Phrase-Based SMT for Finnish with More Data, Better Models and Alternative Alignment and Translation Tools

Jörg Tiedemann
University of Helsinki

Fabienne Cap
Uppsala University

Jenna Kanerva and **Filip Ginter**
University of Turku

Sara Stymne
Uppsala University

Robert Östling
University of Helsinki

Marion Di Marco
University of Stuttgart

Abstract

This paper summarises the contributions of the teams at the University of Helsinki, Uppsala University and the University of Turku to the news translation tasks for translating from and to Finnish. Our models address the problem of treating morphology and data coverage in various ways. We introduce a new efficient tool for word alignment and discuss factorisations, gappy language models and re-inflection techniques for generating proper Finnish output. The results demonstrate once again that training data is the most effective way to increase translation performance.

1 Introduction

In this paper we revisit phrase-based models with and without factors to translate from and into a morphologically-rich language, Finnish. We discuss the impact of training data, the use of factored models and ideas of re-inflection as post-processing. We also introduce the framework of gappy language models within document-level machine translation (without much success in the given task). Our efforts prove the importance of training data once again and demonstrate the use of noisy and out-of-domain data sets as well as the possibility of integrating synthetic training data based on back-translation in phrase-based SMT.

2 Data and Tools

This section discusses data sets and tools that we applied in our models. We focus on non-standard resources but also summarise the basic setup of our training procedures.

Training Data: Our submissions include constrained and unconstrained systems. The constrained systems apply all the data provided by WMT and also the English Giga-Word corpus that is distributed by the LDC. Our best systems include additional parallel data sets coming from OPUS (Tiedemann, 2012) and syntactically analysed monolingual data from the Finnish Internet Parsebank (Luotolahti et al., 2015). Additional to the parallel data we used in our submission last year (Tiedemann et al., 2015a), we include the new version of the OpenSubtitle corpus (Lison and Tiedemann, 2016) with its 18.6 million aligned translation units in English and Finnish. Furthermore, we make use of alternative subtitle translations that have been aligned monolingually in the same collection (Tiedemann, 2016). Expanding the parallel corpus with alternative translations extends the subtitle corpus by roughly 350,000 translation units with about 6.8 million tokens (counting both languages together). The contribution is quite small compared to the original corpus with its 107 million Finnish tokens and 167 million English tokens, but, nevertheless, it contributes to the overall collection especially by providing additional variation of the translation examples, which is very valuable for the resulting system. The final training corpus contains 27.7 million translation units comprising 353 million English tokens and 244 million tokens in the Finnish part.

For Finnish, we also increased the coverage of our language model by further 4.9 billion tokens compared to our last year submission. The data comes from an extensive web-crawl and amounts to 9.5 billion tokens of text, deduplicated on document level. Five-gram language models are trained using KenLM (Heafield et al., 2013). The English language model based on the provided Common Crawl data is limited to trigrams.

Pre-Processing Tools: For processing Finnish, we apply the Finnish parsing pipeline developed at the University of Turku (Haverinen et

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 391–398,
Berlin, Germany, August 11-12, 2016.

al., 2013). It integrates all the necessary pre-processing steps including tokenisation, morphological analyses and part-of-speech tagging, and produces dependency analyses according to the Universal Dependencies scheme.[1] The morphological component relies on OMorFi - an open-source finite-state toolkit with a large-coverage morphology for modern Finnish (Lindén et al., 2009). The readings given by OMorFi are combined with predictions of the *MarMoT* CRF-based tagger (Mueller et al., 2013), and the data is subsequently parsed using the *mate-tools* data-driven dependency parser (Bohnet, 2010). The labeled attachment score of the parsing pipeline is 82.7% and the pipeline is robust and reliable even for large data sets and long sentences (Pyysalo et al., 2015).

We also apply various pre-processing tools provided by the Moses toolbox. In particular, we make use of tokenisers (especially for English), punctuation and Unicode normalisers.

For the factored models of English, we built our own pre-processing pipeline mainly adapted from the Finnish pipeline but adjusted for processing English. They include tools for handling long sentences and keeping track of sentence alignment points when parsing parallel data sets. We use the English models for sentence boundary detection and tokenisation provided by OpenNLP,[2] which is compatible with the Penn Treebank style of tokenisation. This is important for the subsequent tagging and parsing steps, which we trained on the Universal Dependencies treebank for English using MarMoT and mate-tools.

MT Tools: Most of our systems are based on Moses (Koehn et al., 2007) and common components for training and tuning models. We apply KenLM (Heafield et al., 2013) and SRILM (Stolcke, 2002) for estimating language model parameters and MERT (Och, 2003) and batch-MIRA (Cherry and Foster, 2012) for parameter tuning. Most of our models are based on lowercased training data. All language models use order five with modified Kneser-Ney smoothing if not stated otherwise. All MT systems apply the phrase-based paradigm, some of them with factored representations and generation models if necessary.

For word alignment we experiment with different tools. We apply standard tools like GIZA++ (Och and Ney, 2003) and fast_align (Dyer et al., 2013) but also the recently proposed Bayesian word aligner **efmaral** (Östling, 2015). Efmaral is an efficient implementation of a Markov-Chain aligner using Gibbs sampling with a Bayesian extension of the IBM alignment models. It is both fast and accurate and works as a straightforward plug-in replacement for standard tools in the SMT training pipeline. The aligner is faster than fast_align but more accurate in terms of alignment error rate in various benchmark tests. The advantage of using Gibbs sampling rather than the Expectation-Maximisation algorithm (as do both fast_align and GIZA++) is that inference remains quadratic with respect to sentence length even when word order and fertility models are added, which enables the efficient use of higher-order models. This is the first time that the performance of this tool is reported in the setting of statistical machine translation.

Besides Moses, we also apply another phrase-based machine translation decoder, **Docent** (Hardmeier et al., 2013), which implements a stochastic local search decoder that is able to incorporate features with long-distance dependencies even across sentence boundaries. Docent emphasises document-level decoding but includes standard local features that make the decoder comparable with standard phrase-based SMT. The decoding algorithm applies randomly selected state-change operations to complete translation hypotheses (covering the whole document) that may be accepted by a strict hill-climbing procedure or a simulated annealing schedule. The main motivation for using Docent in our setup is to introduce non-local dependencies that may improve, for example, agreement problems in morphologically-rich languages such as Finnish. However, the experiments are very initial and, unfortunately, do not show the desired effect yet.

3 Translating English into Finnish

Our main efforts went into the development of translation models for the direction from English to Finnish. Four types of experiments were conducted: (1) Changing word alignment and data sets; (2) Factored models with morphological features; (3) Re-inflection models with robust generation from underspecified representations; and (4) Gappy language models for long-distance dependencies.

[1] http://universaldependencies.org
[2] https://opennlp.apache.org

3.1 Changing Alignment and Adding Data

Our first series of experiments considered three different word alignment tools that can be used in the training pipeline of standard phrase-based SMT. We use the well-known IBM alignment models (up to model 4) implemented in GIZA++, the modified IBM model 2 implemented in fast_align and the above introduced Bayesian word aligner based on fertility-enhanced HMM models implemented in efmaral. Table 1 summarises the results when applied in the constrained setup and tested on the news test set from WMT 2015. The three models use the same feature weights and the same symmetrisation and phrase extraction/scoring parameters to make the scores comparable with each other. The results indicate that efmaral is comparable and even better than GIZA++ in this setup even though it is magnitudes faster than the IBM model 4 training and Viterbi alignment. Efmaral is also considerably faster than fast_align, which makes it a valuable drop-in replacement of these standard tools. The processing times in Table 1 illustrate the significant gains when using efmaral making it possible to quickly align large amounts of bitexts. The advantage over fast_align can mainly be seen in CPU time with a speed-up of almost a factor of 10. fast_align, however, has the advantage to naturally run multithreaded over many cores whereas the collapsed Gibbs sampler of efmaral is not as easily parallelised. This can also be seen in our experiments which we ran on a 16 core machine with alignment in both directions in parallel. GIZA++ is by far the slowest option and does not lead to better translations either. The figure also excludes word clustering which is another time-consuming process that is necessary for running IBM model 4.

Part of the experiment is also the inclusion of additional training data. All those runs use efmaral, demonstrating that the software is capable to cope with large data sets. Note, however, that memory requirements grows with the size of the data ($\sum_{e,f}(|e| \times |f|)$) making it possible to run efficiently. The results of our experiments show that the additional data is useful even though it is coming from inappropriate domains. Especially striking is the gain by including alternative subtitle translations – a rather small part of the data. Apparently, those examples introduce necessary variations to push the quality of the models. Another impressive improvement can be seen with the introduction of the large language model based on a diverse set of data. This Finnish language model is estimated on the Finnish Internet Parsebank (Luotolahti et al., 2015), totaling 9.5 billion tokens of text. The data is obtained from a large-scale Internet crawl, seeded from all Finnish pages in CommonCrawl.[3] However, actual CommonCrawl data is only a small fraction of the total, roughly 1.5B tokens, the remainder originating from an independent crawl. The data is heavily filtered, only preserving clean, parseable text comprising of complete sentences.

		time for word align	
newstest 2015	BLEU	real	CPU
GIZA++	13.65	38,514s	–
fast_align	13.56	682s	8,344s
efmaral	14.10	370s	895s
+ OPUS	14.81	–	–
+ alternatives	15.55	2,630s	6,599s
+ WWW-LM	16.98	–	–
retuned	18.11	–	–
back-translated	14.78	954s	2,606s
+ OPUS, ...	18.22	2,758s	7,187s

Table 1: Lower-cased BLEU scores for standard-phrase based SMT on development test data (newstest 2015). The first three and the second-to-last rows represent constrained settings whereas the other rows refer to systems with additional resources. Efmaral is used in all cases except for the two models at the top. The last two systems include back-translated news data. Running time is given for some aligners in terms of walltime (real) and CPU time (user+sys).

Even the models with additional data use the same feature weights and only replace the indicated component to enable comparisons between them. The system denoted by "retuned", however, shows the importance of proper tuning when replacing system components.

The final part of Table 1 shows additional results with back-translated news data in the constrained and unconstrained setup. We used our Finnish-English model to translate approximately 1.25 million sentences of the Finnish shuffled monolingual news data from 2014 and 2015 to enhance the parallel training data. The result in terms of BLEU significantly improves when these noisy data sets are included in the standard train-

[3] www.commoncrawl.org

ing pipeline. Note that the models are retuned from scratch in both cases.

3.2 Factored Models

The factored models we developed use features extracted from dependency trees coming out of the Finnish and English pre-processing pipelines. We include separate translation models for translating between English surface word forms and Finnish lemmas and for translating morphosyntactic features between the two languages. The latter includes dependency relations besides part-of-speech labels (on both sides) and detailed morphological information (in Finnish only). Table 2 summarises the results of these models.

newstest 2015	BLEU
(a) surface form	14.10
(b) morph	5.45
(c) constructions	10.89
combined (a) + (c)	14.17
+back-translated	14.70

Table 2: Lower-cased BLEU scores for factored SMT models on development test data (newstest 2015). System (a) is the same as the constrained model in Table 1. System (b) uses a factored model that translates surface words to target lemmas and morphosyntactic features separately. System (c) keeps closed-class words in the translation table of morphosyntactic features. (b) and (c) include a generation model trained on large monolingual parsed training data to generate surface word forms from lemmas and morphosyntactic features.

The morphologically enhanced factored model underperforms significantly when used in isolation. Therefore, we used a variant of the setup that replaces morphosyntactic features with surface words for all closed-class words in the training corpus. The assumption is that there is sufficient evidence for those word types even in morphologically-rich languages such as Finnish. Using this type of lexicalisation helps to find construction-like mappings between the two languages which seems to be beneficial for the system according to the scores in our experiments (system (c) in Table 2). In combination with the surface-oriented translation model this also leads to a slight improvement over the non-factored model (without back-translated news), which is also evident in the final scores of our submitted systems at least in the constrained setup (see Table 4).

3.3 Re-inflection Models

Furthermore, we also investigated re-inflection models. These experiments require a different representation of the training data for each variant and are, therefore, not directly comparable with the other systems. The underlying idea of what we call re-inflection models in our submission is that we reduce all Finnish training data to an underspecified representation, where words are reduced to their lemmas and noun and adjective compounds are split into their component parts. Then, we train models and translate from English into this underspecified representation of Finnish and in a post-processing step we then merge compounds and predict morphological features for Finnish. This approach has been successfully applied to Russian and Arabic (Toutanova et al., 2008) and to German (Fraser et al. (2012), Cap et al. (2014)). Note however, that for example Fraser et al. (2012) relied on German prepositions to predict case-markers on underspecified German SMT output. In contrast to many other languages, Finnish only has a limited number of stand-alone pre- and postpositions. Instead, the prepositional meaning is encoded by case-marking. We thus adapt an approach by Tiedemann et al. (2015b) and introduce *place-holder prepositions* in the Finnish training data, which are likely to correspond to the prepositions used on the English side and thus improve word alignment quality.

Place-holder Prepositions: In contrast to Tiedemann et al. (2015b), we do not apply factored models (with both, lemmatised and surface forms) here but strip the case-markers from those words and only keep the underspecified representation. Moreover, we apply the approach in the opposite translation direction, which requires a generation component. The place-holder prepositions will not only lead to improved word alignments, but we will also use them to predict case-markers after translation. Overall, we follow the processing pipeline of (Cap et al., 2014): we use a rule-based morphological analyser (Pirinen, 2015) to split compounds (using the Finnish parsing pipeline to disambiguate multiple analyses) and lemmatise all Finnish training data. Compound modifiers are reduced to their lemmas and marked with a symbol that distinguishes them from other words. Sim-

ilar to Tiedemann et al. (2015b), we introduce place-holder prepositions at the beginning of noun phrases bearing the corresponding case-marker in order to support word alignment.

Prediction of Case-Markers After translation, we apply CRF models to predict the case markers of Finnish. Besides the occurrences of placeholder prepositions, these take some more local context, both on lemma and POS level into account. Clean-data experiments have shown that our CRF models for re-inflection are very accurate. We reduce all compounds of the CRF training data to their heads and train the models on this representation. As we are using the words and lemmas as features for the CRFs, the reduction of compounds to their heads reduces data sparsity and allows the model to better generalise over all occurrences. For the translation output we remove all compound modifiers before case prediction.

Morphological Generation The predicted case-markers are then fed into the morphological generation automaton (Pirinen, 2015) in order to get fully inflected forms. In cases where this generation failed, we used a supervised machine learning approach as a backoff (Durrett and DeNero, 2013).

Compound Processing In a final step, we merge compounds using a POS-matching strategy (Stymne et al., 2008). We merge the marked compound modifiers with the following word if it is a noun or adjective, and add hyphens for modifiers in coordinated compounds. Compounding forms of modifiers are restored based on corpus frequencies. Like Stymne et al. (2008) and Cap et al. (2014), we also merge compounds in every iteration of the tuning process before the translations are scored against the reference.

All re-inflection systems are constrained systems. We used Europarl and Wikipedia as parallel resources and all of the Finnish data available from WMT to train five-gram language models with SRILM (Stolcke, 2002) and KENLM (Heafield, 2011). No particular cleaning or preprocessing of the data has happened. This makes the re-inflection systems differ from all other systems in this paper. Otherwise, we trained a conventional phrase-based Moses system with default settings, tuned weights using batch-MIRA with "safe-hope" (Cherry and Foster, 2012) and used an underspecified representation of the tuning reference set to derive BLEU scores. The final result of our system is listed in Table 4.

3.4 Gappy Language Models

Tiedemann (2015) introduces the use of language models over selected words in the framework of document-level SMT using Docent applied to the pronoun-aware translation task of DiscoMT (Hardmeier et al., 2015). We extended this idea by developing a general framework for what we call *gappy language models* that refer to monolingual or bilingual n-gram language models over selected words and their alignments. We can use different factors attached to the source and target language tokens to filter for word sequences that we would like to consider. Given word alignments are used to establish the link between source and target tokens. Gappy language models may cross sentence-boundaries but may also stop at those borders. Regular expressions can be used to make the selection more flexible. Multi-word alignments can be concatenated into single tokens and empty alignments can be represented as a special token to avoid the length-penalising effect of N-gram models. Word selection based on the source language also helps as this is given and fixed. However, word alignment is noisy and may negatively influence the use of the extracted target item sequence. Therefore, the selection can also be done on target language properties only and an additional penalty feature is then used to control the length of the generated strings. Bilingual models add both source and aligned target tokens whereas monolingual models only use target language tokens. Items are always sorted in the order of the target language.

We experimented with various selections and bilingual models to see the effect of these additional features functions. Five-gram Language model parameters are estimated using KenLM (Heafield et al., 2013). Our main selection criteria are part-of-speech patterns (matching coarse universal POS labels) and dependency relations:

- nouns and their alignments (sentence-internal only and even document-wide)
- verbs and their alignments (sentence-internal only and even document-wide)
- subject-predicate sequences (including negation particles) and their alignments
- closed-class words and their alignments

Gappy language models are fully integrated in Docent but one unsolved problem is the tuning of their weights. Currently, we do not have a stable

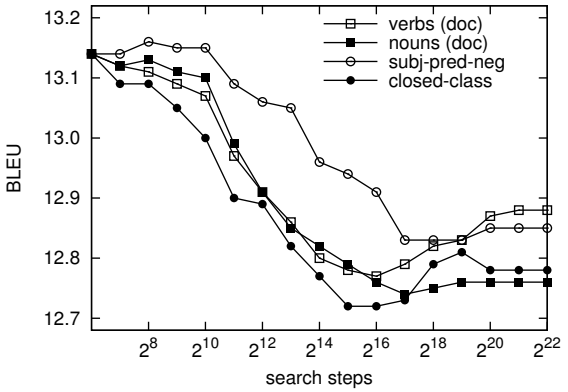

Figure 1: Adding gappy LM features and testing on development test data (newstest 2015).

framework for finding appropriate values for them and, hence, we needed to set them to a quite arbitrary value (0.1 in our case). The disappointing results of our extended models are shown in Figure 1. In general all of them seem to hurt performance in the current setup.

4 Translating Finnish into English

The Finnish–English models re-use the factored setup with pseudo-tokens that we introduced last year (Tiedemann et al., 2015a). The main differences to the previous systems are (i) the use of completely parsed bitexts even with the extended data sets (last year we only parsed Europarl from the constrained data), (ii) the large language model coming from the provided common crawl data (trigram model), and, (iii) improved compound splitting of surface words based on the morphological analyses and the analysed lemma information. For the latter, we use additional string matching heuristics to properly split compounds even if modifying components are inflected and cannot be matched with the lemmatised analyses in a straightforward way. Furthermore, we also add morphological information to the modifying compound components by looking up the most frequent analyses of the given form in a large analysed monolingual corpus. The scores for our factored models in the constrained and unconstrained settings are listed in Table 3.

Again, we can see the substantial impact of additional out-of-domain training data. Alternative subtitle translations contribute marginally in this translation direction. The common crawl data is useful but slows down decoding quite significantly.

newstest 2015	BLEU
basic	19.02
+ OPUS	21.42
+ alternatives	21.46
+ CC LM	22.09
basic + CC LM	19.33

Table 3: Lower-cased BLEU scores for factored SMT models for Finnish-to-English on development test data (newstest 2015).

5 Final Results and Discussions

Table 4 summarises the final scores when applying our models to the news test set from this year's evaluation campaign. A major, but not very surprising effect is the reduction of unknown words when adding more data. The factored model leads to slight improvements in the constrained setting but this does not carry over to the unconstrained setup. A significant difference is the number of unknown tokens which is much higher in the factored model. This may look surprising but when inspecting the data, we could identify the reason for this difference, which is due to the tokenisation applied in the factored setup. The models applied in this approach make different decisions, for example, when keeping numeric and monetary expressions together. This increases the number of unknown units without causing much harm in most cases. Other cases are clearly tokenisation errors. Some examples are listed below:

```
200k|ADJ|JJ|dep
228.89|NUM|CD|num
$22million|NOUN|NN|adpobj
2.5bn|NUM|CD|num
"wrestle|VERB|VB|xcomp
(yet|NOUN|NN|dobj
```

Note that the re-inflection model uses different data pre-processing pipelines and, therefore, the scores are not comparable with the others. In a contrastive run we could see modest improvements over the baseline models without re-inflection. Finally, we can also see that Finnish-English suffers more from unknown tokens even though we apply proper morphological analyses and compound splitting. This is something that we need to address in future work.

References

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceed-*

English – Finnish	BLEU lower	BLEU cased	TER	unknown words #tokens	unknown words #types
constrained - basic	13.3	12.7	0.782	1,582	862
constrained - factored	13.5	12.8	0.784	1,659	1,233
constrained - basic + back-translated	14.2	13.6	0.770	**1,024**	**649**
constrained + factored + back-translated	**14.3**	**13.6**	**0.765**	1,103	890
constrained - re-inflection	12.2	11.6	0.793		
unconstrained - basic	17.0	16.2	0.746	**124**	**60**
unconstrained - factored	16.6	15.7	**0.744**	804	593
unconstrained - basic + back-translated	**17.1**	**16.4**	0.752	544	305
Finnish – English	**BLEU lower**	**BLEU cased**	**TER**	**unknown words #tokens**	**unknown words #types**
constrained - factored	20.5	19.3	0.706	2,655	2,004
unconstrained - factored	**23.3**	**22.1**	**0.670**	**1,128**	**842**

Table 4: Official results for the WMT 2016 news test set. The systems including the back-translated news data were submitted after the deadline and will not be listed as official submissions. The system in *italics* are marked for manual evaluation at WMT.

ings of the 23rd International Conference on Computational Linguistics, COLING '10, pages 89–97. Association for Computational Linguistics.

Fabienne Cap, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014. How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT. In *EACL'14: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 579–587.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *HLT-NAACL'12: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, volume 12, pages 34–35. Association for Computational Linguistics.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *HLT-NAACL*, pages 1185–1195.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of NAACL*, pages 644–648.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word formation in SMT. In *EACL'12: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674. Association for Computational Linguistics.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria, August. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal, September. Association for Computational Linguistics.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2013. Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation*, pages 1–39. In press. Available online.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, pages 690–696.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK, July. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*, pages 177–180.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. HFST tools for morphology — an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, volume 41 of *Communications in Computer and Information Science*, pages 28–47. Springer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2015: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, Portorož, Slovenia.

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards universal web parsebanks. In *Proceedings of the International Conference on Dependency Linguistics (Depling'15)*, pages 211–220. Uppsala University.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

Robert Östling. 2015. *Bayesian Models for Multilingual Word Alignment*. Ph.D. thesis, Stockholm University. software at `https://github.com/robertostling/efmaral`.

Tommi A. Pirinen. 2015. Omorfi —free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT.

Andreas Stolcke. 2002. SRILM – an extensible language modelling toolkit. In *ICSLN'02: Proceedings of the international conference on spoken language processing*, pages 901–904.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT'08)*, pages 135–138, Columbus, Ohio, USA.

Jörg Tiedemann, Filip Ginter, and Jenna Kanerva. 2015a. Morphological segmentation and OPUS for Finnish-English machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 177–183, Lisbon, Portugal, September. Association for Computational Linguistics.

Jörg Tiedemann, Filip Ginter, and Jenna Kanerva. 2015b. Morphological segmentation and OPUS for Finnish-English machine translation. In *WMT'15: Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 177–183.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Jörg Tiedemann. 2015. Baseline models for pronoun prediction and pronoun-aware translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 108–114, Lisbon, Portugal, September. Association for Computational Linguistics.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, Portorož, Slovenia.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *ACL'08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 514–522. Association for Computational Linguistics.