Does syntactic context fare better than positional context?

NCLT/CNGL Internal Workshop

24 July 2008

Ankit Kumar Srivastava

NON PARALLEL CORPORA
Learning a Translation Lexicon from non Parallel Corpora

- Motivation
- Methodology
- Implementation
- Experiments
- Conclusion

Master’s Project
AT
University of Washington
Seattle, USA
JUNE 2008
Word – to – word mapping between 2 languages

Invaluable resource in multilingual applications like CLIR, CL resource, CALL, etc.

<table>
<thead>
<tr>
<th>Wahl</th>
<th>Sheridan &amp; Ballerini 1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>election</td>
<td>0.85</td>
</tr>
<tr>
<td>ballot</td>
<td>0.10</td>
</tr>
<tr>
<td>option</td>
<td>0.02</td>
</tr>
<tr>
<td>selection</td>
<td>0.02</td>
</tr>
<tr>
<td>choice</td>
<td>0.01</td>
</tr>
</tbody>
</table>

McCarley 1999

Yarowsky & Ngai 2001

Cucerzanz & Yarowsky 2002

Nerbonne et al. 1997
Parallel, comparable, non-comparable text
More monolingual text than bitext
5 dimensions of nonparallelness
Most statistical clues no longer applicable
Given any two pieces of text in any two languages…

...Can we extract word translations?
If two words are mutual translations, then their more frequent collocates (context window) are likely to be mutual translations as well.

Counting co-occurrences within a window of size $N$ is less precise than counting co-occurrences within local syntactic contexts [Harris 1985].

2 types of context windows – Positional (window size 4) and Syntactic (head, dependent)
Vinken will join the board as a nonexecutive director Nov 29.

**POSITIONAL:**
Vinken will join the board as a **nonexecutive** director Nov 29.

**SYNTACTIC:**
Vinken will join the board as a **nonexecutive** director Nov 29.
For each unknown word in the SL & TL, define the context in which that word occurs.

Using an initial seed lexicon, translate as many source context words into the target language.

Use a similarity metric to compute the translation of each unknown source word. It will be the target word with the most similar context.

Rapp 1995, 1999
Fung & Yee 1998
Koehn & Knight 2002
Otero & Campos 2005
IMPLEMENTATION

{ system }

1. CORPUS CLEANING
2. PCFG Parsing
1 Raw Text Corpora:

<table>
<thead>
<tr>
<th></th>
<th>ENGLISH</th>
<th>GERMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>Wall Street Journal (WSJ)</td>
<td>Deutsche Presse Agentur (DPA)</td>
</tr>
<tr>
<td>COVERAGE</td>
<td>446 days of news text</td>
<td>530 days of news text</td>
</tr>
</tbody>
</table>

2 Phrase Structures:

Stanford Parser (Lexicalized PCFG) for English and German

http://nlp.stanford.edu/software/lex-parser.shtml

[Klein & Manning 2003]
IMPLEMENTATION

1. CORPUS CLEANING
2. PCFG PARSING
3. PS TO DS CONVERSION
4. DATA SETS

{ system }
3 Dependency Structures:

Head Percolation Table [Magerman 1995; Collins 1997] was used to extract head-dependent relations from each parse tree.

4 Data Sets:

<table>
<thead>
<tr>
<th></th>
<th>ENGLISH</th>
<th>GERMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXT</td>
<td>1,521,998 sentences</td>
<td>808,146 sentences</td>
</tr>
<tr>
<td>TOKENS</td>
<td>36,251,168 words</td>
<td>14,311,788 words</td>
</tr>
<tr>
<td>TYPES</td>
<td>276,402 words</td>
<td>388,291 words</td>
</tr>
</tbody>
</table>
{ system }

1. CORPUS CLEANING
2. PCFG PARSING
3. PS TO DS CONVERSION
4. DATA SETS
5. CONTEXT GENERATOR

SEED LEXICON
PARSED TEXT
RAW TEXT

SYN VECTORS
POS VECTORS
Seed lexicon obtained from a dictionary, identically spelled words, spelling transformation rules.
Context vectors have dimension values (co-occurrence of word with seed) normalized on seed frequency.

5 Context Vectors:

<table>
<thead>
<tr>
<th></th>
<th>ENGLISH</th>
<th>GERMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIMENSION</td>
<td>2,376 words</td>
<td></td>
</tr>
<tr>
<td>SEED</td>
<td>2,350 words</td>
<td>2,376 words</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>74,434 words</td>
<td>106,366 words</td>
</tr>
</tbody>
</table>
IMPLEMENTATION

1. CORPUS CLEANING
2. PCFG PARSING
3. PS TO DS CONVERSION
4. DATA SETS
5. CONTEXT GENERATOR
6. VECTOR SIMILARITY
7. RANK
8. TRANS. LIST

{ system }
Vector similarity metrics used are city block [Rapp 1999] and cosine. Translations sorted in descending order of scores. Evaluation data extracted from online bilingual dictionaries (364 translations).

Ranked Translations Predictor:

<table>
<thead>
<tr>
<th></th>
<th>CITY BLOCK</th>
<th>COSINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSITIONAL CONTEXT</td>
<td>63 out of 364</td>
<td>148 out of 364</td>
</tr>
<tr>
<td>SYNTACTIC CONTEXT</td>
<td>301 out of 364</td>
<td>216 out of 364</td>
</tr>
</tbody>
</table>
CONCLUSION

- Extraction from non parallel corpora useful for compiling lexicon from new domains.
- Syntactic context helps in focusing the context window, more impact on longer sentences.
- Non parallel corpora involves more filtering, search heuristics than in parallel.
- Future directions include using syntactic only on one side, extending coverage through stemming.

CONCLUSION
{ thanks }