the WMT16 Edinburgh NMT systems (Sennrich et al., 2016a) with the following differences: We use the data selection algorithm described in Section 2.3 yielding approximately 5 million additional lines of backtranslated data.

In order to produce this backtranslated data we performed the following steps: 1) We first used Edinburgh's backtranslated data from WMT16 to produce a Nematus-based Russian–English system. 2) Once trained, we used the Amun decoder to translate the 2.8 million lines of subselected monolingual Commoncrawl Russian data into English. 3) The resulting data was then used to train an English–Russian Marian system that then used the Amun decoder to translate the 8.9 million lines of subselected English data to Russian. 4) Following this decoding, a final Russian–English system was trained using Marian with this backtranslated data. Three separate Marian training runs were performed with this final data set. Additionally, a Nematus system was trained for rescoring purposes where the English target data was reversed in word order. The combination of these final inputs was optimized with Drem (Erdmann and Gwinnup, 2015) to determine feature weights.

3.1.2 AFRL OpenNMT Systems

We trained four OpenNMT systems. Two systems employed the backtranslated data used in last year's University of Edinburgh NMT systems (Sennrich et al., 2017). The other two systems employed the subselected data as described in Section 2.3. All systems used 1000 hidden units and 600 unit word embeddings.

The two WMT16-based systems were each fine-tuned with `newstest2012-2015` data. One system was also incrementally trained with the same newstest data. The subselected systems had cased BLEU scores of 30.04 and 30.67 on `newstest2017` while the WMT16-based systems had BLEU scores of 32.16 and 32.78. They were all single systems.

Since OpenNMT currently does not support ensemble decoding, we decided to try doing system combination on the last four epochs of training. Taking the best system from the subselected data then gave a BLEU score of 33.95 while the best

WMT16-based systems increased to 33.23. Combining the four ensembles of each of those systems resulted in a score of 34.45 BLEU. This last ensemble system combination was done after the submission deadline.

3.1.3 MITLL Phrase-Based System

While similar to last year's phrase-based system (Gwinnup et al., 2016), this year's system differs in a few key ways: 1) We use Moses truecased training data, to make our tokenization scheme uniform; 2) We rescore using systems built from data made available by Edinburgh's WMT16 System (Sennrich et al., 2016a); 3) We updated our language models with the new monolingual data sources, and finally 4) We add an additional 4 million lines from the UN v1.0 corpus (Ziemski et al., 2016) into the parallel training data.

For the last item, we used Moore-Lewis (Moore and Lewis, 2010) filtering on the English side of the training data. The in-domain language model was trained on `news.2015.shuffled.en` using a single layer LSTM language model developed in-house. The out-of-domain language model (trained on UNv1.0) used the same vocabulary. We compared word vs character-level language model results, and noted that character-level language modeling did a good job of data cleanup (giving bad scores to personnel records and poorly formatted data). We swept data selection sizes of two, four, and eight million, and found the middle size consistently the best. Our phrase-based system results can be summarized in Table 2.

System	Cased BLEU
Baseline	24.95
Rescore	27.32
Rescore + UN ML-words	27.80
Rescore + UN ML-chars	27.86
Rescore + UN ML-both	28.05
Resc. + UN ML-both + new LMs	28.41

Table 2: MITLL phrase-based system scores on `newstest2016` measured in cased BLEU.

3.1.4 MITLL OpenNMT Systems

We trained an OpenNMT system with the same in-domain data as our phrase-based system, using the default 9 epochs at learning rate 1.0, and reducing the learning rate by 0.7 each epoch thereafter. This yielded a system with 29.07 BLEU on `newstest2016`. Creating an n-best list from the epoch 13 model and rescoring that n-best list with the models from epoch 11 and 12, combined with equal weight, yielded 29.55 BLEU.

3.1.5 AFRL Phrase-Based Systems

In order to provide diversity for system combination, we trained a Moses system with the provided parallel data and the subselected, backtranslated data as outlined in Section 3.1.1. We trained a 5-gram, BPE'd language model from the data used to train the BigLM used in our WMT15 (Gwinnup et al., 2015) systems.

3.2 English–Russian

Due to the surprising effectiveness of the Marian English–Russian translation system used to produce backtranslated data, we decided to enter this system in the English–Russian translation task. This system was used in Step 2 of the Russian–English training process detailed in Section 3.1.1. Results of decoding `newstest2017` are listed as entry 3 in Table 7.

3.3 Turkish–English

We apply the techniques employed in building our Russian–English systems to build Turkish–English translation systems.

3.3.1 AFRL Nematus/Marian Systems

For the Turkish–English task, the only provided parallel data was the SETimes corpus (Tyers and Alperen, 2010) of approximately 220,000 parallel lines. This presented a challenge for our goal of training a neural-based system similar to our Russian–English system (Section 3.1.4). We adopted a multiple step approach as before, but first starting with a Turkish–English Moses (Koehn et al., 2007) system built on the SETimes corpus with BPE applied. An order-5 KenLM (Heafield, 2011) language model was built on a BPE'd version of the BigLM employed in our WMT15 system(Gwinnup et al., 2015). Hierarchical lexicalized reordering (Galley and Manning, 2008) and an order-5 Operation Sequence Model (Durrani et al., 2011) were also employed in this system. Drem (Erdmann and Gwinnup, 2015) was used to optimize system feature weights using the Expected Corpus Bleu (ECB) metric.

In the interest of speed, Moses2 (Hoang et al., 2016) was used to decode the subselected Turkish corpus. An English–Turkish Marian system

was then trained (with default parameters) with the provided parallel data and the backtranslated data from the previous step. This system was then used to decode the English subselected corpus. Finally, our non-combination submission system was trained using both the parallel provided data and the data generated from the previous back-translation step. This final Marian system was trained with a source vocabulary of 70k, target vocabulary of 50k, a 2048-unit RNN hidden layer and a 512-unit word embedding layer. A Nematus system was trained with reversed target sentences to provide right-to-left(r2l) rescoring. Two Marian left-to-right (l2r) and one Nematus r2l training instances were run. Each of the 3 final models are an average of the 8 best-scoring model checkpoints for each distinct training run. These resulting l2r averaged models were used to ensemble decode the test set, with the averaged r2l model rescoring the resulting n-best lists. Finally, the one-best was output and submitted as System 5 in Table 7.

3.3.2 MITLL OpenNMT Systems

In the final week of the evaluation, to produce a diverse system, we attempted backtranslation, iteratively. We began with a Moses system trained on the SETimes corpus. We then took 800K sentences from `news.2016.shuffled` for either language. In training a Turkish to English MT system, we backtranslated the English news data into Turkish using the current best English–Turkish MT system. We then repeated the process in the other direction. In the interest of time, we used a small network with 256 sized word embeddings, 512 sized rnn, and learning rate decay starting at epoch 6. Each pass took one day. Perplexities converged after 3 iterations. See Table 3 below.

Iter	Forward ppl.	Backward ppl.
1	26.92	31.72
2	22.65	27.74
3	16.75	27.88

Table 3: MITLL OpenNMT Turkish–English system perplexities on `newsdev2016`.

3.3.3 AFRL Moses Phrase-Based Systems

For contrast, a phrase-based system was built in the same manner as described in Step 1 of Section 3.3.1, but using the provided and backtranslated data used in the final step. This system contributed to the system combination listed as entry 4 of Table 7.

3.4 Chinese–English

3.4.1 MITLL Nematus and OpenNMT Systems

As in our other systems, we used Moore-Lewis filtering (on characters only here due to time constraints) to sort the data. In this case, we used the entire parallel training corpora provided (25M lines), and filtered it, since we had no prior knowledge of which corpora were useful. For our Nematus system, we took the top 20 million lines, using the subselection method as a form of data "cleanup". Since this system took a month to train, for our OpenNMT system we instead extracted the top 5M sentences, and this system trained in one week. The Nematus system trained to a BLEU score of 16.39 on `newstest2016`, ensembled to 18.59, and the single-best OpenNMT system trained to 18.30. (OpenNMT did not have ensemble decoding implemented at the time of the evaluation.) We also rescored the Nematus ensembled n-best list with our OpenNMT system. We used an n-best list size of 12, and achieved a score of 20.06 (+.06) on `newstest2017`.

3.4.2 AFRL OpenNMT Systems

Similarly to the Chinese–English systems in the previous section, we down-sampled the available parallel data using the algorithms described in (Gwinnup et al., 2016) resulting in a 5 million line parallel training set. OpenNMT systems were trained in the same manner described in Section 3.1.2. The outputs of the 8 best-scoring epochs were ensembled using system combination again in the same manner as the Russian–English systems. This resulting system is listed as entry 6 in Table 7.

3.4.3 AFRL Marian Systems

Again for contrast, we experimented using 5 million lines of down-selected data from the parallel UN corpus as in Section 3.4.2. We character-segmented all Chinese characters on the source side of the data, then applied a BPE model to any remaining non-Chinese words. This BPE model is the same as the one learned from and applied to the target side of the parallel training data. Interestingly, this approach limited the source vocabulary to only 22,000 terms. The target vocabulary is a more typical 40K due to the application of BPE.

Marian was used to train models with 1024, 2048, and 3072 hidden units in the RNN layer. We saw a performance gain when increasing the number of units from 1024 to 2048, but not from 2048 to 3072 (at least for this experiment). These scores are shown in the Table 4.

RNN width	cased BLEU
1024	17.75
2048	18.81
3072	18.84

Table 4: Chinese–English Marian systems with different RNN hiddenunit widths decoding `newstest2017` measured in cased BLEU.

3.5 System Combination

Jane System Combination (Freitag et al., 2014) was used to combine a variety of systems for our Russian–English and Turkish–English combination submissions. We show the individual system combination inputs and final scores for Russian–English in Table 5 and Turkish–English in Table 6. It is important to note that our single-best Russian–English submission did not contribute to the system-combination entry as this system was a late addition at the end of the evaluation period.

For each system combination, five experiment replicates were run to account for variance in the combination process. The resulting best replicate was submitted. Results are shown in Table 7.

4 Conclusion

We present a series of improvements to our Russian–English systems and apply these lessons learned to creating Turkish–English and Chinese–English systems.

While researchers in recent years have been searching for principled methods to combine the strengths of statistical and neural MT, we find that carefully devised system combination and ensembling provides provides aggregate improvement. Thus, "borrowing" the Jane system combination technique allows one to combine old and new for better BLEU.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 22 June 2017. Originator reference number RH-17-117218. Case number 88ABW-2017-3080.

References

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Columbus, Ohio, pages 224–232.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*. Portland, Oregon, pages 1045–1054.

Grant Erdmann and Jeremy Gwinnup. 2015. Drem: The AFRL submission to the WMT15 tuning task. In *Proc. of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pages 422–427.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open source machine translation system combination. In *Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden, pages 29–32.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08, pages 848–856.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Katherine Young, Michaeel Kazi, Elizabeth Salesky, and Brian Thompson. 2016. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Association for Computational Linguistics, chapter The AFRL-MITLL WMT16 News-Translation Task Systems, pages 296–302. https://doi.org/10.18653/v1/W16-2313.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Katherine Young, Christina May, Michaeel Kazi, Elizabeth Salesky, and Brian Thompson. 2015. The AFRL-MITLL WMT15 system: There's more than one way to decode it! In *Proc. of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pages 112–119.

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

Hieu Hoang, Nikolay Bogoychev, Lane Schwartz, and Marcin Junczys-Dowmunt. 2016. Fast, scalable phrase-based smt decoding. In *Proceedings of the 12th Biennial Conference of the Association for Machine Translation in the Americas (AMTA2016)*.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proc. of the 13th International Workshop on Spoken Language Translation (IWSLT'16)*. Seattle, Washington.

#	Description	cased BLEU
1	3x Marian Train ens, 8-best avg per, 1024 unit hidden + 1x Nematus r2l rescore, 8-best avg	31.44
2	2x Marian Train ens., 8-best avg per, 2048 unit hidden + 1x Nematus r2l rescore, 8-best avg	31.86
3	3x Marian Train ens., 8-best avg per, 2048 unit hidden + 1x Nematus r2l rescore, 8-best avg	32.25
4	Moses w/ BPE data, BPE'd BigLM from WMT15	30.53
5	OpenNMT with subsel, backtrans data, single decode	30.04
6	OpenNMT with backtrans data, finetune, inc. train	32.16
7	OpenNMT with backtrans data, finetune only	32.78
8	OpenNMT with UN data + Nematus rescore	33.24
9	OpenNMT with backtrans data, ensemble-syscomb 4-best	33.23
10	OpenNMT with subsel, backtrans data, ensemble syscomb 4-best	31.91
comb	Submitted System Combination	34.71

Table 5: Russian–English System Combination Inputs decoding `newstest2017` measured in cased BLEU.

#	Description	cased BLEU
1	2x Marian Train ens, 8-best avg per, 2048 unit hidden + 1x Nematus r2l rescore, 8-best avg	17.12
2	2x Marian Train ens, 8-best avg per, 2048 unit hidden + 1x Nematus r2l rescore, 8-best avg - alt. model	17.54
3	OpenNMT, iterative backtrans	15.98
4	Moses w/BPE data, BPE'd BigLM from WMT15	13.60
comb	Submitted System Combination	18.05

Table 6: Turkish–English System Combination Inputs decoding `newstest2017` measured in cased BLEU.

#	Lang	System	Cased BLEU
1	ru-en	System Combo	34.7
2	ru-en	OpenNMT-best	34.0
3	en-ru	Marian Backtrans	25.4
4	tr-en	System Combo	18.1
5	tr-en	Marian/Nematus	17.5
6	zh-en	Single-Best	21.3

Table 7: Submission system scores on `newstest2017` measured in cased BLEU.

Michaeel Kazi, , Elizabeth Salesky, Brian Thompson, Jonathan Taylor, Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Eric Hansen, Brian Ore, Katherine Young, and Michael Hutt. 2016. The MITLL-AFRL IWSLT-2016 systems. In *Proc. of the 13th International Workshop on Spoken Language Translation (IWSLT'16)*. Seattle, Washington.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics, Volume 32, Number 4, December 2006* .

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *CoRR* abs/1701.02810. http://arxiv.org/abs/1701.02810.

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 *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL '07, pages 177–180.

X. Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Eval-*

uation (LREC'06). European Language Resources Association (ELRA).

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 220–224.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*. Philadelphia, Pennsylvania, pages 311–318.

Lane Schwartz, Timothy Anderson, Jeremy Gwinnup, and Katherine Young. 2014. Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT'14)*. Baltimore, Maryland, USA, pages 186–194.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dwmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain.

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*. Association for Computational Linguistics, Berlin, Germany, pages 371–376.

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 (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*. Sofia, Bulgaria, pages 1374–1383.

Francis M. Tyers and Murat Sedar Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Proceedings of the MultiLR Workshop at the Language Resources and Evaluation Conference, LREC2010*.

WMT. 2017. Findings of the 2017 Conference on Statistical Machine Translation. In *Proc. of the Second Conference on Statistical Machine Translation (WMT '17)*. Copenhagen, Denmark.

Katherine Young, Jeremy Gwinnup, and Lane Schwartz. 2016. A taxonomy of weeds: A field guide for corpus curators to winnowing the parallel text harvest. In *Proceedings of the 12th Biennial Conference of the Association for Machine Translation in the Americas (AMTA2016)*.

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)*.