A fully Unsupervised approach for mining parallel data from comparable corpora

Thi-Ngoc-Diep Do\textsuperscript{1,2}, Laurent Besacier\textsuperscript{1}, Eric Castelli\textsuperscript{2}

\textsuperscript{1}\textsuperscript{1}LIG Laboratory, CNRS/UMR-5217, Grenoble, France
\textsuperscript{2}\textsuperscript{2}MICA Center, CNRS/UMI2954, HUT, Hanoi, Vietnam

thi-ngoc-diep.do@imag.fr
A parallel bilingual corpus

- Statistical machine translation (SMT): a large parallel bilingual text corpus.
- To build parallel corpora:
  - Collect from parallel document pairs (Resnik and Smith, 2003; Kilgarriff and Grefenstette, 2003)
  - Apply alignment methods at document level, sentence level for the source and target monolingual corpora (Koehn, 2005; Gale and Church, 1993, Patry and Langlais, 2005)
  - Mine a comparable corpus (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2006)
  - etc.
Mining a comparable corpus

- A comparable corpus:
  - “closely related by conveying the same information” (Zhao and Vogel, 2002)
  - “mostly bilingual translations of the same document” (Fung and Cheung, 2004)
  - “various levels of parallelism, such as words, phrases, clauses, sentences, and discourses…” (Kumano et al., 2007).

- Source: News domain

- “comparable”\textsubscript{IR} “noisy parallel”

  Advanced IR approaches are outside of the scope of this paper
Mining a comparable corpus

- bilingual dictionary
- human translation pairs
- parallel corpus

Ref: Zhao and Vogel (2002), Munteanu and Marcu (2006), Abdul-Rauf and Schwenk (2009), Sarikaya et al. (2009)

- a translation lexicon model
- a proper statistical machine translation system
- maximum likelihood criterion
- evaluation metric
Mining a comparable corpus

- bilingual dictionary
- human translation pairs
- parallel corpus

Does a fully unsupervised method, starting with a comparable corpus, allow us to overcome the problem of lacking parallel data?

Ref: Zhao and Vogel (2002), Munteanu and Marcu (2006), Abdul-Rauf and Schwenk (2009), Sarikaya et al. (2009)
Our unsupervised learning method
Our unsupervised learning method

- **Translation module**
  - A statistical machine translation system
  - Start with a simple noisy comparable corpus (named C2), without using additional parallel data
Our unsupervised learning method

Filtering module:

- Use evaluation metric estimated for each sentence pair
- Which one? Bleu, Nist, Ter, Per* (based on the similarity of two sentences)
  \[
  \text{PER}^* = \frac{2 \times \text{number of identical words}}{\text{length of hypothesis} + \text{length of reference}}
  \]

- A pair is parallel if score > threshold (for Bleu, Nist, Per*) or < threshold (for Ter)
Our unsupervised learning method

Iterative scheme:
- Combine the extracted pairs with the translation module => new one
- Re-translate D → re-calculate score → re-filter data → re-combine ...

Different combinations at iteration i:
- **W1**: $S_0$ is retrained on $C_2$ and $E_{i-1}$
- **W2**: $S_0$ is retrained on $C_2$ and $E_0 + E_1 + ... + E_{i-1}$
- **W3**: $E_{j-1}$ → a new separate phrase-table. Decode using phrase-table of $S_0$ and this new one (log-linear model) without weighting them.
- **W4**: the same combination as W3, but the phrase-table of $S_0$ and the new one are weighted, e.g. 1:2.
Experiments for French-English SMT
Compare the semi- and un-supervised methods
Data preparation

- Two systems were constructed (using the Moses toolkit (Koehn et al., 2007)) to mine a comparable corpus D:
  - semi-supervised method (Sys1)
  - unsupervised method (Sys2)
- Create “simulated” noisy parallel corpus:
  - C1: 50K parallel sentence pairs from the Europarl v.3
  - C2: 25K correct parallel sentence pairs (withdrawn from C1) and 25K wrong sentence pairs
  - D: 10K parallel sentence pairs from the Europarl v.3 (marked) and 10K wrong sentence pairs, which were different from sentence pairs of C1 and C2
Experiments

- Whether Sys2 can be used to filter the input data in the same fashion as Sys1 does?
  - Translate the French side of corpus D by Sys1 and Sys2
  - Calculate the scores BLEU, NIST, TER and PER* for the translated output with the English side of the corpus D
  - Display the distributions of evaluation scores for correct parallel sentence pairs and wrong sentence pairs
The distributions of scores have the same shape between Sys1 and Sys2. In particular, the distributions of scores for the wrong pairs were nearly identical in both systems.

PER* can be considered as the most suitable score.
Iterations

- The iterations of the unsupervised method
  - improve the quality of the translation system
  - increase the number of correctly extracted sentence pairs
- Combined the extracted sentence pairs in 4 ways: $w_1, w_2, w_3, w_4$
- Chose the score $\text{PER}^*$ and the threshold $= 0.3$
Iterations

The number of correct extracted pairs was increased in all cases.
W2 brought the largest number of correct extracted sentence pairs.
Iterations

A test set: 400 French-English parallel sentence pairs from Europarl corpus.

Use one reference.

The quality of the translation system was increased quickly during the first few iterations, but decreased after that.

The quality of the translation systems
APPLICATION FOR
FRENCH-VIETNAMESE LANGUAGE PAIR
A truly comparable corpus
Preparing the data

- Vietnamese daily news website, the Vietnam News Agency¹ (VNA): tends to contain parallel sentences or rough translations of sentences on the same topics
  - 20,884 French documents (from 12 April 2006 to 14 August 2008)
  - 54,406 Vietnamese documents
  - 10 sentences per document
  - 30 words per sentence

A noisy comparable corpus

- A noisy comparable corpus
  - Apply a publishing date filter
  - Merge sentence: a m-sentence Vietnamese document and a n-sentence French document => m x n pairs of sentences.
  - From VNA => 1,442,448 pairs of sentences: really noisy parallel
  - Filter by the ratio of the French sentence’s length to the Vietnamese sentence’s length = 0.8 ÷ 1.3

=> 345,575 pairs of sentences (named Call).
The initial translation system

- A cross-filtering process to extract C2 and D

### Sub corpus statistics

<table>
<thead>
<tr>
<th>Sub corpus</th>
<th># pairs</th>
<th># C2</th>
<th># D</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC1</td>
<td>85,011</td>
<td>2916</td>
<td>82,095</td>
</tr>
<tr>
<td>SC2</td>
<td>85,008</td>
<td>3495</td>
<td>81,513</td>
</tr>
<tr>
<td>SC3</td>
<td>86,529</td>
<td>3820</td>
<td>82,709</td>
</tr>
<tr>
<td>SC4</td>
<td>89,027</td>
<td>3892</td>
<td>85,135</td>
</tr>
</tbody>
</table>

C2: 14,123 pairs
D: 331,452 pairs
Applying the unsupervised method

<table>
<thead>
<tr>
<th>SMT iter.</th>
<th>#extracted pairs</th>
<th>Bleu</th>
<th>Nist</th>
<th>Ter</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>14,123</td>
<td>30.67</td>
<td>6.45</td>
<td>0.59</td>
</tr>
<tr>
<td>1</td>
<td>26,517</td>
<td>32.18</td>
<td>6.70</td>
<td>0.57</td>
</tr>
<tr>
<td>2</td>
<td>37,210</td>
<td>32.42</td>
<td>6.75</td>
<td>0.56</td>
</tr>
<tr>
<td>3</td>
<td>38,530</td>
<td>32.45</td>
<td>6.77</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>39,254</td>
<td>32.14</td>
<td>6.73</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>39,758</td>
<td>31.85</td>
<td>6.68</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Test set: 400 manually extracted Vietnamese-French parallel sentence pairs

- The number of extracted sentence pairs increased with each iteration.
- The quality of the translation system was increased quickly during the first few iterations, but decreased after that.
The former method

- Method1 (Do et al. 2009):
  - Mining method:
    - Filter possible parallel document pairs by publishing date and special words (numbers, attached symbols, named entities).
    - Align sentences in a possible parallel document pair using lexical information (lexemes, stop words, a bilingual dictionary, etc.).
    - Extract sentence pairs based on the sentence alignment information, which combines document length information and lexical information.
  - From VNA => extracted 50,322 “parallel” sentence pairs
## Compare unsupervised method and Method1

<table>
<thead>
<tr>
<th>Mining method</th>
<th>#extracted pairs</th>
<th>Bleu</th>
<th>Nist</th>
<th>Ter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical info. + Heuristics (Method1)</td>
<td>50,322</td>
<td>32.74</td>
<td>6.78</td>
<td>0.55</td>
</tr>
<tr>
<td>Unsupervised method</td>
<td>38,530</td>
<td>32.45</td>
<td>6.77</td>
<td>0.56</td>
</tr>
</tbody>
</table>

The same test set of 400 manually extracted Vietnamese-French parallel sentence pairs

- The number of extracted sentence pairs is lower than that in the Method1
- The quality of the SMT systems are comparable
Conclusion and perspectives

- An unsupervised method for extracting parallel sentence pairs from a comparable corpus based on a comparable corpus, instead of a parallel corpus using iterative scheme.

- The quality of the translation system can be improved during the first iterations, but it becomes worse later because of adding the noisy data into the statistical models.

- This method may be applied successfully even in those cases where parallel data are lacking.

- The quality of the translation system is comparable with that of another method which requires better quality data for bootstrapping (bilingual dictionary, etc.).
Conclusion and perspectives

- Our future works:
  - deeper analysis of the filtering and data inclusion techniques
  - experiments at a larger scale
  - human evaluations to confirm improvements obtained with our unsupervised method
Thank you!
References

References

- Munteanu, D.S. and D. Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. 44th annual meeting of the Association for Computational Linguistics.
- Stolcke, Andreas. 2002. SRILM an extensible language modeling toolkit. Intl. Conf. on Spoken Language Processing.
Our unsupervised learning method

- a proper statistical machine translation system
- evaluation metrics (Bleu, Nist, Ter, Per*)

Iterative scheme with different combinations:
- W1: $S_0$ at step $i$ is retrained on $C_2$ and $E_{i-1}$
- W2: $S_0$ at step $i$ is retrained on $C_2$ and $E_0 + E_1 + \ldots + E_{i-1}$
- W3: at iteration $i$, a new separate phrase-table is built based on the extracted data $E_{i-1}$. System decodes using both phrase-table of $S_0$ and this new one (log-linear model) without weighting them.
- W4: the same combination as W3, but the phrase-table of $S_0$ and the new one are weighted, e.g. 1:2.

$PER^* = \frac{2 \times \text{number of identical words}}{\text{length of hypothesis} + \text{length of reference}}$