Open Questions in Sequence Tagging

Yoav Goldberg

What do we do when we do not have (enough) training data?

Out of domain tagging

- How do we accurately tag legal text? biomedical text? emails? tweets?
- These are hard for various reasons:
  - New words / old words with new behaviors
  - Different syntactic preferences
  - New linguistic phenomena! (what's the tag of ;) ?)

Out of domain tagging

- We have some training data in English News.
- We have plenty of text (without annotation) in Legal text.
- Can we build a good tagger for Legal?
- And what if we had **a few** annotations for Legal?

Tagging outside of English

- Maybe we do not have whitespace-separated words.

- Maybe we do not have large training data.

Tagging outside of English

- We have plenty of **English** data.
- We have text with no annotation in **Language X**.
- Can we improve on Language X?

Tagging outside of English

- We have plenty of **English** data + data for some other languages (Y, Z, U, W,...).

- Can we improve on Language X?

Guessing the tags of unknown words

- We have never seen a word before. What can we say about it?

- Can we learn from **inflections** of it that we see in the corpus?

Dictionary Based Tagging

- We have a dictionary ("lexicon") mapping each word to a list of possible tags (without probabilities)

- We also have un-annotated text. How can we use the lexicon and the text to improve tagging?