Improving neural tagging with lexical information

Benoît Sagot – Héctor Martínez Alonso
Inria (ALMAnaCH), Paris (France)

IWPT 2017 — 21st September 2017
Previous work and motivations

- State-of-the-art approaches to PoS tagging = machine-learning-based approaches relying on annotated corpora for training
  - Statistical approaches
  - Neural architectures — cf. (Plank et al. 2016)
- **Lexical information** helps to improve tagging accuracy
- Different types of lexical information
  - External lexicons as source of constraints or additional features in statistical approaches
  - (word-level) word embeddings extracted from large volumes of text and/or learned while training neural architectures such as LSTMs (Ling et al. 2015, Ballesteros et al. 2015, Plank et al. 2016)
  - Character-level (word) embeddings also capture lexical information, in a more “compositional” way, and have been shown to help dealing with low-frequency/unknown words (Plank et al. 2016)
- Motivation: **how do these different types of lexical information can contribute to tagging accuracy?**
  - How can we take into account external lexicons in a neural architecture?
  - Do external lexicons provide new, useful information w.r.t. word embeddings and character-level embeddings?
Architecture
Starting point: Plank et al.’s (2016) LSTM architecture
Integrating lexical information
Experimental setup
Data: corpora, word embeddings

• Corpora: **Universal Dependencies** dataset, v. 1.3 (Nivre *et al.* 2016)
  • Covers several dozen typologically diverse languages with annotated corpora of various sizes
• Pre-computed word embeddings: following Plank *et al.* (2016), we used Polyglot pre-computed embeddings (Al-Rfou *et al.* 2013)
  • Not available for all languages
Data: lexicons

• Two main sources

1. **Apertium and Giellatekno** projects
   • For languages for which only a morphological analyser (vs. lexicon) is available:
     • we used the corresponding monolingual part of OPUS’s OpenSubtitles2016
     • we tokenised it, extracted the 1 million most frequent tokens, and retrieved all their morphological analyses to create a “lexicon”
   • Rule-based conversion to UD PoS / UD Morph. Feats.
   • 2 lexicon variants: “coarse” (tag = UD PoS) + “full” (tag = UD PoS + UD Morph. Feats.)

2. Other existing lexicon, in particular **Alexina lexicons** (Sagot 2010), using only main categories, with a few language-specific adaptations

• **We only used the “best” lexicon for each language**
  • Selected based on tagging accuracy on dev sets
  • The “best” lexicon is almost never a “full” variant
Experimental setup

• **Implementation**: Extension of Plank et al.’s (2016) freely available source code (bilty)
  
  • standard configuration
    
    • 1 bi-LSTM layer
    • character-level embeddings size = 100
    • word embedding size = 64 (same as Polyglot embeddings)
    • no multitask learning
    • 20 iterations for training

• **Experimental settings**
  
  • with vs. without initialisation of the word embedding layer with pre-computed Polyglot word embeddings (when available)
  • with vs. without character-level embeddings
  • with vs. without external lexical information
Results
Overall results

• **Consistent improvements when using information from an external lexicon**
  
  • Greatest improvements = without character-level embeddings
    
    *Macro-average gain: +2.56, vs. +0.57 when also using character-based embeddings*
  
  • When also using pre-computed Polyglot embeddings, improvements are smaller
    
    *Macro-average gain: +0.21 (restricted to languages with Polyglot embeddings)*
Influence of corpus size

Accuracy gain when using an external lexicon as a function of the training corpus size

- **word embeddings**
- **word embeddings + character-level embeddings**
- **word embeddings w/ initialisation with pre-trained embeddings (Polyglot) + character-level embeddings**
Influence of type/token ratio

Accuracy gain when using an external lexicon as a function of the token/type ratio
Influence of unknown word rate

Accuracy gain w.r.t. tokens in the test set unseen in the training set but known to the lexicon

-1% +5%

-0.01 0.01 0.02 0.03 0.04 0.05

0 5 10 15 20 25

-1% +5%

-1% +5%

word embeddings only

word embeddings + character-level embeddings

word embeddings w/ initialisation with pre-trained embeddings (Polyglot) + character-level embeddings
A surprising result

Accuracy gain on all unknown words vs. on unknown words known to the corpus (configuration with character-based and Polyglot-initialised embeddings)
Conclusion and perspectives
Conclusion and perspectives

• Lexical information from morphological lexicons is helpful for neural tagging
  • Information provided by character-level embeddings and word embeddings is only partially the same

• Future work
  • Compare learning curves for the different neural configuration and non-neural (statistical) taggers
    • Preliminary experiments tend to show that a neural tagger does not perform significantly better on average than a MEMM tagger, provided external lexical information is used
  • Better understand what information is really helpful in the external lexicon, and what information is redundant with the different types of embeddings
    • Character-level embeddings capture regular morphology, for instance
Thank you