Pre- and Postprocessing for Statistical Machine Translation into Germanic Languages

Sara Stymne
Department of Computer and Information Science
Linköping University, Linköping, Sweden
sara.stymne@liu.se

Abstract

In this thesis proposal I present my thesis work, about pre- and postprocessing for statistical machine translation, mainly into Germanic languages. I focus my work on four areas: compounding, definite noun phrases, reordering, and error correction. Initial results are positive within all four areas, and there are promising possibilities for extending these approaches. In addition I also focus on methods for performing thorough error analysis of machine translation output, which can both motivate and evaluate the studies performed.

1 Introduction

Statistical machine translation (SMT) is based on training statistical models from large corpora of human translations. It has the advantage that it is very fast to train, if there are available corpora, compared to rule-based systems, and SMT systems are often relatively good at lexical disambiguation. A large drawback of SMT systems is that they use no or little grammatical knowledge, relying mainly on a target language model for producing correct target language texts, often resulting in ungrammatical output. Thus, methods to include some, possibly shallow, linguistic knowledge seem reasonable.

The main focus for SMT to date has been on translation into English, for which the models work relatively well, especially for source languages that are structurally similar to English. There has been less research on translation out of English, or between other language pairs. Methods that are useful for translation into English have problems in many cases, for instance for translation into morphologically rich languages. Word order differences and morphological complexity of a language have been shown to be explanatory variables for the performance of phrase-based SMT systems (Birch et al., 2008). German and the Scandinavian languages are a good sample of languages, I believe, since they are both more morphologically complex than English to a varying degree, and the word order differ to some extent, with mostly local differences between English and Scandinavian, and also long distance differences with German, especially for verbs.

Some problems with SMT into German and Swedish are exemplified in Table 1. In the German example, the translation of the verb *welcome* is missing in the SMT output. Missing and misplaced verbs are common error types, since the German verb should appear last in the sentence in this context, as in the reference, *begrüßen*. There is also an idiomatic compound, *redebeitrag* (*speech+contribution; intervention*) in the reference, which is produced as the single word *beitrag* in the SMT output. In the Swedish example, there are problems with a definite NP, which has the wrong gender of the definite article, *den* instead of *det*, and is missing a definite suffix on the noun *synsätt(et)* (*(the) approach*).

In this proposal I outline my thesis work which aims to improve statistical machine translation, particularly into Germanic languages, by using pre- and postprocessing on one or both language sides, with an additional focus on error analysis. In section 2 I present a thesis overview, and in section 3 I briefly overview MT evaluation techniques, and discuss my work on MT error analysis. In section 4 I describe my work on pre- and postprocessing, which is focused on compounding, definite noun phrases, word order, and error correction.

Proceedings of the ACL-HLT 2011 Student Session, pages 12–17,
Portland, OR, USA 19-24 June 2011.

En source	I too would like to welcome Mr Prodi's forceful and meaningful intervention.
De SMT	Ich möchte auch herrn Prodis energisch und sinnvollen Beitrag.
De reference	Ich möchte meinerseits auch den klaren und substanziellen Redebeitrag von Präsident Prodi begrüßen.
En source	So much for the scientific approach.
Se SMT	Så mycket för den vetenskapliga synsätt.
Se reference	Så mycket för den vetenskapliga infallsvinkeln.

Table 1: Examples of problematic PBSMT output

2 Thesis Overview

My main research focus is how pre- and postprocessing can be used to improve statistical MT, with a focus on translation into Germanic languages. The idea behind preprocessing is to change the training corpus on the source side and/or on the target side in order to make them more similar, which makes the SMT task easier, since the standard SMT models work better for more similar languages. Postprocessing is needed after the translation when the target language has been preprocessed, in order to restore it to the normal target language. Postprocessing can also be used on standard MT output, in order to correct some of the errors from the MT system. I focus my work about pre- and postprocessing on four areas: compounding, definite noun phrases, word order, and error correction. In addition I am making an effort into error analysis, to identify and classify errors in the MT output, both in order to focus my research effort, and to evaluate and compare systems.

My work is based on the phrase-based approach to statistical machine translation (PBSMT, Koehn et al. (2003)). I further use the framework of factored machine translation, where each word is represented as a vector of factors, such as surface word, lemma and part-of-speech, rather than only as surface words (Koehn and Hoang, 2007). I mostly utilize factors to translate into both words and (morphological) part-of-speech, and can then use an additional sequence model based on part-of-speech, which potentially can improve word order and agreement. I take advantage of available tools, such as the Moses toolkit (Koehn et al., 2007) for factored phrase-based translation.

I have chosen to focus on PBSMT, which is a very successful MT approach, and have received much research focus. Other SMT approaches, such as hierarchical and syntactical SMT (e.g. Chiang (2007), Zhang et al. (2007a)) can potentially overcome some language differences that are problematic for PBSMT, such as long-distance word order differences. Many of these models have had good results, but they have the drawback of being more complex than PBSMT, and some methods do not scale well to large corpora. While these models at least in principle address some of the drawbacks of the flat structure in PBSMT, Wang et al. (2010) showed that a syntactic SMT system can still gain from preprocessing such as parse-tree modification.

3 Evaluation and Error Analysis

Machine translation systems are often only evaluated quantitatively by using automatic metrics, such as Bleu (Papineni et al., 2002), which compares the system output to one or more human reference translations. While this type of evaluation has its advantages, mainly that it is fast and cheap, its correlation with human judgments is often low, especially for translation out of English (Callison-Burch et al., 2009). In order to overcome these problems to some extent I use several metrics in my studies, instead of only Bleu. Despite this, metrics only give a single score per sentence batch and system, which even using several metrics gives us little information on the particular problems with a system, or about what the possible improvements are.

One alternative to automatic metrics is human judgments, either absolute scores, for instance for adequacy or fluency, or by ranking sentences or segments. Such evaluations are a valuable complement to automatic metrics, but they are costly and time-consuming, and while they are useful for comparing systems they also fail to pinpoint specific problems. I mainly take advantage of this type of evaluation as part of participating with my research group in MT shared tasks with large evaluation campaigns such as WMT (e.g. Callison-Burch et al. (2009)).

To overcome the limitation of quantitative evaluations, I focus on error analysis (EA) of MT output in my thesis. EA is the task of annotating and classifying the errors in MT output, which gives a qualitative view. It can be used to evaluate and compare systems, but is also useful in order to focus the research effort on common problems for the language pair in question. There have been previous attempts of describing typologies for EA for MT, but they are not unproblematic. Vilar et al. (2006) suggested a typology with five main categories: missing, incorrect, unknown, word order, and punctuation, which have also been used by other researchers, mainly for evaluation. However, this typology is relatively shallow and mixes classification of errors with causes of errors. Farrús et al. (2010) suggested a typology based on linguistic categories, such as orthography and semantics, but their descriptions of these categories and their subcategories are not detailed. Thus, as part of my research, I am in the progress of designing a fine-grained typology and guidelines for EA. I have also created a tool for performing MT error analysis (Stymne, 2011a). Initial annotations have helped to focus my research efforts, and will be discussed below. I also plan to use EA as one means of evaluating my work on pre- and postprocessing.

4 Main Research Problems

In this section I describe the four main problem areas I will focus on in my thesis project. I summarize briefly previous work in each area, and outline my own current and planned contributions. Sample results from the different studies are shown in Table 2.

4.1 Compounding

In most Germanic languages, compounds are written without spaces or other word boundaries, which makes them problematic for SMT, mainly due to sparse data problems. The standard method for treating compounds for translation from Germanic languages is to split them in both the training data and translation input (e.g. (Nießen and Ney, 2000; Koehn and Knight, 2003; Popović et al., 2006)). Koehn and Knight (2003) also suggested a corpus-based compound splitting method that has been much used for SMT, where compounds are split based on corpus frequencies of its parts.

If compounds are split for translation into Germanic languages, the SMT system produces output with split compounds, which need to be postprocessed into full compounds. There has been very little research into this problem. For this process to be successful, it is important that the SMT system produces the split compound parts in a correct word order. To encourage this I have used a factored translation system that outputs parts-of-speech and uses a sequence model on parts-of-speech. I extended the part-of-speech tagset to use special part-of-speech tags for split compound parts, which depend on the head part-of-speech of the compound. For instance, the Swedish noun *päronträd* (*pear tree*) would be tagged as *päron|N-part träd|N* when split. Using this model the number of compound parts that were produced in the wrong order was reduced drastically compared to not using a part-of-speech sequence model for translation into German (Stymne, 2009a).

I also designed an algorithm for the merging task that uses these part-of-speech tags to merge compounds only when the next part-of-speech tag matches. This merging method outperforms reimplementations and variations of previous merging suggestions (Popović et al., 2006), and methods adapted from morphology merging (Virpioja et al., 2007) for translation into German (Stymne, 2009a). It also has the advantage over previous merging methods that it can produce novel compounds, while at the same time reducing the risk of merging parts into non-words. I have also shown that these compound processing methods work equally well for translation into Swedish (Stymne and Holmqvist, 2008). Currently I am working on methods for further improving compound merging, with promising initial results.

4.2 Definite Noun Phrases

In Scandinavian languages there are two ways to express definiteness in noun phrases, either by a definite article, or by a suffix on the noun. This leads to problems when translating into these languages, such as superfluous definite articles and wrong forms of nouns. I am not aware of any published research in this area, but an unpublished

Language pair	Corpus	Corpus size	Testset size	In article	System	Bleu	NIST
En-De	Europarl	439,513	2,000	Stymne (2008)	BL	19.31	5.727
					+Comp	19.73	5.854
En-Se	Europarl	701,157	2,000	Stymne and Holmqvist (2008)	BL	21.63	6.109
					+Comp	22.12	6.143
En-Da	Automotive	168,046	1,000	Stymne (2009b)	BL	70.91	8.816
					+Def	76.35	9.363
En-Se	Europarl	701,157	1,000	Stymne (2011b)	BL	21.63	6.109
					+Def	22.03	6.178
En-De	Europarl	439,513	2,000	Stymne (2011c)	BL	19.32	5.901
					+Reo	19.59	5.936
En-Se	Europarl	701,157	335	Stymne and Ahrenberg (2010)	BL	19.44	5.381
					+EC	22.12	5.447

Table 2: A selection of results for the four pre- and postprocessing strategies. Corpus sizes are given as number of sentences. BL is baseline systems, +Comp with compound processing, +Def with definite processing, +Reo with iterative reordering and alignment and monotone decoding, +EC with grammar checker error correction. The test set for error correction only contains sentences that are affected by the error correction.

report shows no gain for a simple pre-processing strategy for translation from German to Swedish (Samuelsson, 2006). There is similar work on other phenomena, such as Nießen and Ney (2000), who move German separated verb prefixes, to imitate the English phrasal verb structure.

I address definiteness by preprocessing the source language, to make definite NPs structurally similar to target language NPs. The transformations are rule-based, using part-of-speech tags. Definite NPs in Scandinavian languages are mimicked in the source language by removing superfluous definite articles, and/or adding definite suffixes to nouns. In an initial study, this gave very good results, with relative Bleu improvements of up to 22.1% for translation into Danish (Stymne, 2009b). In Swedish and Norwegian, the distribution of definite suffixes is more complex than in Danish, and the basic strategy that worked well for Danish was not successful (Stymne, 2011b). A small modification to the basic strategy, so that superfluous English articles were removed, but no suffixes were added, was successful for translation from English into Swedish and Norwegian. A planned extension is to integrate the transformations into a lattice that is fed to the decoder, in the spirit of (Dyer et al., 2008).

4.3 Word Order

There has been a lot of research on how to handle word order differences between languages. Preprocessing approaches can use either hand-written rules targeting known language differences (e.g. Collins et al. (2005), Li et al. (2009)), or automatically learnt rules (e.g. Xia and McCord (2004), Zhang et al. (2007b)), which are basically language independent.

I have performed an initial study on a language independent word order strategy where reordering rule learning and word alignment are performed iteratively, since they both depend on the other process (Stymne, 2011c). There were no overall improvements as measured by Bleu, but an investigation of the reordering rules showed that the rules learned in the different iterations are different with regard to the linguistic phenomena they handle, indicating that it is possible to learn new information from iterating rule learning and word alignment. In this study I only choose the 1-best reordering as input to the SMT system. I plan to extend this by presenting several reorderings to the decoder as a lattice, which has been successful in previous work (see e.g. Zhang et al. (2007b)).

My preliminary error analysis has shown that there are two main word order difficulties for translation between English and Swedish, adverb placement, and V2 errors, where the verb is not placed in the correct position when it should be placed before the subject. I plan to design a preprocessing scheme to tackle these particular problems for English-Swedish translation.

4.4 Error Correction

Postprocessing can be used to correct MT output that has not been preprocessed, for instance in order to improve the grammaticality. There has not been much research in this area. A few examples are Elming (2006), who use transformation-based learning for word substitution based on aligned human post-edited sentences, and Guzmán (2007) who used regular expression to correct regular Spanish errors. I have applied error correction suggestions given by a grammar checker to the MT output, showing that it can improve certain types of errors, such as NP agreement and word order, with a high precision, but unfortunately with a low recall (Stymne and Ahrenberg, 2010). Since the recall is low, the positive effect on metrics such as Bleu is small on general test sets, but there are improvements on test sets which only contains sentences that are affected by the postprocessing. An error analysis showed that 68–74% of the corrections made were useful, and only around 10% of the changes made were harmful. I believe that this approach could be even more useful for similar languages, such as Danish and Swedish, where a spell-checker might also be useful.

The initial error analysis I have performed has helped to identify common errors in SMT output, and shown that many of them are quite regular. A strategy I intend to pursue is to further identify common and regular problems, and to either construct rules or to train a machine learning classifier to identify them, in order to be able to postprocess them. It might also be possible to use the annotations from the error analysis as part of the training data for such a classifier.

5 Discussion

The main focus of my thesis will be on designing and evaluating methods for pre- and postprocessing of statistical MT, where I will contribute methods that can improve translation within the four areas discussed in section 4. The effort is focused on translation into Germanic languages, including German, on which there has been much previous research, and Swedish and other Scandinavian languages, where there has been little previous research. I believe that both language-pair dependent and independent methods for pre- and postprocessing can be useful. It is also the case that some language-pair dependent methods carry over to other (similar) language pairs with no or little modification. So far I have mostly used rule-based processing, but I plan to extend this with investigating machine learning methods, and compare the two main approaches.

I strongly believe that it is important for MT researchers to perform qualitative evaluations, both for identifying problems with MT systems, and for evaluating and comparing systems. In my experience it is often the case that a change to the system to improve one aspect, such as compounding, also leads to many other changes, in the case of compounding for instance because of the possibility of improved alignments, which I think we lack a proper understanding of.

My planned thesis contributions are to design a detailed error typology, guidelines, and a tool, targeted at MT researchers, for performing error annotation, and to improve statistical machine translation in four problem areas, using several methods of pre- and postprocessing.

References

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of EMNLP*, pages 745–754, Honolulu, Hawaii, USA.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of WMT*, pages 1–28, Athens, Greece.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):202–228.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*, pages 531–540, Ann Arbor, Michigan, USA.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of ACL*, pages 1012–1020, Columbus, Ohio, USA.

Jakob Elming. 2006. Transformation-based correction of rule-based MT. In *Proceedings of EAMT*, pages 219–226, Oslo, Norway.

Mireia Farrús, Marta R. Costa-jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of EAMT*, pages 52–57, Saint Raphaël, France.

Rafael Guzmán. 2007. Advanced automatic MT post-editing using regular expressions. *Multilingual*, 18(6):49–52.

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

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of EACL*, pages 187–193, Budapest, Hungary.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*, pages 48–54, Edmonton, Alberta, Canada.

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 ACL, demonstration session*, pages 177–180, Prague, Czech Republic.

Jin-Ji Li, Jungi Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2009. Chinese syntactic reordering for adequate generation of Korean verbal phrases in Chinese-to-Korean SMT. In *Proceedings of WMT*, pages 190–196, Athens, Greece.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *Proceedings of CoLing*, pages 1081–1085, Saarbrücken, Germany.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Maja Popović, Daniel Stein, and Hermann Ney. 2006. Statistical machine translation of German compound words. In *Proceedings of FinTAL – 5th International Conference on Natural Language Processing*, pages 616–624, Turku, Finland. Springer Verlag, LNCS.

Yvonne Samuelsson. 2006. Nouns in statistical machine translation. Unpublished manuscript: Term paper, Statistical Machine Translation.

Sara Stymne and Lars Ahrenberg. 2010. Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *Proceedings of LREC*, pages 2175–2181, Valetta, Malta.

Sara Stymne and Maria Holmqvist. 2008. Processing of Swedish compounds for phrase-based statistical machine translation. In *Proceedings of EAMT*, pages 180–189, Hamburg, Germany.

Sara Stymne. 2008. German compounds in factored statistical machine translation. In *Proceedings of GoTAL – 6th International Conference on Natural Language Processing*, pages 464–475, Gothenburg, Sweden. Springer Verlag, LNCS/LNAI.

Sara Stymne. 2009a. A comparison of merging strategies for translation of German compounds. In *Proceedings of EACL, Student Research Workshop*, pages 61–69, Athens, Greece.

Sara Stymne. 2009b. Definite noun phrases in statistical machine translation into Danish. In *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, pages 4–9, Odense, Denmark.

Sara Stymne. 2011a. Blast: A tool for error analysis of machine translation output. In *Proceedings of ACL, demonstration session*, Portland, Oregon, USA.

Sara Stymne. 2011b. Definite noun phrases in statistical machine translation into Scandinavian languages. In *Proceedings of EAMT*, Leuven, Belgium.

Sara Stymne. 2011c. Iterative reordering and word alignment for statistical MT. In *Proceedings of the 18th Nordic Conference on Computational Linguistics*, Riga, Latvia.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *Proceedings of LREC*, pages 697–702, Genoa, Italy.

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

Wei Wang, Jonathan May, Kevin Knight, and Daniel Marcu. 2010. Re-structuring, re-labeling, and re-aligning for syntax-based machine translation. *Computational Linguistics*, 36(2):247–277.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of CoLing*, pages 508–514, Geneva, Switzerland.

Min Zhang, Hongfei Jiang, Ai Ti Aw, Jun Sun, Sheng Li, and Chew Lim Tan. 2007a. A tree-to-tree alignment-based model for statistical machine translation. In *Proceedings of MT Summit XI*, pages 535–542, Copenhagen, Denmark.

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007b. Improved chunk-level reordering for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 21–28, Trento, Italy.