Tuning Phrase-Based Segmented Translation for a Morphologically Complex Target Language

Stig-Arne Grönroos
Department of Signal Processing and Acoustics
Aalto University, Finland
stig-arne.gronroos@aalto.fi

Sami Virpioja
Department of Computer Science
Aalto University, Finland
sami.virpioja@aalto.fi

Mikko Kurimo
Department of Signal Processing and Acoustics
Aalto University, Finland
mikko.kurimo@aalto.fi

Abstract

This article describes the Aalto University entry to the English-to-Finnish shared translation task in WMT 2015. The system participates in the constrained condition, but in addition we impose some further constraints, using no language-specific resources beyond those provided in the task. We use a morphological segmenter, Morfessor FlatCat, but train and tune it in an unsupervised manner. The system could thus be used for another language pair with a morphologically complex target language, without needing modification or additional resources.

1 Introduction

In isolating languages, such as English, suitable smallest units of translation are easy to find using whitespace and punctuation characters as delimiters. This approach of using words as the smallest unit of translation is problematic for synthetic languages with rich inflection, derivation or compounding. Such languages have very large vocabularies, leading to sparse statistics and many out-of-vocabulary words.

A synthetic language uses fewer words than an isolating language to express the same sentence, by combining several grammatical markers into each word and using compound words. This difference in granularity is problematic in alignment, when a word in the isolating language properly aligns with only a part of a word in the synthetic language.

In order to balance the number of tokens between target and source, it is often possible to segment the morphologically richer side. Oversegmentation is detrimental, however, as longer windows of history need to be used, and useful phrases become more difficult to extract. It is therefore important to find a balance in the amount of segmentation. A linguistically accurate segmentation may be oversegmented for the task of translation, if some of the distinctions are either unmarked or marked in a similar way in the other language.

An increase in the number of tokens means that the distance spanned by dependencies becomes longer. Recurrent Neural Network (RNN) based language models have been shown to perform well for English (Mikolov et al., 2011). Their strength lies in being theoretically capable of modeling arbitrarily long dependencies.

Moreover, a huge vocabulary is particularly detrimental for neural language models due to their computationally heavy training and need to marginalize over the whole vocabulary during prediction. As morphological segmentation can reduce the vocabulary size considerably, using RNN language models seems even more suitable for this approach.

Our system is designed for translation in the direction from a morphologically less complex to a more complex language. The opposite direction – simplifying morphology – has received more attention, especially with English as the target language.

Of the target languages in this year's task, Finnish is the most difficult to translate into, shown by Koehn (2005) and reconfirmed by the evaluations of this shared task. Even though the use of supervised linguistic tools

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 105–111,
Lisboa, Portugal, 17-18 September 2015.

(such as taggers, parsers, or morphological analyzers) was allowed in the constrained condition, our method does not use them. It is therefore applicable to other morphologically complex target languages.

1.1 Related work

The idea of transforming morphology to improve statistical machine translation (SMT) is well established in the literature. An early example is Nießen and Ney (2004), who apply rule-based morphological analysis to enhance German→English translation.

In particular, many efforts have focused on increasing the symmetry between languages in order to improve alignment. Lee (2004) uses this idea for Arabic→English translation. In this translation direction, symmetry is increased through morphological simplification.

It has been shown that a linguistically correct segmentation does not coincide with the optimal segmentation for purposes of alignment, both using rule-based simplification of linguistic analysis (Habash and Sadat, 2006), and through the use of statistical methods (Chung and Gildea, 2009).

Using segmented translation with unsupervised statistical segmentation methods has yielded mixed results. Virpioja et al. (2007) used Morfessor Categories-MAP in translation between three Nordic languages, including Finnish, while Fishel and Kirik (2010) used Morfessor Categories-MAP in English↔Estonian translation. In these studies, segmentation has in many cases worsened BLEU compared to word-based translation. The main benefit of segmentation has been a decrease in the ratio of untranslated words.

Salameh et al. (2015) translate English→Arabic, and find that segmentation is most useful when the extracted phrases are morphologically productive, and that using a word-level language model reduces this productivity (albeit increasing the BLEU score).

The desegmentation process, and the effect of different strategies for marking the word-internal token boundaries, have mostly been examined in recombining split compound words. Stymne and Cancedda (2011) explore different marking strategies, including use of part-of-speech tags, in order to allow the translation system to produce compounds unseen in the training data.

2 System overview

An overview of the system is shown in Figure 1. The four main contributions of this work are indicated by numbered circles:

1. Use of unsupervised Morfessor FlatCat (Grönroos et al., 2014) for morphological segmentation,
2. Tuning the morphological segmentation directly to balance the number of translation tokens between source and target,
3. A new marking strategy for morph boundaries,
4. Rescoring n-best lists with RNNLM (Mikolov et al., 2010).

Our system extends an existing phrase-based SMT system to perform segmented translation, by adding pre-processing and post-processing steps, with no changes to the decoder. As translation system to be extended, we used the Moses release 3.0 (Koehn et al., 2007). We used GIZA++ alignment, and a 5-gram LM with modified-KN smoothing. Many Moses settings were left at their default values: phrase length 10, grow-diag-final-and alignment symmetrization, msd-bidirectional-fe reordering, and distortion limit 6.

The standard pre-processing steps not specified in Figure 1 consist of normalization of punctuation, tokenization, and statistical truecasing. All three of these were performed with the tools included in Moses.

In addition, the parallel data was cleaned and duplicate sentences were removed. Cleaning was performed after morphological segmentation, as the segmentation can increase the length in tokens of a sentence.

The post-processing steps are the reverse of the pre-processing steps: desegmentation, detruecasing, and detokenization. Rescoring of the n-best list was done before post-processing.

The feature weights were tuned using MERT (Och, 2003), with BLEU (Papineni et al., 2002) of the post-processed hypothesis

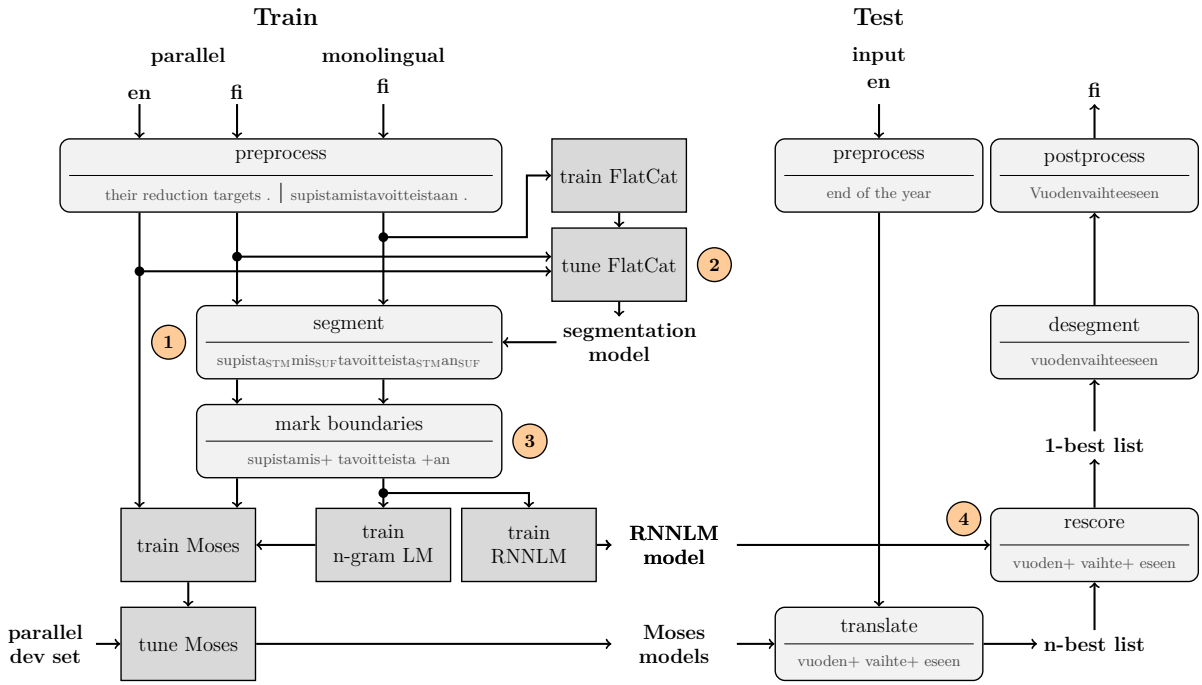

Figure 1: A pipeline overview of training and testing of the system. Main contributions are hilighted with numbers 1-4.

against a tuning set as the metric. 20 random restarts per MERT iteration were used, with iterations repeated until convergence.

A similar MERT procedure was also used for choosing the interpolation weights for rescoring, with 100 random restarts in a single iteration. A single-iteration approach was chosen, as there was no need to translate a new n-best list during the MERT for rescoring.

2.1 Morphological segmentation

For morphological segmentation, we use the latest Morfessor variant, FlatCat (Grönroos et al., 2014). Morfessor FlatCat is a probabilistic method for learning morphological segmentations, using a prior over morph lexicons inspired by the Minimum Description Length principle (Rissanen, 1989).

Morfessor FlatCat applies a Hidden Markov model for morphotactics. Compared to Morfessor Baseline, it provides morph category tags (stem, prefix, suffix) and has superior consistency especially in compound word splitting. In contrast to Categories-MAP (Creutz and Lagus, 2005), used for statistical machine translation e.g. by Clifton and Sarkar (2011), it supports semi-supervised learning and hyper-parameter tuning.

No annotated data was used in the training of Morfessor FlatCat, neither in training nor parameter tuning. Instead of aiming for a linguistic morphological segmentation, our goal was to balance the number of translation tokens between source and target languages.

In order to bring the number of tokens on the Finnish target side closer to the English source side, we segmented the Finnish text with an unsupervised Morfessor FlatCat model, tuned specifically to achieve this balance. The corpus weight hyper-parameter α was chosen by minimizing the sentence-level difference in token counts between the English and the segmented Finnish sides of the parallel corpus

$$\alpha = \arg\min_{\alpha} \sum_{(e,f)\in(E,F)} \Big|\#(e) - \#\big(M(f;\alpha)\big)\Big|, \tag{1}$$

where # gives the number of tokens in the sentence, and $M(f;\alpha)$ is the segmentation with a particular α.

Numbers and URLs occurring in the parallel corpus were passed through Morfessor unseg-

mented, but translated by Moses without any special handling.

2.2 Morph boundary marking strategy

In the desegmentation step, consecutive tokens are concatenated either with or without an intermediary space. Morph boundaries must be distinguished from word boundaries, so that the desegmentation step can reconstruct the words correctly. There are various ways to mark the boundaries, some of them shown in Table 1.

A common way is to attach a symbol to all morphs on the right (or left) side of the morph boundary. We call this strategy *right-only*.

Alternatively *both-sides* of the boundary can be marked. In this strategy, a decision must be made whether to be aggressive or conservative in joining morphs, if the translation system outputs an incorrect sequence where the markers do not match up on both sides. For these experiments we chose the conservative approach, removing the unmatched marker from a half-marked boundary, and treating it as a word boundary.

A downside of the *right-only* and *both-sides* strategies is that a stem is marked differently depending on whether it has a prefix attached or not, even if the surface form of the stem does not change.

The morph categories produced by FlatCat can be used for marking boundaries according to the structure of the word. We can mark affixes from the side that points towards the stem, leaving stems unmarked regardless of the presence of affixes. However, this would leave the boundaries between compound parts indistinguishable from word boundaries, making some additional marking necessary.

Marking affixes by category and compound boundaries with a special linking token is called the *compound-symbol* strategy. Instead marking the last morpheme in the compound modifiers (non-final compound parts), results in the *compound-left* strategy.

After initial unimpressive results with the compound marking strategies, we concluded that segmenting the compound modifiers does not lead to productive translation phrases, in contrast to boundaries between compound parts and boundaries separating inflective affixes. In response, we formulated the *advanced* marking strategy, which goes beyond boundary marking to modify the segmentation, by rejoining the morphs in the modifier parts of compounds.

Strategy	Example
Surface form	supistamistavoitteistaan
Segmentation	$\text{supista}_{\text{STM}}\text{mis}_{\text{SUF}}\text{tavoitteista}_{\text{STM}}\text{an}_{\text{SUF}}$
Translation	of their reduction targets
right-only	supista +mis +tavoitteista +an
both-sides	supista+ +mis+ +tavoitteista+ +an
compound-sym	supista +mis +@+ tavoitteista +an
compound-left	supista +mis@ tavoitteista +an
advanced	supistamis+ tavoitteista +an

Table 1: Morph boundary marking strategies.

The sequence of morph categories is used for grouping the morphs into compound parts. A word consists of one or more compound parts. Each compound part consists of exactly one stem, and any number of preceding prefixes and following suffixes.

$$\begin{aligned} \textsc{CompoundPart} &= \textsc{Pre}^* \ \textsc{Stm} \ \textsc{Suf}^* \\ \textsc{Word} &= \textsc{CompoundPart}^+ \end{aligned} \quad (2)$$

For all compound parts except the last one, the affixes are rejoined to their stem. Morphs of length 5 or above were treated as stems, regardless of the category assigned to them by FlatCat.

Prefixes and compound modifiers are marked with a trailing '+', suffixes are marked with a leading '+', and the stems of the word-final compound parts are left unmarked.

2.3 Rescoring n-best lists

Segmentation of the word forms increases the distances spanned by dependencies that should be modeled by the language model. To compensate this, we apply a strong recurrent neural network language model (RNNLM) (Mikolov et al., 2010). The additional language model is used in a separate rescoring step, to speed up translation, and for ease of implementation.

The RNNLM model was trained on morphologically segmented data. Morphs occurring only once were removed from the vocabulary, and replaced with <UNK>. The parameters were set to 300 nodes in the hidden layer, 500 vocabulary classes, 2M direct connections of

	Monolingual data	Parallel data			
Purpose	news2014 v2	europarl v8	wikititles	newsdev2015	test2006
Training Morfessor	fi	fi	fi		
Training LMs	fi	fi	fi		
Training Moses		en – fi	en – fi		
Tuning Morfessor		en – fi			
Tuning RNNLM				fi	
Tuning Moses				en – fi	
Development testing					en – fi
Sentences	1378582	1926114	153728	1500	2000

Table 2: The data sets used for different purposes. "en–fi" signifies that parallel data was used, "fi" signifies monolingual data, or using only the Finnish side of parallel data.

order 4, backpropagation through 5 time steps, with blocksize 25.

At translation time, 1000-best lists of morph segmented hypotheses produced by Moses were scored using the RNNLM.

The Moses features were extended by including the RNNLM score as an additional feature. A new linear combination of the features was optimized with MERT, and used for the final hypothesis ranking. For the BLEU measurement in MERT the segmented hypothesis was post-processed (including desegmentation) and compared to an un-preprocessed reference.

3 Data

The data sets used in training and tuning are shown in Table 2. Both *europarl v8* and *wikititles* were used as parallel training data, but only europarl was used for tuning the hyperparameter α, as the titles do not follow a typical sentence structure.

The Finnish side of the parallel sets was used to extend the monolingual training data. The monolingual data were concatenated for LM training, instead of interpolating different n-gram models.

After cleaning, the combined parallel training data contained 2,004,450 sentences. The parallel set used for testing during development is *test2006*, a europarl subset of 2000 sentences sampled from three last months of 2000.[1]

[1] `http://matrix.statmt.org/test_sets/list`

Configuration	dev-test test2006 BLEU	test newstest2015 BLEU
advanced, $\alpha = 0.7$	**.147**	.112
+rescoring	**.147**	**.116**
advanced, $\alpha = 0.4$	.145	.112
both-sides	.141	.114
compound-left	.140	.113
compound-sym	.139	.111
right-only	.139	.111
(word)	.146	.100

Table 3: Results of evaluation.

4 Results

Table 3 shows cased BLEU scores on the in-domain development set and out-of-domain test set, for various configurations. The entry marked *word* is a baseline system without segmentation.

When evaluating on the in-domain development set, most configurations that use segmentation achieve worse BLEU compared to the word baseline. Only the best configurations, using the *advanced* strategy, are able to achieve slightly higher BLEU.

Switching domains to the test corpus leads to a larger difference, in favor of the segmenting methods. The choice of morph boundary marking strategy and the sentence-based tuning of the segmentation had a moderate effect on BLEU. The addition of rescoring did not improve BLEU on the in-domain dev-test corpus, but resulted in a slight improvement on

the out-of-domain test corpus.

The proportion of word tokens that were segmented into at least two parts was 19.8%. The joining of compound modifiers did not have a large effect on the total number of tokens, causing a reduction from 49,524,520 to 49,475,291 (0.1%).

Using the sentence-level balancing, the optimal value for the corpus weight hyperparameter α was 0.7. The change in the number of tokens caused by the joining of compound modifiers did not affect the optimum. Balancing the token count of the whole corpus yielded a much lower α of 0.4, leading to oversegmentation and lower BLEU.

The weight of the RNNLM in the final linear combination was 0.092, compared to 0.119 of the n-gram LM. This indicates that it is able to complement the n-gram model, but does not dominate it.

In the human evaluation of WMT15, the system with advanced morph boundary marking strategy and RNNLM rescoring was ranked in tied second place of five methods participating in the constrained condition.

5 Conclusions

To improve English-to-Finnish translation in a phrase-based machine translation system, we tuned an unsupervised morphological segmentation preprocessor to balance the token count between source and target languages. Appropriate choice of morph boundary marking strategy and amount of segmentation brought the BLEU score slightly above a word-based baseline, in contrast to some previous work with unsupervised segmentation (Virpioja et al., 2007; Fishel and Kirik, 2010).

To compensate for the need of longer contexts, we added a recurrent neural network language model as a rescoring step. It did not help for the in-domain development corpus, but improved results on the out-of-domain test corpus.

Possible directions for future work include Minimum Bayes Risk combination of translation hypotheses from systems trained with different segmentations and marking strategies (De Gispert et al., 2009), using morphology generation instead of segmented translation (Clifton and Sarkar, 2011), and improving the alignment directly in addition to balancing of token counts (Snyder and Barzilay, 2008).

Acknowledgments

This research has been supported by the Academy of Finland under the Finnish Centre of Excellence Program 2012–2017 (grant n°251170), and LASTU Programme (grants n°256887 and 259934). Computer resources within the Aalto University School of Science "Science-IT" project were used.

References

[Chung and Gildea2009] Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 718–726. Association for Computational Linguistics.

[Clifton and Sarkar2011] Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of ACL-HLT*, pages 32–42. Association for Computational Linguistics.

[Creutz and Lagus2005] Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In Timo Honkela, Ville Könönen, Matti Pöllä, and Olli Simula, editors, *Proceedings of AKRR'05*, pages 106–113, Espoo, Finland, June. Helsinki University of Technology, Laboratory of Computer and Information Science.

[De Gispert et al.2009] Adrià De Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of HLT-NAACL 2009: Short Papers*, pages 73–76. Association for Computational Linguistics.

[Fishel and Kirik2010] Mark Fishel and Harri Kirik. 2010. Linguistically motivated unsupervised segmentation for machine translation. In *LREC*.

[Grönroos et al.2014] 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*, pages 1177–1185. Association for Computational Linguistics.

[Habash and Sadat2006] Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of HLT-NAACL*. Association for Computational Linguistics.

[Koehn et al.2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

[Koehn2005] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

[Lee2004] Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 57–60. Association for Computational Linguistics.

[Mikolov et al.2010] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

[Mikolov et al.2011] Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukas Burget, and Jan Honza Cernocky. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *Interspeech*. ISCA, August.

[Nießen and Ney2004] Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics*, 30(2):181–204.

[Och2003] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Rissanen1989] Jorma Rissanen. 1989. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, Singapore.

[Salameh et al.2015] Mohammad Salameh, Colin Cherry, and Grzegorz Kondrak. 2015. What matters most in morphologically segmented SMT models? *Syntax, Semantics and Structure in Statistical Translation*, page 65.

[Snyder and Barzilay2008] Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *ACL*, pages 737–745.

[Stymne and Cancedda2011] Sara Stymne and Nicola Cancedda. 2011. Productive generation of compound words in statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 250–260. Association for Computational Linguistics.

[Virpioja et al.2007] Sami Virpioja, Jaakko J Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007:491–498.