

Statistical Machine Translation Output Improvement using Morphological Templates

Abstract. We introduce a novel approach to improving statistical machine translation output. The main idea is to train a new MT system to translate into lemmas, morphological descriptions and parts-of-speech and to use this output as a template for modifying the common word-form output. The focus of this work is on evaluating different levels of detail for the templates. This is done empirically by training statistical MT systems on each combination of the levels in both the source and target language. We used parallel corpora containing legislation texts in Estonian and English for the task. The experiment results show that none of the selected levels of detail cause significantly higher accuracy than the baseline. However there is a slight improvement, correlating with the amount of information in the reduced word-forms. The results call for further evaluation of the approach with additional levels of detail, on other language pairs and corpora.

1 Introduction

Statistical approach to machine translation provides an opportunity for acquiring working translation systems with minimum human efforts, provided that there is large enough parallel corpora for the selected language pair. However, the approach works better for related languages (like French, Italian, Spanish) and not as well for languages with different structural characteristics (like differing morphology, word order, compound word formation, etc) – e.g. judging by the scores in [1].

There also exist several techniques that complement statistical MT so that additional language aspects are taken into consideration. For example, word order differences are most efficiently solved by reordering the input sentence words so that the resulting word order resembles the one of the target language more – then it is easier for the distortion model of the statistical MT system to solve the otherwise unfeasible task of reordering the words or phrases in translation [2]. However, many such techniques require syntactical parsing, phrase boundary detection and other higher level linguistic processing tools. The main problem for less studied languages is that the availability of the appropriate tools is limited.

This paper introduces a novel approach to statistical MT output improvement which only involves lower level language processing, such as morphological analysis and part-of-speech tagging. Such tools are available for a much wider selection of languages on one hand, and perform with accuracies close to 100% on the other [3–5].

The approach is introduced in section 2. Section 3 describes the experiments, conducted on Estonian-English statistical MT, aimed at evaluating the approach

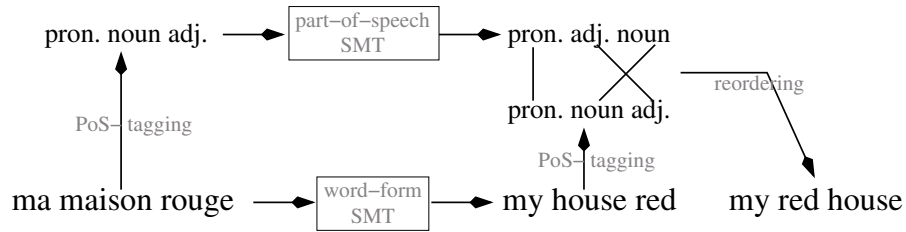


Fig. 1. Example of using morphological templates for reordering SMT output

and determining the optimal parameters for this language pair. Section 4 offers a discussion of the experiment results. The paper is concluded in section 5.

2 Part-of-speech and Morphological Templates

2.1 Main Idea

The main idea of the introduced approach is the following. The baseline system for the suggested improvement is a usual statistical MT system (or instance word- or phrase-based) with no pre- or post-processing added. Assumingly there is a parallel corpus that was used for training the system. Let us reduce the amount of information in each word-form of the corpus by replacing it with a more general category, like the word lemma, morphological description or part-of-speech. Using this new corpus a new statistical MT system can be trained, which would translate between the corresponding reduced word-forms.

Hypothetically such a system should show much better performance than the base system, since several typical problems with a common parallel corpus are removed this way. First of all, language inflection doesn't cause a bigger amount of potential word-forms – thus the probability that the available corpus allows to extrapolate the trained models to the whole application domain is higher. In other words, the sparse data effect is heavily reduced. Second, the variety in the present word-form sequences is much lower. This hopefully causes richer and more reliable statistics – since each word-form sequence is potentially represented by more samples in the corpus.

Assuming that the accuracy of the new system is higher, its output can be used as a template, according to which the output of the common statistical MT system can be post-processed. If the same reduction is applied to the common statistical MT output, it can be compared to the output of the new system, and altered according to it. For instance, the word order can be changed to resemble the template more (see figure 1 for an example). Alternatively, the whole n-best list of the common system outputs can be re-ranked according to similarity to the template.

However the main question is the optimal amount of information that can be removed from the corpus word-forms. First let us consider the input sentences.

On one hand, the more information there is in the input, the more competent the resulting models are in generating the output. On the other hand the variety that results from more information is higher, and this requires a much larger corpus to cover it. Therefore translation accuracy might benefit from reducing the informativeness of the source language sentences.

The same uncertainty applies to the output. Presumably, it is easier to generate more general word-forms – since less information has to be generated. However, more detailed output would serve as a more specific template for the common output to be compared with.

For instance, let us focus on word-form alignment with the commonly used expectation-maximization (EM) approach of the IBM-1 model [6]. In case it is applied to a corpus with detailed word-forms (starting from the ones without any reduction applied to them), many of them occur a small number of times. As a result, in the generated word alignment several rare word-forms are confused with other rare ones, occurring in the same one or two sentences with them. On the other hand, if the level of detail is too low, like with part-of-speech tags replacing the word-forms, the most frequent tags (e.g. nouns and verbs) will occur together in the majority of the corpus sentences – in that case the word alignment algorithm also cannot tell the difference between them.

To conclude, it is necessary to find the optimal levels of detail for both source and target languages, which would result in best translation accuracy between the corresponding reduced word-forms.

In this paper we focus on translation from Estonian into English. Experiments of applying statistical MT to Estonian are described in [7]. Automatic processing of Estonian has most of the problems referred to in the introduction – the language is highly inflective, has heterogeneous word order, a complex system of joining and splitting compound nouns and other features that make the task harder. In addition the amount of corpora and tools for Estonian is much more limited than, for instance, for English or German. In particular a syntactical parser isn't yet available for usage (although it is in development stage [8]).

In the experiments of this work different levels of detail in the source and target languages are evaluated empirically, by training statistical MT systems on the appropriately preprocessed corpus. We compare each detail level of the source and target languages to each other; thus, the resulting translation accuracies can be grouped into a matrix, where the top-left corner corresponds to the baseline system (trained on a corpus with no reduction applied to it).

2.2 Related Work

Lioma and Ounis [9] describe testing the usage of part-of-speech patterns for modifying the phrase table probabilities. The authors report no significant improvement in terms of the BLEU metric [10].

In addition to [2], mentioned in the introduction, methods of improving statistical MT output include replacing each word-form with its lemma and mor-

Table 1. Levels of detail for Estonian and English with corresponding number of distinct tags

Level of detail	Nr. of distinct tags	
	Estonian	English
Original word-form	166914	49105
Lemma	77843	44496
Part-of-speech + morphological description	230	67
Morphological description	74	-
Part-of-speech	17	56
Generalized part-of-speech	-	12

phological description to cope with morphological differences between languages [11], splitting the compound words [12], and many others.

3 Experiments

The experimental part of this work consists of applying phrase-based statistical MT to a parallel Estonian-English corpus with various amounts of information reduced in both languages.

The levels of detail for information reduction are summarized in table 1, along with the corresponding numbers of distinct tags for both languages. The utilized morphological analysis and part-of-speech tagging tools are described in [5] and [3] in case of English, and in [13] – in case of Estonian.

Both languages include the original word-form, word lemma and part-of-speech levels. Estonian levels also included the morphological description, which consist of case (nominative, genitive, partitive, etc; a total of 14) and number (singular/plural) of nouns, adjectives and numerals, verb tense (present/past), person (1st/2nd/3rd), etc.

Morphological description as a separate level of detail was left out of the English part, since it only included 5 tags (no modification, -ing, -ed, -en verbs and -s for plural nouns, adjectives, etc) and was considered not descriptive enough for the current task. However, both languages included an additional level where part-of-speech tags were combined with morphological descriptions into single tags. Finally, the English part included a level with 12 generalized part-of-speech tags. For instance, all noun tags (singular NN, plural NNS, singular personal NP, plural personal NPS) were grouped into one, similarly with verbs, adjectives, etc. The generalized part-of-speech tagging was added as a more comparable level to the Estonian parts-of-speech, as the original English tagset was much more specific.

For example, the part-of-speech+morphological level of Estonian included such tags as **sg-n-S** (singular noun in nominative), **vad-V** (verb, plural, 3rd person, present tense), etc. The level of English lemmas included sentences like

‘this directive be address to the member state’ (the original sentence being ‘this directive is addressed to the member states’).

3.1 Experiment Setup

We used Moses (available online¹) as the phrase-based statistical MT system software. This included the SRI language modeling toolkit [14], GIZA++ word alignment software [15], the Moses decoder and the included script for inducing phrase-based translation models from word-based ones. Factored translation models were not used – i.e. transfer between the 0-th factors and no generation steps were included.

The used parallel corpus was created at the University of Tartu and is available online along with descriptive information². It consists of texts from the Estonian and EU legislation. The corpus contains 7.8 million words in Estonian and 5.0 million – in English. It is aligned on sentence level, the total number of sentence pairs is 435 692.

The corpus was randomly split into two parts, the smallest part (0.3% of the corpus, 1308 sentences) was used for evaluation. The remaining part was filtered, so that sentence pairs with one of the sentences longer than 50 words, and the ones with the ratio of the word numbers exceeding 9 (including pairs with one of the sentences empty) were left out.

We used the BLEU metric for evaluating and comparing the trained systems. Although it was shown not to correlate with human judgement in all cases [16], in our case the amount of testing data from all the experiments makes human evaluation unfeasible. In addition, the output in this case isn’t always composed of words, but also of lemmas and parts of speech, which makes human evaluation even harder. However, in order to provide an opportunity of re-evaluating the results with a different metric, we put all of the results online³.

Since the number of required model training sessions is relatively high (25), it was decided to run them in parallel. This was accomplished using GRID technologies, with the resources of the BalticGrid project⁴.

3.2 Results

The results of the experiments are presented in table 2; five highest scores are printed in bold. The best result is achieved by translating from Estonian word-forms into reduced English parts-of-speech; however, compared to the the baseline (translation between word-forms, top-left corner of the table) the gain is hardly significant.

Generally, no other setup produced a significantly higher score than the baseline. Reducing the informativeness of the target language causes improvement of the scores; however, with word-forms as input the improvement is only slight.

¹ <http://www.statmt.org/moses>

² <http://www.cl.ut.ee/korpused/paralleel>

³ <http://ats.cs.ut.ee/smt/posmorph>

⁴ <http://balticgrid.org>

Table 2. The BLEU scores for each level of detail in the source and target language

Source Language Lev. of Detail	Target Language Level of Detail				
	Word Form	Lemma	PoS + Morphology	PoS	Reduced PoS
Word Form	44.08	43.96	45.46	45.08	46.32
Lemma	43.54	44.46	43.35	42.75	45.15
PoS+Morphology	11.71	11.51	29.44	29.30	25.91
Morphology	7.70	7.35	16.05	18.33	15.87
PoS	7.76	8.05	20.30	19.85	20.51

4 Discussion

The experiment results show no improvement from reducing the word-forms of the source language, despite the complications associated with the specifics of the Estonian language, mentioned in section 2. One of the possible explanations is that the sparse data effect is not as strong with the used corpus (although by size the corpus is smaller than, e.g., Europarl [1]).

On the other hand, reducing information in the target language causes moderate improvement in the output score (i.e. comparing the accuracy of translation into English word-forms and into reduced parts-of-speech). This supports the hypothesis that generating a language with some information reduced is easier. But since the improvement is only slight (2.24 BLEU points over the baseline) this conclusion requires further testing.

Another interesting fact is that morphological description as the only input produces lower results for all levels of detail in the target language than parts-of-speech. This shows the parts-of-speech to be a relevant piece of information, despite the fact that the number of distinct tags is lower than in case of Estonian morphology.

One of the probable reasons for the lack of significant improvement with reduced word-forms is the difference in word order in the Estonian and English parts of the used corpus; previous experiments with Estonian-English statistical MT show that one of the main problems for this language pair is the failure of the distortion model to adapt to the corpus [7]. At the same time, in the current work, word order information was preserved in all levels of detail, unlike lexical and morphological information. Therefore the task of learning the word order transfer almost as difficult as with the baseline.

To sum up, according to the experiment results the currently tested levels of detail for information reduction in the target language cause no significant improvement on the used corpus. This leaves several directions for development in future work:

- The approach can be tested on other language pairs. In particular, the Finnish part of the Europarl corpus has to be considered. Finnish is very

similar to Estonian from the point of view of grammar, and the challenges might be comparable.

- Also, other domains and corpora have to be explored, the Europarl corpus being the first candidate.
- In addition, the approach can be tested with different levels of detail. For instance, looking at table 1, there is a large gap between lemmas and parts-of-speech combined with morphology in terms of number of distinct forms. Possibly, reducing only part of lexical information by replacing each word with a more general word class (with different levels of generalization) can cause improvement.

5 Conclusions

This paper introduced the idea of using morphological templates for improving statistical MT output. The templates are to be obtained by preprocessing parallel corpora via replacing each word-form with a more general category (like its part-of-speech) and training a new SMT system with it. The approach is potentially beneficial for languages with limited availability of corpora and language processing tools. In particular, there is no requirement for syntactic analysis or language processing of higher levels.

The focus of this work is on evaluating different categories for replacing the source and target language words. This has been done empirically for Estonian-English translation. Phrase-based statistical MT systems were trained on each combination of levels of detail in the source and target language. The output was scored using the BLEU metric.

The experiment results show that the selected levels of detail do not cause significant improvement with the used corpus. We suggest that the main reason is that in Estonian-English translation word order difference is much more influential than morphology difference, which was the target of the current work. The introduced approach has to be tested on different language pairs, corpora, domains and with additional levels of detail for information reduction.

Acknowledgements

We would like to thank Joakim Nivre, Markus Saers (Uppsala University) and Mare Koit (University of Tartu) for their helpful advice, and Ilya Livenson (BalticGrid, University of Tartu) for technical support.

References

1. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit X, Phuket, Thailand (2005)
2. Nießen, S., Ney, H.: Morpho-syntactic analysis for reordering in statistical machine translation. In: Proceedings of MT Summit VIII, Santiago de Compostela, Galicia, Spain (2001) 1081–1085

3. Minnen, G., Carroll, J., Pearce, D.: Applied morphological processing of English. *Natural Language Engineering* **7**(3) (2001) 207–223
4. Van den Bosch, A., Daelemans, W.: Memory-based morphological analysis. In: *Proceedings of the 37th Annual Meeting of the ACL*, College Park, MD, USA (1999) 285–292
5. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK (1994)
6. Brown, P.F., Pietra, S.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19**(2) (1994) 263–311
7. Fishel, M., Muischnek, K., Kaalep, H.J.: Estonian-English statistical machine translation: the first results (in print). In: *Proceedings of NODALIDA 2007*, Tartu, Estonia (2007)
8. Muischnek, K., Müürisep, K., Puolakainen, T.: Parsing of Estonian: morphological disambiguation and determination of syntactic functions. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V* (2001) 411–417
9. Lioma, C.A., Ounis, I.: Deploying part-of-speech patterns to enhance statistical phrase-based machine translation resources. In: *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. (2005) 163–166
10. Papieni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, PA, USA (2001) 311–318
11. Bojar, O., Matusov, E., Ney, H.: Czech-English phrase-based machine translation. In: *Proceedings of the 5th International Conference on NLP, FinTAL 2006*, Turku, Finland (2006) 214–224
12. Koehn, P., Knight, K.: Empirical methods for compound splitting. In: *Proceedings of the 10th Conference European Chapter of the Association for Computational Linguistics EACL*, Budapest, Hungary (2003) 187–193
13. Kaalep, H.J., Vaino, T.: Complete morphological analysis in the linguist’s toolbox. In: *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, Tartu, Estonia (2001) 9–16
14. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: *Proceedings of the International Conference on Spoken Language Processing. Volume 2.*, Denver, Colorado, USA (2002) 901–904
15. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. *Computational Linguistics* **29**(1) (2003) 19–51
16. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of BLEU in machine translation research. In: *Proceedings of the 11th Conference European Chapter of the Association for Computational Linguistics EACL*, Trento, Italy (2006) 249–256