# LING439/539 - Statistical NLP
# Chapter 10. Part-of-speech tagging

Tuesday, September 13 2016

The ultimate goal of research on natural language processing is to *parse* and *understand* language.

→ Still far from achieving the goal

Much research in NLP has focused on **intermediate** tasks...

# Part-of-speech tagging

POS tagsets:

- Brown POS tagset
- Penn POS tagset
- Universal POS tagset
- ...

# Universal POS tags (2012)

A set of 12 universal part-of-speech tags:

| | |
|---|---|
| VERB | - verbs (all tenses and modes) |
| NOUN | - nouns (common and proper) |
| PRON | - pronouns |
| ADJ | - adjectives |
| ADV | - adverbs |
| ADP | - adpositions (prepositions and postpositions) |
| CONJ | - conjunctions |
| DET | - determiners |
| NUM | - cardinal numbers |
| PRT | - particles or other function words |
| X | - other: foreign words, typos, abbreviations |
| . | - punctuation |

Slav Petrov, Dipanjan Das and Ryan McDonald (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, `https://github.com/slavpetrov/universal-pos-tags`

# Universal POS tags (2016) from Universal Dependencies

A set of 17 universal part-of-speech tags:

| | |
|---|---|
| VERB | - verbs (all tenses and modes) |
| **AUX** | - auxiliary verb |
| NOUN | - nouns (common) |
| **PROPN** | - proper noun |
| PRON | - pronouns |
| ADJ | - adjectives |
| ADV | - adverbs |
| ADP | - adpositions (prepositions and postpositions) |
| CONJ | - conjunctions |
| DET | - determiners |
| NUM | - cardinal numbers |
| PRT → **PART** | - particles or other function words |
| **INTJ** | - interjection |
| **SCONJ** | - subordinating conjunction |
| **SYM** | - symbol |
| X | - other: foreign words, typos, abbreviations |
| . → **PUNCT** | - punctuation |

See `http://universaldependencies.org/u/pos`

# POS tagging approaches

- rule-based tagging
- transformation-based tagging (Brill's tagger)
- any sequence labeling algorithms...
  - HMM
  - ME
  - CRFs

# POS tagging resources

- Scottish Gaelic http://datashare.is.ed.ac.uk/handle/10283/2011
- Tamil http://au-kbc.org/nlp/corpusrelease.html (requires license agreement by email)
- Afrikaans http://rma.nwu.ac.za/index.php/resource-catalogue/afribooms.html
- Turkish http://ii.metu.edu.tr/corpus (requires sending a digital copy of a signed license agreement)
- Persian http://stp.lingfil.uu.se/~mojgan/UPDT.html
- Norwegian http://www.nb.no/sprakbanken/show?serial=sbr-10
- BrazPortugese Newswire http://www.nltk.org/nltk_data
- Dutch Alpino https://www.let.rug.nl/vannoord/trees
- Spanish https://www.iula.upf.edu/recurs01_tbk_uk.htm
- Italian-TurinTree/Parallel http://www.di.unito.it/~tutreeb/treebanks.html
- Polish National Corpus http://nkjp.pl/index.php?page=14&lang=1
- Icelandic-Historical Corpus http://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)
- Icelandic http://www.malfong.is/index.php?lang=en&pg=mim
- Slovene-English Parallel Corpus http://nl.ijs.si/elan/
- Finnish Treebank http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/
- German Tiger http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
- German Hamburg Treebank https://corpora.uni-hamburg.de/drupal/en/islandora/object/treebank:hdt
- Russian Open Corpus http://opencorpora.org/?page=downloads
- Italian-Pisa http://www.corpusitaliano.it/en/contents/description.html
- English https://corpling.uis.georgetown.edu/gum/
- Coptic https://github.com/CopticScriptorium/corpora
- French https://deep-sequoia.inria.fr/corpus/
- French https://perso.limsi.fr/pap/free_multitag.tgz
- Danish https://code.google.com/p/copenhagen-dependency-treebank/
- Croatian http://nlp.ffzg.hr/resources/corpora/setimes-hr/
- Swedish Talbanken http://stp.lingfil.uu.se/%7Emojgan/UPDT.html
- English Ted Talk Treebank http://ahclab.naist.jp/resource/tedtreebank

Multi Universial Dependencies http://universaldependencies.org

# Rule-based part-of-speech tagging

1. assign each word a list of potential POS labels using the **dictionary**
2. winnow down the list to a single POS label for each word using **large lists of hand-written disambiguation rules**

# Adverbial-that rule

```
Given input: “that”
if
    (+1 A/ADV/QUANT); /* if next word is adj, adverb, or quantifier */
    (+2 SENT-LIM);    /* and following which is a sentence boundary. */
    (NOT-1 SVOC/A);   /* and the previous word is not a verb like */
           /* ‘consider’ which allows adjs as object complements */
then eliminate non-ADV tags
else eliminate ADV tag
```

# Transformation-based tagging

Transformation-based learning (TBL)