The Impact of Morphological Errors in Phrase-based Statistical Machine Translation from English and German into Swedish

Oscar Täckström
Swedish Institute of Computer Science
SE-16429, Kista, Sweden
oscar@sics.se

Abstract

We have investigated the potential for improvement in target language morphology when translating into Swedish from English and German, by measuring the errors made by a state of the art phrase-based statistical machine translation system. Our results show that there is indeed a performance gap to be filled by better modelling of inflectional morphology and compounding; and that the gap is not filled by simply feeding the translation system with more training data.

Keywords: Machine Translation, Swedish Morphology

1. Introduction

The state of the art in statistical machine translation has evolved rapidly since the advent of the noisy-channel inspired models in the early nineties (Brown et al., 1993). A leap was taken when the field departed from using only word(s)-to-word (M:1) alignments and instead began to use phrase-to-phrase (M:N) alignments (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004; Chiang, 2005). This endowed the models with capacity for modelling, for example, idiomatic expressions and local agreement, which are difficult to model with only word alignments.

Recently, *factored* translation models that generalise the original noisy-channel models and the phrase-based models, were proposed by Koehn and Hoang (2007) and Koehn et al. (2007). In these models, different factors can be used to model different aspects of the translation process, such as word order, inflectional morphology, agreement, and compounding.

However, the increased capacity of these models come at a rather high price with respect to both computational complexity and systems complexity. Vilar et al. (2006) argues that before one embarks on designing and implementing a more complex model, one should evaluate the potential improvement on translation quality that one could hope to obtain from this model.

In this vein we have investigated the potential for improvement of target language morphology when translating from English and German into Swedish, when taking into account the option of adding more training data to the standard phrase-based model.

2. Statistical machine translation

Machine translation is a *structured prediction* task in which given a source text f_1^J, consisting of a sequence of tokens $(f_j)_1^J$, the aim is to predict a target text e_1^I, such that a score function $S(e_1^I; f_1^J)$ is maximised. Ideally the score function should emulate the judgements made by humans, but in practice some measure that is hoped to correlate well with at least some aspects of human judgements is used. Commonly, *fluency* and *adequacy*, discussed in more depth later, are taken into account. Usually translation is only modelled on the sentence level and other pragmatic aspects, such as task-oriented utility or post-editing costs are typically left out, together with stylistic aspects.

Statistical machine translation is different from rule based translation in that probabilistic translation "rules" are induced from training data, rather than being devised by human experts. The training data usually consists in a combination of bilingual and monolingual text, the latter typically being much more abundant.

2.1. Noisy-channel models

Structured prediction is a general machine learning problem studied by the machine learning research community for quite some time; see, e.g., Bakır et al. (2007) for a comprehensive overview. The models used in statistical machine translation are still very much inspired by the early *noisy-channel* approach pioneered by Brown et al. (1993). In this framework, the objective is to find the most probable translation of a source sentence, given a generative probabilistic model of the translation process. After application of Bayes' theorem and dropping the a priori probability of a source language sentence, the following score function is derived:

$$S(e_1^I; f_1^J) = \mathrm{P}(e_1^I)\,\mathrm{P}(f_1^J|e_1^I),$$

It is noticeable that there is no reference to either fluency or adequacy in this score function, the model makes no commitments to what humans would consider to be a good translation, just an intuition that translation is a non-deterministic process in which sequences of tokens in one language is mapped to sequences of tokens in another language. It is up to the modeller to define a proper probabilistic model that captures the aspects of the translation process, that are assumed to be important.

In the original IBM models, based on *word alignments*, the language model $\mathrm{P}(e_1^I)$ is a standard n-gram language model, while the translation model is factored in terms of word-to-word alignments:

$$\mathrm{P}(f_1^J|e_1^I) = \sum_{a_1^J} \mathrm{P}(f_1^J, a_1^J|e_1^I),$$

where a_1^J is the alignment variable, with $a_j = i$ indicating that word f_j is aligned with word e_i. To account for words

with no translation in the target language, the *null* word is included as e_0. The alignment variable, a_1^J, is constrained to only allow M:1 alignments.

2.2. Phrase-based models

The constraint of only allowing M:1 translations was relieved with the advent of *phrase-based* alignment models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), in which M:N alignments are allowed. This affords the models with capacity to model constructions that are abundant in real texts, but hard to model with only words-to-word alignments, such as local agreement and idioms.

Phrase-based systems typically model the translation process according to the following probabilistic model: first the target sentence e_1^I is segmented into phrases $\bar{e}_1^L$, then these phrases are reordered according to a reordering model $d(i, j)$, which commonly penalises long range movements; finally each phrase $\bar{e}_l$ is translated in isolation to a source language phrase $\bar{f}_m$ according to a translation model $\mathrm{P}(\bar{f}_1^M|\bar{e}_1^L)$.

The phrase alignments are typically created by performing the M:1 alignment in both source-target and target-source directions. This gives one set of M:1 alignments and one set of 1:N alignments, which are combined into one M:N alignment, using a heuristic procedure.

2.3. Factored models

Recently a more principled move beyond the simple noisy-channel formulation was taken with *factored* translation models (Koehn and Hoang, 2007; Koehn et al., 2007). In these models, the score function of the phrase-based models is generalised to a log-linear formulation:

$$S(e_1^I; f_1^J, \lambda_1^K) = \frac{1}{Z} \exp \sum_k \lambda_k h_k(e_1^I; f_1^J),$$

where h_k is a feature function, or *factor*, encoding some aspect of the translation, $\lambda_k \in [0, 1]$ is a feature function weight, with $\sum_k \lambda_k = 1$ and Z is a normalisation factor.

Factored models allow the incorporation of more knowledge sources into the translation process. With a modification to the decoder, one can divide the translation process into sets of *translation* and *generation* steps. In this way, one can devise models in which lemmas and parts-of-speech are translated separately, with target morphology generated by a combination of translated lemmas and translated parts-of-speech. The log-linear score function, furthermore, allows one to tune the hyper-parameters of the model, the λ_k:s, to optimise performance on a held out development set, by means of minimum error rate training, MERT (Och, 2003). This is the only step in which one makes use of the score that one really wants to optimise, the factored score function has not any direct relation to this measure.

Factored phrase-based models, combined with MERT, make it possible – in theory – to incorporate a large number of knowledge sources in the process; balancing the influence of these sources, with respect to a specific target domain, in a (locally) optimal way.

3. Morphological issues and factored models

Aspects of Swedish morphology pertaining to machine translation has, to our knowledge, not yet been studied in depth. Such a study is beyond the scope of this work, but the results of our empirical study clearly indicate that handling these issues will be crucial in order to improve on the state of the art. In this evaluation, we focus on inflectional, derivational, and compositional aspects, all in which Swedish differs from English and German. These properties have been studied by Hedlund et al. (2001), in the context of information retrieval. That study showed that proper handling of inflectional morphology and compound words improve results for retrieval in Swedish. It seems reasonable to assume that these results should carry over when translating to and from Swedish.

Differences in inflectional and derivational morphology, may influence phrase-based statistical machine translation primarily in two ways. First, an increased number of word forms may lead to data sparseness problems. This is especially the case for productive morphological processes. Second, the phrase-based nature of these systems makes it difficult to enforce longer range agreement requirements, when these differ between the source and target languages.

By factoring the translation process, the aim is to solve these problems. By using generalised factors, such as lemmas and parts-of-speech, the problem of data sparseness can be conquered. With surface token phrase alignment, using phrases consisting of more than about three tokens seems to be futile, due to data sparseness. The same is true for language models, for which tri- to 5-grams are usually an upper limit. On the other hand, if one uses language models based on parts-of-speech, reliable statistics can be collected for models of significantly higher order. Compound words can be handled by splitting them on the source language side and/or generate them on the target language side, in order to conquer the problem of data sparseness. This has been shown to improve translation quality when translating between German and English (Koehn and Knight, 2003; Stymne et al., 2008). By using a translate and generate approach, the problem of underspecification can be alleviated. For different language pairs and translation directions, some aspects, e.g. morphological, of the target language might be underspecified with respect to the source language. In general, translation is an asymmetric process in which some information present in the source language can be thrown away, while some information is lacking in the source language to sufficiently specify aspects of the target language. This is in a way related to the problem of data sparseness, however depending on the model, even adding infinite amounts of training data would not help the situation. Using factored models, the source language can be enriched by additional information, which can remedy the problem of underspecification.

Even though factored models in theory allows arbitrarily complex translation models, in practice this flexibility comes at the price of computational complexity, during both model estimation and decoding. Minimum error rate train-

ing is a huge bottleneck in this respect; in practice when using more than five or six factors, the optimisation requires a prohibitively long time.

Another problem is that the interaction between different factors, and the decoding scheme, is not easily predicted. Thus, in practice the development of new models requires a very costly trial and error procedure. This makes it even more important to focus on those aspects where the potential for improvement is largest. Thus, before devising more complex models and spending time evaluating them and trying to analyse whether the often small improvements in evaluation scores are significant, we should evaluate how much the translation could be improved on different levels.

4. Evaluating translation quality

How to evaluate translation quality still remains an open field of research. First, there often exist several possible good or acceptable translations of a given source text; it is hard even for humans to judge which of a set of translations is the best one. Second, although ideally one should use human evaluators to assess translation quality, unfortunately this is often much too costly. Instead one has to rely on machine computable evaluation measures. There are a range of possible evaluation measures available. For example, precision based measures, such as BLEU (Papineni et al., 2001) and NIST (Doddington, 2002) or precision / recall focused measures such as *Meteor* (Banerjee and Lavie, 2005).

The machine computable measures are defined with respect to one or more reference translations. Usually, in practice only one reference translation is available, which given the multiplicity of good potential translations is arguably a large drawback. With the *Meteor* measure, the aim is to overcome this, by providing the option of measuring the overlap of lemmatised words or synonym-sets, between a source text and a reference text. This is similar to the morphological normalisation employed in this paper.

4.1. The BLEU measure

The most commonly used machine computable evaluation measure is the BLEU (short for *bilingual evaluation understudy*) measure, which was developed to provide a quick and cheap way of assessing translation quality (Papineni et al., 2001). BLEU is a word n-gram precision oriented measure. Fluency is measured by counting the precision of n-grams in the translation, with respect to a set of reference translations. Adequacy is measured by unigram precision, i.e., by calculating how much the "concepts" used in the translation matches those used in the reference translations. The n-gram counts are modified, so that each occurrence of a particular n-gram is matched only once. In addition, a brevity penalty is introduced to ensure that the translation does not differ too much from the reference translation(s) in length. A cumulative score over different values of n is computed by the geometric mean of the modified n-gram precision. This makes the metric more sensitive to longer n-grams, which usually have much lower absolute precision.

It has been shown that n-gram based measures do not correlate well with human judgements of translation quality (Callison-Burch et al., 2006). This critique is well founded, however, we believe that BLEU is a useful measure for our purposes precisely because of its coarse nature, since this allows us to assess the potential improvement of a single system with respect to an isolated aspect of translation.

4.2. Evaluating the potential for improvement

As argued above, and also by Vilar et al. (2006) and Popovic et al. (2006), it is important to find the most prominent problems in translation, in order to focus research efforts in the right direction. The predominant use of "holistic" measures, as those discussed above, unfortunately does not really support this development, since it is difficult to interpret the measures in terms of a more fine grained error typology. This fact is further aggravated, given that each language pair and each translation direction give rise to different translation problems. For the field to take the next leap, we would argue that it is of fundamental importance to find the real pains to cure, before one starts developing more complex models aimed at curing problems, whose magnitude is not evaluated.

In this work we focus on evaluating the magnitude of errors related to target language morphology. We do this by simply normalising morphology in both the target translation and the reference translation, before computing the standard BLEU measure. The resulting score provides an upper bound on the improvement that could be achieved by perfect handling of target language morphology. Figure 1 shows an example translation with morphological errors and the corresponding normalisation. The fact that the normalised translation correlates better with the normalised gold translation, compared to the original translations, indicates that there were morphological errors in the original translation.

This is, of course, a hack, rather than a principled solution to the problem and our results are thus only a very coarse-grained measure of these errors. Moreover, we focus on aspects on the target side only; no clues are given on how to actually model these phenomena on a bilingual level. Furthermore, it is not clear how much of the performance gap indicated by this kind of a priori error analysis could be filled in practice. Perhaps even more importantly, changes to the model aimed at rectifying one aspect of the translation could affect other aspects of the translation. These shortcomings suggest the potential for using more fine-grained automatic measurements of translation errors, as a form of regression test, analogous to those commonly used in software engineering, for validating system performance after each development cycle.

Automatic measures, coping with a wide range of translation aspects, such as the aspects suggested by Vilar et al. (2006) for manual evaluations, is a prerequisite for a more principled solution to both the problem of finding the most pressing pains to cure, and for a regression test system as that just sketched. We view this as an important challenge for future research.

(i) "med tanke på **den snäva marginaler** för *lånegarantin reserv*"
(ii) "med tanke på den snäv marginal för låne garanti reserv"
(iii) "med tanke på lånegarantireservens snäva marginaler"
(iv) "med tanke på låne garanti reserv snäv marginal"

Figure 1: Translation of the English clause "bearing in mind the narrow margin of the loan guarantee reserve" with errors in inflectional morphology (bold) and compound words (italic). (i) the systems translation; (ii) normalised translation; (iii) gold translation; (iv) normalised gold translation.

5. Experiments

In order to examine the potential for improvement of morphology in translation from English and German into Swedish, we conducted a set of simple experiments aimed at answering two questions:

1. How much would the translation quality be improved, given that the morphology of the target language was handled perfectly?

2. How much could the morphological aspects of the target language be improved by simply adding more training data?

The aim of the first question, is to find out whether morphological errors are at all a significant source of translation errors, when translating from German and English into Swedish. Thus we are in effect trying to estimate whether a factored model, aimed at modelling morphological aspects of the target language, has any potential for improving the quality of translations in these cases.

With the second question, we aim to investigate whether it is the case that the cure for morphological errors is simply to add more training data. For some language pairs a positive answer to this question would entail that developing factors for modelling target language morphology is not worthwhile. For other language pairs, for which more training data is not available, modelling target language morphology could still be of interest.

5.1. Setup

For all experiments, we used the *Moses* SMT system (www.statmt.org/moses), the SRILM n-gram language modeling toolkit (www.speech.sri.com/projects/srilm), and the GIZA++ and *mkcls* word alignment and word class induction tools (code.google.com/p/giza-pp). We used the baseline settings for these tools as provided for the WMT2008 shared task at ACL (www.statmt.org/wmt08).

We used version 3 of the Europarl corpus (www.statmt.org/europarl) to create all training and test sets. For each language pair, English-Swedish and German-Swedish, we created five different sentence-aligned training sets by selecting 10%, 25%, 50%, 75% and 100%, respectively, of the available sentence-aligned data, excluding the fourth quarter of year 2000, which was put aside for use as a test set. The first 2000 sentences of the sentence-aligned texts from the fourth quarter of year 2000 were put aside for testing. All the available Swedish monolingual data (excluding quarter 4 of year 2000) were used in building the language model. This is a reasonable choice, since monolingual data is typically abundant compared to sentence-aligned bilingual data.

Size	Base	Lem	Dec	Dec+Lem
10	24.41	27.43 (*12.4*)	25.67 (*5.2*)	29.01 (*18.8*)
25	25.27	28.17 (*11.5*)	26.54 (*5.0*)	29.70 (*17.5*)
50	25.91	28.83 (*11.3*)	27.23 (*5.1*)	30.41 (*17.4*)
75	26.21	29.01 (*10.7*)	27.52 (*5.0*)	30.58 (*16.7*)
100	26.17	29.01 (*10.9*)	27.48 (*5.0*)	30.58 (*16.9*)

Table 1: Results when translating from English to Swedish, with respect to training set size, as measured with the BLEU measure. *Base*, *Lem*, and *Dec* denote baseline, lemmatised, and de-compounded translation results, respectively. Parenthesised numbers indicate percentage improvement w.r.t. baseline scores.

Size	Base	Lem	Dec	Dec+Lem
10	20.55	22.79 (*10.9*)	21.31 (*3.7*)	23.74 (*15.5*)
25	21.51	23.85 (*10.9*)	22.42 (*4.2*)	24.95 (*16.0*)
50	21.86	24.19 (*10.7*)	22.79 (*4.3*)	25.32 (*15.8*)
75	22.38	24.76 (*10.6*)	23.42 (*4.6*)	26.00 (*16.2*)
100	22.59	24.94 (*10.4*)	23.69 (*4.9*)	26.22 (*16.1*)

Table 2: Results when translating from German to Swedish (see the caption of table 1 for explanation).

After training the baseline systems, one for each of the ten datasets, we applied each trained model to its respective test set. We then created three different morphologically normalised variations of the translated test set and the reference set by lemmatising, splitting compounds, and splitting compounds as well as lemmatising the translations, using the lemmatiser and compound splitter from the *Granska* grammar checking and parts-of-speech tagging tool (www.csc.kth.se/tcs/humanlang/tools.html). Finally we computed BLEU scores for each of the, in total, forty different variants of the translated test sets. By comparing the scores for the normalised reference and translation texts, we achieve an upper bound on the improvement in translation quality, that could be achieved by perfect modelling of inflectional and derivational target language morphology and perfect handling of compounds in the target language.

5.2. Results

Tables 1 and 2 show the translation quality, with respect to language pairs, training set size, and morphological normalisation, as measured with the BLEU evaluation measure. As can be seen, translations from English consistently score about 4 absolute points BLEU higher than translations from German. Not surprisingly, results improve when more training data is used, however, the initial steep improvement flattens rather quickly.

Looking at the effect of lemmatisation, we see that the potential improvement of inflectional and derivational morphology diminishes somewhat when more training data is added. However, especially when translating from German, this effect is rather small; even when using all the available training data, results could be boosted by over ten per cent, if the Swedish morphology was handled perfectly. While

more training data thus enables the model to make better word choices, it does not bring about much improvement with respect to morphological aspects. This clearly shows the need for models that go beyond the standard phrase base model, when translating from these languages into Swedish.

Turning to the impact of compound words, there is also some room for improvement, although less so. Interestingly, the relation between training set size and errors caused by compounds is different between translation from English and German. In the former case, compound split scores are consistently about five per cent higher than when compounds are not split, regardless of training set size. In the latter case, the potential of improving results by handling compounds correctly increases steadily with the addition of more training data.

Looking at the rightmost column of tables 1 and 2, we see that when all available training data is used, if inflectional and derivational morphology, as well as compound words, were handled perfectly, translation quality would increase by over sixteen per cent, compared to the baseline model.

Before concluding, although these results speak in favour of more complex models, one should keep in mind that these figures are upper bounds on the potential improvement. In other words, we should expect less substantial improvements in practice. Furthermore, it might be the case that modeling these aspects will change the model in such a way that some other aspect of the translation is affected in a negative way.

6. Conclusions

We have investigated the potential for improving target language morphology, when translating from English to Swedish and from German to Swedish. The evaluation was performed by estimating the upper bound on the potential improvement, rather than starting experimenting with more elaborate models. This was done by applying the trained system to test data, then lemmatising and compound splitting both a reference translation and the system's translation, before measuring the quality of both the original version and the lemmatised and compound split versions.

Results show that there is indeed room for improvement of target language morphology when translating into Swedish and that this room remains open in the face of more training data. Thus a factored translation model aimed at improving target language morphology could prove fruitful.

References

Bakır, G., T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. Vishwanathan (Eds.) (2007). *Predicting Structured Data*. MIT Press.

Banerjee, S. and A. Lavie (2005). METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72.

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

Callison-Burch, C., M. Osborne, and P. Koehn (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*, pp. 249–256.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*, pp. 263–270.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT*, pp. 138–145.

Hedlund, T., A. Pirkola, and K. Järvelin (2001). Aspects of swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing & Management 37*(1), 147–161.

Koehn, P. and H. Hoang (2007). Factored translation models. In *Proceedings of EMNLP*, pp. 868–876.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens (2007). Moses: Open source toolkit for statistical machine translation. In *Proceeings of ACL*, Volume 45, pp. 2.

Koehn, P. and K. Knight (2003). Empirical methods for compound splitting. In *Proceedings of EACL*, pp. 187–193.

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

Marcu, D. and W. Wong (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceeings of EMNLP*, pp. 133–139.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings ACL*, pp. 160—167.

Och, F. J. and H. Ney (2004). The alignment template approach to statistical machine translation. *Computational Linguistics 30*(4), 417–449.

Papineni, K., S. Roukos, T. Ward, and W. J. Zhu (2001). BLEU: a method for automatic evaluation of MT. *Research Report, Computer Science RC22176 (W0109-022), IBM Research Division, TJ Watson Research Center 17*.

Popovic, M., H. Ney, A. D. Gispert, J. B. Mariño, D. Gupta, M. Federico, P. Lambert, and R. Banchs (2006). Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proceedings of the NAACL Workshop on SMT*, pp. 1–6.

Stymne, S., M. Holmqvist, and L. Ahrenberg (2008). Effects of morphological analysis in translation between german and english. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 135–138.

Vilar, D., J. Xu, L. F. D'Haro, and H. Ney (2006). Error analysis of statistical machine translation output. In *Proceedings of LREC*, pp. 697–702.