Improving alignment for SMT by reordering and augmenting the training corpus

Maria Holmqvist, Sara Stymne, Jody Foo and Lars Ahrenberg
Department of Computer and Information Science
Linköping University, Sweden
{marho,sarst,jodfo,lah}@ida.liu.se

Abstract

We describe the LIU systems for English-German and German-English translation in the WMT09 shared task. We focus on two methods to improve the word alignment: (i) by applying Giza++ in a second phase to a reordered training corpus, where reordering is based on the alignments from the first phase, and (ii) by adding lexical data obtained as high-precision alignments from a different word aligner. These methods were studied in the context of a system that uses compound processing, a morphological sequence model for German, and a part-of-speech sequence model for English. Both methods gave some improvements to translation quality as measured by Bleu and Meteor scores, though not consistently. All systems used both out-of-domain and in-domain data as the mixed corpus had better scores in the baseline configuration.

1 Introduction

It is an open question whether improved word alignment actually improves statistical MT. Fraser and Marcu (2007) found that improved alignments as measured by AER will not necessarily improve translation quality, whereas Ganchev et al. (2008) did improve translation quality on several language pairs by extending the alignment algorithm.

For this year's shared task we therefore studied the effects of improving word alignment in the context of our system for the WMT09 shared task. Two methods were tried: (i) applying Giza++ in a second phase to a reordered training corpus, where reordering is based on the alignments from the first phase, and (ii) adding lexical data obtained as high-precision alignments from a different word aligner. The submitted system includes the first method in addition to the processing of compounds and additional sequence models used by Stymne et al. (2008). Heuristics were used to generate true-cased versions of the translations that were submitted, as reported in section 6.

In this paper we report case-insensitive Bleu scores (Papineni et al., 2002), unless otherwise stated, calculated with the NIST tool, and case-insensitive Meteor-ranking scores, without WordNet (Agarwal and Lavie, 2008).

2 Baseline system

Our baseline system uses compound splitting, compound merging and part-of-speech/morphological sequence models (Stymne et al., 2008). Except for these additions it is similar to the baseline system of the workshop[1].

The translation system is a factored phrase-based translation system that uses the Moses toolkit (Koehn et al., 2007) for decoding and training, GIZA++ for word alignment (Och and Ney, 2003), and SRILM (Stolcke, 2002) for language models. Minimum error rate training was used to tune the model feature weights (Och, 2003).

Tuning was performed on the news-dev2009a set with 1025 sentences. All development testing was performed on the news-dev2009b set with 1026 sentences.

2.1 Sequence model based on part-of-speech and morphology

The translation models were factored with one additional output factor. For English we used part-of-speech tags obtained with TreeTagger (Schmid, 1994). For German we enriched the tags from TreeTagger with morphological information, such as case or tense, that we get from a commercial

[1] http://www.statmt.org/wmt09/baseline.html

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 120–124,
Athens, Greece, 30 March – 31 March 2009. ©2009 Association for Computational Linguistics

dependency parser[2].

We used the extra factor in an additional sequence model which can improve agreement between words, and word order. For German this factor was also used for compound merging.

2.2 Compound processing

Prior to training and translation, compound processing was performed using an empirical method based on (Koehn and Knight, 2003; Stymne, 2008). Words were split if they could be split into parts that occur in a monolingual corpus. We chose the split with the highest arithmetic mean of the corpus frequencies of compound parts. We split nouns, adjectives and verbs into parts that were content words or particles. A part had to be at least 3 characters in length and a stop list was used to avoid parts that often lead to errors, such as *arische* (*Aryan*) in *konsularische* (*consular*). Compound parts sometimes have special compound suffixes, which could be additions or truncations of letters, or combinations of these. We used the top 10 suffixes from a corpus study of Langer (1998), and we also treated hyphens as suffixes of compound parts. Compound parts were given a special part-of-speech tag that matched the head word.

For translation into German, compound parts were merged to form compounds, both during test and tuning. The merging is based on the special part-of-speech tag used for compound parts (Stymne, 2009). A token with this POS-tag is merged with the next token, either if the POS-tags match, or if it results in a known word.

3 Domain adaptation

This year three training corpora were available, a small bilingual news commentary corpus, a reasonably large Europarl corpus, and a very large monolingual news corpus, see Table 1 for details. The bilingual data was filtered to remove sentences longer than 60 words. Because the German news training corpus contained a number of English sentences, this corpus was cleaned by removing sentences containing a number of common English words.

Based on Koehn and Schroeder (2007) we adapted our system from last year, which was focused on Europarl, to perform well on test data from the news domain. We used the possibility to include several translation models in the Moses decoder by using multiple alternative decoding paths. We first trained systems on either bilingual news data or Europarl. Then we trained a mixed system, with two translation models one from each corpus, a language model from the bilingual news data, and a Europarl reordering model. The mixed system was slightly better than the Europarl only system. All sequence models used 5-grams for surface form and 7-grams for part-of-speech. All scores are shown in Table 2.

Corpus	German	English
news-commentary09	81,141	
Europarl	1,331,262	
news-train08	9,619,406	21,215,311

Table 1: Number of sentences in the corpora (after filtering)

Corpus	En⇒De		De⇒En	
	Bleu	Meteor	Bleu	Meteor
News com.	12.13	47.01	17.21	36.08
Europarl	12.92	47.27	18.53	37.65
Mixed	12.91	47.96	18.76	37.69
Mixed+	**14.62**	**49.48**	**19.92**	**38.18**

Table 2: Results of domain adaptation

We wanted to train sequence models on the large monolingual corpora, but due to limited computer resources, we had to use a lower order for this, than on the small corpus. Thus our sequence models on this data has lower order than those trained on bilingual news or Europarl, with 4-grams for surface form and 6-grams for part-of-speech. We also used the entropy-based pruning included in the SRILM toolkit, with 10^{-8} as a threshold. Using these sequence models in the mixed model, called mixed+, improved the results drastically, as shown in Table 2.

The other experiments reported in this paper are based on the mixed+ system.

4 Improved alignment by reordering

Word alignment with Giza++ has been shown to improve from making the source and target language more similar, e.g., in terms of segmentation (Ma et al., 2007) or word order.

We used the following simple procedure to improve alignment of the training corpus by reordering the words in one of the texts according to the

[2] Machinese syntax, from Connexor Oy http://www.connexor.eu

Corpus	En⇒De		De⇒En	
	Bleu	Meteor	Bleu	Meteor
Mixed+	14.62	49.48	19.92	38.18
Re-Src	**14.63**	**49.80**	**20.54**	**38.86**
Re-Trg	14.51	48.62	20.48	38.73

Table 3: Results of reordering experiments

word order in the other language:

1. Word align the corpus with Giza++.
2. Reorder the German words according to the order of the English words they are aligned to. (This is a common step in approaches that extract reordering rules for translation. However, this is not what we use it for here.)
3. Word align the reordered German and original English corpus with Giza++.
4. Put the reordered German words back into their original position and adjust the alignments so that the improved alignment is preserved.

After this step we will have a possibly improved alignment compared to the original Giza++ alignment. A phrase table was extracted from the alignment and training was performed as usual. The reordering procedure was carried out on both source (Re-Src) and target data (Re-Trg) and the results of translating devtest data using these alignments are shown in Table 3.

Compared with our baseline (mixed+), Bleu and Meteor increased for the translation direction German–English. Both source reordering and target reordering resulted in a 0.6 increase in Bleu.

For translation into German, source reordering resulted in a somewhat higher Meteor score, but overall did not seem to improve translation. Target reordering in this direction resulted in lower scores.

It is not clear why reordering improved translation for German–English and not for English–German. In all experiments, the heuristic symmetrization of directed Giza++ alignments was performed in the intended translation direction [3].

5 Augmenting the corpus with an extracted dictionary

Previous research (Callison-Burch et al., 2004; Fraser and Marcu, 2006) has shown that including word aligned data during training can improve translation results. In our case we included a dictionary extracted from the news-commentary corpus during the word alignment.

Using a method originally developed for term extraction (Merkel and Foo, 2007), the news-commentary09 corpus was grammatically annotated and aligned using a heuristic word aligner. Candidate dictionary entries were extracted from the alignments. In order to optimize the quality of the dictionary, dictionary entry candidates were ranked according to their Q-value, a metric specifically designed for aligned data (Merkel and Foo, 2007). The Q-value is based on the following statistics:

- Type Pair Frequencies (TPF), i.e. the number of times where the source and target types are aligned.
- Target types per Source type (TpS), i.e. the number of target types a specific source type has been aligned to.
- Source types per Target type (SpT), i.e. the number of source types a specific target type has been aligned to.

The Q-value is calculated as $Q{-}value{=}\frac{TPF}{TpS+SpT}$. A high Q-value indicates a dictionary candidate pair with a relatively low number of translation variations. The candidates were filtered using a Q-value threshold of 0.333, resulting in a dictionary containing 67287 entries.

For the experiments, the extracted dictionary was inserted 200 times into the corpus used during word alignment. The added dictionary entries were removed before phrase extraction. Experiments using the extracted dictionary as an additional phrase table were also run, but did not result in any improvement of translation quality.

The results can be seen in Table 4. There was no evident pattern how the inclusion of the dictionary during alignment (DictAl) affected the translation quality. The inclusion of the dictionary produced both higher and lower Bleu scores than the

[3] Our experiments show that symmetrization in the wrong translation direction will result in lower translation quality scores.

Corpus	En⇒De		De⇒En	
	Bleu	Meteor	Bleu	Meteor
Mixed+	14.62	49.48	19.92	38.18
DictAl	14.73	49.39	18.93	37.71

Table 4: Results of domain adaptation

Corpus	En⇒De	De⇒En
Mixed+	13.31	17.47
with OOV	13.74	17.96

Table 5: Case-sensitive Bleu scores

baseline system depending on the translation direction. Meteor scores were however consistently lower than the baseline system.

6 Post processing of out-of-vocabulary words

In the standard systems all out-of-vocabulary words are transferred as is from the translation input to the translation output. Many of these words are proper names, which do not get capitalized properly, or numbers, which have different formatting in German and English. We used post-processing to improve this.

For all unknown words we capitalized either the first letter, or all letters, if they occur in that form in the translation input. For unknown numbers we switched between the German decimal comma and the English decimal point for decimal numbers. For large numbers, English has a comma to separate thousands, and German has a period. These were also switched. This improved case-sensitive Bleu scores in both translation directions, see Table 5.

7 Submitted system

For both translation directions De-En and En-De we submitted a system with two translation models trained on bilingual news and Europarl. The alignment was improved by using the reordering techniques described in section 4. The systems also use all features described in this paper except for the lexical augmentation (section 5) which did not result in significant improvement. The results of the submitted systems on devtest data are boldfaced in Table 3.

Corpus	En⇒De	De⇒En
All	14.63	20.54
En-De orig.	19.93	26.82
Other set	11.66	16.17

Table 6: Bleu scores for the reordered systems on two sections of development set news-dev2009b. NIST scores show the same distribution.

8 Results on two sections of devtest data

Comparisons of translation output with reference translations on devtest data showed some surprising differences, which could be attributed to corresponding differences between source and reference data. The differences were not evenly distributed but especially frequent in those sections where the original language was something other than English or German. To check the homogeneity of the devtest data we divided it into two sections, one for documents of English or German origin, and the other for the remainder. It turned out that scores were dramatically different for the two sections, as shown in Table 6.

The reason for the difference is likely to be that only the En-De set contains source texts and translations, while the other section contains parallel translations from the same source. This suggests that it would be interesting to study the effects of splitting the training corpus in the same way before training.

9 Conclusion

The results of augmenting the training corpus with an extracted lexicon were inconclusive. However, the alignment reordering improved translation quality, especially in the De–En direction. The result of these reordering experiments indicates that better word alignment quality will improve SMT. The reordering method described in this paper also has the advantage of only requiring two runs of Giza++, no additional resources or training is necessary to get an improved alignment.

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.

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting of ACL*, pages 175–182, Barcelona, Spain.

Alexander Fraser and Daniel Marcu. 2006. Semi-supervised training for statistical word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of ACL*, pages 769–776, Sydney, Australia.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Kuzman Ganchev, João de Almeida Varelas Graça, and Ben Taskar. 2008. Better alignments = better translations? In *Proceedings of the 46th Annual Meeting of ACL*, pages 986–993, Columbus, Ohio.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the tenth conference of 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 ACL, demonstration session*, Prague, Czech Republic.

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

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Boostrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of ACL*, pages 304–311, Prague, Czech Republic.

Magnus Merkel and Jody Foo. 2007. Terminology extraction and term ranking for standardizing term banks. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*, pages 349–354, Tartu, Estonia.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of ACL*, pages 160–167, Sapporo, Japan.

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 ACL*, pages 311–318, Philadelphia, Pennsylvania.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *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 (ICSLP)*, pages 901–904, Denver, Colorado.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 135–138, Columbus, Ohio.

Sara Stymne. 2008. German compounds in factored statistical machine translation. In Aarne Ranta and Bengt Nordström, editors, *Proceedings of GoTAL, 6th International Conference on Natural Language Processing*, LNCS/LNAI Volume 5221, pages 464–475.

Sara Stymne. 2009. A comparison of merging strategies for translation of German compounds. In *Proceedings of the EACL09 Student Research Workshop*, Athens, Greece.