Experiments in morphosyntactic processing for translating to and from German

Alexander Fraser
Institute for Natural Language Processing
University of Stuttgart
`fraser@ims.uni-stuttgart.de`

Abstract

We describe two shared task systems and associated experiments. The German to English system used reordering rules applied to parses and morphological splitting and stemming. The English to German system used an additional translation step which recreated compound words and generated morphological inflection.

1 Introduction

The Institute for Natural Language Processing (IfNLP), Stuttgart, participated in the WMT-2009 shared tasks for German to English and English to German translation with constrained systems which employed morphological and syntactic processing techniques. The systems were based on the open source Moses docoder (Koehn et al., 2007). We combined IfNLP tools for syntactic and morphological analysis (which are publicly available and widely used) with preprocessing techniques that were successfully used by other groups in WMT-2008, and extended these. For English to German translation, we additionally performed a step which recreated compound words and generated morphological inflection.

1.1 Baseline

The baseline is the standard system supplied for the shared task. We used the default parameters of the Moses toolkit, except for a small difference in the generation of the word alignments, see section 3.

2 Improvements

2.1 Character Normalization

We normalize both the English and German by converting all characters to their nearest equivalent in Latin-1 (ISO 8859-1) encoding[1], except for the euro sign, which is handled specially. We did not modify the SGML files used for calculating BLEU and METEOR scores in any way.

2.2 German Writing Reform

German underwent a writing reform from the *alte Rechtschreibung* (old spelling rules/orthography) to the *neue Rechtschreibung* (gloss: new spelling rules/orthography) recently. Early Europarl data are written using the *alte Rechtschreibung* and hence need to be converted to the *neue Rechtschreibung* in order to match the news data, which is in the new form.

We began the process by mapping all cased variants of a particular word to a single class (such as by mapping two words which are written with *ue* and *ü*, but are otherwise identical, to a single class). We then tried to automatically identify the correct variant under the writing reform for each class. Initially we tried the linux tool *aspell* but found that its coverage (the recall of its lexicon) was poor.

We used a simple technique for finding the best variant. We separated the Europarl corpus into portions written using the old and new forms. We used the incidence of the word *dass* (the complementizer meaning *that*) and its old rules variant *daß*. We used a chunk size of 70 sentences to segment Europarl into old and new by counting whether there were more instances of *daß* or *dass*, respectively, in each chunk. We added the news corpora to the new portion. For each variant we counted the number of times it occurred in the new data and subtracted the number of times it occurred in the old data; the variant with the highest adjusted count was selected.

---

[1] Latin-1 is an 8-bit encoding which has the common accented characters used in Western European languages. A reviewer pointed out that ISO 8859-15 has superseded ISO 8859-1.

2.3 Reordering German

German word order differs from English substantially. Preprocessing approaches involving the use of a syntactic parse of the source sentence to change the word order to more closely match the word order of the target language have been studied by Niessen and Ney (2004), Xia and McCord (2004), Drábek and Yarowsky (2004), Collins et al. (2005), Popović and Ney (2006), Wang et al. (2007) and many others.

To obtain a parse of each German sentence in the training, dev and test corpora, we employed the IfNLP BitPar probabilistic parser (Schmid, 2004), using models learned from the Tiger Treebank for German.

Dealing with morphological productivity is important in the syntactic parsing of German. BitPar has been designed with this in mind. IfNLP's SMOR analyzer is used for morphological analysis (Schmid et al., 2004). SMOR is run over a list of types in each German sentence, and outputs a list of analyses for each type, each of which corresponds to a POS tag. BitPar is limited to choosing one of these POS tags for this type. Words which SMOR fails to analyze are allowed to occur with any POS tag.

We reimplemented the syntactic preprocessing approach of Collins et al. (2005), with modifications. Reordering rules are applied to a German parse tree (generated by BitPar), and focus on reordering the words in the German clause structure to more closely resemble English clause structure. The rules are applied to both the training data for the SMT system, and the input (the dev and test sets). We previously performed an error analysis of this approach and for the work described here we addressed some of the shortcomings identified through the analysis. The analysis was performed on the Europarl dev2006 set.

The first error that we noticed occurring frequently was that some large clausal units which were labeled as subjects were being moved forward in the sentence. We modified the rule moving subjects forward to not apply to the constituents *S*, *CS*, *VP* and *CVP*. See the first part of table 3 for an example. The phrase "dass der Balkan ist kein Gebiet" is moved under the original rules, and with the modification is no longer moved[2].

---

[2]Note that there is an unrelated reordering error at the end of the sentence for both BEFORE and AFTER, *gibt* (gloss: gives) should have moved to follow *das* (gloss: that).

System	BLEU	METEOR	LR
no processing	18.91	49.50	1.0097
c+w	19.37	49.69	1.0067
c+w, s/s	19.18	51.13	1.0035
c+w, old reordering	19.61	50.44	1.0092
c+w, new reordering	19.91	50.84	1.0059
c+w, new reordering, s/s (submitted, bug)	19.65	51.57	1.0093
* c+w, new reordering, s/s	19.73	51.59	1.0062
as * IRSTLM quantized	19.52	51.33	1.0003
as * IRSTLM	19.75	51.61	1.0013
as * IRSTLM 21.2 quantized	19.52	51.51	1.0095
as * RANDLM	19.67	51.73	1.0067
as * RANDLM 21.2	21.03	51.96	1.0111

Table 1: German to English, dev-2009b (case sensitive), c+w = char+word normalization, s/s = splitting/stemming, 21.2 = larger LM

System	BLEU	METEOR	LR
no processing	13.55	38.31	0.9910
c+w (no second step)	14.11	38.27	0.9991
c+w, s/s, second step (submitted, bug)	12.34	37.89	1.0338
c+w, s/s, second step	13.05	37.94	1.0157

Table 2: English to German, dev-2009b (case sensitive), c+w = char+word normalization, s/s = splitting/stemming

The second error that we handled was that *S-RC* constituents which do not have a complementizer are reordered incorrectly. We modified the original verb 2nd rule, so that if there is no complementizer in a *S-RC* constituent, then the head is moved to the second position, see the second part of table 3 for an example. Using the original rules, the verb 2nd rule fails to fire, incorrectly leaving *haben* (gloss: have) at the end of the clause.

2.4 Morphological Decomposition

We implemented the frequency-based word splitting approach of Koehn and Knight (2003), and made modifications, including some similar to those described by Stymne et al. (2008). This well-known technique splits compound words. In addition, we performed simple suffix elimination, aimed at removing inflection marking features such as gender and case that are not necessary for translation to English. We took the stem combination with the highest geometric mean of the frequencies of the stems, but following Stymne et al. (2008), we restricted stems to minimum length 4, and we allowed an extended list of infixes: *s*, *n*, *en*, *nen*, *es*, *er* and *ien*. For suffixes, we allowed: *e*, *en*, *n*, *es*, *s*, *em* and *er*, which is more aggressive

116

INPUT	Mir ist bewusst , dass der Balkan kein Gebiet ist , das Anlass zu Optimismus gibt .
gloss	me is clear , that the Balkans not area is , that opportunity for optimism gives .
BEFORE	Mir dass der Balkan ist kein Gebiet ist bewusst , , das Anlass zu Optimismus gibt .
gloss	me that the Balkans is not area is clear , that opportunity for optimism gives .
AFTER	Mir ist bewusst , dass der Balkan ist kein Gebiet , das Anlass zu Optimismus gibt .
gloss	me is clear , that the Balkans is not area , that opportunity for optimism gives .
REF	I am aware that the Balkans are not the most promising area for optimism .
INPUT	Am 23. November 1999 hat ein Partner-schaftstag stattgefunden , an dem viele von uns teilgenommen haben .
gloss	on 23 November 1999 have a partnership-day took-place , in which many of us participated have .
BEFORE	Am 23. November 1999 ein Partner-schaftstag hat stattgefunden , an dem teilgenommen viele von uns haben .
gloss	on 23 November 1999 a partnership-day have took-place , in which participated many of us have .
AFTER	Am 23. November 1999 ein Partner-schaftstag hat stattgefunden , an dem viele von uns haben teilgenommen .
gloss	on 23 November 1999 a partnership-day have took-place , in which many of us have participated .
REF	A partnership day was held on 23 November 1999 , in which many of us participated .

Table 3: Differences in reordering: BEFORE is reordering using rules in (Collins et al., 2005), AFTER is our modified reordering

than used in previous work (and therefore generalizes more but at the same time causes some erroneous conflation). We stripped *e*, *en* and *n* from all stems (but remembered the most frequent variant, so that applying the procedure to *Kirchturm* results in *Kirche Turm* (gloss: church tower)). We store an alignment from the original German to the simplified German which we will use in the next section.

2.5 Morphological Generation

For translation from English to German, we first translated from English to the simplified German presented in the previous section, and then performed an independent translation step from simplified German to fully inflected German.

Two processes are handled by this step. First, series of stems corresponding to compound words are recomposed (along with infixes which are not present in the simplified German form) into compound words. Second, inflection is added (e.g., case and gender agreement is handled). Both of these processes are implemented using a Moses system trained on a parallel corpus where the source language is simplified German and the target language is fully inflected German. The alignment is error-free as it was generated as a side effect of the splitting and stemming process described in the previous section. In translation, reordering is not allowed, but we otherwise use standard Moses settings.

3 Experiments

3.1 German to English

We trained our German to English system on the constrained parallel data. The English data was processed using character normalization. The German data was first processed using character and word (writing reform) normalization. We then parsed the German data using BitPar and applied the modified reordering rules. After this the splitting and stemming process was applied. Finally, we lowercased the data.

Word alignments were generated using Model 4 (Brown et al., 1993) using the multi-threaded implementation of GIZA++ (Och and Ney, 2003; Gao and Vogel, 2008). We first trained Model 4 with English as the source language, and then with German as the source language, resulting in two Viterbi alignments[3]. The resulting Viterbi alignments were combined using the *Grow Diag Final And* symmetrization heuristic (Koehn et al., 2003). We estimated a standard Moses system using default settings. MERT was run until convergence using dev-2009a (separately for each experiment).

One limitation of our German to English system is that we were unable to scale to the full language modeling data using SRILM (Stolcke, 2002), 5-grams and modified Kneser-Ney with no singleton deletion[4]. The language model in our submitted system is based on all of the available English data, but news-train08 is truncated to the first 10193376 lines, meaning that we did not train on the remaining 11038787 lines, so we used a little less than half of the data. We converted the lan-

---

[3]We used 5 iterations of Model 1, 4 iterations of HMM (Vogel et al., 1996) and 4 iterations of Model 4.

[4]SRILM failed when trained on the full data, even when a machine with 32 GB RAM and 48 GB swap was used.

guage model trained using SRILM to the binary format using IRSTLM.

Experiments are presented in table 1, using BLEU (Papineni et al., 2001) and METEOR[5] (Banerjee and Lavie, 2005), and we also show the length ratio (ratio of hypothesized tokens to reference tokens). For translation into English METEOR had superior correlation with human rankings to BLEU at WMT 2008 (Callison-Burch et al., 2008). Our submitted system had a bug where the environment variable *LC_ALL* was set to *en_US* when creating the binarized filtered lexicalized reordering table for the test set (and for the blindtest set, but not for the dev set used for MERT). This caused minor degradation, see the system marked (*) for the system with the bug corrected.

Each system increases in both BLEU and METEOR as improvements are added. An exception is that splitting/stemming decreases BLEU somewhat. However, we trust the METEOR results more due to their better correlation with human judgements.

We also compared using a different language model instead of the SRILM model (the bottom half of table 1). These used either the reduced English language modeling data or the full data (21.2 M segments, marked *21.2* in the results). RANDLM (Talbot and Osborne, 2007) performs well and scaled to the full data with improvement (resulting in our best overall system). IRSTLM (Federico and Cettolo, 2007) also performs well, but the quantized model on the 21.2 data did not improve over the smaller quantized model[6]. IRSTLM uses an approximation of Witten-Bell smoothing, our results support that this is competitive.

3.2 English to German

We trained our English to German system on the constrained parallel data. The first SMT system translates from lowercased English to lowercased simplified German, which is then recased. The syntactic reordering process is not used, but otherwise the German data is processed identically. The alignment from simplified German to English is generated as described in the previous section. We used all of the German data to train the language

model on simplified German. The second SMT system translates mixed case simplified German to mixed case unsimplified German. The translation model is built only on the simplified German from the parallel text, and the language model is trained on all German data.

We present the results in table 2. METEOR[7] did not correlate as well as BLEU for translation out of English in WMT 2008. The BLEU score of our final system is worse than the baseline. We had chosen to submit this system as we found it more interesting than submitting a vanilla system. In addition, the system of Stymne et al. (2008) received a good human evaluation despite having a relatively low BLEU score, and we hoped we were performing similar morphological generalization. We expect to be able to improve this system through error analysis. In an initial inspection we found case mismatching problems between step one and step two.

4   Conclusion

We presented our German to English system which employed character normalization, compensated for problems caused by the German writing reform, used modified syntactic reordering rules (in combination with morphologically aware parsing), and employed substring-based morphological analysis. Our best system improves by 2.46 METEOR and 1.12 BLEU over a standard Moses system. Our English to German system used the same two normalizations and the substring-based morphological analysis, and additionally implemented a second translation step for recreating compound words and generating case and gender inflection. We will improve this system in future work.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter

---

[5]METEOR used default weights, stemming and Wordnet synsets.

[6]After speaking with the authors, we plan to try IRSTLM on the full data using memory mapping for binarization.

[7]METEOR for this task is calculated using default weights but no Wordnet synsets.

estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *ACL Third Workshop on Statistical Machine Translation*, Columbus, Ohio.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL*, pages 531–540, Ann Arbor, MI.

Elliott F. Drábek and David Yarowsky. 2004. Improving bitext word alignments via syntax-based reordering of English. In *The Companion Volume to the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 146–149, Barcelona, Spain.

Marcello Federico and Mauro Cettolo. 2007. Efficient handling of n-gram language models for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 88–95, Prague, Czech Republic.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL*, pages 187–193, Morristown, NJ.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*, pages 127–133, Edmonton, Canada.

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 *ACL, Demonstration Program*, Prague, Czech Republic.

Sonja Niessen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

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

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY.

Maja Popović and Hermann Ney. 2006. POS-based word reorderings for statistical machine translation. In *LREC*, pages 1278–1283, Genoa, Italy.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: a German computational morphology covering derivation, composition, and inflection. In *LREC*, pages 1263–1266, Lisbon, Portugal.

Helmut Schmid. 2004. Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *COLING*, Geneva, Switzerland.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Intl. Conf. Spoken Language Processing*, 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.

David Talbot and Miles Osborne. 2007. Randomised language modelling for statistical machine translation. In *ACL*, pages 512–519, Prague, Czech Republic.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING*, pages 836–841, Copenhagen, Denmark.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *EMNLP-CONLL*, pages 737–745.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *COLING*.