Vs and OOVs: Two Problems for Translation between German and English

Sara Stymne, Maria Holmqvist, Lars Ahrenberg
Linköping University
Sweden
{sarst,marho,lah}@ida.liu.se

Abstract

In this paper we report on experiments with three preprocessing strategies for improving translation output in a statistical MT system. In training, two reordering strategies were studied: (i) reorder on the basis of the alignments from Giza++, and (ii) reorder by moving all verbs to the end of segments. In translation, out-of-vocabulary words were preprocessed in a knowledge-lite fashion to identify a likely equivalent. All three strategies were implemented for our English↔German system submitted to the WMT10 shared task. Combining them lead to improvements in both language directions.

1 Introduction

We present the Liu translation system for the constrained condition of the WMT10 shared translation task, between German and English in both directions. The system is based on the 2009 Liu submission (Holmqvist et al., 2009), that used compound processing, morphological sequence models, and improved alignment by reordering.

This year we have focused on two issues: translation of verbs, which is problematic for translation between English and German since the verb placement is different with German verbs often being placed at the end of sentences; and OOVs, out-of-vocabulary words, which are problematic for machine translation in general. Verb translation is targeted by trying to improve alignment, which we believe is a crucial step for verb translation since verbs that are far apart are often not aligned at all. We do this mainly by moving verbs to the end of sentences previous to alignment, which we also combine with other alignments. We transform OOVs into known words in a post-processing step, based on casing, stemming, and splitting of hyphenated compounds. In addition, we perform general compound splitting for German both before training and translation, which also reduces the OOV rate.

All results in this article are for the development test set newstest2009, on truecased output. We report Bleu scores (Papineni et al., 2002) and Meteor ranking (without WordNet) scores (Agarwal and Lavie, 2008), using percent notation. We also used other metrics, but as they gave similar results they are not reported. For significance testing we used approximate randomization (Riezler and Maxwell, 2005), with $p < 0.05$.

2 Baseline System

The 2010 Liu system is based on the PBSMT baseline system for the WMT shared translation task[1]. We use the Moses toolkit (Koehn et al., 2007) for decoding and to train translation models, Giza++ (Och and Ney, 2003) for word alignment, and the SRILM toolkit (Stolcke, 2002) to train language models. The main difference to the WMT baseline is that the Liu system is trained on truecased data, as in Koehn et al. (2008), instead of lowercased data. This means that there is no need for a full recasing step after translation, instead we only need to uppercase the first word in each sentence.

2.1 Corpus

We participated in the constrained task, where we only trained the Liu system on the news and Europarl corpora provided for the workshop. The translation and reordering models were trained using the bilingual Europarl and news commentary corpora, which we concatenated.

We used two sets of language models, one where we first trained two models on Europarl and news commentary, which we then interpolated

[1] http://www.statmt.org/wmt10/baseline.html

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 189–194,
Uppsala, Sweden, 11-16 July 2010.

with more weight given to the news commentary, using weights from Koehn and Schroeder (2007). The second set of language models were trained on monolingual news data. For tuning we used every second sentence, in total 1025 sentences, of news-test2008.

2.2 Training with Limited Computational Resources

One challenge for us was to train the translation sytem with limited computational resources. We trained all systems on one Intel Core 2 CPU, 3.0Ghz, 16 Gb of RAM, 64 bit Linux (RedHat) machine. This constrained the possibilities of using the data provided by the workshop to the full. The main problem was training the language models, since the monolingual data was very large compared to the bilingual data.

In order to train language models that were both fast at runtime, and possible to train with the available memory, we chose to use the SRILM toolkit (Stolcke, 2002), with entropy-based pruning, with 10^{-8} as a threshold. To reduce the model size we also used lower order models for the large corpus; 4-grams instead of 5-grams for words and 6-grams instead of 7-grams for the morphological models. It was still impossible to train on the monolingual English news corpus, with nearly 50 million sentences, so we split that corpus into three equal size parts, and trained three models, that were interpolated with equal weights.

3 Morphological Processing

We added morphological processing to the baseline system, by training additional sequence models on morphologically enriched part-of-speech tags, and by compound processing for German.

We utilized the factored translation framework in Moses, to enrich the baseline system with an additional target sequence model. For English we used part-of-speech tags obtained using TreeTagger (Schmid, 1994), enriched with more fine-grained tags for the number of determiners, in order to target more agreement issues, since nouns already have number in the tagset. For German we used morphologically rich tags from RFTagger (Schmid and Laws, 2008), that contains morphological information such as case, number, and gender for nouns and tense for verbs. We used the extra factor in an additional sequence model on the target side, which can improve word order and agreement between words. For German the factor was also used for compound merging.

System	Bleu	Meteor
Baseline	13.42	48.83
+ morph	13.85	49.69
+ comp	14.24	49.41

Table 1: Results for morphological processing, English→German

System	Bleu	Meteor
Baseline	18.34	38.13
+ morph	18.39	37.86
+ comp	18.50	38.47

Table 2: Results for morphological processing, German→English

Prior to training and translation, compound processing was performed, using an empirical method (Koehn and Knight, 2003; Stymne, 2008) that splits words if they can be split into parts that occur in a monolingual corpus, choosing the splitting option with the highest arithmetic mean of its part frequencies in the corpus. We split nouns, adjectives and verbs, into parts that are content words or particles. We imposed a length limit on parts of 3 characters for translation from German and of 6 characters for translation from English, and we had a stop list of parts that often led to errors, such as *arische* (*Aryan*) in *konsularische* (*consular*). We allowed 10 common letter changes (Langer, 1998) and hyphens at split points. Compound parts were given a special part-of-speech tag that matches the head word.

For translation into German, compound parts were merged into full compounds using a method described in Stymne and Holmqvist (2008), which is based on matching of the special part-of-speech tag for compound parts. A word with a compound POS-tag were merged with the next word, if their POS-tags were matching.

Tables 1 and 2 show the results of the additional morphological processing. Adding the sequence models on morphologically enriched part-of-speech tags gave a significant improvement for translation into German, but similar or worse results as the baseline for translation into English. This is not surprising, since German morphology is more complex than English morphology. The addition of compound processing significantly improved the results on Meteor for translation into

English, and it also reduced the number of OOVs in the translation output by 20.8%. For translation into German, compound processing gave a significant improvement on both metrics compared to the baseline, and on Bleu compared to the system with morphological sequence models. Overall, we believe that both compound splitting and morphology are useful; thus all experiments reported in the sequel are based on the baseline system with morphology models and compound splitting, which we will call *base*.

4 Improved Alignment by Reordering

Previous work has shown that translation quality can be improved by making the source language more similar to the target language, for instance in terms of word order (Wang et al., 2007; Xia and McCord, 2004). In order to harmonize the word order of the source and target sentence, they applied hand-crafted or automatically induced reordering rules to the source sentences of the training corpus. At decoding time, reordering rules were again applied to input sentences before translation. The positive effects of such methods seem to come from a combination of improved alignment and improved reordering during translation.

In contrast, we focus on improving the word alignment by reordering the training corpus. The training corpus is reordered prior to word alignment with Giza++ (Och and Ney, 2003) and then the word links are re-adjusted back to the original word positions. From the re-adjusted corpus, we create phrase tables that allow translation of non-reordered input text. Consequently, our reordering only affects the word alignment and the phrase tables extracted from it.

We investigated two ways of reordering. The first method is based on word alignments and the other method is based on moving verbs to similar positions in the source and target sentences. We also investigated different combinations of reorderings and alignments. All results for the systems with improved reordering are shown in Tables 3 and 4.

System	Bleu	Meteor
base	14.24	49.41
reorder	14.32	49.58
verb	13.93	49.22
base+verb	14.38	49.72
base+verb+reorder	14.39	49.39

Table 3: Results for improved alignment, English→German

System	Bleu	Meteor
base	18.50	38.47
reorder	18.77	38.53
verb	18.61	38.53
base+verb	18.66	38.61
base+verb+reorder	18.73	38.59

Table 4: Results for improved alignment, German→English

4.1 Reordering Based on Alignments

The first reordering method does not require any syntactic information or rules for reordering. We simply used symmetrized Giza++ word alignments to reorder the words in the source sentences to reflect the target word order and applied Giza++ again to the reordered training corpus. The following steps were performed to produce the final word alignment:

1. Word align the training corpus with Giza++.
2. Reorder the source words according to the order of the target words they are aligned to (store the original source word positions for later).
3. Word align the reordered source and original target corpus with Giza++.
4. Re-adjust the new word alignments so that they align source and target words in the original corpus.

The system built on this word alignment (reorder) had a significant improvement in Bleu score over the unreordered baseline (*base*) for translation into English, and small improvements otherwise.

4.2 Verb movement

The positions of finite verbs are often very different in English and German, where they are often placed at the end of sentences. In several cases we noted that finite verbs were misaligned by Giza++. To improve the alignment of verbs, we moved all verbs in both English and German to the end of the sentences prior to word alignment. The reordered sentences were word aligned with Giza++ and the

resulting word links were then re-adjusted to align words in the original corpus.

The system created from this alignment (verb) resulted in significantly lower scores than *base* for translation into German, and similar scores as *base* for translation into English.

4.3 Combination Systems

The alignment based on reordered verbs did not produce a better alignment in terms of Bleu scores of the resulting translations, which led us to the conclusion that the alignment was noisy. However, it is possible that we did correctly align some words that were misaligned in the baseline alignment. To investigate this issue we concatenated first the baseline and verb alignments, and then all three alignments, and extracted phrase tables from the concatenated training sets.

All scores for both combined systems significantly outperformed the unfactored baseline, and were slightly better than *base*. For translation into German it was best to use the combination of only verb and *base*, which was significantly better than *base* on Meteor. This shows that even though the verb alignments were not good when used in a single system, they still could contribute in a combination system.

5 Preprocessing of OOVs

Out-of-vocabulary words, words that have not been seen in the training data, are a problem in statistical machine translation, since no translations have been observed for them. The standard strategy is to transfer them as is to the translation output, which, naive as it sounds, actually works well in some cases, since many OOVs are numbers or proper names (Stymne and Holmqvist, 2008). However, it still results in incomprehensible words in the output in many cases. We have investigated several ways of changing unknown words into similar words that have been seen in the training data, in a preprocessing step.

We also considered another OOV problem, number formatting, since it differs between English and German. To address this, we swapped decimal points/commas, and other delimeters for unknown numbers in a post-processing step.

In the preprocessing step, we applied a number of transformations to each OOV word, accepting the first applicable transformation that led to a known word:

Type	German	English
total OOVs	1833	1489
casing	124	26
stemming	270	72
hyphenated words	230	124
end hyphens	24	–

Table 5: Number of affected words by OOV-preprocessing

1. Change the word into a known cased version (since we trained a truecased system, this handles cased variations of words)
2. Stem the word, and if we know the stem, choose the most common realisation of that stem (using a Porter stemmer)
3. For hyphenated words, split at the hyphen (if any of the resulting parts are OOVs, they are recursively treated as well)
4. Remove hyphens at the end of German words (that could result from compound splitting)

The first two steps were based on frequency lists of truecased and stemmed words that we compiled from the monolingual training corpora.

Inspection of the initial results showed that proper names were often changed into other words in English, so we excluded them from the preprocessing by not applying it to words with an initial capital letter. This happened to a lesser extent for German, but here it was impossible to use the same simple heuristic for proper names, since German nouns also have an initial capital letter.

The number of affected words for the baseline using the final transformations are shown in Table 5. Even though we managed to transform some words, we still lack a transformation for the majority of OOVs. Despite this, there is a tendency of small improvements on both metrics in the majority of cases in both translation directions, as shown in Tables 6 and 7.

Figure 1 shows an example of how OOV processing affects one sentence for translation from German to English. In this case splitting a hyphenated compound gives a better translation, even though the word *opening* is chosen rather than *jack*. There is also a stemming change, where the adjective *ausgereiftesten* (*the most well-engineered*), is changed form superlative to positive. This results in a more understandable trans-

DE original	Die besten und technisch *ausgereiftesten* Telefone mit einer *3,5-mm-Öffnung* für normale Kopfhörer kosten bis zu fünfzehntausend Kronen.
DE preprocessed	die besten und technisch *ausgereifte* Telefone mit einer *3,5 mm Öffnung* für normale Kopf Hörer kosten bis zu fünfzehntausend Kronen .
base+verb+reorder	The best and technically *ausgereiftesten* phones with a *3,5-mm-Öffnung* for normal earphones cost up to fifteen thousand kronor.
base+verb+reorder +OOV	The best and technologically *advanced* phones with a *3.5 mm opening* for normal earphones cost up to fifteen thousand kronor.
EN reference	The best and most technically *well-equipped* telephones, with a *3.5 mm jack* for ordinary headphones, cost up to fifteen thousand crowns.

Figure 1: Example of the effects of OOV processing for German→English

System	Bleu	Meteor
base	14.24	49.41
+ OOV	14.26	49.43
base+verb	14.38	49.72
+ OOV	14.42	49.75
+ MBR	**14.41**	**49.77**

Table 6: Results for OOV-processing and MBR, English→German.

System	Bleu	Meteor
base	18.50	38.47
+ OOV	18.48	38.59
base+verb+reorder	18.73	38.59
+ OOV	18.81	38.70
+ MBR	**18.84**	**38.75**

Table 7: Results for OOV-processing and MBR, German→English.

lation, which, however, is harmful to automatic scores, since the preceding word, *technically*, which is identical to the reference, is changed into *technologically*.

This work is related to work by Arora et al. (2008), who transformed Hindi OOVs by using morphological analysers, before translation to Japanese. Our work has the advantage that it is more knowledge-lite, as it only needs a Porter stemmer and a monolingual corpus. Mirkin et al. (2009) used WordNet to replace OOVs by synonyms or hypernyms, and chose the best overall translation partly based on scoring of the source transformations. Our OOV handling could potentially be used in combination with both these strategies.

6 Final Submission

For the final Liu shared task submission we used the base+verb+reorder+OOV system for German→English and the base+verb+OOV system for English→German, which had the best overall scores considering all metrics. To these systems we added minimum Bayes risk (MBR) decoding (Kumar and Byrne, 2004). In standard decoding, the top suggestion of the translation system is chosen as the system output. In MBR decoding the risk is spread by choosing the translation that is most similar to the N highest scoring translation suggestions from the system, with $N = 100$, as suggested in Koehn et al. (2008). MBR decoding gave hardly any changes in automatic scores, as shown in Tables 6 and 7. The final system was significantly better than the baseline in all cases, and significantly better than *base* on Meteor in both translation directions, and on Bleu for translation into English.

7 Conclusions

As in Holmqvist et al. (2009) reordering by using Giza++ in two phases had a small, but consistent positive effect. Aligning verbs by co-locating them at the end of sentences had a largely negative effect. However, when output from this method was concatenated with the baseline alignment before extracting the phrase table, there were consistent improvements. Combining all three alignments, however, had mixed effects. Combining reordering in training with a knowledge-lite method for handling out-of-vocabulary words led to significant improvements on Meteor scores for translation between German and English in both directions.

References

Abhaya Agarwal and Alon Lavie. 2008. METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio, USA.

Karunesh Arora, Michael Paul, and Eiichiro Sumita. 2008. Translation of unknown words in phrase-based statistical machine translation for languages of rich morphology. In *Proceedings of the 1st International Workshop on Spoken Languages Technologies for Under-Resourced Languages*, pages 70–75, Hanoi, Vietnam.

Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. 2009. Improving alignment for SMT by reordering and augmenting the training corpus. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 120–124, Athens, Greece.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the EACL*, pages 187–193, Budapest, Hungary.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, 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 45th Annual Meeting of the ACL, demonstration session*, pages 177–180, Prague, Czech Republic.

Philipp Koehn, Abhishek Arun, and Hieu Hoang. 2008. Towards better machine translation quality for the German-English language pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio, USA.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the 2004 Human Language Technology Conference of the NAACL*, pages 169–176, Boston, Massachusetts, USA.

Stefan Langer. 1998. Zur Morphologie und Semantik von Nominalkomposita. In *Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, pages 83–97, Bonn, Germany.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 791–799, Suntec, Singapore.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the ACL*, pages 57–64, Ann Arbor, Michigan, USA.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22th International Conference on Computational Linguistics*, pages 777–784, Manchester, UK.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

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

Sara Stymne and Maria Holmqvist. 2008. Processing of Swedish compounds for phrase-based statistical machine translation. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, 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.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 737–745, Prague, Czech Republic.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514, Geneva, Switzerland.