Effects of Morphological Analysis in Translation between German and English

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

Abstract

We describe the LIU systems for German-English and English-German translation submitted to the Shared Task of the Third Workshop of Statistical Machine Translation. The main features of the systems, as compared with the baseline, is the use of morphological pre- and post-processing, and a sequence model for German using morphologically rich parts-of-speech. It is shown that these additions lead to improved translations.

1 Introduction

Research in statistical machine translation (SMT) increasingly makes use of linguistic analysis in order to improve performance. By including abstract categories, such as lemmas and parts-of-speech (POS), in the models, it is argued that systems can become better at handling sentences for which training data at the word level is sparse. Such categories can be integrated in the statistical framework using factored models (Koehn et al., 2007). Furthermore, by parsing input sentences and restructuring based on the result to narrow the structural difference between source and target language, the current phrase-based models can be used more effectively (Collins et al., 2005).

German differs structurally from English in several respects (see e.g. Collins et al., 2005). In this work we wanted to look at one particular aspect of restructuring, namely splitting of German compounds, and evaluate its effect in both translation directions, thus extending the initial experiments reported in Holmqvist et al. (2007). In addition, since German is much richer in morphology than English, we wanted to test the effects of using a sequence model for German based on morphologically subcategorized parts-of-speech. All systems have been specified as extensions of the Moses system provided for the Shared Task.

2 Part-of-speech and Morphology

For both English and German we used the part-of-speech tagger TreeTagger (Schmid, 1994) to obtain POS-tags.

The German POS-tags from TreeTagger were refined by adding morphological information from a commercial dependency parser, including case, number, gender, definiteness, and person for nouns, pronouns, verbs, adjectives and determiners in the cases where both tools agreed on the POS-tag. If they did not agree, the POS-tag from TreeTagger was chosen. This tag set seemed more suitable for SMT, with tags for proper names and foreign words which the commercial parser does not have.

3 Compound Analysis

Compounding is common in many languages, including German. Since compounding is highly productive it increases vocabulary size and leads to sparse data problems.

Compounds in German are formed by joining words, and in addition filler letters can be inserted or letters can be removed from the end of all but the last word of the compound (Langer, 1998). We have chosen to allow simple additions of letter(s) (*-s, -n, -en, -nen, -es, -er, -ien*) and simple truncations (*-e,*

Proceedings of the Third Workshop on Statistical Machine Translation, pages 135–138,
Columbus, June 2008.

-en, -n). Example of compounds with additions and truncations can be seen in (1).

(1) a. Staatsfeind (Staat + Feind)
public enemy

b. Kirchhof (Kirche + Hof)
graveyard

3.1 Splitting compounds

Noun and adjective compounds are split by a modified version of the corpus-based method presented by Koehn and Knight (2003). First the German language model data is POS-tagged and used to calculate frequencies of all nouns, verbs, adjectives, adverbs and the negative particle. Then, for each noun and adjective all splits into these known words from the corpus, allowing filler additions and truncations, are considered, choosing the splitting option with the highest arithmetic mean[1] of the frequencies of its parts.

A length limit of each part was set to 4 characters. For adjectives we restrict the number of parts to maximum two, since they do not tend to have multiple parts as often as nouns. In addition we added a stop list with 14 parts, often mistagged, that gave rise to wrong adjective splits, such as *arische* ('Aryan') in *konsularische* ('consular').

As Koehn and Knight (2003) points out, parts of compounds do not always have the same meaning as when they stand alone, e.g. *Grundrechte* ('basic rights'), where the first part, *Grund*, usually translates as *foundation*, which is wrong in this compound. To overcome this we marked all compound parts but the last, with the symbol '#'. Thus they are handled as separate words. Parts of split words also receive a special POS-tag, based on the POS of the last word of the compound, and the last part receives the same POS as the full word.

We also split words containing hyphens based on the same algorithm. Their parts receive a different POS-tag, and the hyphens are left at the end of all but the last part.

[1]We choose the arithmetic mean over the geometric mean used by Koehn and Knight (2003) in order to increase the number of splits.

3.2 Merging compounds

For translation into German, the translation output contains split compounds, which need to be merged. An algorithm for merging has been proposed by Popović et al. (2006) using lists of compounds and their parts. This method cannot merge unseen compounds, however, so instead we base merging on POS. If a word has a compound-POS, and the following word has a matching POS, they are merged. If the next POS does not match, a hyphen is added to the word, allowing for coordinated compounds as in (2).

(2) Wasser- und Bodenqualität
water and soil quality

4 System Descriptions

The main difference of our system in relation to the baseline system of the Shared Task[2] is the pre- and post-processing described above, the use of a POS factor, and an additional sequence model on POS. We also modified the tuning to include compound merging, and used a smaller corpus, 600 sentences picked evenly from the dev2006 corpus, for tuning. We use the Moses decoder (Koehn et al., 2007) and SRILM language models (Stolcke, 2002).

4.1 German ⇒ English

We used POS as an output factor, as can be seen in Figure 1. Using additional factors only on the target side means that only the training data need to be POS-tagged, not the tuning data or translation input. However, POS-tagging is still performed for German as input to the pre-processing step. As Figure 1 shows we have two sequence models. A 5-gram language model based on surface form using Kneser-Ney smoothing and in addition a 7-gram sequence model based on POS using Witten-Bell[3] smoothing.

The training corpus was filtered to sentences with 2–40 words, resulting in a total of 1054688 sentences. Training was done purely on Europarl data, but results were submitted both on Europarl and

[2]`http://www.statmt.org/wmt08/baseline.html`

[3]Kneser-Ney smoothing can not be used for the POS sequence model, since there were counts-of-counts of zero. However, Witten-Bell smoothing gives good results when the vocabulary is small.

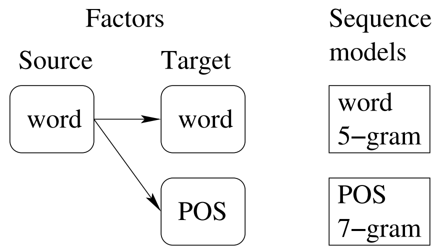

Figure 1: Architecture of the factored system

News data. The news data were submitted to see how well a pure out-of-domain system could perform.

In the pre-processing step compounds were split. This was done for training, tuning and translation. In addition German contracted prepositions and determiners, such as *zum* from *zu dem* ('to the'), when identified as such by the tagger, were split.

4.2 English ⇒ German

All features of the German to English system were used, and in addition more fine-grained German POS-tags that were sub-categorized for morphological features. This was done for training, tuning and sequence models. At translation time no pre-processing was needed for the English input, but a post-processing step for the German output is required, including the merging of compounds and contracted prepositions and determiners. The latter was done in connection with uppercasing, by training an instance of Moses on a lower cased corpus with split contractions and an upper-cased corpus with untouched contractions. The tuning step was modified so that merging of compounds were done as part of the tuning.

4.3 Baseline

For comparison, we constructed a baseline according to the shared-task description, but with smaller tuning corpus, and the same sentence filtering for the translation model as in the submitted system, using only sentences of length 2-40.

In addition we constructed a factored baseline system, with POS as an output factor and a sequence model for POS. Here we only used the original POS-tags from TreeTagger, no additional morphology was added for German.

	De-En	En-De
Baseline	26.95	20.16
Factored baseline	27.43	20.27
Submitted system	27.63	20.46

Table 1: Bleu scores for Europarl (test2007)

	De-En	En-De
Baseline	19.54	14.31
Factored baseline	20.16	14.37
Submitted system	20.61	14.77

Table 2: Bleu scores for News Commentary (nc-test2007)

5 Results

Case-sensitive Bleu scores[4] (Papineni et al., 2002) for the Europarl devtest set (test2007) are shown in table 1. We can see that the submitted system performs best, and that the factored baseline is better than the pure baseline, especially for translation into English.

Bleu scores for News Commentary[5] (nc-test2007) are shown in Table 2. Here we can also see that the submitted system is the best. As expected, Bleu is much lower on out-of-domain news text than on the Europarl development test set.

5.1 Compounds

The quality of compound translations were analysed manually. The first 100 compounds that could be found by the splitting algorithm were extracted from the Europarl reference text, test2007, together with their English translations[6].

System translations were compared to the annotated compounds and classified into seven categories: correct, alternative good translation, correct but different form, part of the compound translated, no direct equivalent, wrong and untranslated. Out of these the first three categories can be considered good translations.

We performed the error analysis for the submitted and the baseline system. The result can be seen in

[4] The %Bleu notation is used in this report

[5] No development test set for News test were provided, so we present result for the News commentary, which can be expected to give similar results.

[6] The English translations need not be compounds. Compounds without a clear English translation were skipped.

	De ⇒ En		En ⇒ De	
	Subm	Base	Subm	Base
Correct	50	46	40	39
Alternative	36	26	32	29
Form	5	7	6	8
Part	2	5	10	15
No equivalent	6	2	8	5
Wrong	1	7	1	1
Untranslated	–	7	3	3

Table 3: Results of the error analysis of compound translations

Table 3. For translation into English the submitted system handles compound translations considerably better than the baseline with 91% good translations compared to 79%. In the submitted system all compounds have a translation, compared to the baseline system which has 7% of the compounds untranslated. In the other translation direction the difference is smaller, the biggest difference is that the submitted system has fewer cases of partial translation.

5.2 Agreement in German NPs

To study the effects of using fine-grained POS-tags in the German sequence model, a similar close study of German NPs was performed. 100 English NPs having at least two dependents of the head noun were selected from a randomly chosen subsection of the development test set. Their translations in the baseline and submitted system were then identified. Translations that were not NPs were discarded. In about two thirds (62 out of 99) of the cases, the translations were identical. For the remainder, 12 translations were of equal quality, the submitted system had a better translation in 17 cases (46%), and a worse one in 8 cases (22%). In the majority of cases where the baseline was better, this was due to word selection, not agreement.

6 Conclusions

Adding morphological processing improved translation results in both directions for both text types. Splitting compounds gave a bigger effect for translation from German. Marking of compound parts worked well, with no untranslated parts left in the sample used for evaluation. The mini-evaluation of German NPs in English-German translation indicates that the morphologically rich POS-based sequence model for German also had a positive effect.

Acknowledgement

We would like to thank Joe Steinhauer for help with the evaluation of German output.

References

M. Collins, P. Koehn, and I. Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 531–540, Ann Arbor, Michigan.

M. Holmqvist, S. Stymne, and L. Ahrenberg. 2007. Getting to know Moses: Initial experiments on German-English factored translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 181–184, Prague, Czech Republic. Association for Computational Linguistics.

P. Koehn and K. Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the tenth conference of EACL*, pages 187–193, Budapest, Hungary.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, demonstration session*, Prague, Czech Republic.

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

K. Papineni, S. Roukos, T. Ward, and W.-J. 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.

M. Popović, D. Stein, and H. 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.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Preoceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, Colorado.