The RWTH Machine Translation System for WMT 2009

Maja Popović, David Vilar, Daniel Stein, Evgeny Matusov and Hermann Ney
RWTH Aachen University
Aachen, Germany

Abstract

RWTH participated in the shared translation task of the Fourth Workshop of Statistical Machine Translation (WMT 2009) with the German-English, French-English and Spanish-English pair in each translation direction. The submissions were generated using a phrase-based and a hierarchical statistical machine translation systems with appropriate morpho-syntactic enhancements. POS-based reorderings of the source language for the phrase-based systems and splitting of German compounds for both systems were applied. For some tasks, a system combination was used to generate a final hypothesis. An additional English hypothesis was produced by combining all three final systems for translation into English.

1 Introduction

For the WMT 2009 shared task, RWTH submitted translations for the German-English, French-English and Spanish-English language pair in both directions. A phrase-based translation system enhanced with appropriate morpho-syntactic transformations was used for all translation directions. Local POS-based word reorderings were applied for the Spanish-English and French-English pair, and long range reorderings for the German-English pair. For this language pair splitting of German compounds was also applied. Special efforts were made for the French-English and German-English translation, where a hierarchical system was also used and the final submissions are the result of a system combination. For translation into English, an additional hypothesis was produced as a result of combination of the final German-to-English, French-to-English and Spanish-to-English systems.

2 Translation models

2.1 Phrase-based model

We used a standard phrase-based system similar to the one described in (Zens et al., 2002). The pairs of source and corresponding target phrases are extracted from the word-aligned bilingual training corpus. Phrases are defined as non-empty contiguous sequences of words. The phrase translation probabilities are estimated using relative frequencies. In order to obtain a more symmetric model, the phrase-based model is used in both directions.

2.2 Hierarchical model

The hierarchical phrase-based approach can be considered as an extension of the standard phrase-based model. In this model we allow the phrases to have "gaps", i.e. we allow non-contiguous parts of the source sentence to be translated into possibly non-contiguous parts of the target sentence. The model can be formalized as a synchronous context-free grammar (Chiang, 2007). The model also included some additional heuristics which have shown to be helpful for improving translation quality, as proposed in (Vilar et al., 2008).

The first step in the hierarchical phrase extraction is the same as for the phrased-based model. Having a set of initial phrases, we search for phrases which contain other smaller sub-phrases and produce a new phrase with gaps. In our system, we restricted the number of non-terminals for each hierarchical phrase to a maximum of two, which were also not allowed to be adjacent. The scores of the phrases are again computed as relative frequencies.

2.3 Common models

For both translation models, phrase-based and hierarchical, additional common models were used: word-based lexicon model, phrase penalty, word penalty and target language model.

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 66–69,
Athens, Greece, 30 March – 31 March 2009.

The target language model was a standard n-gram language model trained by the SRI language modeling toolkit (Stolcke, 2002). The smoothing technique we apply was the modified Kneser-Ney discounting with interpolation. In our case we used a 4-gram language model.

3 Morpho-syntactic transformations

3.1 POS-based word reorderings

For the phrase-based systems, the local and long range POS-based reordering rules described in (Popović and Ney, 2006) were applied on the training and test corpora as a preprocessing step.

Local reorderings were used for the Spanish-English and French-English language pairs in order to handle differences between the positions of nouns and adjectives in the two languages. Adjectives in Spanish and French, as in most Romanic languages, are usually placed after the corresponding noun, whereas for English it is the other way round. Therefore, for these language pairs local reorderings of nouns and adjective groups in the source language were applied. The following sequences of words are considered to be an adjective group: a single adjective, two or more consecutive adjectives, a sequence of adjectives and coordinate conjunctions, as well as an adjective along with its corresponding adverb. If the source language is Spanish or French, each noun is moved behind the corresponding adjective group. If the source language is English, each adjective group is moved behind the corresponding noun.

Long range reorderings were applied on the verb groups for the German-English language pair. Verbs in the German language can often be placed at the end of a clause. This is mostly the case with infinitives and past participles, but there are many cases when other verb forms also occur at the clause end. For the translation from German into English, following verb types were moved towards the beginning of a clause: infinitives, infinitives+zu, finite verbs, past participles and negative particles. For the translation from English to German, infinitives and past participles were moved to the end of a clause, where punctuation marks, subordinate conjunctions and finite verbs are considered as the beginning of the next clause.

3.2 German compound words

For the translation from German into English, German compounds were split using the frequency-based method described in (Koehn and Knight, 2003). For the other translation direction, the English text was first translated into the modified German language with split compounds. The generated output was then postprocessed, i.e. the components were merged using the method described in (Popović et al., 2006): a list of compounds and a list of components are extracted from the original German training corpus. If the word in the generated output is in the component list, check if this word merged with the next word is in the compound list. If it is, merge the two words.

4 System combination

For system combination we used the approach described in (Matusov et al., 2006). The method is based on the generation of a consensus translation out of the output of different translation systems. The core of the method consists in building a confusion network for each sentence by aligning and combining the (single-best) translation hypothesis from one MT system with the translations produced by the other MT systems (and the other translations from the same system, if n-best lists are used in combination). For each sentence, each MT system is selected once as "primary" system, and the other hypotheses are aligned to this hypothesis. The resulting confusion networks are combined into a signle word graph, which is then weighted with system-specific factors, similar to the approach of (Rosti et al., 2007), and a trigram LM trained on the MT hypotheses. The translation with the best total score within this word graph is selected as consensus translation. The scaling factors of these models are optimized using the Condor toolkit (Berghen and Bersini, 2005) to achieve optimal BLEU score on the dev set.

5 Experimental results

5.1 Experimental settings

For all translation directions, we used the provided EuroParl and News parallel corpora to train the translation models and the News monolingual corpora to train the language models. All systems were optimised for the BLEU score on the development data (the "dev-a" part of the 2008 evaluation data). The other part of the 2008 evaluation set ("dev-b") is used as a blind test set. The results reported in the next section will be referring to this test set. For the tasks including a system combination, the parameters for the system combination

were also trained on the "dev-b" set. The reported evaluation metrics are the BLEU score and two syntax-oriented metrics which have shown a high correlation with human evaluations: the PBLEU score (BLEU calculated on POS sequences) and the POS-F-score PF (similar to the BLEU score but based on the F-measure instead of precision and on arithmetic mean instead of geometric mean). The POS tags used for reorderings and for syntactic evaluation metrics for the English and the German corpora were generated using the statistical n-gram-based TnT-tagger (Brants, 2000). The Spanish corpora are annotated using the FreeLing analyser (Carreras et al., 2004), and the French texts using the TreeTagger[1].

5.2 Translation results

Table 1 presents the results for the German-English language pair. For translation from German into English, results for the phrase-based system with and without verb reordering and compound splitting are shown. The hierarchical system was trained with split German compounds. The final submission was produced by combining those five systems. The improvement obtained by system combination on the unseen test data 2009 is similar, i.e. from the systems with BLEU scores of 17.0%, 17.2%, 17.5%, 17.6% and 17.7% to the final system with 18.5%.

German→English	BLEU	PBLEU	PF
phrase-based	17.8	31.6	39.7
+reorder verbs	18.2	32.6	40.3
+split compounds	18.0	31.9	40.0
+reord+split	18.4	33.1	40.7
hierarchical+split	18.5	33.5	40.1
system combination	19.2	33.8	40.9

English→German	BLEU	PBLEU	PF
phrase-based	13.6	31.6	39.7
+reorder verbs	13.7	32.4	40.2
+split compounds	13.7	32.3	40.1
+reord+split	13.7	32.3	40.1
system combination	14.0	32.7	40.3

Table 1: Translation results [%] for the German-English language pair, News2008_dev-b.

The other translation direction is more difficult and improvements from morpho-syntactic transformations are smaller. No hierarchical system was trained for this translation direction. The combination of the four phrase-based systems leads to further improvements (on the unseen test set as well: contrastive hypotheses have the BLEU scores in the range from 12.7% to 13.0%, and the final BLEU score is 13.2%).

The results for the French-English language pair are shown in Table 2. For the French-to-English system, we submitted the result of the combination of three systems: a phrase-based with and without local reorderings and a hierarchical system. For the unseen test set, the BLEU score of the system combination output is 24.4%, whereas the contrastive hypotheses have 23.2%, 23.4% and 24.1%. For the other translation direction we did not use the system combination, the submission is produced by the phrase-based system with local adjective reorderings.

French→English	BLEU	PBLEU	PF
phrase-based	20.9	37.1	43.2
+reorder adjectives	21.3	38.2	43.6
hierarchical	20.3	36.7	42.6
system combination	21.7	38.5	43.8

English→French	BLEU	PBLEU	PF
phrase-based	20.2	39.5	45.9
+reorder adjectives	20.7	40.6	46.4

Table 2: Translation results [%] for the French-English language pair, News2008_dev-b.

Table 3 presents the results for the Spanish-English language pair. As in the English-to-French translation, the phrase-based system with adjective reorderings is used to produce the submitted hypothesis for both translation directions.

Spanish→English	BLEU	PBLEU	PF
phrase-based	22.1	38.5	44.1
+reorder adjectives	22.5	39.2	44.6

English→Spanish	BLEU	PBLEU	PF
phrase-based	20.6	29.3	35.7
+reorder adjectives	21.1	29.7	35.9

Table 3: Translation results [%] for the Spanish-English language pair, News2008_dev-b.

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

The result of the additional experiment, i.e. for the multisource translation int English is presented in Table 4. The English hypothesis is produced by the combination of the three best systems for each language pair, and it can be seen that the translation performance increases in all measures. This suggests that each language pair poses different difficulties for the translation task, and the combination of all three can improve performance.

F+S+G→English	BLEU	PBLEU	PF
system combination	25.1	41.0	46.4

Table 4: Multisource translation results [%]: the English hypothesis is obtained as result of a system combination of all language pairs, News2008_dev-b.

6 Conclusions

The RWTH system submitted to the WMT 2009 shared translation task used a phrase-based system and a hierarchical system with appropriate morpho-syntactic extensions, i.e. POS based word reorderings and splitting of German compounds were used. System combination produced gains in BLEU score over phrasal-system baselines in the German-to-English, English-to-German and French-to-English tasks.

Acknowledgments

This work was realised as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

Frank Vanden Berghen and Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175.

Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, pages 224–231, Seattle, WA.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings 4th International Conference on Language Resources and Evaluation (LREC)*, pages 239–242, Lisbon, Portugal, May.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, (33):201–228.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 347–354, Budapest, Hungary, April.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proceedings of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, pages 33–40, Trento, Italy, April.

Maja Popović and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1278–1283, Genoa, Italy, May.

Maja Popović, Daniel Stein, and Hermann Ney. 2006. Statistical machine translation of german compound words. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL)*, pages 616–624, Turku, Finland, August. Lecture Notes in Computer Science, Springer Verlag.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007. Combining Outputs from Multiple Machine Translation Systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York, April. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

David Vilar, Daniel Stein, and Hermann Ney. 2008. Analysing soft syntax features and heuristics for hierarchical phrase based machine translation. *International Workshop on Spoken Language Translation 2008*, pages 190–197, October.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, and G. Lakemeyer, editors, *25th German Conference on Artificial Intelligence (KI2002)*, volume 2479 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 18–32, Aachen, Germany, September. Springer Verlag.