Extended tokenizer for Polish

Tomasz Bartosiak
Konrad Gołuchowski
Katarzyna Krasnowska

14 March 2014

Stage 1: token splitting

- Text is split on spaces.
- Additionally, leading and trailing punctuation marks are separated:
    - exception: dot preceded by non-punctuation,
    - exception from exception: three consecutive dots.

Stage 2: token tagging (and further splitting)

- Cascade of tag filters.
- Regular expression-based, e.g.:
  - `rom`, `ara`, `e-mail`, `www`
  - dates in formats 14.03.2014, 14.03.2014.
- More complicated, e.g.:
  - abbreviations,
  - *I*/*i* conjunction,
  - hyphen-separated tokens.
- Helper tags:
  - `int` for arabic integers,
  - `date` for dates as above,
  - `m-i`, ..., `m-xii` for month names.

Stage 2: abbreviations

- A list of about 1300 abbreviations is used.
- Dot-ended abbreviations identical with some other word's inflected form:
  - e.g., *giełd.*, *gwar.*, *ul.*
  - heuristic: only tag as `abbrev` in the middle of a sentence.
- Other dot-ended abbreviations:
  - e.g., *dot.*, *egip.*, *popr.*
- Mutli-part abbreviations:
  - e.g., *m.in.*, *p.n.e.*
- Abbreviations without the dot:
  - e.g., *mjr*, *s-ka*, *EUR*, *MB*
  - can be inflected: *dra*, *OSiR-u*.

Stage 3: date parsing

- Straightforward for `date`-tagged tokens.
- Look for specific token/tag sequences, e.g.:
  - tag=`int` – tag=`rom` – tag=`int`
  - tag=`int` – tag=`m-*` – tag=`int` – tok=`"r"` – tok=`"."`
- Check day and month range.
- Merge tokens into one and assign them appropriate tag.
- Replace remaining `int` and `m-*` tags with `ara` and `word` respectively.