KernelTagger – a PoS tagger for very small amount of training data

Pavel Rychlý

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

December 2, 2017

Part of speech tagging

- Part of speech = Noun, Verb, Adjective, Adverb, Preposition, ...
- Assign a PoS to each token (word) in a text (sentence)
- Tagger is usually trained on an annotated corpus
- 100,000 tokens needed for training
- Manual annotation is expensive

Ready to use resources

- Large corpora from web
- Statistics/models derived from corpora
  - frequencies of words
  - frequencies of prefixes, suffixes, ...
  - word embedings (word → number[300])
  - similarity of words

HaBiT project

- corpora and tools for less resorced languages
- 4 Ethiopian languages, Czech, Norwegian
- annotation by native speakers
- 100–1200 annotated sentences
- need for a tagger based on such corpora

MiniTag

- created in 2016 by Lukas Banic
- word embedings using fasttext
- neural network (Keras) trained on annotated sentences

MiniTag

- created in 2016 by Lukas Banic
- word embedings using fasttext
- neural network (Keras) trained on annotated sentences
- best results with *context window size* $= 0$

KernelTagger

- most probable PoS tag for annotated words
- derive a PoS tag from 5 most similar words (kernel trick)
- word similarities from a big corpus

Word Similarity Computation

- Sketch Engine thesaurus

Word Similarity Computation

- Sketch Engine thesaurus

```
=r1/l1
    1:[] 2:[]
```

- context: one preceding and one following word
- logDice salience $D(w_a, c)$ of word w_a and context c.
- count only contexts with $D(w_a, c) > 0$
- similarinty of words w_a and w_b:

$$sim(w_a, w_b) = \frac{\sum_c min(D(w_a, c), D(w_b, c))}{\sum_c D(w_a, c) + \sum_c D(w_b, c)}$$

Evaluation

- DESAM corpus (1 mil tokens)
- only the main PoS, no grammatical atributes (12 tags)
- accuracy depending on the size of training data

train tokens	DESAM (1 mil.)	czTenTen (33 mil.)
1,000	70.7	72.9
10,000	78.8	81.7
100,000	87.7	88.5
980,000	92.9	92.8

Conclusion

- competitive results for very small annotated texts
- no dependency on magic (fastetext) or commercial software (Sketch Engine)