

Universal Dependencies

- [A Universal Part-of-Speech Tagset \(Petrov, Das, McDonald\)](#)
- [Universal Dependencies v1: A Multilingual Treebank Collection](#)

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 from 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 collaboration and refinement from researchers 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-grained 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 learn 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 constrain a 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)

Universal Dependencies v1: A Multilingual Treebank Collection

(Nivre, Marneffe, Ginter,
Goldberg, Hajic, Manning,
McDonald, Petrov, Pyysalo,
Silveira, Tsarfaty, Zeman)

Introduction

- When looking at three different, related languages (Swedish, Danish and English) that represent parallel sentences using parallel structure, the shared dependency relations have only 40% overlap
- Goals of UD:
 -

“Develop cross-linguistically consistent treebank annotation for many languages

“ capture similarities as well as idiosyncracies among typologically different languages

- support the following research activities:
 - comparative evaluation
 - cross-lingual learning
 - *not sure if this means human language learning or machine learning*
 - multilingual natural language processing
 - comparative linguistic studies
- This work is a fusion of several other initiatives (Stanford dependencies, Google universal dependencies, Intersect morphosyntactic tag sets)

History

UD today is dependent upon prior research

- Morphological layer
 - Google universal tagset grew from cross-lingual error analysis (McDonald & Nivre 2007)
 - Intersect (Zeman 2008) started as a tool for converting between the morphological tagsets of different languages
- Dependencies (syntactic layer)
 - Stanford dependencies developed for English in 2005
 - Adapted to several other languages

What other UD-like projects

existed?

- Google UDT project (McDonald et al 2013) was first to combine google POS tags and Stanford dependencies
- HamleDT v2 "provided Stanford/Google annotation for 30 languages by automatically harmonizing treebanks with different native annotations"
- Universal Stanford Dependencies revised stanford dependencies for cross-linguistic use

Annotation guideline principles

- Based upon dependencies
- based upon lexicalism
 - “ words are the basic units of grammatical annotation
- syntactic wordhood != orthographic wordhood
- Recoverability principle
 - “ there should be a transparent relation between the original textual representation and the linguistically motivated word segmentation
- maximize the parallelism between languages
 - ensuring the same construction is annotated in the same way across languages
 - don't want to annotate things that do not exist in a language simply because that's how they work in other languages *this seems to conflict with the annotation of Korean stative verbs as Adj for the Universal POS tagset paper*
- “ use a universal pool of structural and functional categories that languages

select from

- “possible to refine the analysis by adding language-specific subtypes

Word segmentation

- Clitics split off
- contractions are undone *seems like a strange decision. why not split up compounds too if you're undoing contractions*
- “UD currently does not allow words with spaces

Morphology

Lemma

No guidelines provided for what the lemmas should look like. E.g. should lemmas include derivational morphemes, what should you do for suppletives etc.

Part of speech tag

- 17 part of speech tags, a fixed set for all languages to draw from but not all tags need to be present in all languages

Morphological features

- Based on the interset system
- Each feature is associated with a set of possible values

Syntax

- 40 different grammatical relations for version 1.0
- 3 types of structure:
 - nominals
 - clauses
 - modifier words
- Distinction between core arguments and other dependents which is different from complements vs adjuncts.
 - Core arguments are subjects and objects, *other arguments are non-core even if they are required by the verb*
- The attachment point of a relation is crucial
 - For example, an adverbial clause that modifies a noun is **acl**, an adverbial clause that modifies a predicate is **advcl**
- Rich collection of noun dependents
- Relations for non-edited/informal text also included
 - e.g. reparandum
 - goeswith
- compounding
 - mwe for fixed expressions containing function words *largely corresponds to* **fixed** in UD v2
 - name for names consisting of multiple proper nouns *largely corresponds to* **flat** in UD v2
 - compound is used for any kind of lexical compounding *still* **compound** in UD v2

mwe and name are both left headed with a flat structure (e.g. all are connected to the left-most part of the name or mwe). *This is carried over to* **fixed** *and* **flat** *in UD v2 which means I need to fix some of my names that I've annotated*

Relations between content words

- Priority is given for dependency relations between content words
 - Increases chances of parallel structure between languages because functional words

can just be indicated using morphology or other non-syntactic means

“ The UD view is that we need to recognize both lexical and functional heads, but in order to maximize parallelism across languages, only lexical heads are inferable from the topology of our tree structures

- Very close to the view of Tesnière (1959) *the OG dependency grammar*

Language-specific relations

- UD allows the use of language-specific relations to capture extra stuff