Pipeline Approaches: Post-processing

Rule-based Translation With Statistical Phrase-based Post-editing

Katherine Munro and Theresa Bertele
June 6, 2018
Table of contents

1. Introduction
2. System description
3. Experimental Results
4. Conclusions
Introduction
Simard et al. (2007)

• shown how statistical phrase-based machine translation system can be used as an automatic post-editing (APE) layer on top of a RBMT system

• motivation: repetitive nature of the typical errors of a rule-based system

• given appropriate training → SMT can be trained to correct errors → reducing those post-editing effort
  • SMT source language: output of rule-based system
  • SMT target language: human translations

• training material for APE layer domain-specific → process of automatically adapting a rule-based system to a specific application domain
Results

• **large improvements** in performance over the baseline rule-based system and over similar statistical phrase-based MT system used *in standalone mode*

• **reduction of post-editing effort up to a third** in comparison to input rule-based translation

• 5 BLEU points improvement over direct SMT approach
Criticism

Impressive results were obtained in very specific and somewhat unusual context:

- training and test corpora were extracted from a collection of manually post-edited MTs
- two corpora (E<>F), each contained three parallel "views" of the same data
  1. source language text
  2. MT of text into target language (as produced by commercial RBMT system)
  3. final target-language version of text (produced by manually post-editing the MT)
- corpus was very small (by SMT standards): 500K words of source-language data F>E, 350K words E>F
Two questions arose

1. How would the results scale up to much larger quantities of training data?

2. Are the results related to the dependent nature of translations, i.e. is the automatic post-editing approach still effective when the machine and human translations are produced independently of one another?
Goal

Participation in the shared task of the second Workshop on SMT with automatic post-editing strategy

1. translate input text into target-language using RBMT (SYSTRAN)
2. automatically post-edit output with statistical phrase-based system (PORTAGE), which performs domain-specific corrections and adaptations to the output
System description
• SYSTRAN, version 6
• French-to-Englisch and English-to-French configurations
• basic "out-the-box" configuration
PORTAGE - Statistical Phrase-based Post-Editing

• output of SYSTRAN fed into post-editing layer that performs domain-specific corrections/adaptions

PORTAGE

• state-of-the-art statistical phrase-based machine translation system developed at NRC (National Research Council of Canada)
Configurations/Training for post-editing similar to corresponding translation system:

- use of **two distinct phrase tables** (phrase pairs from Europarl & NewsCommentary training corpora)
Configurations/Training for post-editing similar to corresponding translation system:

- **use of two distinct phrase tables** (phrase pairs from Europarl & NewsCommentary training corpora)

- **multiple phrase-probability feature functions** in the log-linear models (joint probability estimate, standard frequency-based conditional probability estimate, variants thereof based on different smoothing methods)
Configurations/Training for post-editing similar to corresponding translation system:

- use of **two distinct phrase tables** (phrase pairs from Europarl & NewsCommentary training corpora)
- **multiple phrase-probability feature functions** in the log-linear models (joint probability estimate, standard frequency-based conditional probability estimate, variants thereof based on different smoothing methods)
- **4-gram language model & 5-gram truecasing model**, both trained on combined Europarl & NewsCommentary target-language corpora respectively
Configurations/Training for post-editing similar to corresponding translation system:

- **use of two distinct phrase tables** (phrase pairs from Europarl & NewsCommentary training corpora)
- **multiple phrase-probability feature functions** in the log-linear models (joint probability estimate, standard frequency-based conditional probability estimate, variants thereof based on different smoothing methods)
- **4-gram language model & 5-gram truecasing model**, both trained on combined Europarl & NewsCommentary target-language corpora respectively
- **3-gram adapted language model**: trained on mini-corpus of test-relevant target-language sentences (extracted from training material)
Ideal training data

Ideal:

- training material for post-editing layer consists in a corpus of text in two parallel versions
  1. raw machine translation output
  2. manually post-edited versions of these translations

→ seldomly available (was used in Simard et al. 2007)
Actual training data

Instead:

- training material derived directly from standard source-target parallel corpora → translate source portion of parallel corpus into target language with RBMT component → ”source” for training material
- existing target portion of parallel corpus → ”target” for training material

=> ”source” & ”target” produced independently by man and machine, in contrast to Simard et al. (2007) → initial motivation of the paper
• concentrated on language pair English-French

<table>
<thead>
<tr>
<th></th>
<th>Europarl</th>
<th>NewsCommentary</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentence pairs</td>
<td>&gt; 1.3 million</td>
<td>&gt; 50K</td>
</tr>
<tr>
<td>words/languages</td>
<td>&gt; 32 million</td>
<td>1 million</td>
</tr>
<tr>
<td>configuration</td>
<td>same</td>
<td>same</td>
</tr>
<tr>
<td>adapted language model</td>
<td>dev2006/devtest2006</td>
<td>nc-dev2007</td>
</tr>
</tbody>
</table>

• optimization procedure gives higher weights to Europarl-trained phrase tables for the Europarl domain systems and vice versa
Experimental Results
## BLEU score results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>en → fr</th>
<th>fr → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl (&gt;32M words/language)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYSTRAN</td>
<td>23.06</td>
<td>20.11</td>
</tr>
<tr>
<td>PORTAGE</td>
<td>31.01</td>
<td>30.90</td>
</tr>
<tr>
<td>SYSTRAN+PORTAGE</td>
<td>31.11</td>
<td>30.61</td>
</tr>
<tr>
<td>News Commentary (1M words/language)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYSTRAN</td>
<td>24.41</td>
<td>18.09</td>
</tr>
<tr>
<td>PORTAGE</td>
<td>25.98</td>
<td>25.17</td>
</tr>
<tr>
<td>SYSTRAN+PORTAGE</td>
<td>28.80</td>
<td>26.79</td>
</tr>
</tbody>
</table>

**Figure 1:** BLEU scores for all 4 systems on 2006 test data
Increasing amount of training data

**Figure 2:** series of APE & SMT systems trained on Europarl data
Conclusions
• experiments confirm conclusions of earlier studies:
  • phrase-based post-editing can significantly improve output of a
    RBMT system (in terms of BLEU score)
  • training data for post-editing component, does not need to be
    manually post-edited translations (generated from parallel
    corpora)
  • while post-editing is most effective when little training data is
    available, it remains competitive with phrase-based translation
    even with much larger amounts of data
Outlook

• verify how a phrase-based APE as form of automatic domain-adaption for rule-based methods compares to the standard "lexical customization" method
• modify the phrase-based system so as to better adapt it to the APE task
• Which part of the rule-based processing is really making things easier for the phrase-based post-editing layer?