Evaluating Data Sources in a Large Czech-English Corpus CzEng 0.9

Ondřej Bojar, Adam Liška, Zdeněk Žabokrtský
{bojar,zabokrtsky}@ufal.mff.cuni.cz
adam.liska@gmail.com
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

May 19, 2010
Outline

• CzEng 0.9 overview
• Our contribution:
  – Evaluating CzEng 0.9 filters.
  – Implementing and evaluating new filters.
• Utility of data sources.
CzEng 0.9

- large parallel Czech-English corpus
- various sources to include as much material as possible

Number of tokens

- 8 million parallel sentences
- 93 million English tokens, 82 million Czech tokens
Common Processing Pipeline

All documents go through the same processing pipeline:

- conversion to UTF-8 encoded plain text
- segmentation
- sentence alignment using Hunalign
- only 1-1 aligned sentences are kept (82%)
- heuristic filters filter out mis-aligned/malformed pairs
- automatic analyses at the morphological, analytical (surface syntactic) and tectogrammatical (deep syntactic) layers

TectoMT platform, following Functional Generative Description and the Prague Dependency Treebank (PDT, Hajič et al. (2006))
Filters Used in CzEng 0.9

- the Czech and English sentences identical
- the lengths of the sentences are too different
- no Czech word on the Czech side or English word on the English side
- suspicious character
- clearly suspicious segmentation or tokenization
- outstanding HTML entities or tags
- relicts of metainformation

The filters were not empirically evaluated!
New filters

- applied on segments included in CzEng 0.9
- non-ASCII characters on the English side that are not present in the Czech sentence
- use of numbers in the Czech and English sentences are different
- word-alignment score of each sentence pair is below a given threshold
Typical problem:

“English”  Koupě zboží za účelem jeho dalšího prodeje a prodej .
(The purchase of goods for the purposes of re-selling and selling.)

Czech  Specialista na osobní a nákladní vozidla .
(The specialist for cars and lorries.)

Causes: incorrect document/sentence alignment, non-parallel input

English segments with non-ASCII characters that are not present in the Czech segment are filtered out.
New Filter: Use of Numbers

- Filter looks for numerical and written equivalents of the numbers found in the English segment
- Filters out a wide range of mistakes:

<table>
<thead>
<tr>
<th>English</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours must be reported in .25 increments.</td>
<td>Hodiny je nutné zadat v intervalech po 0</td>
</tr>
<tr>
<td>(Hours have to be entered in increments of 0)</td>
<td>(Hours have to be entered in increments of 0)</td>
</tr>
</tbody>
</table>
New Filter: Word-alignment Score

- Filter considers alignment probabilities in both directions
- GIZA++: Hidden Markov Model, IBM Model 1, IBM Model 3 and IBM Model 4 trained on lemmas

\[
Score (e_1^J, f_1^I) = \frac{1}{J} \log (p (e, a | f)) + \frac{1}{I} \log (p (f, a | e)) \tag{1}
\]
Overall Evaluation

- Evaluated on two sets of 1000 sentence pairs:
  - CzEng filters: sent. pairs selected from aligned plaintext files
  - new filters: first 1000 segments from CzEng (randomized at the level of short sequences of sentences)

- overall precision: any filter fires ⇒ was it indeed a bad segment?

\[
\frac{|\text{segments marked by both human and at least one filter}|}{|\text{segments marked by at least one filter}|} \quad (2)
\]

- overall recall: how many bad segments are found?

\[
\frac{|\text{segments marked by both human and at least one filter}|}{|\text{segments marked by human}|} \quad (3)
\]
Evaluation of the Filters

- Extended sets of sentence pairs:
  - CzEng filters: 200 segments where the filter fired
  - new filters: 500 segments where the filter fired

- filter precision: the filter fires ⇒ was it indeed a bad segment?
  \[ \frac{\text{segments marked by both human and the filter}}{\text{segments marked by the filter, i.e. 200 or 500}} \] (4)

- filter recall: how many bad segments are found?
  \[ \frac{\text{segments marked by both human and the filter}}{\text{segments marked by human}} \] (5)
Evaluation of Czeng Filters

<table>
<thead>
<tr>
<th>Selected CzEng Filters</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not enough letters</td>
<td>94%</td>
<td>7%</td>
</tr>
<tr>
<td>Mismatching lengths</td>
<td>91%</td>
<td>11%</td>
</tr>
<tr>
<td>Repeated character</td>
<td>88%</td>
<td>2%</td>
</tr>
<tr>
<td>No English word</td>
<td>80%</td>
<td>11%</td>
</tr>
<tr>
<td>Suspicious char.</td>
<td>75%</td>
<td>1%</td>
</tr>
<tr>
<td>Identical</td>
<td>72%</td>
<td>26%</td>
</tr>
<tr>
<td>No Czech word</td>
<td>67%</td>
<td>2%</td>
</tr>
<tr>
<td>Too long sentence</td>
<td>12%</td>
<td>0%</td>
</tr>
<tr>
<td>Extra header</td>
<td>2%</td>
<td>0%</td>
</tr>
</tbody>
</table>

| Overall (all filters)                  | 57%       | 42%    |
| Overall (evaluated filters only)       | 57%       | 41%    |

- Surprisingly low precision of many filters.
- Large margin for recall improvement.
# Evaluation of New Filters

<table>
<thead>
<tr>
<th>Filter</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-ASCII characters in English</td>
<td>100%</td>
<td>4%</td>
</tr>
<tr>
<td>Number</td>
<td>88%</td>
<td>6%</td>
</tr>
<tr>
<td>Word-alignment scores</td>
<td>77%</td>
<td>33%</td>
</tr>
<tr>
<td>Overall</td>
<td>79%</td>
<td>40%</td>
</tr>
</tbody>
</table>

- Applied on top of original CzEng 0.9 filtering.
- Word-alignment can be tuned for precision/recall.
Prec/Rec for Alignment Filters

Word-alignment score:
- 100k
Prec/Rec for Hunalign Scores

Hunalign quality:
--- higher --- lower is better

⇒ Hunalign scores not suitable for filtering.
<table>
<thead>
<tr>
<th>Bad 1-1 Segments [%]</th>
<th>Most Frequent Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>subtitles 4.6</td>
<td>Mismatching lengths (42.0%),</td>
</tr>
<tr>
<td>eu 33.3</td>
<td>Identical (39.9%),</td>
</tr>
<tr>
<td>techdoc 10.2</td>
<td>Identical (37.9%),</td>
</tr>
<tr>
<td>paraweb 59.5</td>
<td>Identical (61.7%),</td>
</tr>
<tr>
<td>fiction 3.1</td>
<td>Mismatching lengths (54.9%),</td>
</tr>
<tr>
<td>news 3.8</td>
<td>Identical (54.1%),</td>
</tr>
<tr>
<td>navajo 11.9</td>
<td>Identical (40.9%),</td>
</tr>
</tbody>
</table>

- Large share of Parallel Web and EU texts filtered out
- Fiction, news and subtitles show high utility
## Utility of Data Sources

<table>
<thead>
<tr>
<th>Bad 1-1 Segments [%]</th>
<th>Most Frequent Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>subtitles 6.8</td>
<td>Alignment score (94.5%),</td>
</tr>
<tr>
<td>eu 3.3</td>
<td>Alignment score (68.7%),</td>
</tr>
<tr>
<td>techdoc 3.4</td>
<td>Alignment score (93.7%),</td>
</tr>
<tr>
<td>paraweb 17.6</td>
<td>ASCII (51.2%),</td>
</tr>
<tr>
<td>fiction 7.4</td>
<td>Alignment score (86.0%),</td>
</tr>
<tr>
<td>news 2.2</td>
<td>Alignment score (55.3%),</td>
</tr>
<tr>
<td>navajo 1.9</td>
<td>Alignment score (57.1%),</td>
</tr>
</tbody>
</table>

- Cleanest source: news
- Original filtering still insufficient for Parallel Web segments
Conclusion

• Original CzEng 0.9 filters insufficient.
  – Overall recall $\sim\!40\%$, precision 57% only.

• New filters on top of CzEng 0.9 ones:
  – Overall recall $\sim\!40\%$, precision 79%.

• Most reliable sources of data: fiction, news and subtitles.

Future:

• Merge sets of filters.

• Ensemble of many high-precision filters to achieve high recall.

Download: http://ufal.mff.cuni.cz/czeng
Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0. LDC, Philadelphia.