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

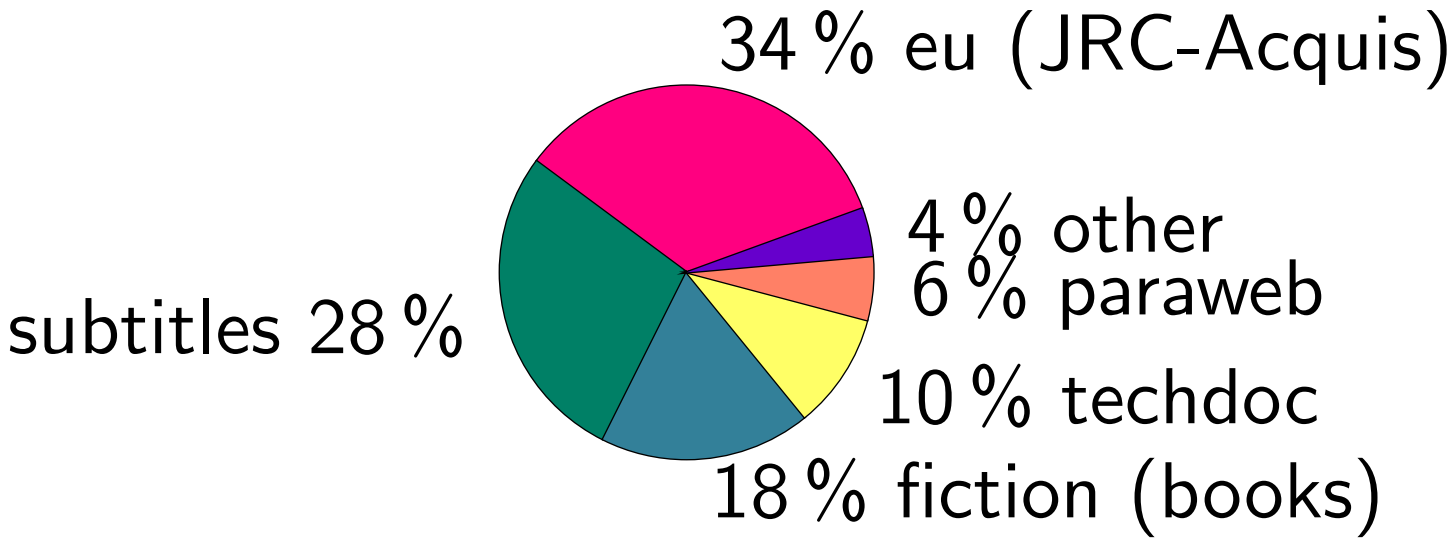
Ondřej Bojar, Adam Liška, Zdeněk Žabokrtský
{bojar,zabokrtsky}@ufal.mff.cuni.cz
adam.liska@gmail.com
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

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-aligment score of each sentence pair is below a given threshold

New Filter: Non-ASCII characters

- 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:

English	Hours must be reported in . 25 increments .
Czech	Hodiny je nutné zadat v intervalech po 0 (Hours have to be entered in increments of 0)

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

$$Score\left(e_1^J, f_1^I\right) = \frac{1}{J}\log\left(p\left(\boldsymbol{e}, a \mid \boldsymbol{f}\right)\right) + \frac{1}{I}\log\left(p\left(\boldsymbol{f}, a \mid \boldsymbol{e}\right)\right) \quad (1)$$

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 $\Rightarrow$ was it indeed a bad segment?

$$\left|\begin{array}{c}\text{segments marked by both human} \\ \text{and at least one filter}\end{array}\right| \Big/ \left|\text{segments marked by at least one filter}\right| \quad (2)$$

- overall recall: how many bad segments are found?

$$\left|\begin{array}{c}\text{segments marked by both human} \\ \text{and at least one filter}\end{array}\right| \Big/ \left|\text{segments marked by human}\right| \quad (3)$$

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 $\Rightarrow$ was it indeed a bad segment?

$$\left|\begin{array}{c}\text{segments marked by both human} \\ \text{and the filter}\end{array}\right| \Big/ \left|\begin{array}{c}\text{segments marked by the filter,} \\ \text{i.e. 200 or 500}\end{array}\right| \qquad (4)$$

- filter recall: how many bad segments are found?

$$\left|\begin{array}{c}\text{segments marked by both human} \\ \text{and the filter}\end{array}\right| \Big/ |\text{segments marked by human}| \qquad (5)$$

Evaluation of Czeng Filters

Selected CzEng Filters	Precision	Recall
Not enough letters	94%	7%
Mismatching lengths	91%	11%
Repeated character	88%	2%
No English word	80%	11%
Suspicious char.	75%	1%
Identical	72%	26%
No Czech word	67%	2%
Too long sentence	12%	0%
Extra header	2%	0%
Overall (all filters)	57%	42%
Overall (evaluated filters only)	57%	41%

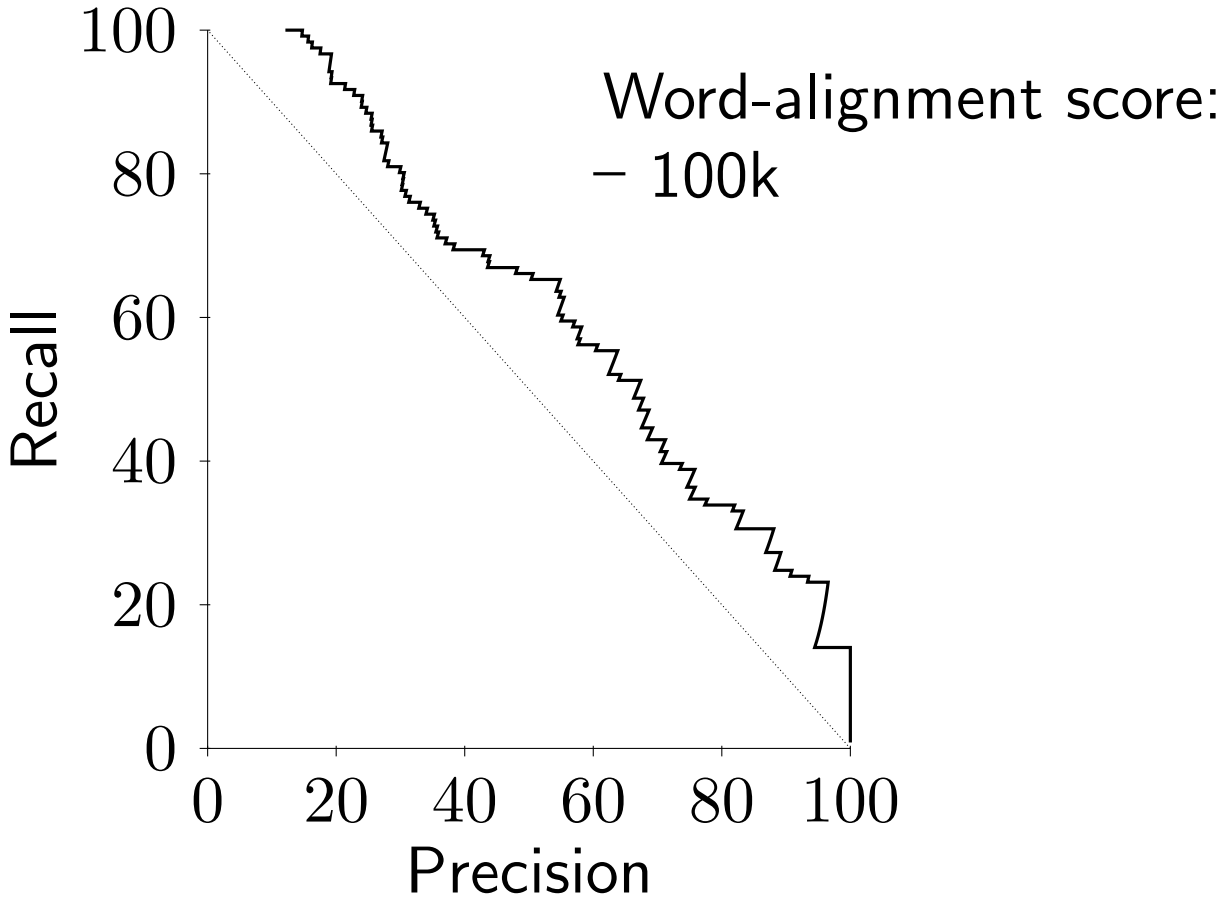
- Surprisingly low precision of many filters.
- Large margin for recall improvement.

Evaluation of New Filters

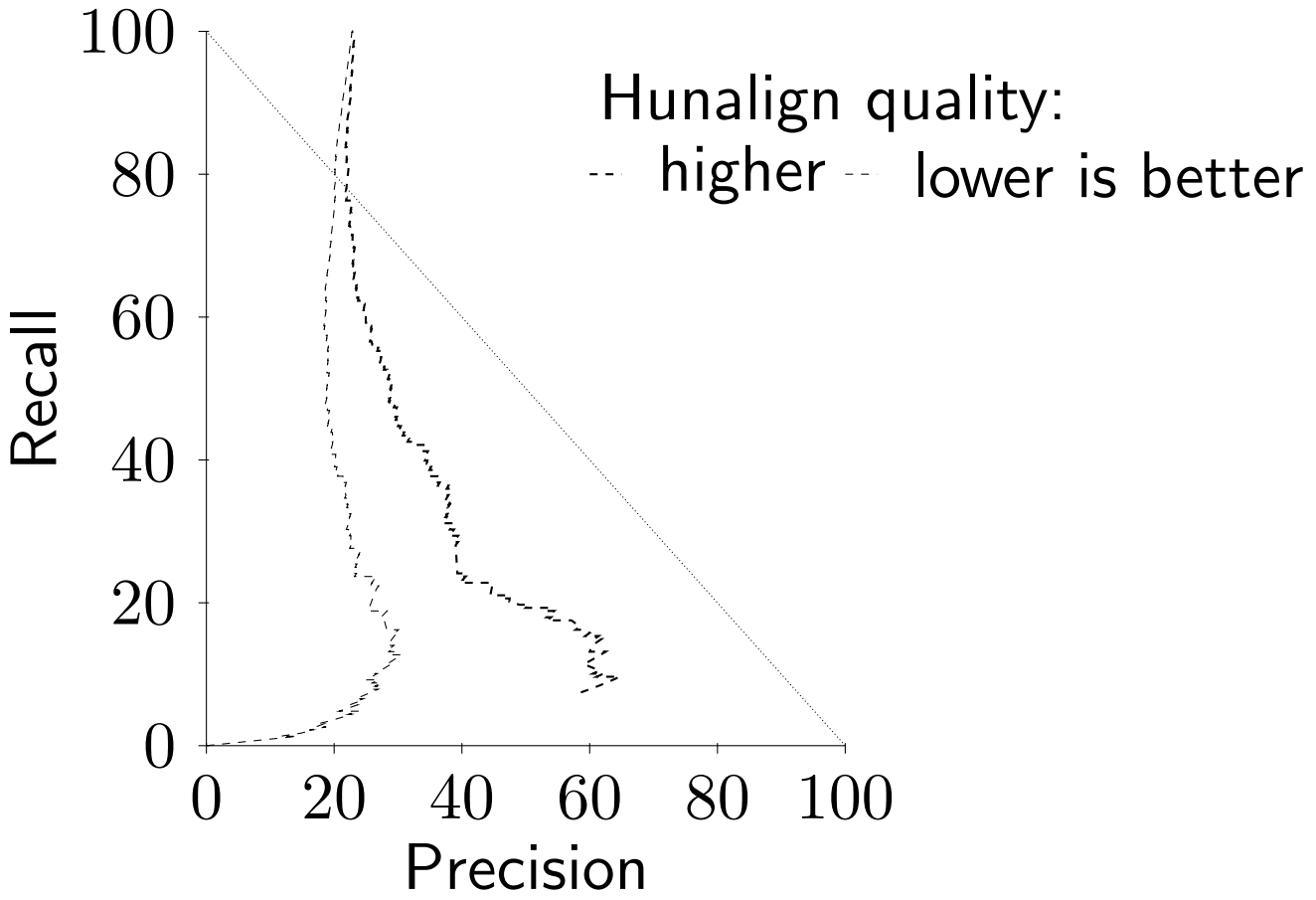
Filter	Precision	Recall
Non-ASCII characters in English	100%	4%
Number	88%	6%
Word-alignment scores	77%	33%
Overall	79%	40%

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

Prec/Rec for Alignment Filters



Prec/Rec for Hunalign Scores



⇒ Hunalign scores not suitable for filtering.

Utility of Data Sources 1

Bad 1-1 Segments [%]		Most Frequent Error
subtitles	4.6	Mismatching lengths (42.0%),
eu	33.3	Identical (39.9%),
techdoc	10.2	Identical (37.9%),
paraweb	59.5	Identical (61.7%),
fiction	3.1	Mismatching lengths (54.9%),
news	3.8	Identical (54.1%),
navajo	11.9	Identical (40.9%),

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

Utility of Data Sources 2 - CzEng

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

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

Conclusion

- Original CzEng 0.9 filters insufficient.
  - Overall recall ∼40%, precision 57% only.
- New filters on top of CzEng 0.9 ones:
  - Overall recall ∼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.