Linguistically Motivated Unsupervised Segmentation for Machine Translation

Mark Fishel, Harri Kirik

University of Tartu
Tartu, Estonia
{fishel, harts}@ut.ee

Abstract

In this paper we use statistical machine translation and morphology information from two different morphological analyzers to try to improve translation quality by linguistically motivated segmentation. The morphological analyzers we use are the unsupervised Morfessor morpheme segmentation and analyzer toolkit and the rule-based morphological analyzer T3. Our translations are done using the Moses statistical machine translation toolkit with training on the JRC-Acquis corpora and translating on Estonian to English and English to Estonian language directions. In our work we model such linguistic phenomena as word lemmas and endings and splitting compound words into simpler parts. Also lemma information was used to introduce new factors to the corpora and to use this information for better word alignment or for alternative path back-off translation. From the results we find that even though these methods have shown previously and keep showing promise of improved translation, their success still largely depends on the corpora and language pairs used.

1. Introduction

This work studies an approach to statistical machine translation were we use unsupervised and supervised morphology information to try to improve the output quality of statistical machine translation for Estonian $\rightarrow$ English and English $\rightarrow$ Estonian translations. One language from the pair, the Estonian language is highly inflectional and has a rich morphology and thus offers good ground for morphology-related experiments.

Several works (see next section for related work) have suggested ways to use morphological analysis to improve the quality of machine translation when (at least) one of the languages is morphologically rich and therefore processing it is additionally challenging. Here we explore ways to achieve the same effect without using linguistic morphological analysis by replacing it with unsupervised morphology.

Although unsupervised segmentation has already been suggested for machine translation (see next section on related work), here we also use classification of the resulting segments to model such morphological phenomena as word lemmas, endings and compound words. This is utilized prior to training a translation model to either segment the corpus or generate additional factors for factored machine translation (Koehn and Hoang, 2007).

We use the unsupervised statistical toolkit Morfessor (Creutz and Lagus, 2005), which classifies the word segments into stems, suffixes and prefixes. We test several methods of handling the complex morphology of Estonian, comparing the usage of Morfessor and linguistic morphological analyzer (Kaalep and Vaino, 2001), which will be further referred to as T3.

2. Related Work

As mentioned in the introduction, a lot of work has been done on using morphological analysis to improve statistical machine translation. In many cases this is done similarly to the present work by segmenting the word forms into morphemes, or just the lemma and ending, or stem and morphological infliction identifier: (Nießen and Ney, 2004), (Badr et al., 2008), (Lee, 2004).

In an earlier work (Kirik and Fishel, 2008) we studied similar concepts but used a significantly smaller corpora consisting mostly of spoken language. We found that although only one language pair and a small training data set was used, the unsupervised morphology seemed to have potential at improving SMT output quality. The biggest improvement came for using the lemmas for creating word alignments, and also the segmenting of the input language gave moderate improvements on BLEU scores and a more favorable untranslated word ratio.

Kirik (2008) studied ways to improve factored alignment and reordering using Morfessor. No improvement for reordering was shown; alignment got the most improvement with all the suffix morphemes removed from the end of each word form.

Virpioja et al. (2007) used the output of Morfessor to segment the words into separate morphemes in both source and target language and treated the morphemes as separate input strings during translation. Although all resulting BLEU scores were below the baseline, there was improvement in terms of the number of partially translated sentences and untranslated words.

Other examples of unsupervised segmentation used for SMT are (Sereewattana, 2003) and (Bojar et al., 2008).

3. Corpus and Tools

Experiments described in this paper were carried out using several statistical machine translation tools. For training language and translation models, we used Giza++ (Och and Ney, 2003) and SRILM (Stolcke, 2002) toolkits, for translation we used the state-of-art Moses SMT decoder (Koehn et al., 2007).

To conduct the experiments we used the the Estonian-English part of the JRC-Acquis corpus (Steinberger et al., 2006) for training and testing the systems. JRC-Acquis is a parallel text corpus in 22 official EU languages which contains documents on political objectives, treaties, declarations, resolutions, agreements, EU legislation, etc.

We applied standard preprocessing to the corpus: all text was lower-cased, XML entities were removed, sentences longer than 100 words and pairs with word number ratio

higher than 9 were also removed. Furthermore, all sentences without at least 4 letters in sequence were removed. Finally the corpus was split into training, development and testing sets and the test and development sets were filtered to ensure statistical independence, resulting in 1 088 389 training, 2 500 development and 2 500 testing sentence pairs. The training set was used to train language and translation models, the development set – to optimize Moses configuration with minimum error-rate training and finally, the test set was used to evaluate the hypotheses.
Creation of morphological annotation was done by unsupervised morphological analyzer Morfessor and by Estonian morphological analyzer T3. The following two subsections give more information about the two aforementioned tools.

3.1. Morfessor

Morfessor (Creutz and Lagus, 2005) is an unsupervised morpheme segmentation and analyzer toolkit developed at the Helsinki University of Technology. Morfessor is trained on an unannotated corpora and as a result of the training accomplishes two tasks. Firstly, it uses the Morfessor Baseline algorithm to create a lexicon of morphemes for the input corpora, so that it is possible to form any words in the corpora by the concatenation of some morphemes from this lexicon. Secondly, based on the lexicon it creates the Morfessor Categories-ML model which categorized all the morphemes into three categories: prefixes, stems, and suffixes. The segmentation that the Morfessor produces is as "(prefix* stem suffix*)+".
For example, for the Estonian word "läbivaatamiseks" one might get (depending on the training corpora) the following segmentation and categorization: "läbi/PREF + vaata/STEM + mise/SUF + ks/SUF", where "PREF" designates a prefix, "STM" designates word stem and "SUF" - word suffix.

3.2. T3

T3 (Kaalep and Vaino, 2001) is a rule-based morphological analyzer for the Estonian language developed at the University of Tartu. T3 first uses it's built-in rule set to create a morphological analysis for the input text and then tries to remove all of the ambiguity in the word categorization by using a 2nd order Hidden Markov Model (HMM).
For example for the Estonian word "läbivaatamiseks" it will produce the following analysis: "läbi_vaatamine+ks //_S_ sg tr, //", which means that this is a singular compound noun consisting of words "läbi" and "vaatamine".

4. Experiments

In the subsections below we describe all of the experiments we conducted. Every section contains experiment description, scores recorded with different metrics, and a short explanation.
Results were evaluated using two metrics: the BLEU (Papieni et al., 2001) and the NIST (NIST, 2002). In addition we present the percent of unknown words present in the decoder output (UNK).
As we used morphological information only for the Estonian language then depending on the nature of the experiment there may be results for both translation directions or only for one of them.

Model	BLEU	NIST	UNK
`Baseline et-en`	43.51%	9.4133	1.42%
`Baseline en-et`	31.59%	7.9369	0.45%

Table 1: Primary baseline system (Estonian → English and English → Estonian) - no segmentation or factors generated.

4.1. Baseline systems

Our first task was to set up a baseline system which we could use as a base of comparison for the later experiments. In this paper we use two different kind of baseline systems. The primary baseline system is the usual non-factored and non-segmented translation with default reordering and alignment settings for the Moses SMT toolkit. This system was built for both translation directions, and the resulting scores for both translation directions are presented in the table 1.

Model	BLEU	NIST	UNK
`Morfessor`	38.35%	8.9535	0.38%
`T3`	40.89%	9.2443	0.36%

Table 2: Secondary baseline system (Estonian → English experiments) - default segmentation with Morfessor or T3 applied to Estonian.

Model	BLEU	NIST	UNK
`Morfessor`	17.39%	5.4456	0.42%
`T3`	17.36%	5.6269	0.36%

Table 3: Secondary baseline system (English → Estonian experiments) - default segmentation with Morfessor or T3 applied to Estonian.

The secondary baseline we are using was segmented using either Morfessor or T3 to create the segmentation. This was motivated by (Virpioja et al., 2007) and allows us to compare our results with basic segmentation experiments. These results are for Estonian → English and English → Estonian translation directions.
Secondary baseline results are presented in the tables 2 and 3 . From the results one can see that usage of T3 tends to yield better results on the Estonian → English direction but on the opposite direction the results are very similar.
When comparing the results of the two baseline systems we can see that the system with segmented Estonian language scored worse on BLEU and NIST than primary baseline system. But the out-of-vocabulary rates are lower for the segmented corpus (table 4) than for the primary baseline.
Also, we can see that the experiments on the English → Estonian translation direction of the secondary baseline have significantly lower scores from Estonian → English direction than when comparing the same experiments from the

primary baseline. We attribute this difference to the fact that for the SMT models generating the segmented output is a lot harder task than translation from segmented input.

Corpus	OOV
Baseline	0.963%
Full segmentation:	
Morfessor	0.047%
T3	0.145%
Compound splitting:	
Morfessor	0.873%
T3	0.340%
Ending splitting:	
Morfessor, one-suf	0.717%
Morfessor, many-suf	0.258%
Morfessor, all-suf	0.707%
T3	0.378%

Table 4: The out-of-vocabulary rates of the test sets, before and after segmentation

4.2. Lemma-based alignment

The first group of experiments tested using the word lemma for factored alignment, as suggested by (Koehn and Hoang, 2007). Thus translation was conducted on the original word forms, but word-alignment was generated using the lemmas generated with Morfessor or with T3. For Morfessor two different kinds of lemmas were constructed: in the one-suf experiment the word form was split into lemma and ending by treating the last suffix-morpheme as the ending; in the all-suf experiment by treating all the final suffix-morphemes as the ending. Experiments with T3 segmentation is marked with T3.

The results are presented in table 5. It appears that the tested lemmas have the potential to be effective for word-alignments, but the results depend largely on the corpora used. In current case only the T3 lemma-based experiment score is higher than the primary baseline, but this was shown to be statistically insignificant (with a p-value of 0.056, obtained with paired bootstrap resampling (Riezler and Maxwell, 2005)). But when comparing to the secondary baseline then all of the results have better scores.

4.3. Using lemmas for alternative path back-off translation

The second group of experiments uses lemmas for doing alternative path back-off translation. For doing this experiment the translation model was configured to use the word forms as the primary path for translation. But when dealing with previously unseen word forms it was able to use lemmas as a back-off translation option in hope that this is better than not translating the word at all. Again, like in the previous experiment, we have experiments with using the Morfessor (one-suf & all-suf) and with using the T3 (T3).

The results from this back-off translation experiment are only for the Estonian → English direction and presented in table 6. From the results one can see that this approach gives worse scores when compared to the primary baseline system, but better scores when compared to the secondary baseline system.

Model	BLEU	NIST	UNK
one-suf	42.70%	9.3088	1.42%
all-suf	43.41%	9.4031	1.43%
T3	43.39%	9.0177	1.35%

Table 6: Using lemmas for alternative path back-off translation

4.4. Segmenting into lemmas and endings

In the third group of the experiments we segmented the Estonian language into lemmas and endings and use this segmented corpora in hopes that in this way the SMT models are better able to learn to translate from and to the morphologically rich Estonian language. Splitting for the experiments that use Morfessor annotation are again done with different ways. In some experiments the word form was split into lemma and ending by treating either the last suffix-morpheme as the ending (referred to as one-suf) or all last suffix morphemes; in the latter case the suffix morphemes were either treated as a single ending (all-suf) or several separate endings (many-suf).

The results from this experiment are presented in table 7. As can be seen from the table the results are below the primary baseline, but almost all of them are better than the secondary baseline system. The larger score drop with the English → Estonian many-suf experiment can probably be attributed to the fact that in the case of Estonian language there are a lot less prefixes than stems or suffixes (in the training corpus: there are about 4.1 times less prefixes than suffixes and about 7.4 times less prefixes than stems) and so this segmentation is very similar to segmentation experiment of the secondary baseline and to the score drop there.

4.5. Segmenting into composite parts

In the final group of experiments we segmented the Estonian language into composite parts by separating each stem-morpheme with its adjacent prefixes and suffixes using Morfessor and T3.

The results from this experiment are presented in table 8. Again, the results are below the primary baseline but better than the secondary baseline scores.

5. Conclusions

We described a series of experiments on using unsupervised morphological analysis to improve the output quality of machine translation. Unsupervised morphology was used to model such linguistic phenomena as word lemmas and endings and splitting compound words into simpler parts. Also lemma information was used to introduce new factors to the corpora and to use this information for better word alignment or for alternative path back-off translation.

In this work we showed that even if linguistically motivated unsupervised segmentation showed great promise in our

	et-en			en-et		
	BLEU	NIST	UNK	BLEU	NIST	UNK
one-suf	43.46	9.4336	1.23%	28.32	7.5135	0.40%
all-suf	43.46	9.4157	1.16%	28.12	7.4947	0.38%
T3	43.83	9.1653	1.10%	29.33	8.9721	0.35%

Table 5: Using lemmas for alignment

	et-en			en-et		
	BLEU	NIST	UNK	BLEU	NIST	UNK
one-suf	42.37	9.2756	1.23%	28.38	7.6977	0.42%
many-suf	38.79	8.9331	0.83%	19.69	6.0367	0.44%
all-suf	42.29	9.2660	1.28%	28.58	7.7523	0.44%
T3	41.18	9.0778	0.79%	22.75	7.5955	0.38%

Table 7: Segmenting into lemmas and endings

earlier works, the performance of these methods seems to largely be dependent on the corpora and specific translation directions used. In this regard there is plenty of room for future work, whether it is applying this approach to other corpora and languages or introducing new kinds of splitting schemes.

6. Acknowledgments

This research was supported by the Estonian Science Foundation grant no 7503 and the European Social Funds Doctoral Studies and Internationalization Program DoRa.

7. References

Ibrahim Badr, Rabih Zbib, and James Glass. 2008. Segmentation for english-to-arabic statistical machine translation. In *Proceedings of ACL'08*, pages 153–156, Columbus, Ohio.

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. 2008. English-Hindi Translation in 21 Days. In *Proceedings of the ICON-2008 NLP Tools Contest*, Pune, India.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland.

Heiki-Jaan Kaalep and Tarmo Vaino. 2001. Complete morphological analysis in the linguists toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pages 9–16.

Harri Kirik and Mark Fishel. 2008. Modelling linguistic phenomena with unsupervised morphology for improving statistical machine translation. In *Proceedings of the SLTC'08 Workshop on Unsupervised Methods in NLP*, Stockholm, Sweden.

Harri Kirik. 2008. Juhendamata morfoloogia statistilises masintolkes (Unsupervised Morphology in Statistical Machine Translation). Bachelor thesis, University of Tartu, Dept. of Computer Science.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL'07*, pages 868–876, Prague, Czech Republic.

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 *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of NAACL-HLT'04*, pages 57–60, Boston, Massachusetts, USA.

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

NIST. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Technical report.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papieni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL'01*, pages 311–318, Philadelphia, PA, USA.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, USA.

Siriwan Sereewattana. 2003. *Unsupervised segmentation for statistical machine translation*. Ph.D. thesis, University of Edinburgh.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel

	et-en			en-et		
	BLEU	NIST	UNK	BLEU	NIST	UNK
Morfessor	43.33	9.3509	1.44%	31.19	7.95	0.45%
T3	42.73	9.3515	0.71%	27.18	7.9518	0.41%

Table 8: Segmenting into composite parts

corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, page 2142, Genoa, Italy.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, USA.

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