Abstract

This paper presents the systems submitted by the Abu-MaTran project to the English-to-Finnish language pair at the WMT 2016 news translation task. We applied morphological segmentation and deep learning in order to address (i) the data scarcity problem caused by the lack of in-domain parallel data in the constrained task and (ii) the complex morphology of Finnish.

We submitted a neural machine translation system, a statistical machine translation system reranked with a neural language model and the combination of their outputs tuned on character sequences. The combination and the neural system were ranked first and second respectively according to automatic evaluation metrics and tied for the first place in the human evaluation.

1 Introduction

This paper presents the machine translation (MT) systems submitted by the Abu-MaTran project to the WMT 2016 news translation task. We participated in the English-to-Finnish constrained task.

English-to-Finnish is a particularly challenging language pair for corpus-based MT because of the lack of in-domain parallel data (the only available parallel corpus in the shared task is Europarl) and the complex morphology of Finnish. The fact that the same root can be inflected in many different ways and that nouns can be joined together in order to build compound words exacerbates the aforementioned lack of parallel data problem.

As in our last year’s submission (Rubino et al., 2015), we used morphological segmentation (Pirinen, 2015) on the Finnish side in order to deal with data scarcity and reduce the size of the Finnish vocabulary. We also used character-level evaluation metrics during the development of our systems, which correlate better than word-based ones with human judgements according to the results of last year’s metrics shared task (Stanojević et al., 2015) for English-to-Finnish.

When a Finnish sentence is morphologically segmented, it becomes much longer (number of tokens) than its English counterpart. This results in the distance between the Finnish tokens that depend on each other to produce a correct translation increasing too.¹ We addressed this potential issue by introducing deep learning in our systems: we submitted a neural MT (NMT) system and a phrase-based statistical MT (SMT) system enhanced with a neural language model (LM). In the latter, we reduced the length of the Finnish segmented sentences by joining the most frequent sequences of morphs. We also submitted a system that combines the outputs of our best NMT and SMT systems and is tuned on character sequences.

The paper is organised as follows: the data and tools used are described in Section 2, while our NMT, SMT and combined submissions are presented respectively in sections 3, 4 and 5. The paper ends with some concluding remarks.

2 Datasets and Tools

We preprocessed the training corpora with scripts included in the Moses toolkit (Koehn et al., 2007). We performed the following operations: punctuation normalisation, tokenisation, true-casing and escaping of problematic characters. The true-caser is lexicon-based and it was trained on all the monolingual data. In addition, we removed sentence pairs from the parallel corpora where either side is longer than 80 tokens.

¹For instance, the distance between the morph that represents the case of an adjective and the morph that represents the case of the noun being modified by the adjective is increased. Morphs are the segments in which a word is split after applying morphological segmentation (see Section 3.1).
Since the Common Crawl Finnish monolingual corpus was obtained by crawling websites, we applied a set of additional preprocessing steps in order to remove as much noisy data as possible: (i) detecting sentences with an incorrect character encoding and re-encoding them with the right one; (ii) replacing XML entities with the characters they represent; (iii) removing sentences with a low proportion of alphabetic characters (less than 50%); (iv) removing short sentences (less than 3 alphabetic tokens); and (v) removing sentences whose first 18 tokens are equal to those in another sentence. The last filtering is necessary because it is relatively common in the corpus to find the same sentence with some segment missing at the end. If these lines were kept, \( n \)-gram counts from which LM probabilities are estimated would be less reliable. As a result of these preprocessing steps, around 43 million sentences were removed.

Table 1 shows the Finnish monolingual corpora we used together with their size and Table 3 shows the same information for the parallel corpora. We used an additional synthetic parallel corpus to train our NMT system, which was obtained by backtranslating the Finnish News Crawl corpora into English with an SMT system (see Section 3).\(^2\) The monolingual corpora used for training its LM are listed in Table 2.

Throughout the paper we evaluate the systems we build in terms on three automatic evaluation metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and chrF1 (Popović, 2015). As the performance obtained in the development (newsdev2015) and validation (newstest2015) sets guides our decisions, we believe it is sensible to use three metrics with different underlying methodologies and that work on different elements (words and characters). Statistical significance of the difference between systems is computed with paired bootstrap resampling (Koehn, 2004) \((p \leq 0.05, 1,000 \text{ iterations})\).

\(^2\)The number of sentences in News Crawl displayed in Tables 1 and 3 do not match because, due to time constraints, we did not backtranslate a few tens of thousands of sentences.

\(^3\)We used the code available at: https://github.com/sebastien-j/LV_groundhog/tree/master/experiments/nmt

| Corpus              | Sentences (k) | Words (M) | Corpus              | Sentences (k) | Words (M) |
|---------------------|--------------|-----------|---------------------|--------------|-----------|
| Europarl v7         | 2,121        | 39.5      | Europarl v8         | 1901         | 50.9      |
| Common Crawl        | 113,995      | 2,416.7   | English             | Finnish      |
| News Crawl 2014–15  | 6,741        | 83.1      | backtranslated      |              |
| News Crawl 2007–15  | 117,446      | 2,713.2   | (only for NMT)      |              |
| News Discussions    | 57,804       | 983.2     |                     |              |

Table 3: Parallel data, after preprocessing, used to train our SMT and NMT systems.
followed the set-up of the rule-based morphologically segmented system from our last year’s constrained submission (Rubino et al., 2015). It was trained on Europarl and the concatenation of the English monolingual corpora listed in Table 2.

Most of the NMT architectures in the literature can only operate with a fixed TL vocabulary (that ranges from 30,000 to 80,000 words, according to Jean et al. (2015)), since training and decoding computational complexity grows with its size. Although Jean et al. (2015) proposed an to reduce that complexity and hence use larger vocabularies, Sennrich et al. (2015b) showed that segmenting words into smaller units can also reduce complexity, increase effective vocabulary size and even improve translation quality. We followed the latter strategy. The evaluation of character-based NMT approaches (Ling et al., 2015; Costa-Jussà and Fonollosa, 2016; Chung et al., 2016b) was left as future work.

In the remainder of this section, we present the segmentation approach we followed together with the alternatives we evaluated and we describe the training and decoding set-up of our NMT system, including the strategy followed to translate out-of-vocabulary words (OOVs).

### 3.1 Word Segmentation

Existing word segmentation approaches for NMT (Sennrich et al., 2015b) rely on frequencies of sequences of characters in the training corpus. We studied whether using linguistic information to segment the training corpus allows the neural network to generalise better: we applied the rule-based morphological segmentation provided by Omorfi (Pirinen, 2015) for Finnish. It splits words into morphs, that is, minimal segments carrying semantic or syntactic meaning.

We evaluated the segmentation schemes listed below.\(^4\) Table 4 depicts an example of the effect they produce on a Finnish sentence.

- No segmentation at all.
- Byte pair encoding (BPE) on both the source language (SL) and the TL. This is one of the best performing strategies proposed by Sennrich et al. (2015b). It consists of initially segmenting each word in characters, and iteratively joining the most frequent pair of segments in the training corpus. We applied it independently to the SL and TL sides of the parallel corpus. We performed 60,000 join operations on each language.
- BPE only on the TL side of the parallel corpus, since Finnish is morphologically more complex than English.
- Morphological segmentation with Omorfi on the TL.
- BPE on the TL using the morphs produced by Omorfi as the starting point. We evaluated the effect of performing 1,000, 10,000, 25,000 and 50,000 join operations.
- Morphological segmentation produces an average sentence length significantly higher than that of the English side of the parallel corpus. After performing 1,000 operations, average sentence lengths are similar: we reduce vocabulary size without significantly increasing sentence length. As the number of operations increases, average sentence length is closer to that of the unsegmented approach.

For each of these segmentation schemes, we trained an NMT system on Europarl during 5 days (a model was saved every 3 hours of training), we chose the model that achieved the highest translation quality on newsdev2015\(^5\) and evaluated it on newstest2015. The remainder of the training and decoding parameters were the same ones we used in our submission (described in Section 3.2).

Table 5 depicts the results of the evaluation together with the vocabulary size of the NMT system\(^6\) and the proportion of tokens in the training corpus that belong to the vocabulary. Results show that, despite the fact that the BPE-based systems have full coverage of the training corpus, their performance is below that of the unsegmented alternative. These results are probably related to the fact that domains of the training and testing corpora do not match, and words in the test set that do not contain subsegments observed in the training

\(^4\)We did not include unsupervised morphological segmentation (Virpioja et al., 2013; Grönroos et al., 2014) in our evaluation since the results in our last year’s submission (Rubino et al., 2015, Table 4) showed that it was outperformed by rule-based morphological segmentation.

\(^5\)Translation quality was measured by chrF1 in the segmentation alternatives that included BPE, since segments were joined before performing the evaluation and this metric is reported to correlate better than BLEU with human judgments. For the evaluation of the segmentation scheme based solely on Omorfi, we chose the best model according to BLEU, as the evaluation was performed before joining the morphs (the TL side of the development corpus was also segmented with Omorfi).

\(^6\)This size may represent words or subword units, depending on whether word segmentation was performed.
Table 4: Example of the application of the different segmentation schemes described in Section 3.1 to a Finnish sentence. Arrows represent boundaries between the morphs in which a word is split. Note how the compound word *perusasian* is segmented by the different schemes: Omorfi splits it into *perus* (“basic”), *asia* (“thing, affair”) and the case marker *n*, while the application of BPE over it joins the marker to the second noun. The pure BPE scheme, however, fails to segment *perusasian* correctly.

### 3.3 Dealing with Unknown Words

In order to translate OOVs, we followed an enhanced version of the approach by Jean et al. (2015, Sec. 3.3). OOVs in the training corpus were replaced with the special token UNK, as were those in the SL sentences to be translated by the NMT system. As a result, the output contained some UNK tokens.

In order to replace the UNK tokens generated by the model, we identified the most likely SL word to which the unknown TL word was aligned. If the SL word started with an uppercase letter, we copied it to the output. Otherwise, we replaced the UNK token with its translation according to a bilingual dictionary obtained from the parallel corpus with fast align (Dyer et al., 2013).

For each UNK token, Jean et al. (2015) selected the SL word with the highest alignment probability according to the attention mechanism, while our enhanced approach combines the attention mechanism and a heuristic that aims at preserving the named entities in the SL sentence. We considered the top 5 SL words with the highest attention alignment probability for each UNK token, and, for each sentence, we chose the set of SL words that ensured that the maximum number of words that start with an uppercase letter in the SL sentence were included in the translation. Ta-
Table 5: Results of the evaluation of different word segmentation schemes on an NMT system trained on Europarl. The vocabulary size of the NMT system is depicted, as well as the proportion of tokens covered in the training corpus. Scores displayed correspond to the evaluation on newstest2015. The best score for each metric is shown in bold. An arrow pointing upwards (↑) means that the corresponding system outperforms the system without segmentation by a statistically significant margin, while an arrow pointing downwards (↓) means the opposite: the system without segmentation wins.

Table 6: Results of the evaluation on newstest2016 of our NMT submission (in bold), the simpler strategy for translating unknown words by Jean et al. (2015, Sec. 3.3) (labelled as most probable SL word) and our best individual NMT model. The best score for each metric is shown in bold. An arrow pointing upwards (↑) means that the corresponding system outperforms the system in the previous row by a statistically significant margin.

4 Statistical Machine Translation

Our work on SMT systems built upon our last year’s best constrained individual system (Rubino et al., 2015). This was a phrase-based SMT system where the Finnish data was segmented to morphs with Omorfi (Pirinen, 2015). It also used two additional models: an Operation Sequence Model (Durrani et al., 2011) and a Bilingual Neural Language Model (Devlin et al., 2014), as well as three reordering models: word- and phrase-based and hierarchical (Koehn et al., 2005; Galley and Manning, 2008).

This year’s SMT systems used the same models and datasets, except for the LMs, which this time were log-linearly interpolated and used the additional corpus available (Common Crawl, cf. Table 1). We built three SMT systems, which share the same models and data, with the only difference being the segmentation used in the Finnish data:

- No segmentation.
- Segmentation on morphs (Omorfi).
- Segmentation on morphs followed by joining the most frequent sequences (Omorfi + BPE).

In the latter we joined the most frequent sequences (1 000 operations) so that the length of the Finnish side (measured in number of tokens) becomes similar to that of the English side. As previously mentioned in Section 3.1, this is a trade-off to avoid both having a big vocabulary (as is the case without segmentation), and having to deal with long-distance phenomena (as is the case with Omorfi).

Table 7 shows the results of these three SMT systems. We corroborate the results found out last year, i.e. morphological segmentation outperforms the unsegmented system by a statistically significant margin across all the automatic metrics. We also observe that joining the most frequent morphs results in a further improvement on BLEU (2.3% relative), and small changes in TER (−0.5%) and chrF1 (−0.3%).
Table 7: Results of the evaluation on newstest2016 of the SMT systems built. The best score for each metric is shown in bold. An arrow pointing upwards (↑) means that the corresponding system outperforms the system without segmentation by a statistically significant margin.

| System         | BLEU   | TER    | chrF1 |
|----------------|--------|--------|-------|
| No segmentation| 0.1444 | 0.7775 | 49.63 |
| Omorfi         | 0.1501 ↑| 0.7717 ↑| 51.13 ↑|
| Omorfi + BPE   | 0.1536 ↑| 0.7679 ↑| 50.99 ↑|

4.1 Reranking

We reranked the n-best list (top 500 distinct translations) produced by our best SMT system (Omorfi + BPE) using two neural LMs: left-to-right (i.e. trained in the same direction as the LMs included in the SMT system) and right-to-left (i.e. reverse direction). We hypothesise that the latter LM might bring a higher improvement as the sequences this LM is trained on have not been used by the SMT decoder.\(^\text{12}\)

Both neural LMs were trained on in-domain data (a subset of 4 million sentences\(^\text{13}\) randomly selected from News Crawl) with the rwthlm toolkit (Sundermeyer et al., 2014). The main parameters we used are as follows: vocabulary limited to the 50,000 most frequent tokens, 2 layers (linear and LSTM), both of size 200 and 1,000 word classes, generated with mkcls.

Table 8 shows the results of reranking using left-to-right and right-to-left neural LMs on their own and jointly (row bidirectional). Reranking with left-to-right or the right-to-left LMs on their own does not result in a substantial improvement. However, when both LMs are used jointly we observe better scores for all the metrics: 1.7% relative improvement for BLEU, −0.5% for TER and 0.1% for chrF1.

| System       | BLEU   | TER    | chrF1 |
|--------------|--------|--------|-------|
| Without reranking| 0.1536 | 0.7679 | 50.99 |
| Left-to-right| 0.1536 | 0.7671 | 50.96 |
| Right-to-left| 0.1536 | 0.7707 | 50.94 |
| Bidirectional| 0.1562 ↑| 0.7644 ↑| 51.04 |

5 System Combination

As we have seen in the previous two sections, our best NMT system outperforms by a wide margin our best SMT system. These two systems are typologically different, and thus, despite the gap in performance, we might expect them to have complementary strengths. We therefore explored combining both systems in order to answer the following question: whether SMT, despite the gap in performance, can still be useful, used jointly with NMT, to improve upon NMT on its own.

We combined the outputs produced by the best NMT and SMT systems with MEMT (Heafield and Lavie, 2010). We used default settings, except for radius (5), following empirical results obtained on newsdev2015. The LM used in the combination was built on the concatenation of all the Finnish monolingual corpora available, cf. Table 1.

As the systems combined use different segmentations (Omorfi in NMT and Omorfi followed by BPE in SMT), we joined the morphs before combining them. Therefore the tuning of the system combination was performed without segmentation. Since chrF1 was found to correlate well with human evaluation for Finnish last year (Stanojević et al., 2015), we explored tuning on this metric, alongside tuning on BLEU.

Finally, we reranked the n-best list of the system combination (top 500 translations) with the same procedure used to rerank the best SMT system (cf. Section 4.1). While the best SMT system was reranked on segmented data (Omorfi + BPE), the output of the system combination is not segmented. Therefore, similarly to what we did for system combination, we explored tuning the reranking on chrF1.

Table 9 shows the results of system combination and its rerankings. In system combination, we observe that tuning on character sequences results in considerably better scores compared to tuning on BLEU. That said, the output produced by the best system combination system without reranking (i.e. tuned on chrF1) is still worse than the one produced by the NMT system alone according the automatic metrics (−3.4% relative on BLEU and −0.1% on chrF1) except for TER (2.3% relative improvement).

Overall, reranking the system combination\(^\text{14}\)

---

\(^{12}\)Because of the way SMT decoders work they can use left-to-right LMs but not reverse LMs.

\(^{13}\)Due to time constraints.

\(^{14}\)We reranked the system combination that performed
Table 9: Results of the system combination experiments on newstest2016. The best score for each metric is shown in bold, as is the system submitted. An arrow pointing upwards (↑) means that the corresponding system outperforms the best NMT system by a statistically significant margin.

| System              | BLEU | TER  | chrF1 |
|---------------------|------|------|-------|
| Best SMT            | 0.1562 | 0.7644 | 51.04 |
| Best NMT            | 0.1830 | 0.7411 | 52.43 |
| Combo (BLEU)        | 0.1638 | 0.7298 | 51.75 |
| Combo (chrF1)       | 0.1767 | 0.7241 | 52.37 |
| Reranked (BLEU)     | 0.1791 | 0.7257 | 52.38 |
| Reranked (chrF1)    | 0.1845 | 0.7290 | 52.65 |

Tuning on character sequences (chrF1 metric), used for system combination, resulted in better performance than tuning on the de facto standard BLEU, corroborating the results seen in human evaluation, i.e. better correlation.

Our combined and NMT submissions were ranked first and second respectively (both in terms of BLEU and TER) in the English-to-Finnish news translation task automatic evaluation\(^{16}\) and they tied for the first place in the human evaluation.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (AbuMaTran). We would like to thank the Irish Centre for High-End Computing (www.ichec.ie) for providing computational infrastructure and Kazuki Irie for his help to our questions re RWTHLM.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016a. A character-level decoder without explicit segmentation for neural machine translation. arXiv preprint arXiv:1603.06147.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016b. A character-level decoder without explicit segmentation for neural machine translation. arXiv preprint arXiv:1603.06147.

Junyoung Chung, Çağlar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Marta R. Costa-Jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. arXiv preprint arXiv:1603.00810.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland, pages 1370–1380.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A Joint Sequence Translation...
Model with Integrated Reordering. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA, pages 1045–1054.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia, pages 644–648.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, pages 848–856.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 1177–1185.

Kenneth Heafield and Alon Lavie. 2010. Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme. The Prague Bulletin of Mathematical Linguistics 93:27–36.

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

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China, pages 1–10.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, volume 4, pages 388–395.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In Proceedings of the International Workshop on Spoken Language Translation. pages 68–75.

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. Prague, Czech Republic, pages 177–180.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. arXiv preprint arXiv:1511.04586.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, pages 1412–1421.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA, pages 311–318.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In Sanjoy Dasgupta and David Mcallester, editors, Proceedings of the 30th International Conference on Machine Learning (ICML-13). JMLR Workshop and Conference Proceedings, volume 28, pages 1310–1318.

Tommi A. Pirinen. 2015. Omorfi —free and open source morphological lexical database for finnish. In Proceedings of the 20th Nordic Conference of Computational Linguistics (NODAL-IDA 2015). Vilnius, Lithuania, pages 313–315.

Maja Popović. 2015. chrf: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, Portugal, pages 392–395.

Raphael Rubino, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas,
Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. 2015. Abu-matran at wmt 2015 translation task: Morphological segmentation and web crawling. In Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, Portugal, pages 184–191.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving Neural Machine Translation Models with Monolingual Data. arXiv preprint arXiv:1511.06709.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas. Cambridge, USA, pages 223–231.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, Portugal, pages 256–273.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2014. rwthlm – the RWTH Aachen university neural network language modeling toolkit. In Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore, pages 2093–2097.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Matthew D Zeiler. 2012. Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701.