Getting to know Moses: Initial experiments on German–English factored translation

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

Abstract

We present results and experiences from our experiments with phrase-based statistical machine translation using Moses. The paper is based on the idea of using an off-the-shelf parser to supply linguistic information to a factored translation model and compare the results of German–English translation to the shared task baseline system based on word form. We report partial results for this model and results for two simplified setups. Our best setup takes advantage of the parser's lemmatization and decompounding. A qualitative analysis of compound translation shows that decompounding improves translation quality.

1 Introduction

One of the stated goals for the shared task of this workshop is "to offer newcomers a smooth start with hands-on experience in state-of-the-art statistical machine translation methods". As our previous research in machine translation has been mainly concerned with rule-based methods, we jumped at this offer.

We chose to work on German-to-English translation for two reasons. Our primary practical interest lies with translation between Swedish and English, and of the languages offered for the shared task, German is the one closest in structure to Swedish. While there are differences in word order and morphology between Swedish and German, there are also similarities, e.g., that both languages represent nominal compounds as single orthographic words. We chose the direction from German to English because our knowledge of English is better than our knowledge of German, making it easier to judge the quality of translation output. Experiments were performed on the Europarl data.

With factored statistical machine translation, different levels of linguistic information can be taken into account during training of a statistical translation system and decoding. In our experiments we combined syntactic and morphological factors from an off-the-shelf parser with the factored translation framework in Moses (Moses, 2007). We wanted to test the following hypotheses:

- Translation models based on lemmas will improve translation quality (Popovič and Ney, 2004)
- Decompounding German nominal compounds will improve translation quality (Koehn and Knight, 2003)
- Re-ordering models based on word forms and parts-of-speech will improve translation quality (Zens and Ney, 2006).

2 The parser

The parser, *Machinese Syntax*, is a commercially available dependency parser from Connexor Oy [1]. It provides each word with lemma, part-of-speech, morphological features and dependency relations (see Figure 1). In addition, the lemmas of compounds are marked by a '#' separating the two parts of the compound. For the shared task we only used shallow linguistic information: lemma, part-of-speech and morphology. The compound boundary identification was used to split noun com-

[1] Connexor Oy, http://www.connexor.com.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 181–184,
Prague, June 2007. ©2007 Association for Computational Linguistics

pounds to make the German input more similar to English text.

```
1 Mit   mit   pm>2     @PREMARK PREP
2 Blick blick advl>10 @NH N MSC SG DAT
3 auf   auf   pm>5     @PREMARK PREP
```

Figure 1. Example of parser output

We used the parser's tokenization as given. Some common multiword units, such as 'at all' and 'von heute', are treated as single words by the parser (cf. Niessen and Ney, 2004). The German parser also splits contracted prepositions and determiners like 'zum' – 'zu dem' ("to the").

3 System description

For our experiments with Moses we basically followed the shared task baseline system setup to train our factored translation models. After training a statistical model, minimum error-rate tuning was performed to tune the model parameters. All experiments were performed on an AMD 64 Athlon 4000+ processor with 4 Gb of RAM and 32 bit Linux (Ubuntu).

Since time as well as computer resources were limited we designed a model that we hoped would make the best use of all available factors. This model turned out to be too complex for our machine and in later experiments we abandoned it for a simpler model.

3.1 Pre-processing

In the pre-processing step we used the standard pre-processing of the shared task baseline system, parsed the German and English texts and processed the output to obtain four factors: word form, lemma, part-of-speech and morphology. Missing values for lemma, part-of-speech and morphology were replaced with default values.

Noun compounds are very frequent in German, 2.9% of all tokens in the tuning corpus were identified by the parser as noun compounds. Compounds tend to lead to sparse data problems and splitting them has been shown to improve German-English translation (Koehn and Knight, 2003). Thus we decided to decompund German noun compounds identified as such by our parser.

We used a simple strategy to remove fillers and to correct some obvious mistakes. We removed the filler '-s' that appear before a marked split unless it was one of '-ss', '-urs', '-eis' or '-us'. This applied to 35% of the noun compounds in the tuning corpus. The fillers were removed both in the word form and the lemma (see Figure 2).

There were some mistakes made by the parser, for instance on compounds containing the word 'nahmen' which was incorrectly split as 'stellungn#ahmen' instead of 'stellung#nahmen' ("statement"). These splits were corrected by moving the 'n' to the right side of the split.

We then split noun-lemmas on hyphens unless there were numbers on either side of it and on the places marked by '#'. Word forms were split in the corresponding places as the lemmas.

The part-of-speech and morphology of the last word in the compound is the same as for the whole compound. For the other parts we hypothesized that part-of-speech is Noun and the morphology is unknown, marked by the tag *UNK*.

```
Parser output:
unionsländer unions#land N NEU PL ACC

Factored output:
union|union|N|UNK
länder|land|N|NEU_PL_ACC
```

Figure 2. Compound splitting for 'unionsländer' ("countries in the union")

These strategies are quite crude and could be further refined by studying the parser output thoroughly to pinpoint more problems.

3.2 Training translation models with linguistic factors

After pre-processing, the German–English Europarl training data contains four factors: 0: word form, 1: lemma, 2: part-of-speech, 3: morphology. As a first step in training our translation models we performed word alignment on lemmas as this could potentially improve word alignment.

3.2.1 First setup

Factored translation requires a number of decoding steps, which are either *mapping* steps mapping a source factor to a target factor or *generation* steps generating a target factor from other target factors. Our first setup contained three mapping steps, T0–T2, and one generation step, G0.

T0: 0-0 (word – word)
T1: 1-1 (lemma – lemma)
T2: 2,3-2,3 (pos+morph – pos+morph)
G0: 1,2,3-0 (lemma+pos+morph – word)

With the generation step, word forms that did not appear in the training data may still get translated if the lemma, part-of-speech and morphology can be translated separately and the target word form can be generated from these factors.

Word order varies a great deal between German and English. This is especially true for the placement of verbs. To model word order changes we included part-of-speech information and created two reordering models, one based on word form (0), the other on part-of-speech (2):

0-0.msd-bidirectional-fe
2-2.msd-bidirectional-fe

The decoding times for this setup turned out to be unmanageable. In the first iteration of parameter tuning, decoding times were approx. 6 min/sentence. In the second iteration decoding time increased to approx. 30 min/sentence. Removing one of the reordering models did not result in a significant change in decoding time. Just translating the 2000 sentences of test data with untuned parameters would take several days. We interrupted the tuning and abandoned this setup.

3.2.2 Second setup

Because of the excessive decoding times of the first factored setup we resorted to a simpler system that only used the word form factor for the translation and reordering models. This setup differs from the shared task baseline in the following ways: First, it uses the tokenization provided by the parser. Second, alignment was performed on the lemma factor. Third, German compounds were split using the method described above. To speed up tuning and decoding, we only used the first 200 sentences of development data (dev2006) for tuning and reduced stack size to 50.

T0: 0-0 (word – word)
R: 0-0.msd-bidirectional-fe

3.2.3 Third setup

To test our hypothesis that word reordering would benefit from part-of-speech information we created another simpler model. This setup has two mapping steps, T0 and T1, and a reordering model based on part-of-speech.

T0: 0-0 (word – word)
T1: 2,3-2,3 (pos+morph – pos+morph)
R: 2-2.msd-bidirectional-fe

4 Results

We compared our systems to a baseline system with the same setup as the WMT2007 shared task baseline system but tuned with our system's simplified tuning settings (200 instead of 2000 tuning sentences, stack size 50). Table 1 shows the Bleu improvement on the 200 sentences development data from the first and last iteration of tuning.

System	Dev2006 (200)	
	1st iteration	Last iteration
Baseline	19.56	27.07
First	21.68	-
Second	20.43	27.16
Third	20.72	24.72

Table 1. Bleu scores on 200 sentences of tuning data before and after tuning

The final test of our systems was performed on the development test corpus (devtest2006) using stack size 50. The results are shown in Table 2. The low Bleu score for the third setup implies that reordering on part-of-speech is not enough on its own. The second setup performed best with a slightly higher Bleu score than the baseline. We used the second setup to translate test data for our submission to the shared task.

System	Devtest2006 (NIST/Bleu)
Baseline	6.7415 / 25.94
First	-
Second	6.8036 / 26.04
Third	6.5504 / 24.57

Table 2. NIST and Bleu scores on development test data

4.1 Decompounding

We have evaluated the decompounding strategy by analyzing how the first 75 identified noun compounds of the devtest corpus were translated by our second setup compared to the baseline. The sample excluded doubles and compounds that had no clear translation in the reference corpus.

Out of these 75 compounds 74 were nouns that were correctly split and 1 was an adjective that was split incorrectly: ‘allumfass#ende’. Despite that it was incorrectly identified and split it was translated satisfyingly to ‘comprehensive’.

The translations were grouped into the categories shown in Table 3. The 75 compounds were classified into these categories for our second system and the baseline system, as shown in Table 4. As can be seen the compounds were handled better by our system, which had 62 acceptable translations (C or V) compared to 48 for the baseline and did not leave any noun compounds untranslated.

Category	Example
C-correct	Regelungsentwurf Draft regulation Ref: Draft regulation
V-variant	Schlachthöfen Abattoirs Ref: Slaughter houses
P-partly correct	Anpassungsdruck Pressure Ref: Pressure for adaption
F-wrong form	Länderberichte Country report Ref: Country reports
W-wrong	Erbonkel Uncle dna Ref: Sugar daddy
U-untranslated	Schlussentwurf Schlussentwurf Ref: Final draft

Table 3. Classification scheme with examples for compound translations

		Baseline system						
		C	V	P	W	U	F	Tot
Second system	C	36	1	3		3	1	44
	V	1	9	2	1	5		18
	P			3		2		5
	W				1	2		3
	U							0
	F	1					4	5
	Tot	38	10	8	2	12	5	75

Table 4. Classification of 75 compounds from our second system and the baseline system

Decompounding of nouns reduced the number of untranslated words, but there were still some left. Among these were cases that can be handled such as separable prefix verbs like ‘aufzeigten’ (“pointed out”) (Niessen and Ney, 2000) or adjective compounds such as ‘multidimensionale’ (“multi dimensional”). There were also some noun compounds left which indicates that we might need a better decompounding strategy than the one used by the parser (see e.g. Koehn and Knight, 2003).

4.2 Experiences and future plans

With the computer equipment at our disposal, training of the models and tuning of the parameters turned out to be a very time-consuming task. For this reason, the number of system setups we could test was small, and much fewer than we had hoped for. Thus it is too early to draw any conclusions as regards our hypotheses, but we plan to perform more tests in the future, also on Swedish–English data. The parser's ability to identify compounds that can be split before training seems to give a definite improvement, however, and is a feature that can likely be exploited also for Swedish-to-English translation with Moses.

References

Koehn, Philipp and Kevin Knight, 2003. Empirical methods for compound splitting. In *Proceedings of EACL 2003*, 187-194. Budapest, Hungary.

Moses – a factored phrase-based beam-search decoder for machine translation. 13 April 2007, URL: http://www.statmt.org/moses/ .

Niessen, Sonja and Hermann Ney, 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 181-204 .

Niessen, Sonja and Hermann Ney, 2000. Improving SMT Quality with Morpho-syntactic Analysis. In *Proceedings of Coling 2000*. 1081-1085. Saarbrücken, Germany.

Popovič, Maja and Hermann Ney, 2004. Improving Word Alignment Quality using Morpho-Syntactic Information. In *Proceedings of Coling 2004,* 310-314, Geneva, Switzerland.

Zens, Richard and Hermann Ney, 2006. Discriminative Reordering Models for Statistical Machine Translation. In *HLT-NAACL: Proceedings of the Workshop on Statistical Machine Translation*, 55-63, New York City, NY.