Universal Depedencies 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 langauges

    - 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, Interset 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)
    - Interset (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 teh same way across langauges
  - don't want to annotate thigns 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 ssytem
- 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 attachement 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 propoer 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 Tesniere (1959) *the OG dependency grammar*

Language-specific relations

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

---

Revision #6

Created Thu, Apr 16, 2020 1:50 PM by kenneth

Updated Thu, Apr 16, 2020 8:53 PM by kenneth