Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi

Helsinki University of Technology
Adaptive Informatics Research Centre
P.O.Box 5400, 02015 TKK
Finland
{sami.virpioja, jaakko.j.vayrynen, mathias.creutz, markus.sadeniemi}@tkk.fi

Abstract

In this paper, we apply a method of unsupervised morphology learning to a state-of-the-art phrase-based statistical machine translation (SMT) system. In SMT, words are traditionally used as the smallest units of translation. Such a system generalizes poorly to word forms that do not occur in the training data. In particular, this is problematic for languages that are highly compounding, highly inflecting, or both. An alternative way is to use sub-word units, such as morphemes. We use the Morfessor algorithm to find statistical morpheme-like units (called morphs) that can be used to reduce the size of the lexicon and improve the ability to generalize. Translation and language models are trained directly on morphs instead of words. The approach is tested on three Nordic languages (Danish, Finnish, and Swedish) that are included in the Europarl corpus consisting of the Proceedings of the European Parliament. However, in our experiments we did not obtain higher BLEU scores for the morph model than for the standard word-based approach. Nonetheless, the proposed morph-based solution has clear benefits, as morphologically well motivated structures (phrases) are learned, and the proportion of words left untranslated is clearly reduced.

1. Introduction

Statistical machine translation was applied to the direct translation between eleven European languages, all those present in the Europarl corpus, by Koehn (2005). An impressive number of 110 different translation systems were created, one for each language pair. Koehn discovered that the most difficult language to translate from or to is Finnish. Finnish is a non-Indo-European language and is well known for its extremely rich morphology. As verbs and nouns can, in theory, have hundreds and even thousands of word forms, data sparsity and out-of-vocabulary words present a huge problem even when large corpora are available.

It appears that especially translating into a morphologically rich language poses an even bigger problem than translating from such a language. The study also showed that English, which has almost exclusively been used as the target language, was the easiest language to translate into. Thus it is natural to suspect that English as a target language has biased SMT research.

In this paper, we examine the possibility of using morphological information found in an *unsupervised* manner in SMT. We test the approach with the three Nordic languages: Finnish, Danish and Swedish. Danish and Swedish are closely related languages but differ considerably from Finnish. Danish and Swedish are grammatically very close and much of the vocabulary is shared except for some differences in pronunciation and orthography. The translation task should here be easier than between many other languages, but it is interesting to observe how similar morphological segmentation, on one hand, and phrase structure, on the other, resemble each other.

Recently, many SMT systems have been enhanced with syntactic or semantic elements in the model. Morphological analysis has often been seen as part of this. We like to point out three issues from the previous work on this topic. Our work is not novel for any single of them, but the combination is such that it has not been studied before.

Need of morphological analyzers. Nearly all of the previous studies apply morphological analyzers crafted just for the used languages. Some tools, such as commonly used TreeTagger (Schmid, 1994), are adaptable to many languages, but still need training material tagged by a human. For less resourced languages such analyzers or material may not exist, and even when they do, a more universal way of handling morphology could be preferred. With an unsupervised, language-independent approach, it would be straightforward to build the same 110 SMT systems that Koehn (2005) did. In addition, the unsupervised approach to morphological analysis has been found to work very well on a related task, automatic speech recognition.

It seems that before us, only Sereewattana (2003) has used unsupervised segmentation to enhance SMT. However, she used only small training corpora, and studied only translations from German and French to English.

Choice of the target language. When a morphologically rich language is involved, it has almost exclusively been the source language with English as the target. The two most common source languages seem to be German (Nießen and Ney, 2004; Corston-Oliver and Gamon, 2004) and Arabic (Lee, 2004; Zollmann et al., 2006). There are also studies for translating from Czech (Goldwater and McClosky, 2004), Finnish (Yang and Kirchhoff, 2006), and Spanish, Catalan and Serbian (Popović and Ney, 2004).

A recent exception to the direction of the translation is the English-Turkish translation system by Oflazer and El-Kahlout (2007). With a morphological analyzer for Turkish and TreeTagger for English, they do the translation at the morpheme-level, just as we do. With additional tweaking, such as selective morpheme-grouping for Turkish and augmenting the training data with samples containing only the content words, they improve significantly the translation results.

Size of the training corpora. Usage of morphology has often been seen as a way to manage with scarce resources (Nießen and Ney, 2004). Also those that do not explicitely point this out, obtain larger improvements compared to the baseline the smaller the training corpus is (Sereewattana, 2003; Lee, 2004; Yang and Kirchhoff, 2006). Only Lee (2004) and Yang and Kirchhoff (2006) have used more than half a million sentence pairs for training, and still outperformed the word-based approach.

Like Yang and Kirchhoff (2006), we use the Europarl corpus. For each studied language, we first train the model for morpheme segmentation, and then for each language pair, we train the translation system on a corpora containing more than 800 000 sentences. The discussion to follow focuses on both quantitative and qualitative aspects of the performance of the systems.

2. Methodology

Since the introduction of the so-called IBM model (Brown et al., 1993), a standard statistical machine translation system divides into two parts: the translation model and the language model. Given a text S in the source language, we want to find the text T in the target language that is the most probable translation of S. Bayes' Theorem states that the probability $P(T|S)$ is maximized when the product of the prior probability $P(T)$ and the translation probability $P(S|T)$ is maximized. The former is defined by the language model, and the latter by the translation model. The texts S and T consist of *tokens* separated by whitespace characters. The tokens are the smallest parts that can be translated as such, and typically words are used as tokens. In our morph-based approach, the tokens are morphs instead of words.

In the following subsections, we describe the methods and software used as components of our machine translation system. First, we introduce the Morfessor algorithm for inducing morpheme-like units in an unsupervised manner. Then, we discuss the language models that are used to assign probabilities to the sentences in the target language. The third subsection describes the applied framework for phrase-based machine translation and how morphs are used in the translation.

2.1. Morphological model for words

Morfessor (Creutz and Lagus, 2007) is a method for finding morpheme-like units (morphs) of a language in an unsupervised manner. Morfessor can cope with languages where words can consist of multiple prefixes, stems, and suffixes concatenated together. This distinguishes Morfessor from other algorithms that pose harder restrictions on the possible structures of words, such that each word is assumed to consist of one stem optionally followed by a suffix; see, e.g., Goldsmith (2001). Using morph-based rather than word-based vocabularies has been shown to result in better performance in automatic speech recognition for highly inflecting and agglutinative languages (Hirsimäki et al., 2006; Kurimo et al., 2006).

There exist a few different versions of Morfessor, which correspond to chronological development steps of the algorithm.[1] In this work, we use the Morfessor Categories-MAP algorithm (Creutz and Lagus, 2005), which is formulated in a maximum a posteriori (MAP) framework. Morfessor Categories-MAP has a better segmentation accuracy with respect to a morphological gold standard than the baseline algorithm, and it can treat words not seen in the training data (so called out-of-vocabulary, or OOV, words) in a more convenient manner. For instance, if we encounter a new name with a known suffix (*Pietra's*), the Categories-MAP algorithm can usually separate the suffix (*'s*) and leave the actual name (*Pietra*) intact, whereas Morfessor Baseline would segment it into morphs that are known from other words (*Pie t ra 's*).

[1] Variants of the Morfessor algorithm can be downloaded for free at `http://www.cis.hut.fi/projects/morpho/`. An online demo is also available.

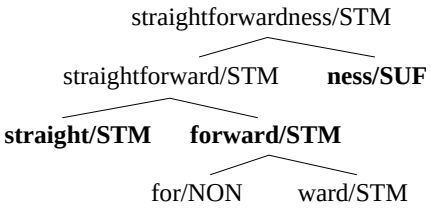

Figure 1: Example segmentation of *straightforwardness* with Morfessor Categories-MAP. The surface form of the segmentation is *straight/STM + forward/STM + ness/SUF*.

Morfessor Categories-MAP tags each discovered morph with one out of four categories. Three are surface categories: prefix (PRE), stem (STM) and suffix (SUF). The fourth one is a special non-morpheme category (NON), which is used only in the internal representation of the model. Each word is assumed to consist of the surface morphemes in a way that is captured by the following regular expression:

```
WORD = ( PRE* STM SUF* )+        (1)
```

The internal structure of the model is hierarchical. For instance, a possible four level segmentation tree of the word *straightforwardness* is shown in Figure 1. The categories of the morphs are estimated based on their length and context. The surface segmentation is selected to be the finest resolution that does not contain non-morphemes.

2.2. Language models

N-gram models are traditional methods of language modeling. They are based on the assumption that the probability of the word in a word sequence depends only on a fixed number $(n-1)$ of previous words. The probabilities are based on maximum likelihood estimates. In order to give non-zero probabilities to unseen words and n-grams, estimates are smoothed. The state-of-the-art smoothing technique is modified Kneser–Ney interpolation (Chen and Goodman, 1999).

Word-based n-gram models are unsuitable for languages of rich morphology. If the modeling is based on words, their number is too large and the data too sparse. A practical solution is to use sub-word units, such as morphs, as the basic units of the language model. However, this forces us to use longer n-grams to model the same context lengths as in word-based models.

Including all n-grams up to the same large n requires much space. Well-known solutions for getting smaller n-gram models without restricting n too much are count cutoffs, model pruning and model growing. In pruning, first a full n-gram model is estimated, and then some of the n-grams are removed using a criterion that is usually based on the change in likelihood of the training data. A more efficient solution is to grow the model incrementally, including longer contexts only when they are really needed. These kind of models are usually called variable n-gram or *varigram* models. Efficient methods for growing and pruning Kneser–Ney smoothed models are presented by Siivola et al. (2007).

We use three types of language models to model the target language in our translation tasks. The two base-

line models, 3-gram and 4-gram models, are trained with the SRI Language Modeling toolkit (Stolcke, 2002). The third is a varigram model trained with the VariKN Language Modeling toolkit based on the algorithms given by Siivola et al. (2007).[2]

2.3. Phrase-based statistical machine translation

The first SMT models (Brown et al., 1993) estimated translation probabilities $P(S|T)$ for pairs of words in the source and target languages. When the framework of phrase-based statistical machine translation was proposed (Koehn et al., 2003), it was observed that the translation quality could be improved by translating sequences of words, *phrases*, at a time. The phrases are collected in an unsupervised manner from the training data.

Moses is an open-source toolkit for phrase-based statistical machine translation (Koehn et al., 2007). Moses automatically trains and applies translation models for any language pair. The system needs a parallel corpus (the same text in two languages) for training the models. After training, Moses can be used for translating new sentences of the source language into the target language.

We use the Moses system to demonstrate how a phrase-based framework can be generalized to be morphologically aware by segmenting words with the Morfessor Categories-MAP model and using the resulting morphs as tokens instead of words. This approach makes it possible to use morphs with Moses without any modifications to the system. For comparison, we also report results on standard word-based translation systems. Hybrid solutions are possible, such that the morph segmentation would only be performed on one of the languages, but such systems are not studied in the current work.

Examples of phrase-based translation using words and morphs as tokens are shown in Figure 2. Similar phrases are constructed from morphs as from words, but the morph phrases are additionally suitable for translating, for instance, compound words in parts. Morph category information (prefix, stem or suffix) is part of the morph label (shown in subscript in the figure) and the plus signs in superscript signifies that the morph is not the last morph of the word. The latter information is necessary in order to reconstruct words from the morphs in the final output.

3. Experiments

In this section, we compare our morphology-aware phrase-based statistical machine translation framework to the more traditional word-based framework and analyze the differences in the two approaches. All main experiments are run on the Moses systems on all six language pairs and with both word tokens and morph tokens. Quantitative evaluation is provided with BLEU scores (Bilingual Evaluation Understudy); see Papineni et al. (2002). The BLEU calculations are based on words regardless of the token type used in the translation, and (wholly or partially) untranslated words are included.

We will first introduce our data sets, which comprise the three Nordic languages present in Europarl, and then we report on the experiments conducted. Language models with different n-gram orders are compared, since morphs are shorter than words and thus a higher order model may be needed in order to span a history of sufficient length. We also test different phrase lengths for morphs. Finally, we compare the performance of the morph-based system to that of the word-based baseline.

languages	token type	train	test
da-fi	word	877 944	1 000
	morph	863 454	1 000
da-sv	word	877 683	1 000
	morph	862 146	1 000
fi-sv	word	888 668	1 000
	morph	876 600	1 000

Table 1: Number of aligned sentences in training and test data sets for each language pair and token type.

3.1. Data

Our data consists of the proceedings of European Parliament from 1996 to 2001 in 11 languages (Koehn, 2005), of which the Nordic languages Danish (da), Finnish (fi) and Swedish (sv) were selected for our experiments. All three pairs of the sentence-aligned bi-texts were preprocessed by removing XML-tags, conversion of some special characters and lowercasing all characters.

The corpora were divided into training, development and test sets. The test set consisted of the last three months of the year 2000, the development set consisted of the sessions of September 2000, and the rest was included in the training set. Morph segmentations were trained with Morfessor using the training sets. The segmentation models produced were utilized to segment the development and test sets. At this point, two data sets were created for each alignment pair: one with the original word tokens and the other with morph tokens. The training sets were used for language model training, and the development sets for parameter tuning.

Additional filtering for the training data was performed by the Moses cleaning script, which removed sentence alignments when either part had no tokens or too many tokens or the ratio of tokens in the two languages was not appropriate. Table 1 shows the number of aligned sentences for each data set after the filtering.

Such sentence pairs were selected into the test set in which both sentences had at least 5 words and at most 15 words. Depending on the language pair, the filtered test set had 10 700–12 900 sentences. Of this set, we used only the 1 000 first sentences for the evaluation.

3.2. Results

The standard approach is to use 3-gram language models for estimating the prior probabilities of sentences in SMT. That may well be enough for word-based translation, but as morphs are shorter than words, we need longer n-grams to cover the same amount of context information. The BLEU scores in Table 2 show that most of the scores are improved if 4-gram models are used instead of 3-gram models. We tested the statistical significance of the differences with the Wilcoxon signed-rank test.[3] In the tables of this section, all statistically significant differences are marked with boldface fonts.

[2] VariKN toolkit is available at `http://varikn.forge.pascal-network.org/`.

[3] One-sided right-tailed test with 10 observed scores each based on 100 translated sentences. The significance level is 0.05.

a	flera reglerande åtgärder behöver införas .			
b	flera	reglerande åtgärder	behöver införas	.
c	eräitä	sääntelytoimia	on toteutettava	.
d	eräitä sääntelytoimia on toteutettava .			

e	flera reglerande åtgärder behöver införas .				
f	$\text{flera}_{\text{STM}}$ $\text{reglera}^{+}_{\text{STM}}$ nde_{SUF} $\text{åtgärd}^{+}_{\text{STM}}$ er_{SUF} $\text{behöv}^{+}_{\text{STM}}$ er_{SUF} $\text{in}^{+}_{\text{PRE}}$ $\text{föra}^{+}_{\text{STM}}$ s_{SUF} $\cdot_{\text{STM}}$				
g	$\text{flera}_{\text{STM}}$	$\text{reglera}^{+}_{\text{STM}}$ nde_{SUF}	$\text{åtgärd}^{+}_{\text{STM}}$ er_{SUF}	$\text{behöv}^{+}_{\text{STM}}$ er_{SUF} $\text{in}^{+}_{\text{PRE}}$ $\text{föra}^{+}_{\text{STM}}$ s_{SUF}	$\cdot_{\text{STM}}$
h	$\text{erä}^{+}_{\text{STM}}$ itä_{SUF}	$\text{sääntely}^{+}_{\text{STM}}$	$\text{toimi}^{+}_{\text{STM}}$ a_{SUF}	on_{STM} $\text{toteute}^{+}_{\text{STM}}$ $\text{tta}^{+}_{\text{SUF}}$ va_{SUF}	$\cdot_{\text{STM}}$
i	$\text{erä}^{+}_{\text{STM}}$ itä_{SUF} $\text{sääntely}^{+}_{\text{STM}}$ $\text{toimi}^{+}_{\text{STM}}$ a_{SUF} on_{STM} $\text{toteute}^{+}_{\text{STM}}$ $\text{tta}^{+}_{\text{SUF}}$ va_{SUF} $\cdot_{\text{STM}}$				
j	eräitä sääntelytoimia on toteutettava .				

Figure 2: Examples of word-based and morph-based Finnish translations for the Swedish sentence "Flera reglerande åtgärder behöver införas ." (*Several regulations need to be implemented* .) The top figure shows the word-based translation process with the source sentence (a), the phrases used (b) and their corresponding translations (c), as well as the final hypothesis (d). The bottom figure illustrates the morph-based translation process with the source sentence as words (e) and as morphs (f), the morph phrases used (g) and their corresponding translations (h), as well as the final hypothesis with morphs (i) and words (j).

	→ da	→ fi	→ sv
da →		+0.59	**+1.19**
fi →	**+1.07**		**+0.93**
sv →	+0.37	-0.18	

Table 2: Absolute changes in BLEU scores for morph-based translations if 4-gram language models are used instead of 3-gram models. Statistically significant differences are highlighted.

	→ da	→ fi	→ sv
da →		+0.23	+0.29
fi →	+0.36		**+0.80**
sv →	-0.25	+0.02	

Table 3: Absolute changes in BLEU scores for word-based translations if 4-gram language models are used instead of 3-gram models.

We performed the same test on the word-based translations. Table 3 shows absolute changes in BLEU scores if 4-gram models are used instead of 3-grams. As the score decreases only in one pair, and increases for the others, we decided to use the 4-gram models also for words.

In addition to longer n-grams in language models, we may need longer phrases in the translation. The default value for the maximum phrase length in the Moses system is 7. Koehn et al. (2003) have shown this to be sufficient for word-based translations. Our preliminary experiments on words supported this: The results were actually worse if the limit was set to 10 instead of 7. But does this hold for morph phrases? Depending on the language, we have 1.3–1.6 morphs per word on average. This means that a maximum phrase length of 7 words would correspond to a maximum phrase length of circa 10 morphs. Table 4 shows that increasing the maximum phrase length to 10 indeed improves the results. In three cases out of six, the increase in the BLEU score is statistically significant.

We also wanted to test the variable n-gram models by Siivola et al. (2007) on morph-based translations, as they have performed well in automatic speech recognition applications. In Table 5, varigram models are compared to

	→ da	→ fi	→ sv
da →		**+0.27**	-0.06
fi →	+0.31		+0.09
sv →	**+0.23**	**+0.20**	

Table 4: Absolute changes in BLEU scores for morph-based translations, if the maximum phrase length is set to 10 instead of 7. The language models were 4-grams in both settings.

	→ da	→ fi	→ sv
da →		-0.55	**-1.09**
fi →	-0.01		+0.20
sv →	-0.07	+0.31	

Table 5: Absolute changes in BLEU scores for morph-based translations, if a varigram model of similar size is used instead of a 4-gram model. The maximum phrase length was 10.

4-gram models. The results are mixed: Two of the scores are somewhat worse, two somewhat better, and two about the same. In one case the decrease in the score is statistically significant.

The results so far are quite interesting as such, but our main result is the comparison of the word and morph-based approaches. For this we have used those language models and maximum phrase lengths that have worked best on average, i.e., 4-gram models for both words and morphs, and a maximum phrase length of 7 for words and 10 for morphs. Table 6 lists the absolute BLEU scores. In Table 7, the differences between the scores are shown, with statistically significant differences highlighted. According to these results, the translations based on morph phrases were slightly worse, but only in two cases the decrease was statistically significant.

4. Analysis of the Results

Although the BLEU scores for word-based and morph-based translation are very close, it is clear that the morphs do not outperform the standard word approach in our ex-

	→ da	→ fi	→ sv
da →		18.26 / 17.66	33.16 / 32.64
fi →	23.63 / 22.40		22.85 / 20.71
sv →	35.95 / 35.49	18.19 / 17.05	

Table 6: BLEU scores for word and morph-based translations. The first value is for the word-based model, the second for the morph-based model. The maximum phrase length was 7 for words and 10 for morphs. 4-gram language models were used for both.

	→ da	→ fi	→ sv
da →		-0.60	-0.52
fi →	-1.23		**-2.14**
sv →	-0.46	**-1.14**	

Table 7: Absolute changes in BLEU scores from word-based translations to morph-based translations. The maximum phrase length was 7 for words and 10 for morphs. 4-gram language models were used for both.

periments. In the following, some further analysis of possible problems and benefits of the morph approach will be discussed.

4.1. BLEU scores

Evaluation of machine translation systems should in the end be dependent on the application that lies behind. Usually there should be human evaluation by several persons to judge the quality of the translations. This, however, is a very expensive method and cannot be used routinely.

As in most of the recent studies, we have used the BLEU scores (Papineni et al., 2002) for quantitative evaluation. BLEU is based on the co-occurrence of n-grams. It counts how many n-grams (usually for $n = 1, \ldots, 4$) the proposed translation has in common with the reference translations and calculates a score based on this. For a realistic evaluation, the calculation of the BLEU scores would need several reference translations made by different persons. Even when such are available, the BLEU score has been criticized, as in some cases human evaluation gives grossly different results (Callison-Burch et al., 2006; Culy and Riehemann, 2003).

Even if we brush aside the criticism and the fact that we have used only one reference translation, BLEU scores have some problematic features for our study. It is clear that for morphologically rich languages, such as Finnish, it is harder to get good scores. Finnish has fewer words for the same text compared to Swedish or Danish, and thus one word includes more information on average. One mistake in one suffix of a word is enough to mark the word as an error. This does not usually prevent understanding the translation, but will drop the scores as much as more "serious" mistakes.

4.2. Untranslated words

In the word-based translation model, only the words that were present in the training data can be translated. The other words are left untranslated. This is the case for all unseen words, even though they may closely resemble known words; for instance, a "new" word may simply be an inflected form of a known word.

language	token type	type count	token count
da	word	226 332	22 714 631
	morph	55 319	29 862 089
fi	word	459 125	17 403 219
	morph	78 222	27 076 855
sv	word	233 217	21 789 747
	morph	59 045	29 370 823

Table 8: Type and token counts for some training data sets with approximately the same number of sentences. The morph have much fewer types and higher token counts than the words. This is especially prominent for Finnish.

word / morph	→ da	→ fi	→ sv
da →		128 / 31	74 / 12
fi →	189 / 41		195 / 44
sv →	76 / 21	132 / 42	

Table 9: Number of sentences not fully translated out of 1 000 with word-based and morph-based phrases. The numbers were the same with all of the tested language models and maximum phrase length combinations.

Our morph-based approach is expected to reduce the problem of unseen tokens, since words that have not been observed before may consist of morphs that are known from the training data. Table 8 shows token and type counts (number of instances vs. number of unique units) for the same data with word and morph tokens. The notably lower type counts with morphs suggests that morphs might produce less untranslated words, since the same vocabulary coverage has been obtained using a smaller number of more frequently occurring units. Table 9 compares the number of sentences not fully translated by the word-based and morph-based systems. It is evident that the morph-based systems are indeed able to translate more words. An examination of the untranslated words reveals that a higher number of compound words and inflected word forms are left untranslated by the word-based systems.

4.3. Performance on the baseforms of words

We noticed that especially when translating into Finnish, both the word and morph models experience difficulties in getting the grammatical endings right. In order to achieve better results it seems that more elaborate models of syntax are needed, or the amount of training data must be increased.

However, since the morph model is capable of translating previously unseen compound words by decomposing them into parts, one may wonder whether the morph model might outperform the word model if the grammatical word endings are disregarded in the evaluation. That is, how do the approaches compare if every word in the proposed translations as well as the reference are restored to their baseforms before the BLEU scores are calculated?

FINTWOL[4], a Finnish morphological analyzer, was used to produce baseforms for each word in the outcome of the Swedish-to-Finnish (sv → fi) translation. A small portion of the words were not recognized by the analyzer and were left unchanged (3.3 % in the word-based trans-

[4]Available from Lingsoft, Inc. (`www.lingsoft.fi`)

	Precision	Recall	F-Measure
da	84.96	64.59	73.39
fi	78.72	52.29	62.84
sv	82.87	64.14	72.31

Table 10: Morpheme segmentation accuracy for the segmentations produced by Morfessor on samples of 500 words.

lation, 2.2 % in the morph-based translation, and 1.8 % in the reference). The BLEU scores obtained for this modified data were circa 5 % higher absolute than the original figures. The morph model improved 0.2 % absolute with respect to the word model, but the word model remained the better of the two. The test was not performed on the other language pairs.

4.4. Quality of the morph segmentations

Since the morph segmentations have been produced using an unsupervised algorithm, Morfessor, the segmentations are not perfectly accurate (if compared to a grammatical, linguistic morpheme segmentation).

To assess how well the Morfessor morphs correspond to linguistic morphs, 500 words were selected by random for each language, and these words were segmented manually. The results of the evaluation of the Morfessor segmentations are shown in Table 10. Precision is the proportion of morph boundaries suggested by Morfessor that are correct according to the linguistic segmentation. Recall corresponds to the proportion of boundaries in the linguistic segmentation that were found by Morfessor. F-measure is the harmonic mean of precision and recall.

Table 10 shows that the segmentation accuracy for Danish and Swedish are very similar. The Finnish morphology is more challenging and consequently the Finnish results are somewhat lower. For all three languages, recall is higher than 50 %, which means that more than half of the correct morpheme boundaries are actually detected. Precision is around 80 %, which implies that 4/5 of the morph boundaries suggested by Morfessor are correct.

It appears that a high precision is to be preferred; that is, the morph boundaries proposed are usually correct. Since recall is not that high, words are undersegmented on average. Compared to a fully linguistic morpheme segmentation, our segmentation is thus more conservative. The difference between a standard word representation and our morph segmentation is smaller than the difference between words and linguistic morphs would be.

4.5. A closer look at some example phrases

Let us take a look at phrases built of morphs. We are particularly interested in phrases that do not span entire words. Although these phrases may contain multiple words, at least one of the phrase boundaries is located at a morph boundary within a word. To the extent that such phrases are beneficial in the translation task, morph models may be considered justified and desirable.

In the following, some true examples of phrases used in the automatic translation are presented. The phrases are marked as |boxes| and morph boundaries are mark+ed with a plus sign.

4.5.1. Productive morphology with the same structure across languages

Compound words are common in the three languages studied. Additionally, inflectional and derivational suffixes exist, to a very high degree in Finnish, and to some extent in Swedish and Danish.

As Swedish and Danish are very closely related, the morphological structure is typically the same in both languages. The Swedish compound word *Köpenhamnskriterierna* (the Copenhagen criteria) has been translated into Danish using four phrases. A literal translation of the phrases is shown on the right:

sv: |Köpenhamn+|s+|kriterier+|na| *(Copenhagen+|'s | criteria | the)*
da: |København+|s+|kriterier+|ne|

Also when translating to or from Finnish, one frequently finds parallel structures, both for nouns (e.g., risk capital markets, of the transition periods) and verbs (e.g., they insulted).

sv: |risk+kapital+|marknad+er| *(risk + capital | market+s)*
fi: |riski+pääoma+|markkina+t|

da: |over+gang+s+|periode+|r+ne+s| *(transition | period+|s_of_the)*
fi: |siirtymä+|jakso+|jen|

fi: |he | herja+|sivat| *(they | insult+|ed)*
sv: |de | förolämpa+|de|

4.5.2. Productive morphology with differing structure across languages

It is pleasant to see that phrases built of morphs often do a successful job, even though the grammatical structure differs across the languages. For instance, in Finnish, the expression *I have* is typically expressed as *by me there is* (minulla on), where *by* is realized as a case ending: *minu+lla* (me + by). A corresponding example of a Swedish-to-Finnish translation is shown below (Lithuania has a ... at its disposal):

sv: |Litauen | för+fogar över en| *(Lithuania | has_at_its_disposal a)*
fi: |Liettua+|lla on käytös+sä+ä+n| *(Lithuania | by there_is at_its_disposal)*

Frequently, a reordering of the phrases must take place. The mood (e.g., would) and person (e.g., we) are marked as verb endings in Finnish, whereas Danish makes use of separate words, which precede the verb:

fi: |reagoi+|si+mme| *(react | would + we)*
da: |vil vi | reagere på| *(will we | react on)*

Where prepositions are commonplace in Danish and Swedish, Finnish utilizes case endings or postpositions. *Into the compartment* is expressed as *lokeroon* (*lokero* being the baseform, and *-on* the illative ending):

fi: |samaa+n | lokero+|on| *(same_into | compartment | into)*
sv: |i | samma | fack| *(in | same | compartment)*

Or to say *between*, one uses a construction akin to *in the between of*, where the postposition *välillä* is used:

da: |mellem | disse | syns+|punkt+er| *(between | these |*

view+|point+s)

fi: | näiden | näkö+|kohti+en | väli+llä | *(of_these | view+| point+s' | between + in)*

Especially in the written language, many a Finn is eager to turn verbs into nouns; instead of saying *to make a decision* one may prefer to say *the making of a decision*. This is yet another case, where phrase reordering is necessary when translating between Finnish on the one hand, and Swedish or Danish on the other hand.

To catch up can be expressed as *the closing of the advantage distance* in Finnish, but has been translated with a verb in the infinitive in Swedish:

fi: | etu+matkan kiinni kuro+|minen | *(advantage + distance's clos+|ing)*

sv: | att | komma i fatt | *(to | catch up)*

No preposition is needed in Finnish to say *with a raise of standards*, since the nominal construction can be used in combination with a case ending: *in the raising of standards*:

fi: | standard+i+en | nosta+|misessa | *(standard+s' | rais+|ing_in)*

da: | med | en for+høje+lse | af standard+er | *(with | a raise of standard+s)*

4.5.3. Lexicalized forms split into phrases

An interesting phenomenon occurs in some translations between Swedish and Danish. Decomposition of words into sub-word phrases takes place even though there is no productive morphological process involved. That is, the whole word is a lexicalized unit, and it is not likely that one would ever need to combine the two neighboring sub-word phrases in any other way.

The words *förståelse/forståelse* (understanding) is one example. Naturally, Danish *for* is a correct translation of Swedish *för*, but it is hard to see that the remainder of the word, *ståelse*, would ever occur in another context:

sv: | för+|stå+else | *(under+|stand+ing)*

da: | for+|ståelse |

Similarly, *uttalande i* has been translated into *udtalelse i* (statement in ...), such that the prefix *ut/ud* (out) has been translated separately:

sv: | ut+|talan+de i | *(out+|speak+ing in = statement in)*

da: | ud+|talelse i |

It seems that the close relationship between the two languages occasionally makes it possible to successfully translate piece by piece, even though the phrases may be very short and not necessarily represent morphologically productive morphemes. This hypothesis is supported by statistics: in the translations between Swedish and Danish (da-sv, sv-da), two thirds of all translated sentences contain at least one phrase boundary within a word, whereas the corresponding is true for only one third of the sentences in translations to and from Finnish.

4.5.4. Questionable phrase segmentations

It is inevitable that the segmentation of words into morphs also gives rise to some errors. Misalignments are one source of errors. For instance, the Swedish word *proportioner* (proportions) has been translated into Finnish as *a sense of proportion*. Curiously enough, the prefix *pro-* is aligned with *suhteelli-* (relative, comparative, proportionate), and *portioner* (portions) is aligned with the farfetched *-suudentaju* (sense of ...-ness).

sv: | pro+|portion+er | *(pro+|portion+s)*

fi: | suhteelli+|suuden+taju | *(proport+|ion's + sense)*

A totally incorrect Swedish translation is produced for the Finnish word *vilpittömänä* (as a sincere ...). A correct translation would be *som en uppriktig*. Apparently, Finnish *vilpi* has been aligned with the Swedish prefix *upp-*. The Finnish suffix *-tön* (and any of its inflections, e.g., *-ttömänä*) denotes a lack of something (e.g., "deceit+*less*"). This often corresponds to the Swedish suffix *-lös* (or some inflection, e.g., *-lösa*). As a result of the translation in parts, we now end up with *upplösa*, a verb meaning *break up* or *dissolve*!

fi: | vilpi+|ttömän+ä | *(deceit+|less_as_a = as a sincere)*

sv: | upp+|lösa | *(up+|loosen = break up / dissolve)*

When languages make different grammatical distinctions, translation gets harder. In Danish and Swedish, the definiteness of noun phrases is marked using suffixes or articles (corresponding to English *a, an, the*). Finnish makes no such distinctions; however, in contrast to Swedish and Danish, Finnish nominals are inflected according to their grammatical case. For instance, the object of the sentence may appear in the genitive, accusative, or nominative case.

To some extent, we have noticed that these different categories may be aligned with each other, more or less consistently. The Swedish definite suffix *-en* is often aligned with the Finnish genitive ending *-n*, for instance in the expression *that the community*:

sv: | att gemenskap+|en | *(that community | the)*

fi: | , että yhteisö+|n | *(that community+|'s)*

5. Discussion and Conclusions

In this paper, we studied how unsupervised morphology learning can be implemented in phrase-based statistical machine translation. Our direct approach applied morph-tokens to SMT in the same way that word-tokens are traditionally used. This requires no changes to the phrase-based framework, only the training of the language and translation models based on morphs. Differences between the methods were analyzed and evaluated in detail using SMT systems trained between three Nordic languages, of which Finnish is clearly separate from Danish and Swedish, which are closely related.

Unfortunately, our morph-based approach resulted in slightly lower BLEU scores than the word baseline; however, only in two systems out of six the drop in performance was statistically significant. Nonetheless, we see several benefits to using morphs: the unsupervised and flexible methodology provides language independence; out-of-vocabulary rates are reduced; and generalization ability is increased through more refined phrases.

We aim to a system that would improve the phrase-based translation with morphology for practically any language pair, and regardless of the size of the training corpus. There are several ways that should improve our current approach. First, there might be some problems in the

morph alignments, as the applied algorithms have been designed for word alignment. Thus word alignments could be used as a starting point for the alignment of morphs. Second, translations based on morphs could be rescored with a word-based language model, as Oflazer and El-Kahlout (2007) have done. Third, translations based on words and morphs could also be combined, e.g., with back-off models (Yang and Kirchhoff, 2006). Fourth, instead of using all the different morph forms, we would like to combine *allomorphs* of the same morphemes into equivalence classes. Factored translation models (Koehn and Hoang, 2007) could help to use them elegantly in the translation. Overall, we are confident that unsupervised morphology learning is useful in the development of the statistical machine translation framework and in improving translation quality across a variety of languages.

6. Acknowledgements

The authors would like to thank Mari-Sanna Paukkeri for her help on the initial experiments, and the anonymous reviewers for useful comments. The second author would like to acknowledge the Nokia Foundation for financial support.

7. References

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the EACL 2006*, Trento, Italy.

Chen, S. F. & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393.

Corston-Oliver, S. & Gamon, M. (2004). Normalizing german and english inflectional morphology to improve statistical word alignment. In *Proceedings of the AMTA 2004*, Washington DC, USA.

Creutz, M. & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the AKRR'05*, Espoo, Finland.

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

Culy, C. & Riehemann, S. Z. (2003). The limits of n-gram translation evaluation metrics. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, USA.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Goldwater, S. & McClosky, D. (2004). Improving statistical mt through morphological analysis. In *Proceedings of the HLT/EMNLP 2005*, Vancouver, Canada.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., & Pylkkönen, J. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541.

Koehn, P. & Hoang, H. (2007). Factored translation models. In *Proceedings of the EMNLP 2007*, Prague, Czech Republic, June.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the HLT-NAACL 2003*, pp. 48–54, Edmonton, Canada.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of ACL, demonstration session*, Czech Republic, June.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pp. 79–86, Phuket, Thailand.

Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., & Saraclar, M. (2006). Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of the HLT-NAACL 2006*, New York, USA.

Lee, Y.-S. (2004). Morphological analysis for statistical machine translation. In *Proceedings of the HLT-NAACL 2004*, pp. 57–60, Boston, USA.

Nießen, S. & Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

Oflazer, K. & El-Kahlout, İ. D. (2007). Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of the Statistical Machine Translation Workshop at ACL 2007*, pp. 25–32, Prague, Czech Republic, June.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pp. 311–318, Morristown, NJ, USA.

Popović, M. & Ney, H. (2004). Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of LREC 2004*, pp. 1585–1588, Lisbon, Portugal.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sereewattana, S. (2003). Unsupervised segmentation for statistical machine translation. Master's thesis, University of Edinburgh, Edinburgh, UK.

Siivola, V., Hirsimaki, T., & Virpioja, S. (2007). On growing and pruning Kneser-Ney smoothed n-gram models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5):1617–1624.

Stolcke, A. (2002). SRILM — an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, pp. 901–904. `http://www.speech.sri.com/projects/srilm/`.

Yang, M. & Kirchhoff, K. (2006). Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the EACL 2006*, pp. 41–48, Trento, Italy.

Zollmann, A., Venugopal, A., & Vogel, S. (2006). Bridging the inflection morphology gap for arabic statistical machine translation. In *Proceedings of the HLT-NAACL 2006*, New York, USA.