# A Universal Part-of-Speech Tagset (Petrov, Das, McDonald)

# Abstract

- 12 universal part of speech tags
- mappings from 25 different treebank tagsets used
- Coverage of 22 different languages
- Show grammar induction for predicted part of speech tags using these "universal" tags

# Introduction

- Recent interest in unsupervised POS tag induction and cross-lingual projection of POS tags.

> “Underlying these studies is the idea that a set of (coarse) syntactic POS categories exist in similar forms across languages"

- When corpora that use a standard tagset are not available, typically a mapping from fine-grained tags to a more universal POS tag set is done.
    - Das and Petrov (2011) was an example of this
- Purposes of constructing this tagset:
    - useful for evaluating unsupervised and cross-lingual taggers
    - allows for meaningful comparisons across languages when looking at supervised taggers *though the size of the corpus used can still fluctuate, at least the tagset size and distribution is roughly consistent*

    - simplifies the development of taggers across multiple languages (less annotation guideline specific information has to be utilized).
- Experiments herein:
    - POS tagging accuracy for 25 different treebanks
    - unsupervised grammar induction system for multiple languages (relying on Das and Petrov (2011) and Naseem et al, (2010).

# Tagset

- Adopt a pragmatic focus, trying to find the POS categories that they expect to be most useful for users of POS taggers. - The focus is on utility for downstream tasks and grammar induction tasks
- > majority of tagsets are very fine-grained and very language specific
- Smith and Eisner (2005) made a set of 17 English POS tags from the conventional 17 *though these did not emphasize the multilingual utility of these tags*
- > McDonald and Nivre (2007) identified eight different coarse POS tags when analyzing the errors of two dependency parsers on the 13 different languages form the CoNLL shared tasks.

# The tags

- NOUN
- VERB
- ADJ
- ADV
- PRON
- DET
- ADP (ADPOSITIONS)
- NUM
- CONJ
- PRT (PARTICLES)
- . (PUNCTUATION)
- X (CATCH ALL)

> we did not rely on intrinsic definitions of the above categories. Instead each category is defined operationally.

  - By this, they mean that they defined these part of speech tags in their relationship to fine-grained POS tags from other treebanks
- Some tags do not occur in all languages *Adjectives don't occur in Wolof if I'm remembering that paper correctly*
  - For Korean, they treated stative verbs that would translate as adjectives in english as adjectives *this seems like a bad, Anglocentric way of doing things*.
- One important thing about these mappings is that they were established to encourage collaboraation and refinement from researcheres working on other languages (using version control etc).
- *The languages considered are very Indo-European, only 7 of the 25 treebanks are non-IE languages. However, this is probably better than most researchers were doing at the time towards including other non-IE languages*

# Experiments

# POS tagging accuracy comparison

- Model: trigram markov model
  - chosen for speed, state of the art accuracy without much tuning
- Using the universal tags reduced the variance in performance across langs from 10.4 to 5.1.
- Still differences across languages
  - Japanese is very good (99% acc), Turkish worse (90.2% acc)
- > The best results are obtained by training on the original fine-graineed tags and then mapping to the UPOS tags at the end

  - > The transition model based on the universal POS tagset is less informative

# Grammar induction

- Previous research on unsupervised grammar induction assumed gold POS tags. They remove this simplification using POS tags that are automatically projected from English

- Das and Petrov (2011) use cross-lingual projection to lear POS taggers without labeled data the target lang, these induced tags are used to learn unsupervised grammar.
- Using Naseem (2010)'s model where a small set of universal syntactic rules constraina bayesian model *I should read that paper if I want to make sense of what was done here*
- Using treebanks from the CoNLL-X shared task (eight indoeuropean languages used by Das and Petrov (2011))
- The method described for the grammar induction experiments in this paper are best with the gold UPOS tags performing a little better (though this wasn't the case for all languages examined, swedish for example did better with the automatically generated tags)

Revision #4

Created Wed, Apr 15, 2020 10:10 PM by kenneth

Updated Thu, Apr 16, 2020 12:59 AM by kenneth