Experiments with word alignment, normalization and clause reordering for SMT between English and German

Maria Holmqvist, Sara Stymne and Lars Ahrenberg
Department of Computer and Information Science
Linköping University, Sweden
firstname.lastname@liu.se

Abstract

This paper presents the LIU system for the WMT 2011 shared task for translation between German and English. For English–German we attempted to improve the translation tables with a combination of standard statistical word alignments and phrase-based word alignments. For German–English translation we tried to make the German text more similar to the English text by normalizing German morphology and performing rule-based clause reordering of the German text. This resulted in small improvements for both translation directions.

1 Introduction

In this paper we present the LIU system for the WMT11 shared task, for translation between English and German in both directions. We added a number of features that address problems for translation between German and English such as word order differences, incorrect alignment of certain words such as verbs, and the morphological complexity of German compared to English, as well as dealing with previously unseen words.

In both translation directions our systems include compound processing, morphological sequence models, and a hierarchical reordering model. For German–English translation we also added morphological normalization, source side reordering, and processing of out-of-vocabulary words (OOVs). For English–German translation, we extracted word alignments with a supervised method and combined these alignments with Giza++ alignments in various ways to improve the phrase table. We experimented with different ways of combining the two alignments such as using heuristic symmetrization and interpolating phrase tables.

Results are reported on three metrics, BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and Meteor ranking scores (Agarwal and Lavie, 2008) based on truecased output.

2 Baseline System

This years improvements were added to the LIU baseline system (Stymne et al., 2010). Our baseline is a factored phrase based SMT system that uses the Moses toolkit (Koehn et al., 2007) for translation model training and decoding, GIZA++ (Och and Ney, 2003) for word alignment, SRILM (Stolcke, 2002) an KenLM (Heafield, 2011) for language modelling and minimum error rate training (Och, 2003) to tune model feature weights. In addition, the LIU baseline contains:

- Compound processing, including compound splitting and for translation into German also compound merging
- Part-of-speech and morphological sequence models

All models were trained on truecased data. Translation and reordering models were trained using the bilingual Europarl and News Commentary corpora that were concatenated before training. We created two language models. The first model is a 5-gram model that we created by interpolating two language

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 393–398,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

models from bilingual News Commentary and Europarl with more weight on the News Commentary model. The second model is a 4-gram model trained on monolingual News only. All models were created using entropy-based pruning with 10^{-8} as the threshold.

Due to time constraints, all tuning and evaluation were performed on half of the provided shared task data. Systems were tuned on 1262 sentences from newstest2009 and all results reported in Tables 1 and 2 are based on a devtest set of 1244 sentences from newstest2010.

2.1 Sequence models with part-of-speech and morphology

To improve target word order and agreement in the translation output, we added an extra output factor in our translation models consisting of tags with POS and morphological features. For English we used tags that were obtained by enriching POS tags from TreeTagger (Schmid, 1994) with additional morphological features such as number for determiners. For German, the POS and morphological tags were obtained from RFTagger (Schmid and Laws, 2008) which provides morphological information such as case, number and gender for nouns and tense for verbs. We trained two sequence models for each system over this output factor and added them as features in our baseline system. The first sequence model is a 7-gram model interpolated from models of bilingual Europarl and News Commentary. The second model is a 6-gram model trained on monolingual News only.

2.2 Compound processing

In both translation directions we split compounds, using a modified version of the corpus-based splitting method of Koehn and Knight (2003). We split nouns, verb, and adjective compounds into known parts that were content words or cardinal numbers, based on the arithmetic mean of the frequency of the parts in the training corpus. We allowed 10 common letter changes (Langer, 1998) and hyphens at split points. Compound parts were kept in their surface form and compound modifiers received a part-of-speech tag based on that of the tag of the full compound.

For translation into German, compounds were merged using the POS-merging strategy of Stymne (2009). A compound part in the translation output, identified by the special part-of-speech tags, was merged with the next word if that word had a matching part-of-speech tag. If the compound part was followed by the conjunction *und* (*and*), we added a hyphen to the part, to account for coordinated compounds.

2.3 Hierarchical reordering

In our baseline system we experimented with two lexicalized reordering models. The standard model in Moses (Koehn et al., 2005), and the hierarchical model of Galley and Manning (2008). In both models the placement of a phrase is compared to that of the previous and/or next phrase. In the standard model up to three reorderings are distinguished, monotone, swap, and discontinuous. In the hierarchical model the discontinuous class can be further subdivided into two classes, left and right discontinuous. The hierarchical model further differs from the standard model in that it compares the order of the phrase with the next or previous block of phrases, not only with the next or previous single phrase.

We investigated one configuration of each model. For the standard model we used the *msd-bidirectional-fe* setting, which uses three orientations, is conditioned on both the source and target language, and considers both the previous and next phrase. For the hierarchical model we used all four orientations, and again it is conditioned on both the source and target language, and considers both the previous and next phrase.

The result of replacing the standard reordering model with an hierarchical model is shown in Table 1 and 2. For translation into German adding the hierarchical model led to small improvements as measured by NIST and Meteor. For translation in the other direction, the differences on automatic metrics were very small. Still, we decided to use the hierarchical model in all our systems.

3 German–English

For translation from German into English we focused on making the German source text more similar to English by removing redundant morphology and changing word order before training translation models.

3.1 Normalization

We performed normalization of German words to remove distinctions that do not exist in English, such as case distinctions on nouns. This strategy is similar to that of El-Kahlout and Yvon (2010), but we used a slightly different set of transformations, that we thought better mirrored the English structure. For morphological tags we used RFTagger and for lemmas we used TreeTagger. The morphological transformations we performed were the following:

- Nouns:
  - Replace with *lemma+s* if plural number
  - Replace with *lemma* otherwise
- Verbs:
  - Replace with *lemma* if present tense, not third person singular
  - Replace with *lemma+p* if past tense
- Adjectives:
  - Replace with *lemma+c* if comparative
  - Replace with *lemma+sup* if superlative
  - Replace with *lemma* otherwise
- Articles:
  - Definite articles:
    - Replace with *des* if genitive
    - Replace with *der* otherwise
  - Indefinite articles:
    - Replace with *eines* if genitive
    - Replace with *ein* otherwise
- Pronouns:
  - Replace with *RELPRO* if relative
  - Replace with *lemma* if indefinite, interrogative, or possessive pronouns
  - Add *+g* to all pronouns which are genitive, unless they are possessive

For all word types that are not mentioned in the list, surface forms were kept.

	BLEU	NIST	Meteor
Baseline	21.01	6.2742	41.32
+hier reo	20.94	6.2800	41.24
+normalization	20.85	6.2370	41.04
+source reordering	21.06	6.3082	41.40
+ OOV proc.	21.22	6.3692	41.51

Table 1: German–English translation results. Results are cumulative.

We also performed those tokenization and spelling normalizations suggested by El-Kahlout and Yvon (2010), that we judged could safely be done for translation from German without collecting corpus statistics. We split words with numbers and letters, such as *40-jährigen* or *40jährigen* (*40 year-old*), unless the suffix indicates that it is a ordinal, such as *70sten* (*70th*). We also did some spelling normalization by exchanging *ß* with *ss* and replacing tripled consonants with doubled consonants. These changes would have been harmful for translation into German, since they change the language into a normalized variant, but for translation from German we considered them safe.

3.2 Source side reordering

To make the word order of German input sentences more English-like a version of the rules of (Collins et al., 2005) were partially implemented using tagged output from the RFTagger. Basically, beginnings of subordinate clauses, their subjects (if present) and final verb clusters were identified based on tag sequences, and the clusters were moved to the beginning of the clause, and reordered so that the finite verb ended up in the second clause position. Also, some common adverbs were moved with the verb cluster and placed between finite and non-finite verbs. After testing, we decided to apply these rules only to subordinate clauses at the end of sentences, since these were the only ones that could be identified with good precision. Still, some 750,000 clauses were reordered.

3.3 OOV Processing

We also added limited processing of OOVs. In a preprocessing step we replaced unknown words with known cased variants if available, removed markup from normalized words if that resulted in an unknown token, and split hyphened words. We also split suspected names in cases where we had a pattern with a single upper-case letter in the middle of a word, such as *ConocoPhillips* into *Conoco Phillips*. In a post-processing step we changed the number formatting of unknown numbers by changing decimal points and thousand separators, to agree with English orthography. This processing only affects a small number of words, and cannot be expected to make a large impact on the final results. Out of 884 OOVs in the devtest, 39 had known cased options, 126 hyphened words were split, 147 cases had markup from the normalization removed, and 13 suspected names were split.

3.4 Results

The results of these experiments can be seen in Table 1 where each new addition is added to the previous system. When we compare the new additions with the baseline with hierarchical reordering, we see that while the normalization did not seem to have a positive effect on any metric, both source reordering and OOV processing led to small increases on all scores.

4 English–German

For translation from English into German we attempted to improve the quality of the phrase table by adding new word alignments to the standard Giza++ alignments.

4.1 Phrase-based word alignment

We experimented with different ways of combining word alignments from Giza++ with alignments created using phrase-based word alignment (PAL) which previously has been shown to improve alignment quality for English–Swedish (Holmqvist, 2010). The idea of phrase-based word alignment is to use word and part-of-speech sequence patterns from manual word alignments to align new texts. First, parallel phrases containing a source segment, a target segment and links between source and target words are extracted from word aligned texts (Figure 1). In the second step, these phrases are matched against new parallel text and if a matching phrase is found, word links from the phrase are added to the corresponding words in the new text. In order to increase the number of matching phrases and improve word alignment recall, words in the parallel segments were replaced by POS/morphological tags from RFTagger.

```
En:    a typical example
De:    ein typisches Beispiel
Links: 0-0 1-1 2-2

En:    a JJ example
De:    ein ADJA Beispiel
Links: 0-0 1-1 2-2

En:    DT JJ NN
De:    ART ADJA N
Links: 0-0 1-1 2-2
```

Figure 1: Examples of parallel phrases used in word alignment.

	BLEU	NIST	Meteor
Baseline	16.16	6.2742	50.89
+hier reo	16.06	6.2800	51.25
+pal-gdfa	16.14	5.6527	51.10
+pal-dual	15.71	5.5735	50.43
+pal-inter	15.92	5.6230	50.73

Table 2: English–German translation results, results are cumulative except for the three alternative *PAL*-configurations.

Alignment patterns were extracted from 1000 sentences in the manually word aligned sample of English–German Europarl texts from Pado and Lapata (2006). All parallel phrases were extracted from the word aligned texts, as when extracting a translation model. Parallel phrases that contain at least 3 words were generalized with POS tags to form word/POS patterns for alignment. A subset of these patterns, with high alignment precision (> 0.80) on the 1000 sentences, were used to align the entire training corpus.

We combined the new word alignments with the Giza++ alignments in two ways. In the first method, we used a symmetrization heuristic similar to grow-diag-final-and to combine three word alignments into one, the phrase-based alignment and two Giza++ alignments in different directions. In the second method we extracted a separate phrase table from the sparser phrase-based alignment using a constrained method of phrase extraction that limited the number of unaligned words in each phrase pair. The reason for constraining the phrase table extraction was that the standard extraction method does not work well for the sparse word alignments that PAL produces, but we think it could still be useful for extracting highly reliable phrases. After some experimentation we decided to allow an unlimited number of internal unaligned words, that is unaligned words that are surrounded by aligned words, but limit the number of external unaligned words, i.e., unaligned words at the beginning or end of the phrase, to either one each in the source and target phrase, or to zero.

We used two ways to include the sparse phrase-table into the translation process:

- Have two separate phrase-tables, the sparse table, and the standard GIZA++ based phrase-table, and use Moses' dual decoding paths.
- Interpolate the sparse phrase-table with the standard phrase-table, using the mixture model formulation of Ueffing et al. (2007), with equal weights, in order to boost the probabilities of highly reliable phrases.

4.2 Results

We evaluated our systems on devtest data and found that the added phrase-based alignments did not produce large differences in translation quality compared to the baseline system with hierarchical reordering as shown in Table 2. The system created with a heuristic combination of PAL and Giza++ (pal-gdfa) had a small increase in BLEU, but no improvement on the other metrics. Systems using a phrase table extracted from the sparse alignments did not produce better results than baseline. The system using dual decoding paths (pal-dual) produced worse results than the system using an interpolated phrase table (pal-inter).

5 Submitted systems

The LIU system participated in German–English and English–German translation in the WMT 2011 shared task. The new additions were a combination of unsupervised and supervised word alignments, spelling normalization, clause reordering and OOV processing. Our submitted systems contain all additions described in this paper. For English-German we used the best performing method of word alignment combination which was the method that uses heuristic combination similar to grow-diag-final-and.

		BLEU	
	System	Devtest	Test
en-de	baseline +hier	16.1	14.5
	submitted	16.1	14.8
de-en	baseline +hier	20.9	19.3
	submitted	21.2	19.9

Table 3: Summary of devtest results and shared task test results for submitted systems and LIU baseline with hierarchical reordering.

The results of our submitted systems are shown in Table 3 where we compare them to the LIU baseline system with hierarchical reordering models. We report modest improvements on the devtest set for both translation directions. We also found small improvements of our submitted systems in the official shared task evaluation on the test set newstest2011.

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.

Michael Collins, Philipp Koehn, and Ivona Kucerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 531–540, Ann Arbor, Michigan.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurence statistics. In *Proceedings of the Second International Conference on Human Language Technology*, pages 228–231, San Diego, California.

İlknur Durgar El-Kahlout and François Yvon. 2010. The pay-offs of preprocessing for German-English statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 251–258, Paris, France.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Honolulu, Hawaii.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth*

Workshop on Statistical Machine Translation, Edinburgh, UK.

Maria Holmqvist. 2010. Heuristic word alignment with parallel phrases. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 744-748, Valletta, Malta.

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, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania.

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*, 177–180, Prague, Czech Republic.

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

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 the ACL*, pages 160–167, Sapporo, Japan.

Sebastian Pado and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 1161–1168, Sydney, Australia.

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 and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 777–784, Manchester, UK.

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*, pages 901–904, Denver, Colorado.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2010. Vs and OOVs: Two problems for translation between German and English. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 189–194, Uppsala, Sweden.

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

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 21(2):77–94.