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

Andreas Eisele¹,², Christian Federmann², Hervé Saint-Amand¹, Michael Jellinghaus¹, Teresa Herrmann¹, Yu Chen¹
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 the 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 Moses¹ (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.

¹see http://www.statmt.org/moses/
The typical process for creating an SMT system with the Moses toolkit includes a tuning step in which...
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 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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td><strong>27.85</strong></td>
<td><strong>20.75</strong></td>
<td><strong>28.12</strong></td>
<td><strong>28.82</strong></td>
<td><strong>33.15</strong></td>
<td><strong>32.31</strong></td>
<td>17.36</td>
<td>13.57</td>
<td>17.66</td>
<td>20.71</td>
<td><strong>22.16</strong></td>
<td><strong>22.55</strong></td>
</tr>
<tr>
<td>RBMT1*</td>
<td>13.34</td>
<td>11.09</td>
<td>—</td>
<td>—</td>
<td>17.19</td>
<td>—</td>
<td>18.63</td>
<td>—</td>
<td>15.11</td>
<td>—</td>
<td>17.13</td>
<td></td>
</tr>
<tr>
<td>RBMT2</td>
<td>16.19</td>
<td>12.06</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>16.66</td>
<td>13.64</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>RBMT3</td>
<td>16.32</td>
<td>10.88</td>
<td>18.18</td>
<td>20.38</td>
<td>19.32</td>
<td>20.89</td>
<td>16.88</td>
<td>12.53</td>
<td>17.20</td>
<td>18.82</td>
<td>19.00</td>
<td>19.98</td>
</tr>
<tr>
<td>RBMT4</td>
<td>15.58</td>
<td>12.09</td>
<td>19.00</td>
<td>22.20</td>
<td>18.99</td>
<td>21.69</td>
<td><strong>17.41</strong></td>
<td><strong>13.93</strong></td>
<td>17.73</td>
<td><strong>20.85</strong></td>
<td>19.14</td>
<td>21.70</td>
</tr>
<tr>
<td>RBMT5</td>
<td>15.58</td>
<td>9.54</td>
<td>21.36</td>
<td>12.98</td>
<td>18.47</td>
<td>20.59</td>
<td>15.99</td>
<td>11.05</td>
<td><strong>18.65</strong></td>
<td>19.49</td>
<td>20.50</td>
<td>20.02</td>
</tr>
<tr>
<td>RBMT6</td>
<td>13.96</td>
<td>9.44</td>
<td>17.16</td>
<td>18.91</td>
<td>18.01</td>
<td>19.18</td>
<td>15.08</td>
<td>10.41</td>
<td>16.86</td>
<td>17.82</td>
<td>18.70</td>
<td>19.97</td>
</tr>
</tbody>
</table>

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

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
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


