The University of Helsinki submissions to the WMT18 news task

Alessandro Raganato, Yves Scherrer, Tommi Nieminen,
Arvi Hurskainen and **Jörg Tiedemann**
Department of Digital Humanities
University of Helsinki

Abstract

This paper describes the University of Helsinki's submissions to the WMT18 shared news translation task for English-Finnish and English-Estonian, in both directions. This year, our main submissions employ a novel neural architecture, the Transformer, using the open-source OpenNMT framework. Our experiments couple domain labeling and fine tuned multilingual models with shared vocabularies between the source and target language, using the provided parallel data of the shared task and additional back-translations. Finally, we compare, for the English-to-Finnish case, the effectiveness of different machine translation architectures, starting from a rule-based approach to our best neural model, analyzing the output and highlighting future research.

1 Introduction

The University of Helsinki participated in the WMT 2018 shared task on news translation with seven primary submissions. While the main focus of our work lay on the English-to-Finnish translation direction, we also participated in the Finnish-to-English, English-to-Estonian and Estonian-to-English translation directions.

In 2017, the University of Helsinki participated in WMT with an in-house implementation of an attentional encoder-decoder architecture based on the Theano framework, called HNMT (Östling et al., 2017). Since then, the development of Theano has stopped, and various open-source Neural Machine Translation (NMT) toolkits based on alternative frameworks have been made available (Klein et al., 2017; Junczys-Dowmunt et al., 2018, *inter alia*). In parallel, a novel neural network architecture for machine translation, called Transformer, has been introduced (Vaswani et al., 2017). The Transformer follows the encoder-decoder paradigm, but does not use any recurrent layers. Instead, its architecture relies primarily on attention mechanisms, stacking on each layer multiple attention components. Preliminary experiments with the Transformer architecture and its implementation in OpenNMT-py (Klein et al., 2017) showed consistent performance improvements compared to our 2017 architecture. Consequently, we used this setup for our main WMT 2018 submissions. For English–Finnish, our submissions also include a rule-based system, an SMT system, and a NMT system making use of a morphological analyzer and generator.

This year's WMT news translation task contains a multilingual sub-track, which includes all models that make use of third language data. We trained a multilingual model with data coming from three languages, English, Finnish and Estonian and then fine-tuned on a single language pair. We also generated synthetic English–Estonian data by pivoting through Finnish.

Additionally, following recent approaches (Johnson et al., 2016; Tars and Fishel, 2018) we added a domain label to each input sentence, according to the data source. For example, each sentence from the Europarl corpus was prepended with the ⟨*EUROPARL*⟩ label. The overall idea of domain labelling is that data coming from different sources are of different quality and represent different genres and writing styles. In this way, the translation model can be informed of the data source without increasing the number of parameters.

2 English→Finnish

2.1 NMT models

We trained our systems on almost all parallel data made available by WMT: Europarl (Koehn, 2005), ParaCrawl[1], Rapid, as well as the WMT 2015 test and development sets. We did not use WikiHeadlines. For development and tuning of the system

[1] https://paracrawl.eu/

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 492–499

https://doi.org/10.18653/v1/W18-6425

parameters, we used the WMT 2016 and 2017 test sets.

A common strategy is to create synthetic training data by back-translation (Sennrich et al., 2016a). For our WMT 2017 submission, we already used SMT to create 5.5M sentences of back-translated data from the Finnish news2014 and news2016 corpora. This year, we created another 5.5M sentences of back-translation from the Finnish news2014-news2017 corpora using our previous NMT system (Östling et al., 2017). The final submissions make use of both resources.

We applied the standard preprocessing pipeline consisting of tokenization,[2] normalization,[3] truecasing and byte-pair encoding (Sennrich et al., 2016b). Following Vaswani et al. (2017), we have used a joint BPE vocabulary of 37 000 units. Regarding domain labeling, we marked the development and test data from WMT 2015, 2016 and 2017 as ⟨*NEWS*⟩. This label is also used for the test sets.

2.2 NMT with morphological analysis and generation

We also submitted a neural machine translation model that uses a morphological analyzer, called TwoStepTransformer. TwoStepTransformer is an English to Finnish transformer-type NMT model trained with the Marian NMT framework (Junczys-Dowmunt et al., 2018), using the default transformer settings (corresponding to Google's original Transformer setup). The model differs from standard NMT models in that the Finnish corpus is analyzed with a morphological analyzer (FINTWOL by Lingsoft Inc.) and segmented into a sequence of interleaved lemmas and morphological tags. The output of the model is converted into surface forms in a separate, deterministic post-processing step.

A similar two-step approach has been found to improve English to Czech NMT (Tamchyna et al., 2017), probably due to alleviating data sparsity caused by morphological complexity. As Finnish is also a morphologically complex language, adapting this approach to Finnish should result in a similar improvement. Finnish is an agglutinative language with a high degree of allomorphy for root, inflectional and derivational morphemes. For instance, the plural affix is expressed as *t*, *i* or *j* depending on the morphological context, and it is common for root lexemes have more than two allomorphs (e.g. the lexeme with the meaning 'hand' has the allomorphs *käsi*, *käde*, *käte* and *kät*). This allomorphy greatly increases data sparsity if segmentation methods based on surface form splitting are used.

The annotation format used differs from the one in Tamchyna et al. (2017) in several aspects, the most important of which is that the morphological tags are not complex, multicategory tags that are interleaved one-to-one with lemmas. Instead, each lemma token can be followed by zero or more morphological tags, each corresponding to a non-default value in a single morphological category:

```
komissio
tiedottaa FINTWOL_PAST
neuvosto FINTWOL_ALL EU FINTWOL_GEN
ja
Marokko FINTWOL_GEN
kalastus LS_PRECOMPOUND
kumppanuus LS_PRECOMPOUND
sopimus FINTWOL_PTV
koskeva FINTWOL_ELA FINTWOL_PL
kahden LS_PRECOMPOUND
välinen FINTWOL_ELA FINTWOL_PL
neuvottelu FINTWOL_ELA FINTWOL_PL
.
```

The first lemma *komissio* is the only one without any morphological tags, the rest of the lemmas are trailed by one or more tags. Tags are only provided if the value of a morphological category differs from the default value, so this means that the lemma *komissio* has the default value for number (singular) and case (nominative). The lemma *tiedottaa* is a verb lemma (lemma form indicates word class so no explicit word class annotation is required), and it has the tag FINTWOL_PAST, indicating that it has the non-default value PAST for the tense category (default is present tense). Several noun lemmas have non-default case and number values, for example *neuvottelu*, which has allative case and plural number. The LS_PRECOMPOUND tag indicates the lemma is part of a compound word.

There are several reasons for using implicit default morphological categories:

1. Explicitly defining each tag would lead to very long target sentences.

2. Having separate tags for each category theoretically allows for more generalization than complex multi-category tags. For instance, case generalizations could be learned from both singular and plural contexts.

3. Languages generally have morphological categories where the most common value has no

[2] We modified the Moses tokenizer to prevent it from splitting word-internal colons that occur regularly in Finnish.

[3] Normalization is applied to English only and consists of resolving common contractions such as *isn't*, *we'll* etc.

explicit morpheme, so segmenting with implicit common values makes the segmented text structurally more similar to natural language.

The morphological segmentation (which includes compound splitting) decreases the amount of token types in the corpus significantly (from over a million to about 300,000 for the bilingual WMT data), but there are still too many token types for efficient NMT training, due to foreign language words, incorrectly spelled words, numbers, codes, character corruption and other out-of-vocabulary tokens. To lower the type count to a manageable level, the annotated corpus is further segmented using BPE. As the model outputs a BPE sequence of lemmas and morphological tags, producing the final translation is more complex than simple concatenation of subword units. First the BPE tags are joined and then the surface forms are generated using the FINTWOL generation functionality, which takes as input lemmas and morphological tags and output all compatible surface forms. The default tags are automatically added for lemmas which do not have explicit tags. Heuristics are used to select a surface form if several possibilities are generated.

The submitted model was trained on the bilingual and back-translated data, as adding the back-translated data greatly improved the quality of the translations.

2.3 Rule-based MT

Hurskainen and Tiedemann (2017) propose a rule-based machine translation system for English–Finnish. During the past year, the rule-based MT system has been developed in several ways. In addition to the usual debugging and rule testing, also some major structural changes have been made. Below we will discuss the latter type of problems.

Translating locative expressions: While English uses prepositions for marking location, Finnish uses locative cases. English has a bewildering number of prepositions for this purpose. At least the following preposition are used: *in, on, at, with, by, to, into, for, of, from, over, through*, and *around*. Finnish uses one of the six locative cases for translating such structures.

Locative cases can be classified into two groups, which are termed as internal (inessive, elative, and illative) and external (adessive, ablative, and allative) locatives. Associating the English locative preposition with one of the Finnish locative cases would require several rules with a varying number of constraints. In the current implementation, the Finnish locative cases are handled in two phases. In the first phase, we only consider what type (no movement, movement from, movement to) the location is, without considering whether it is internal or external.

```
"<he>" "he" { hän } %SUBJ HUM OUT PRON PERS SG3 NOM
"<sent>" "send" { lähetti } %+FMAINV O-ACC O-LOC3 V
    PAST SG
"<letter>" "letter" { kirjeen } %OBJ DEF N SG ACC
"<to>" "to" { M-LOC3 } %ADVL PREP
"<hospital>" "hospital" { sairaalaan } %<P ACE IN
    DEF N SG ILL
"<.>" "." { . }

"<he>" "he" { hän } %SUBJ HUM OUT PRON PERS SG3 NOM
"<sent>" "send" { lähetti } %+FMAINV O-ACC O-LOC3 V
    PAST SG
"<letter>" "letter" { kirjeen } %OBJ DEF N SG ACC
"<to>" "to" { M-LOC3 } %ADVL PREP
"<me>" "i" { minulle } %<P HUM OUT PRON PERS SG ALL
```

We see that in both sentences the preposition *to* has the tag M-LOC3. This stands for illative and allative. The head of the preposition decides which of the cases is selected. If the noun has the tag OUT, then allative is selected. If it has the tag IN or no locative tag, then inessive is selected. The same process applies to the two other locative case pairs (inessive/adessive and elative/ablative).

A special case of using locatives are the Finnish place names. No formal rules can be constructed for producing correct locative inflection. Therefore, we have to tag each place name separately. We use internal inflection as default and provide names using external inflection with the tag OUT.

Translating proper names and acronyms: There are two major problems in dealing with proper names and acronyms. One concerns the question whether the proper name or acronym should be translated or not. The other problem concerns the handling of uppercase letters. The proper names with translation should be listed in the lexicon or handled as an MWE. It is assumed that a non-sentence-inital word with capital initial is a proper name, and possibility to such an interpretation is provided by adding a separate entry with a tag PROP-CAND. If it is not listed in the lexicon, it is interpreted as a proper name. Such words which have also another interpretation in the language are problematic. Many person names belong to this category. Attested cases with both interpretations (i.e. normal translation and proper name) are listed in the rule system. Then, using context sensitive rules, the PROP-CAND interpretation is selected or removed.

Translating subject and object: The default case of the subject in Finnish is nominative, but also other cases, such as adessive, genitive, elative, ablative, and illative, occur. Rules are needed only for the special cases. This is implemented by providing the respective verb with a tag showing the case of the subject. Otherwise the subject case is always nominative. The direct object has three cases, partitive, genitive accusative and nominative accusative. The last one is used in special cases such as the object of imperative verb form and some modal verbs. Partitive and genitive accusative dominate as object case. Part of verbs require always partitive, and some others require the genitive accusative case. However, most verbs are such that they may have either of the object cases. They are not alternatives, however, because the context defines the case in each situation. There is the general trend that if the object is indefinite plural, it is in partitive.

More details of the system are described in Hurskainen (2018a,b,c,d).

2.4 SMT

As a contrastive system, we also reactivated our Statistical Machine Translation (SMT) system submitted at WMT 2016 (Tiedemann et al., 2016). The system was not retrained and it may thus suffer from poor lexical coverage on recent test data. Our main motivation for including this baseline was to have an SMT submission for the Finnish morphology test suite.

3 Finnish→English

We only submitted a standard NMT transformer model with domain labeling for this translation direction. Parallel data and preprocessing steps are identical as for English-to-Finnish. For back-translation, we use 2M sentences from the English news2015 produced with an SMT system, plus another 6.7M sentences from English news2007–news2017 produced with HNMT (Östling et al., 2017).

During the test phase, we discovered that several source lines, in particular in the Finnish test data, consisted of more than one sentence. As our translation systems were trained mostly on single sentences, they tended to stop the translation process after translating the first sentence of the line, leaving the remaining sentences untranslated. In order to avoid this, we applied a simple sentence splitting script to the test set and translated the split sentences separately. According to the output of the sentence splitter, 298 sentences of the Finnish source and 13 sentences of the English source were affected. We applied sentence splitting to both files; while this increased BLEU scores by 0.5 points on Finnish-to-English, it did not affect the BLEU scores of English-to-Finnish translation.

4 English–Estonian

We also participated in the English–Estonian task, in both directions. We used all available parallel data for training: Europarl, ParaCrawl, and Rapid. We used the 2018 dev set for system development and parameter tuning. We applied the same preprocessing steps as for English–Finnish, using again a shared vocabulary of 37 000 BPE units. Regarding domain labeling, no parallel data with the ⟨*NEWS*⟩ label was provided in this setup. Therefore, we labelled the test source data with ⟨*EUROPARL*⟩, which we found to be the most reliable of the three data sources. For comparison, we also tested a model without domain labels (comparative results are given in Section 6).

For our English-to-Estonian submission, we created back-translations using a simpler translation model. This model was based on the Transformer architecture and was trained on a subset of parallel data filtered through a language identification tool, with 20 000 BPE units. We used this model to translate parts of the monolingual BigEst corpus to English; 6.3M back-translations sentences were obtained.

For the Estonian-to-English submission, we also generated back-translations using a simple translation model, as described above, translating parts of the monolingual English news2007–news2017 corpora; 5.2M back-translation sentences were produced in this way.

5 Multilingual models

As Estonian is closely related to Finnish, we experimented with multilingual models containing both languages as well as English. For this experiment, we included all available parallel data in all directions. Following Johnson et al. (2016), we used language labels to indicate the target language coupled with the domain labels, as introduced above. The only other change in preprocessing is the use of 50 000 (instead of 37 000) joint BPE units, as they now need to cover three languages instead of

	Parallel	+Back	+Back +Synth
Et → En	2,178,025	7,356,697	8,942,157
En → Et	2,178,025	8,435,413	10,020,873
Fi → En	3,136,265	11,918,402	–
En → Fi	3,136,265	14,198,188	–

Table 1: Number of training sentences, with and without back-translation (Back) and synthetic data (Synth).

two. In this way, even though Estonian has no parallel news data, the model will see the news label in the Finnish data. Inspired by Zoph et al. (2016), we first train the multilingual model with all languages in all directions, and then fine-tune it on each specific language pair.

5.1 Synthetic data

Another way to take advantage of the close etymological relationship between Estonian and Finnish is to create synthetic training data (Tiedemann, 2012). We explored this option in the following setup:

1. Train a character-level seq2seq system for Finnish-to-Estonian, using the Europarl and EUbookshop (Skadiņš et al., 2014) corpora.
2. Translate the Finnish side of the parallel English–Finnish corpus to Estonian.
3. Combine the Estonian and English parts of the corpus and use this dataset as back-translations to train the final system.

We were able to process 1.5M sentences using this approach. These sentences complemented the other training data, consisting of parallel data and direct English–Estonian back-translations.

6 Experiments

In this section we detail the setup of our experiments. We first describe the size of the training data and the details of the training; we then report and discuss the performance of each model according to the BLEU score as reported on the online evaluation matrix[4].

Table 1 shows the statistics on the number of training sentences. The backtranslations allow us to more than triple the original size of the training data for all the directions. We trained our models

[4] http://matrix.statmt.org/

			Et→En	En→Et
HY-NMT Baseline			21.6	16.7
+Label			20.3	17.6
	+Back		**26.5**	–
+Label	**+Back**		25.4	**21.8***
	+Back	**+Synth**	**26.5***	–
+Label	**+Back**	**+Synth**	25.0	21.0
HY-NMT Multilingual			–	–
+Label			26.4	20.8

Table 2: BLEU-cased scores on newstest2018 for the English–Estonian language pair in various configurations using domain labels (Label), backtranslated data (Back), or synthetic data (Synth). Our primary submissions are marked with *.

for 20 epochs, evaluating each of them on the development set after every epoch, taking the best iteration as final model. As hyper-parameters, we used the *base* version of the Transformer architecture, following the suggestion of the OpenNMT-py tool,[5] including a shared word embedding space between encoder and decoder among others. Unlike last year, we did not include any averaging or ensembling techniques.

English–Estonian results. Table 2 shows the performance of our models for the English–Estonian language pair.[6] In general, the best models include back-translation and synthetic data, improving the BLEU score by around 4 points. The domain labels help when translating into Estonian, while they slightly hurt the performance when translating into English. This behavior could be explained by the different nature of the two languages, Estonian being a morphologically rich language, it could benefit from having a source label indicating good quality translations even if they come from a different domain. As concerns our multilingual model, it achieves results close to our best score, specially for the Estonian-to-English direction. We recall that this model also uses domain labels, and this suggests that, in this case, the Finnish–English data are indeed helpful to achieve a better BLEU score for the Estonian-to-English language pair.

English–Finnish results. Table 3 shows the performance of our models for the English–Finnish language pair. Here, all of our basic Transformer

[5] http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model

[6] HY stands for *Helsingin Yliopisto*, i.e. University of Helsinki, not for *hybrid*.

	Fi→En	En→Fi
Transformer +Label	19.8	15.3
Transformer +Back +Label	**23.3***	**17.8***
Multilingual +Back +Label	20.6	14.9
TwoStep +Back	–	14.5*
Seq2Seq +Back +Label	–	12.1
HY-SMT +Back	–	10.5*
HY-AH (rule based)	–	6.4*

Table 3: BLEU-cased scores on newstest2018 for the English–Finnish language pair for various system architectures. Our primary submissions are marked with *.

models included domain labeling, motivated by the fact that news data are present in the training data, and also by the consistent performance improvements observed in initial experiments. Overall, back-translations again improve the BLEU scores for both directions. The multilingual model achieves lower scores than the standard bilingual model, suggesting that the Estonian data do not provide useful complementary information, in particular because the Estonian data set is rather small compared to its Finnish counterpart and comes from exactly the same source. Finally, we also compare our transformer-based models to other machine translation paradigms. Table 3 reports the BLEU scores of the rule based system described in Section 2.3, the SMT system (Section 2.4) and an additional 2-layer sequence-to-sequence model (Bahdanau et al., 2014) trained on the same data as the Transformer models. Clearly, the Transformer paradigm achieves the best BLEU scores. Overall, our best English-to-Finnish model reaches the second position in the online ranking using automatic evaluation metrics. Finally, in the manual evaluation of the official results of the WMT18 News Translation task (Bojar et al., 2018), our best system shared first place in both English-to-Finnish and Finnish-to-English translation directions.

6.1 English-to-Finnish analysis

To complement the results, we additionally carried out an analysis of the output of our best English-to-Finnish system.

Document knowledge. One of the common mistakes is related to pronouns, especially when their antecedents are located in other sentences. As our systems are trained on isolated sentences, it is hard to predict the right pronoun when it refers to a previous sentence. Moreover, more context would help to better understand the semantics of the sentence. For example, considering the following translation:

(1) EN: "After burying the bodies, the military came looking for me," he says.
FI: "Sotilaat etsivät minut käsiinsä uhrien hautaamisen jälkeen", hän sanoo.

the word *bodies* has been translated as *victims*, which only makes sense if you know the document context where bodies were those of demonstrators.

World knowledge. We found out that some test set translations contain information based on world knowledge" outside of the actual text, and so the system being trained without any external knowledge fails to output the most appropriate translation. For example, in the sentences:

(2) EN: "Americans appreciate this as well as anyone - hence the carefully stage-manged toppling of Saddam Hussein in Firdos square in Baghdad in 2003."
FI: "Amerikkalaiset tietävät sen yhtä hyvin kuin muutkin: irakilaiset kaatoivat yhdessä amerikkalaisten sotilaiden kanssa Saddam Husseinin patsaan Firdosin aukiolla vuonna 2003."

the literal translation of the Finnish sentence would be: *"The Americans know it as well as others: the Iraqi toppled together with American soldiers Saddam Hussein's statue in Firdos square in 2003."*, leaving out Baghdad and introducing Iraqi in this case.

Finally, a number of errors were related to the different structure and ordering of the words of the two languages. It seems like the 2018 test set is translated more freely and document-oriented than in previous years, which explains the overall low BLEU scores compared to the last year's competition.

7 Conclusions

In this paper, we reported the University of Helsinki submissions for the WMT18 news translation task. We participated in the English–Finnish and English–Estonian language pairs, training the novel neural architecture, the Transformer, with the OpenNMT tool, using BPE segmentation, a joint source-target vocabulary and domain labeling. Additionally, we introduced a multilingual model trained

on all our data sets, fine-tuning it on each language pair. Our best systems are trained on the provided parallel data augmented with large amounts of back-translations, achieving top rank results for the English–Finnish language pair. We also carried out further analyses on the English-to-Finnish direction, showing the performance of different machine translation paradigms and highlighting common mistakes that prevented a higher translation quality.

Acknowledgments

The work in this paper is supported by the Academy of Finland through project 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence. We would also like to acknowledge NVIDIA and their GPU grant.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Arvi Hurskainen. 2018a. Comparative and superlative in English to Finnish machine translation. Technical Reports in Language Technology 26, University of Helsinki, http://www.njas.helsinki.fi/salama.

Arvi Hurskainen. 2018b. Implementing location in English to Finnish machine translation. Technical Reports in Language Technology 27, University of Helsinki, http://www.njas.helsinki.fi/salama.

Arvi Hurskainen. 2018c. Proper names and acronyms in English to Finnish machine translation. Technical Reports in Language Technology 28, University of Helsinki, http://www.njas.helsinki.fi/salama.

Arvi Hurskainen. 2018d. Subject and object case in English to Finnish machine translation. Technical Reports in Language Technology 29, University of Helsinki, http://www.njas.helsinki.fi/salama.

Arvi Hurskainen and Jörg Tiedemann. 2017. Rule-based machine translation from English to Finnish. In *Proceedings of the Second Conference on Machine Translation*, pages 323–329, Copenhagen, Denmark. 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. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

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.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Robert Östling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The Helsinki neural machine translation system. In *Proceedings of the Second Conference on Machine Translation*, pages 338–347, Copenhagen, Denmark. 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)*, volume 1, pages 86–96.

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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU Bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ales Tamchyna, Marion Weller-Di Marco, and Alexander M. Fraser. 2017. Modeling target-side inflection in neural machine translation. *CoRR*, abs/1707.06012.

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. *arXiv preprint arXiv:1805.02282*.

Jörg Tiedemann. 2012. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 141–151. Association for Computational Linguistics.

Jörg Tiedemann, Fabienne Cap, Jenna Kanerva, Filip Ginter, Sara Stymne, Robert Östling, and Marion Weller-Di Marco. 2016. Phrase-based SMT for Finnish with more data, better models and alternative alignment and translation tools. In *Proceedings of the First Conference on Machine Translation*, pages 391–398, Berlin, Germany. Association for Computational Linguistics.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.