Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System

Andreas Eisele[1,2], **Christian Federmann**[2], **Hervé Saint-Amand**[1],
Michael Jellinghaus[1], **Teresa Herrmann**[1], **Yu Chen**[1]
1: Saarland University, Saarbrücken, Germany
2: DFKI GmbH, Saarbrücken, Germany

Abstract

Based on an architecture that allows to combine statistical machine translation (SMT) with rule-based machine translation (RBMT) in a multi-engine setup, we present new results that show that this type of system combination can actually increase the lexical coverage of the resulting hybrid system, at least as far as this can be measured via BLEU score.

1 Introduction

(Chen et al., 2007) describes an architecture that allows to combine statistical machine translation (SMT) with one or multiple rule-based machine translation (RBMT) systems in a multi-engine setup. It uses a variant of standard SMT technology to align translations from one or more RBMT systems with the source text and incorporated phrases extracted from these alignments into the phrase table of the SMT system. Using this approach it is possible to employ a vanilla installation of the open-source decoder Moses[1] (Koehn et al., 2007) to find good combinations of phrases from SMT training data with the phrases derived from RBMT. A similar method was presented in (Rosti et al., 2007).

This setup provides an elegant solution to the fairly complex task of integrating multiple MT results that may differ in word order using only standard software modules, in particular GIZA++ (Och and Ney, 2003) for the identification of building blocks and Moses for the recombination, but the authors were not able to observe improvements in terms of BLEU score. A closer investigation revealed that the experiments had suffered from a couple of technical difficulties, such as mismatches in character encodings generated by different MT engines and similar problems. This motivated us to re-do these experiments in a somewhat more systematic way for this year's shared translation task, paying the required attention to all the technical details and also to try it out on more language pairs.

2 System Architecture

For conducting the translations, we use a multi-engine MT approach based on a ”vanilla” Moses SMT system with a modified phrase table as a central element. This modification is performed by augmenting the standard phrase table with entries obtained from translating the data with several rule-based MT systems. The resulting phrase table thus combines statistically gathered phrase pairs with phrase pairs generated by linguistic rules.

Basing its decision about the final translation on the obtained ”combined” phrase table, the SMT decoder searches for the best translation by recombining the building blocks that have been contributed by the different RBMT systems and the original SMT system trained on Europarl data.

A sketch of the overall architecture is given in Fig. 1, where the lighter parts represent the modules and data sets used in purely statistical MT, and the darker parts are the additional modules and data sets derived from the rule-based engines. The last word in the proposed setup is thus given to the SMT decoder, which can recombine (and potentially also tear apart) linguistically well-formed constructs

[1] see http://www.statmt.org/moses/

Proceedings of the Third Workshop on Statistical Machine Translation, pages 179–182,
Columbus, June 2008. ©2008 Association for Computational Linguistics

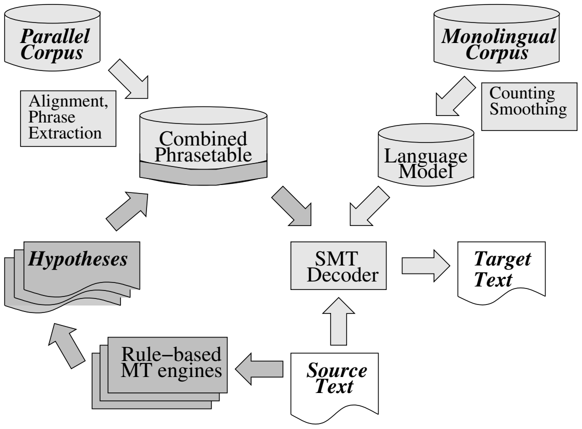

Figure 1: Hybrid architecture of the system

from the rule-based engines' output.

2.1 The Combined Phrase Table

The combined phrase table is built from the original Moses phrase table and separate phrase tables for each of the RBMT systems that are used in our setup. Since the original phrase table is created during the training process of the Moses decoder with the Europarl bilingual corpus as training material, it comprises general knowledge about typical constructions and vocabulary from the Europarl domain. Therefore, a standard Moses SMT system is, in principle, well adapted for input from this domain. However, it will have problems in dealing with vocabulary and structures that did not occur in the training data. The additional phrase tables are generated separately for each RBMT system from the source text and its translation by the respective system. By using a combined phrase table that includes the original Moses phrase table as well as the phrase tables from the RBMT systems, the hybrid system can both handle a wider range of syntactic constructions and exploit knowledge that the RBMT systems possess about the particular vocabulary of the source text.

3 Implementation

3.1 MT Systems and Knowledge Sources

Apart from the Moses SMT system, we used a set of six rule-based MT engines that are partly available via web interfaces and partly installed locally. The web interfaces are provided by Altavista Babelfish (based on Systran), SDL, ProMT and Lucy (a recent offspring of METAL). All of them deliver significantly different output translations. Locally installed systems are OpenLogos (for German↔English, English→Spanish and English→French) and translatePro by lingenio (for German↔English). The language model for our primary setup is based on the Europarl corpus whereas the English Gigaword corpus served as training data for a contrastive setup that was created for the translation direction German→English only.

3.2 Alignment of RBMT output

As already mentioned above, the construction of the RBMT system specific phrase tables is a major part of the overall system architecture. Such an RBMT phrase table is generated from a bilingual corpus consisting of the input text and its translation by the respective RBMT system. Because this corpus has the mere size of the text to be translated, it usually is not big enough to ensure the statistical methods for phrase table building of the Moses system to work. Therefore, we create the alignments between the RBMT input and output with help of another tool (Theison, 2007) that is based on knowledge learned in a previously conducted training phase with an appropriately bigger corpus. On the basis of the alignments created in this manner, the Moses training script provides a phrase table that consists of the source text vocabulary. These steps are carried out for each one of the six RBMT systems leading to six source text specific phrase tables which are then combined with the original Moses phrase table.

3.3 Combination of Phrase Tables

The combination process basically consists of the concatenation of the Moses phrase table and the previously created RBMT phrase tables with one minor adjustment: The phrase table resulting from this combination now also features additional columns indicating which system each phrase table entry originated from. For each new source text, the RBMT phrase tables have to be created from scratch and incorporated into a new combined phrase table.

3.4 Tuning

The typical process for creating an SMT system with the Moses toolkit includes a tuning step in which

	Europarl						NewsCommentary					
	de-en	en-de	fr-en	en-fr	es-en	en-es	de-en	en-de	fr-en	en-fr	es-en	en-es
SMT	22.81	19.78	24.18	21.62	31.68	24.46	14.24	9.75	11.60	12.24	17.27	14.48
Hybrid	**27.85**	**20.75**	**28.12**	**28.82**	**33.15**	**32.31**	17.36	13.57	17.66	20.71	**22.16**	**22.55**
RBMT1*	13.34	11.09	——	17.19	——	18.63	14.90	12.34	——	15.11	——	17.13
RBMT2	16.19	12.06	——	——	——	——	16.66	13.64	——	——	——	——
RBMT3	16.32	10.88	18.18	20.38	19.32	20.89	16.88	12.53	17.20	18.82	19.00	19.98
RBMT4	15.58	12.09	19.00	22.20	18.99	21.69	**17.41**	**13.93**	17.73	**20.85**	19.14	21.70
RBMT5	15.58	9.54	21.36	12.98	18.47	20.59	15.99	11.05	**18.65**	19.49	20.50	20.02
RBMT6	13.96	9.44	17.16	18.91	18.01	19.18	15.08	10.41	16.86	17.82	18.70	19.97

Table 1: Performance of baseline SMT system, our system and RBMT systems (BLEU scores)

the system searches for the best weight configuration for the columns in the phrase table while given a development set to be translated, and corresponding reference translations. In our hybrid setup, it is equally essential to conduct tuning since the combined phrase table we use contains 7 more columns than the original Moses phrase table. All these columns are given the same default weight initially and thus still need be to be tuned to more meaningful values. From this year's Europarl development data the first 200 sentences of each of the data sets dev2006, test2006, test2007 and devtest2006 were concatenated to build our development set. This set of 800 sentences was used for Minimum Error Rate Training (Och, 2003) to tune the weights of our system with respect to BLEU score.

4 Results

In order to be able to evaluate our hybrid approaches in contrast to stand-alone rule-based approaches, we also calculated BLEU scores for the translations conducted by the RBMT systems used in the hybrid setup. Our hybrid system is compared to a SMT baseline and all the 6 RBMT systems that we used. Table 1 shows the evaluation of all the systems in terms of BLEU score (Papineni et al., 2002) with the best score highlighted. The empty cells in the table indicate the language pairs which are not available in the corresponding systems[2]. The SMT system is the one upon which we build the hybrid system. According to the scores, the hybrid system produces better results than the baseline SMT system in all cases. The difference between our system and the baseline is more significant for out-of-domain tests, where gaps in the lexicon tend to be more severe.

Figure 2 illustrates an example of how the hybrid system differs from the baseline SMT system and how it benefits from the RBMT systems. The example lists the English translations of the same German sentence (from News Commentary test set) from different systems involved in our experiment. Neither the word "Pentecost" nor its German translation "Pfingsten" has appeared in the training corpus. Therefore, the SMT baseline system cannot translate the word and chooses to leave the word as it is whereas all the RBMT systems translate the word correctly. The hybrid system appears to have the corresponding lexicon gap covered by the extra entries produced by the RBMT systems. On the other side, these additional entries may not always be helpful. The errors in RBMT outputs can be significant noise that destroys the correct information in the SMT system. In the example translation produced by the hybrid system, there is a comma missing after "in addition", which appears to be frequent in the RBMT outputs.

5 Outlook

The results reported in this paper are still somewhat preliminary in the sense that many possible (including some desirable) variants of the setup could not be tried out due to lack of time. In particular, we think that the full power of our approach on out-of-domain test data can only be exploited with the help of large language models trained on out-of-domain text, but could not yet try this systematically. Furthermore, the presence of multiple instances of

[2] The identities of respective RBMT systems are not revealed in this paper. RBMT1 is evaluated on the partial results produced due to some technical problems.

Source	Darüber hinaus gibt es je zwei Feiertage zu Ostern, Pfingsten, und Weihnachten.
Reference	In addition, Easter, Pentecost, and Christmas are each two-day holidays.
Moses	In addition, there are two holidays, pfingsten to Easter, and Christmas.
Hybrid	In addition there are the two holidays to Easter, Pentecost and Christmas.
RBMT1	Furthermore there are two holidays to Easter, Pentecost and Christmas .
RBMT2	Furthermore there are two holidays each at Easter, Pentecost and Christmas.
RBMT3	In addition there are each two holidays to Easters, Whitsun, and Christmas.
RBMT4	In addition, there is two holidays to Easter, Pentecost, and Christmas.
RBMT5	Beyond that there are ever two holidays to Easter, Whitsuntide, and Christmas.
RBMT6	In addition it gives two holidays apiece to easter, Pentecost, and Christmas.

Figure 2: German-English translation examples

the same phrase pair (with different weight) in the combined phrase table causes the decoder to generate many instances of identical results in different ways, which increases computational effort and significantly decreases the number of distinct cases that are considered during MERT. We suspect that a modification of our scheme that avoids this problem will be able to achieve better results, but experiments in this direction are still ongoing.

The approach presented here combines the strengths of multiple systems and is different from recent work on post-correction of RBMT output as presented in (Simard et al., 2007; Dugast et al., 2007), which focuses on the improvement of a single RBMT system by correcting typical errors via SMT techniques. These ideas are independent and a suitable combination of them could give rise to even better results.

Acknowledgments

This work was supported by the EuroMatrix project funded by the European Commission (6th Framework Programme). We thank Martin Kay, Hans Uszkoreit, and Silke Theison for interesting discussions and practical help, and two anonymous reviewers for hints to improve the paper.

References

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source SMT decoder. In *Proceedings of WMT07*, pages 193–196, Prague, Czech Republic, June. Association for Computational Linguistics.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of WMT07*, pages 220–223, Prague, Czech Republic, June. Association for Computational Linguistics.

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 *Proc. of ACL Demo and Poster Sessions*, pages 177–180, Jun.

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

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007. Combining translations from multiple machine translation systems. In *Proceedings of the Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL'2007)*, pages 228–235, Rochester, NY, April 22-27.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of WMT07*, pages 203–206, Prague, Czech Republic, June. Association for Computational Linguistics.

Silke Theison. 2007. Optimizing rule-based machine translation output with the help of statistical methods. Diploma thesis, Saarland University.