Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation

Holger Schwenk

LIUM, University of Le Mans, France

Holger.Schwenk@lium.univ-lemans.fr

October 20, 2008
Introduction

Training of an SMT system

- Only bilingual sentence-aligned texts ("bitexts") and large monolingual texts are needed
- An SMT system can be developed without the need of any language-specific expertise
- Monolingual data is usually available in large amounts
- But aligned bilingual texts are a sparse resource for many language pairs (too small, out-of-domain, . . .)
Introduction

How to resolve the problem of insufficient bitexts?

• Pay people to produce more bitexts
• Integration of high quality dictionaries
• Try to take better advantage of limited data (factored translation model, . . .)
• Get more bitexts from the Internet:
  • Most of the found bilingual texts are not direct translations of each other that can be easily aligned
    = comparable corpora
    (Wikipedia, international news agencies, . . .)
• How to exploit comparable corpora?
⇒ Try to align some of the sentences
  [Munteanu and Marcu CL’05, Resnik and Smith CL’03, . . .]
Unsupervised Training of an SMT system

Our approach

- Build a baseline SMT system (using a limited amount of bitext or out-of-domain)
- Use this system to translate large amounts of texts in the source language
- Build a new SMT system with these translations together with the source as additional bitexts
- We don’t need a comparable corpus, just texts in the source language

Variants

- Add related translations to the target LM
  ⇒ Light supervision using a comparable corpus
- How good should be the initial SMT system?
Unsupervised Training of an SMT system

Setup

- Try to build a generic news translation system (French → English)
- Lightly-unsupervised training on LDC’s Gigaword corpora:

<table>
<thead>
<tr>
<th>corpus</th>
<th>#Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFP 199x</td>
<td>236M</td>
</tr>
<tr>
<td>AFP 200x</td>
<td>337M</td>
</tr>
<tr>
<td>Europarl</td>
<td>40M</td>
</tr>
</tbody>
</table>
We use the term lightly–supervised training when the target LM data is closely related to the text to be translated.
Related Work

Unsupervised training for domain adaptation

- Several papers by Ueffing et al. [IWSLT’06, ACL’07]
  - translate the test data
  - compute confidence scores and filter the outputs
  - adapt the system (build small additional phrase table, . . .)
- Work by Chen et al [ACL’08]
- Produce bitexts with a Rule-based system [Hu et al, EMNLP’07]
Baseline System (1)

- Based on the Moses toolkit

Dev and Test data

- Newstest2008 from WMT’08 evaluation
- Consists of news texts (politics, health, financial, society, music, . . .) collected from the Internet
- Split randomly into Dev and Test set (about 1000 lines, 22k words each)
- Many spelling errors in the French translations were automatically corrected
Baseline System (2)

Available Bitexts

- News-commentary corpus (1.6M words),
- Europarl corpus (40.1M words),
- Canadian Hansard corpus (72.4M words).
- Bilingual dictionary from SYSTRAN (512k words).

Available LM data

- English part of bitexts
- UN data
- Full English Gigaword
Example Translations (1)

**French source text:**
- La paix exige une direction palestinienne nouvelle et différente, afin que puisse naître un État palestinien. J’appelle le peuple palestinien à élire de nouveaux dirigeants, des dirigeants qui ne soient pas compromis avec le terrorisme.

**Automatic translations:**
- The peace requires a new and different Palestinian leadership, so that we can create a Palestinian state. I call on the Palestinian people to elect new leaders, leaders not compromised by terrorism.
Example Translations (2)

French source text:
- M. Arafat, qui s’est juré de faire de l’année 2000 celle de la proclamation d’un Etat palestinien, a mis un point d’honneur à recevoir les six chefs d’Etat présents.

Automatic translations:
- Mr. Arafat, who has vowed to make the year 2000 the proclamation of a Palestinian state, has made a point of honour to receive the six heads of state present.
Example Translations (3)

French source text:
- Trois heures après, c’était au tour de la Colombie britannique et de Vancouver de célébrer l’arrivée de l’an nouveau.

Automatic translations:
- Three hours later, it was the turn of the British Columbia and Vancouver célébrer the arrival of the new year.
## Performance of Baseline Systems

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Dict.</th>
<th>Words</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big SMT system</strong></td>
<td>+</td>
<td>116M</td>
<td>22.69</td>
<td><strong>22.17</strong></td>
</tr>
<tr>
<td>News-commentary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Eparl + Hansard</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Performance of Baseline Systems

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Dict.</th>
<th>Words</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big SMT system</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News-commentary</td>
<td>+</td>
<td>116M</td>
<td>22.69</td>
<td>22.17</td>
</tr>
<tr>
<td>+ Eparl + Hansard</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Small SMT system</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News-commentary</td>
<td>+</td>
<td>2.4M</td>
<td>20.44</td>
<td>20.18</td>
</tr>
<tr>
<td>News-commentary</td>
<td>-</td>
<td>1.6M</td>
<td>19.41</td>
<td>19.53</td>
</tr>
<tr>
<td>News-commentary + Eparl</td>
<td>+</td>
<td>43.3M</td>
<td>22.27</td>
<td>22.35</td>
</tr>
</tbody>
</table>

The big SMT system was ranked best in the 2008 WMT evaluation.
Some Comments on the Dictionary

- Provides nouns in singular and plural
- Verbs and in all tenses, ...
- These dictionary entries were directly added to the bitexts
  - Can potentially improve the other alignments
  - Words appearing only in the dictionary have bad translation probabilities
  \[\Rightarrow\] We hope to improve these probabilities by lightly-supervised training
- Many of the dictionary entries are likely to appear in the large monolingual texts
Filtering the Automatic Translations

- Try to discard the bad translations
- Some are tables or enumerations of names, places, ...
- We just used the normalized sentence likelihood

⇒ Use up to 150M words of automatic translations
Using the large SMT Baseline System

- Build SMT system with automatic translations only

![Graph showing BLEU score vs. French words for training (M)](image)

- Better than the baseline when using more than 70M words
- Seems to generalize better
- Improved translation probabilities for dictionary words?
Using the large SMT Baseline System

- Build SMT system with **all human and automatic transl.**

- Mainly improves performance on Test data
- Fortunate peak when using a total of 280M words of bitexts: +0.6 BLEU on Test
Using the small SMT Baseline System

Build SMT system with human-provided and translations of afp9x and afp2x.

- Best performance on Dev for a total of about 100M words
- BLEU on Test set is 21.2 (+1 point)
  ⇒ iterate the process?
Retranslating Europarl

- Automatic translations of Europarl seem to be less useful than Gigaword data
- Comparison to the reference translations: third of the improvement with 70% of the data
Conclusions

• Translated up to 300M words from Gigaword news texts from French to English
• Automatic translation directly used as additional bitexts (after simple filtering)
• First application of large-scale lightly-supervised training to SMT
• Improvements in the BLEU score:
  • +0.6 on top of state-of-the-art system
  • +1.1 on top of small SMT system (2.4M words of bitexts)
• Seems to improve generalization behavior
• Method to obtain translation probabilities for dictionary words
• Used several thousands of hours of compute time
Perspectives

• Use a biased LM (with comparable corpora)
• Verify approach when no related texts in the target language are available
• More sophisticated techniques to filter the translations
• Iterate procedure and incrementally improve the system?
• Compare and combine with IR techniques to extract parallel sentences
• Other language pairs
## Result Summary

### Bitexts

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Human-provided</th>
<th>Lightly-supervised</th>
<th>Total Words</th>
<th>BLEU score</th>
<th>Ptable Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nw+dict</td>
<td>2.4M</td>
<td>2.4M</td>
<td>2.4M</td>
<td>20.44</td>
<td>20.18</td>
</tr>
<tr>
<td>Nw+Ep+dict</td>
<td>43M</td>
<td>-</td>
<td>43.3M</td>
<td>22.17</td>
<td>22.35</td>
</tr>
<tr>
<td>Nw+Ep+Hans+dict</td>
<td>116M</td>
<td>-</td>
<td>116M</td>
<td>22.69</td>
<td>22.17</td>
</tr>
</tbody>
</table>

### Translated with the small SMT system:

#### News

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Human-provided</th>
<th>Lightly-supervised</th>
<th>Total Words</th>
<th>BLEU score</th>
<th>Ptable Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>afp9x</td>
<td>28M</td>
<td>2.4M</td>
<td>21.21</td>
<td>21.02</td>
<td>58M</td>
</tr>
<tr>
<td>afp9x</td>
<td>101M</td>
<td>2.4M</td>
<td>21.23</td>
<td>21.18</td>
<td>189M</td>
</tr>
<tr>
<td>afp2x</td>
<td>43M</td>
<td>2.4M</td>
<td>20.98</td>
<td>21.01</td>
<td>77M</td>
</tr>
<tr>
<td>afp2x</td>
<td>102M</td>
<td>2.4M</td>
<td>21.23</td>
<td>21.17</td>
<td>170M</td>
</tr>
<tr>
<td>Eparl</td>
<td>7M</td>
<td>2.4M</td>
<td>20.78</td>
<td>20.65</td>
<td>17M</td>
</tr>
<tr>
<td>Eparl</td>
<td>31M</td>
<td>2.4M</td>
<td>21.14</td>
<td>20.86</td>
<td>67M</td>
</tr>
</tbody>
</table>

### Translated with the big SMT system:

#### News+Eparl

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Human-provided</th>
<th>Lightly-supervised</th>
<th>Total Words</th>
<th>BLEU score</th>
<th>Ptable Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>afp2x</td>
<td>31M</td>
<td>31M</td>
<td>22.23</td>
<td>22.33</td>
<td>55M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
</tbody>
</table>

### News+Eparl

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Human-provided</th>
<th>Lightly-supervised</th>
<th>Total Words</th>
<th>BLEU score</th>
<th>Ptable Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>afp2x</td>
<td>31M</td>
<td>31M</td>
<td>22.23</td>
<td>22.33</td>
<td>55M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Human-provided</th>
<th>Lightly-supervised</th>
<th>Total Words</th>
<th>BLEU score</th>
<th>Ptable Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>afp2x</td>
<td>31M</td>
<td>31M</td>
<td>22.23</td>
<td>22.33</td>
<td>55M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Human-provided</th>
<th>Lightly-supervised</th>
<th>Total Words</th>
<th>BLEU score</th>
<th>Ptable Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>afp2x</td>
<td>31M</td>
<td>31M</td>
<td>22.23</td>
<td>22.33</td>
<td>55M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bitexts</th>
<th>Human-provided</th>
<th>Lightly-supervised</th>
<th>Total Words</th>
<th>BLEU score</th>
<th>Ptable Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>afp2x</td>
<td>31M</td>
<td>31M</td>
<td>22.23</td>
<td>22.33</td>
<td>55M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
<tr>
<td>afp2x</td>
<td>112M</td>
<td>112M</td>
<td>22.56</td>
<td>22.47</td>
<td>180M</td>
</tr>
</tbody>
</table>