Applying morphological decomposition to statistical machine translation

Sami Virpioja and **Jaakko Väyrynen** and **André Mansikkaniemi** and **Mikko Kurimo**
Aalto University School of Science and Technology
Department of Information and Computer Science
PO BOX 15400, 00076 Aalto, Finland
{svirpioj,jjvayryn,ammansik,mikkok}@cis.hut.fi

Abstract

This paper describes the Aalto submission for the German-to-English and the Czech-to-English translation tasks of the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Statistical machine translation has focused on using words, and longer phrases constructed from words, as tokens in the system. In contrast, we apply different morphological decompositions of words using the unsupervised Morfessor algorithms. While translation models trained using the morphological decompositions did not improve the BLEU scores, we show that the Minimum Bayes Risk combination with a word-based translation model produces significant improvements for the German-to-English translation. However, we did not see improvements for the Czech-to-English translations.

1 Introduction

The effect of morphological variation in languages can be alleviated by using word analysis schemes, which may include morpheme discovery, part-of-speech tagging, or other linguistic information. Words are very convenient and even efficient representation in statistical natural language processing, especially with English, but morphologically rich languages can benefit from more fine-grained information. For instance, statistical morphs discovered with unsupervised methods result in better performance in automatic speech recognition for highly-inflecting and agglutinative languages (Hirsimäki et al., 2006; Kurimo et al., 2006).

Virpioja et al. (2007) applied morph-based models in statistical machine translation (SMT) between several language pairs without gaining improvement in BLEU score, but obtaining reductions in out-of-vocabulary rates. They utilized morphs both in the source and in the target language. Later, de Gispert et al. (2009) showed that Minimum Bayes Risk (MBR) combination of word-based and morph-based translation models improves translation with Arabic-to-English and Finnish-to-English language pairs, where only the source language utilized morph-based models. Similar results have been shown for Finnish-to-English and Finnish-to-German in performance evaluation of various unsupervised morpheme analysis algorithms in Morpho Challenge 2009 competition (Kurimo et al., 2009).

We continue the research described above and examine how the level of decomposition affects both the individual morph-based systems and MBR combinations with the baseline word-based model. Experiments are conducted with the WMT10 shared task data for German-to-English and Czech-to-English language pairs.

2 Methods

In this work, morphological analyses are conducted on the source language data, and each different analysis is applied to create a unique segmentation of words into morphemes. Translation systems are trained with the Moses toolkit (Koehn et al., 2007) from each differently segmented version of the same source language to the target language. Evaluation with BLEU is performed on both the individual systems and system combinations, using different levels of decomposition.

2.1 Morphological models for words

Morfessor (Creutz and Lagus, 2002; Creutz and Lagus, 2007, etc.) is a family of methods for unsupervised morphological segmentation. Morfessor does not limit the number of morphemes for each word, making it suitable for agglutinative and compounding languages. An analysis of a single word is a list of non-overlapping segments,

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 201–206,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

morphs, stored in the model lexicon. We use both the Morfessor Baseline (Creutz and Lagus, 2005b) and the Morfessor Categories-MAP (Creutz and Lagus, 2005a) algorithms.[1] Both are formulated in a maximum a posteriori (MAP) framework, i.e., the learning algorithm tries to optimize the product of the model prior and the data likelihood.

The generative model applied by Morfessor Baseline assumes that the morphs are independent. The resulting segmentation can be influenced by using explicit priors for the morph lengths and frequencies, but their effect is usually minimal. The training data has a larger effect on the results: A larger data set allows a larger lexicon, and thus longer morphs and less morphs per word (Creutz and Lagus, 2007). Moreover, the model can be trained with or without taking into account the word frequencies. If the frequencies are included, the more frequent words are usually undersegmented compared to a linguistic analysis, whereas the rare words are oversegmented (Creutz and Lagus, 2005b). An easy way to control the amount of segmentation is to weight the training data likelihood by a positive factor α. If $\alpha > 1$, the increased likelihood results in longer morphs. If $\alpha < 1$, the morphs will be shorter and the words more segmented.

Words that are not present in the training data can be segmented using an algorithm similar to Viterbi. The algorithm can be modified to allow new morphs types to be used by using an approximative cost of adding them into the lexicon (Virpioja and Kohonen, 2009). The modification prevents oversegmentation of unseen word forms. In machine translation, this is important especially for proper nouns, for which there is usually no need for translation.

The Morfessor Categories-MAP algorithm extends the model by imposing morph categories of stems, prefixes and suffixes, as well as transition probabilities between them. In addition, it applies a hierarchical segmentation model that allows it to construct new stems from smaller pieces of "nonmorphemes" (Creutz and Lagus, 2007). Due to these features, it can provide reasonable segmentations also for those words that contain new morphemes. The drawback of the more sophisticated model is the slower and more complex training algorithm. In addition, the amount of the segmentation is harder to control.

Morfessor Categories-MAP was applied to statistical machine translation by Virpioja et al. (2007) and de Gispert et al. (2009). However, Kurimo et al. (2009) report that Morfessor Baseline outperformed Categories-MAP in Finnish-to-English and German-to-English tasks both with and without MBR combination, although the differences were not statistically significant. In all the previous cases, the models were trained on word types, i.e., without using their frequencies. Here, we also test models trained on word tokens.

2.2 Statistical machine translation

We utilize the Moses toolkit (Koehn et al., 2007) for statistical machine translation. The default parameter values are used except with the segmented source language, where the maximum sentence length is increased from 80 to 100 tokens to compensate for the larger number of tokens in text.

2.3 Morphological model combination

For combining individual models, we apply Minimum Bayes Risk (MBR) system combination (Sim et al., 2007). N-best lists from multiple SMT systems trained with different morphological analysis methods are merged; the posterior distributions over the individual lists are interpolated to form a new distribution over the merged list. MBR hypotheses selection is then performed using sentence-level BLEU score (Kumar and Byrne, 2004).

In this work, the focus of the system combination is not to combine different translation systems (e.g., Moses and Systran), but to combine systems trained with the same translation algorithm using the same source language data with with different morphological decompositions.

3 Experiments

The German-to-English and Czech-to-English parts of the ACL WMT10 shared task data were investigated. Vanilla SMT models were trained with Moses using word tokens for MBR combination and comparison purposes. Several different morphological segmentation models for German and Czech were trained with Morfessor. Each segmentation model corresponds to a morph-based SMT model trained with Moses. The word-based vanilla Moses model is compared to each morph-based model as well as to several MBR com-

[1] The respective software is available at `http://www.cis.hut.fi/projects/morpho/`

binations between word-based translation models and morph-based translation models. Quantitative evaluation is carried out using the BLEU score with re-cased and re-tokenized translations.

4 Data

The data used in the experiments consisted of Czech-to-English (CZ-EN) and German-to-English (DE-EN) parallel language data from ACL WMT10. The data was divided into distinct training, development, and evaluation sets. Statistics and details are shown in Table 1.

Aligned data from Europarl v5 and News Commentary corpora were included in training German-to-English SMT models. The English part from the same data sets was used for training a 5-gram language model, which was used in all translation tasks. The Czech-to-English translation model was trained with CzEng v0.9 (training section 0) and News Commentary data. The monolingual German and Czech parts of the training data sets were used for training the morph segmentation models with Morfessor.

The data sets news-test2009, news-syscomb2009 and news-syscombtune2010 from the ACL WMT 2009 and WMT 2010, were used for development. The news-test2008, news-test2010, and news-syscombtest2010 data sets were used for evaluation.

4.1 Preprocessing

All data sets were preprocessed before use. XML-tags were removed, text was tokenized and characters were lowercased for every training, development and evaluation set.

Morphological models for German and Czech were trained using a corpus that was a combination of the respective training sets. Then the models were used for segmenting all the data sets, including development and evaluation sets, with the Viterbi algorithm discussed in Section 2.1. The modification of allowing new morph types for out-of-vocabulary words was not applied.

The Moses cleaning script performed additional filtering on the parallel language training data. Specifically, sentences with over 80 words were removed from the vanilla Moses word-based models. For morph-based models the limit was set to 100 morphs, which is the maximum limit of the Giza++ alignment tool. After filtering with a threshold of 100 tokens, the different morph segmentations for DE-EN training data from combined Europarl and News Commentary data sets ranged from 1 613 556 to 1 624 070 sentences. Similarly, segmented CZ-EN training data ranged from 896 163 to 897 744 sentences. The vanilla words-based model was trained with 1 609 998 sentences for DE-EN and 897 497 sentences for CZ-EN.

5 Results

The details of the ACL WMT10 submissions are shown in Table 2. The results of experiments with different morphological decompositions and MBR system combinations are shown in Table 3. The significances of the differences in BLEU scores between the word-based model (Words) and models with different morphological decompositions was measured by dividing each evaluation data set into 49 subsets of 41–51 sentences, and using the one-sided Wilcoxon signed rank test ($p < 0.05$).

5.1 Segmentation

We created several word segmentations with Morfessor baseline and Morfessor Categories-MAP (CatMAP). Statistics for the different segmentations are given in Table 3. The amount of segmentation was measured as the average number of morphs per word (m/w) and as the percentage of segmented words (s-%) in the training data. Increasing the data likelihood weight α in Morfessor Baseline increases the amount of segmentation for both languages. However, it had little effect on the proportion of segmented words in the three evaluation data sets: The proportion of segmented word tokens was 10–11 % for German and 8–9 % for Czech, whereas the out-of-vocabulary rate was 7.5–7.8 % for German and 4.8–5.6 % for Czech.

Disregarding the word frequency information in Morfessor Baseline (nofreq) produced more morphs per word type and segmented nearly all words in the training data. The Morfessor CatMAP algorithm created segmentations with the largest number of morphs per word, but did not segment as many words as the Morfessor Baseline without the frequencies.

5.2 Morph-based translation systems

The models with segmented source language performed worse individually than the word-based models. The change in the BLEU score was statistically significant in almost all segmentations and

Data set	**Statistics**				**Training**					**Development**	**Evaluation**
	Sentences	**Words per sentence**			**SM**		**LM**	**TM**			
		DE	CZ	EN	DE	CZ	EN	DE-EN	CZ-EN	{DE,CZ}-EN	{DE,CZ}-EN
Europarl v5	1 540 549	23.2		25.2	x		x	x			
News Commentary	100 269	21.9	18.9	21.5	x	x	x	x	x		
CzEng v0.9 (training section 0)	803 286		8.3	9.9		x			x		
news-test2009	2 525	21.7	18.8	23.2						x	
news-syscomb2009	502	19.7	17.2	21.1						x	
news-syscombtune2010	455	20.2	17.3	21.0						x	
news-test2008	2 051	20.3	17.8	21.7							x
news-test2010	2 489	21.7	18.4	22.3							x
news-syscombtest2010	2 034	22.0	18.6	22.6							x

Table 1: Data sets for the Czech-to-English and German-to-English SMT experiments, including the number of aligned sentences and the average number of words per sentence in each language. The data sets used for model training, development and evaluation are marked. Training is divided into German (DE) and Czech (CZ) segmentation model (SM) training, English (EN) language model (LM) training and German-to-English (DE-EN) and Czech-to-English (CZ-EN) translation model (TM) training.

Submission	**Segmentation model for source language**	**BLEU-cased** (news-test2010)
aalto DE-EN WMT10	Morfessor Baseline ($\alpha = 0.5$)	17.0
aalto DE-EN WMT10 CatMAP	Morfessor Categories-MAP	16.5
aalto CZ-EN WMT10	Morfessor Baseline ($\alpha = 0.5$)	16.2
aalto CZ-EN WMT10 CatMAP	Morfessor Categories-MAP	15.9

Table 2: Our submissions for the ACL WMT10 shared task in translation. The translation models are trained from the segmented source language into unsegmented target language with Moses.

all evaluation sets. Morfessor Baseline ($\alpha = 0.5$) was the best individual segmented model for both German and Czech in the sense that it had the lowest number of significant decreases the BLEU score compared to the word-based model. Removing word frequency information with Morfessor Baseline and using Morfessor CatMAP gave the lowest BLEU scores with both source languages.

5.3 Translation system combination

For the DE-EN language pair, all MBR system combinations between each segmented model and the word-based model had slightly higher BLUE scores than the individual word-based model. Nearly all improvements were statistically significant.

The BLEU scores for the MBR combinations in the CZ-EN language pair were mostly not significantly different from the individual word-based model. Two scores were significantly lower.

6 Discussion

We have applied concatenative morphological analysis, in which each original word token is segmented into one or more non-overlapping morph tokens. Our results with different levels of segmentation with Morfessor suggest that the optimal level of segmentation is language pair dependent in machine translation.

Our approach for handling rich morphology has not been able to directly improve the translation quality. We assume that improvements might still be possible by carefully tuning the amount of segmentation. The experiments in this paper with different values of the α parameter for Morfessor Baseline were conducted with the word frequencies. The parameter had little effect on the proportion of segmented words in the evaluation data sets, as frequent words were not segmented at all, and out-of-vocabulary words were likely to be oversegmented by the Viterbi algorithm. Future work includes testing a larger range of values for α, also for models trained without the word frequencies, and using the modification of the Viterbi algorithm proposed in Virpioja and Kohonen (2009).

It might also be helpful to only segment selected words, where the selection would be based on the potential benefit in the translation process. In general, the direct segmentation of words into morphs is problematic because it increases the number of tokens in the text and directly increases both model training and decoding complexity. However, an efficient segmentation decreases the number of types and the out-of-vocabulary rate (Virpioja et al., 2007).

We have replicated here the result that an MBR combination of a morph-based MT system with

Segmentation (DE)	**Statistics (DE)**		**BLEU-cased (DE-EN)**				
			news-test2008		news-test2010	news-syscombtest2010	
	m/w	**s-%**	**No MBR**	**MBR with Words**	**No MBR**	**No MBR**	**MBR with Words**
Words	1.00	0.0%	16.37	-	17.28	13.22	-
Morfessor Baseline ($\alpha = 0.5$)	1.82	72.4%	$\mathbf{15.19}^{-}$	16.47^{+}	$\mathbf{17.04}^{\circ}$	$\mathbf{13.28}^{\circ}$	$\mathbf{13.70}^{+}$
Morfessor Baseline ($\alpha = 1.0$)	1.65	61.0%	15.14^{-}	$\mathbf{16.54}^{+}$	16.87^{-}	11.95^{-}	13.66^{+}
Morfessor Baseline ($\alpha = 5.0$)	1.24	23.7%	15.04^{-}	16.44°	16.63^{-}	11.78^{-}	13.43^{+}
Morfessor CatMAP	2.25	67.5%	14.21^{-}	16.42°	16.53^{-}	11.15^{-}	13.61^{+}
Morfessor Baseline nofreq	2.24	91.6%	13.98^{-}	16.47^{+}	16.36^{-}	10.66^{-}	13.58^{+}

Segmentation (CZ)	**Statistics (CZ)**		**BLEU-cased (CZ-EN)**				
			news-test2008		news-test2010	news-syscombtest2010	
	m/w	**s-%**	**No MBR**	**MBR with Words**	**No MBR**	**No MBR**	**MBR with Words**
Words	1.00	0.0%	14.91	-	16.73	12.75	-
Morfessor Baseline ($\alpha = 0.5$)	1.19	17.7%	13.22^{-}	14.87°	16.01^{-}	$\mathbf{12.60}^{\circ}$	12.53^{-}
Morfessor Baseline ($\alpha = 1.0$)	1.09	8.1%	13.33^{-}	14.88°	$\mathbf{16.10}^{-}$	11.29^{-}	12.84°
Morfessor Baseline ($\alpha = 5.0$)	1.03	2.9%	$\mathbf{13.53}^{-}$	14.83°	15.92^{-}	11.17^{-}	12.85°
Morfessor CatMAP	2.29	71.9%	11.93^{-}	14.86°	15.79^{-}	10.12^{-}	10.79^{-}
Morfessor Baseline nofreq	2.18	90.3%	12.43^{-}	$\mathbf{14.96}^{\circ}$	15.82^{-}	10.13^{-}	$\mathbf{12.89}^{\circ}$

Table 3: Results for German-to-English (DE-EN) and Czech-to-English (CZ-EN) translation models. The source language is segmented with the shown algorithms. The amount of segmentation in the training data is measured with the average number of morphs per word (m/w) and as proportion of segmented words (s-%) against the word-based model (Words). The trained translation systems are evaluated independently (No MBR) and in Minimum Bayes Risk system combination of word-based translation systems (MBR). Unchanged ($^{\circ}$), significantly higher ($^{+}$) and lower ($^{-}$) BLEU scores compared to the word-based translation model (Words) are marked. The best morph-based model for each column is emphasized.

a word-based MT system can produce a BLEU score that is higher than from either of the individual systems (de Gispert et al., 2009; Kurimo et al., 2009). With the DE-EN language pair, the improvement was statistically significant with all tested segmentation models. However, the improvements were not as large as those obtained before and the results for the CZ-EN language pair were not significantly different in most cases. Whether this is due to the different languages, training data sets, the domain of the evaluation data sets, or some problems in the model training, is currently uncertain.

One very different approach for applying different levels of linguistic analysis is factor models for SMT (Koehn and Hoang, 2007), where pre-determined factors (e.g., surface form, lemma and part-of-speech) are stored as vectors for each word. This provides better integration of morphosyntactic information and more control of the process, but the translation models are more complex and the number and factor types in each word must be fixed.

Our submissions to the ACL WMT10 shared task utilize unsupervised morphological decomposition models in a straightforward manner. The individual morph-based models trained with the source language words segmented into morphs did not improve the vanilla word-based models trained with the unsegmented source language. We have replicated the result for the German-to-English language pair that an MBR combination of a word-based and a segmented morph-based model gives significant improvements to the BLEU score. However, we did not see improvements for the Czech-to-English translations.

Acknowledgments

This work was supported by the Academy of Finland in the project *Adaptive Informatics*, the Finnish graduate school in Language Technology, and the IST Programme of the European Community, under the FP7 project EMIME (213845).

References

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, Pennsylvania, USA.

Mathias Creutz and Krista Lagus. 2005a. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the AKRR'05*, Espoo, Finland.

Mathias Creutz and Krista Lagus. 2005b. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1), January.

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 Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73–76, Boulder, USA, June. Association for Computational Linguistics.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the EMNLP 2007*, pages 868–876, Prague, Czech Republic, June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of ACL, demonstration session*, pages 177–180, Czech Republic, June.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the HLT-NAACL 2004*, pages 169–176.

Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumäe, and Murat Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of the HLT-NAACL 2006*, pages 487–494, New York, USA.

Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. 2009. Overview and results of Morpho Challenge 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September.

K. C. Sim, W. J. Byrne, M. J. F. Gales, H. Sahbi, and P. C. Woodl. 2007. Consensus network decoding for statistical machine translation system combination. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*.

Sami Virpioja and Oskar Kohonen. 2009. Unsupervised morpheme analysis with Allomorfessor. In *Working notes for the CLEF 2009 Workshop*, Corfu, Greece.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*, pages 491–498, Copenhagen, Denmark, September.