All Roads Lead to UD: Converting Stanford and Penn Parses to English Universal Dependencies with Multilayer Annotations

SIYAO PENG & AMIR ZELDES
{sp1184, amir.zeldes}@georgetown.edu
GEORGETOWN UNIVERSITY
C@GU

LAW-MWE-CxG-2018 @ COLING, Santa Fe, NM

Overview

- Universal Dependencies (UD) provides treebanks in 50+ languages with a unified scheme (Nivre et al. 2017)
- UD still being revised (now v2.2), older Stanford Dependencies (SD) frozen
- SD2UD conversion more reliable than gold constituent2UD by around 10%
- Head/label error rates: 1.73%/1.38% for pure SD & 0.45%/0.42% for multilayer
- Annotating in SD and converting into latest UD allows stable corpus annotation; access to additional annotations almost eliminates conversion errors

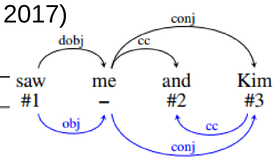
Corpora

	Georgetown University Multilayer Corpus (GUM) (Zeldes 2017)	English Web Treebank (EWT) (Bies et al. 2012, Silveira et al. 2014)
Documents	101	1,174
Tokens	85k	250k
Genres	(8) news, interviews, how-to, travel, academic, bios, fiction & forums	(5) blogs, e-mail, newsgroups, online answers & reviews
Scheme	SD, coref, entities, discourse parsing & more	Constituent trees

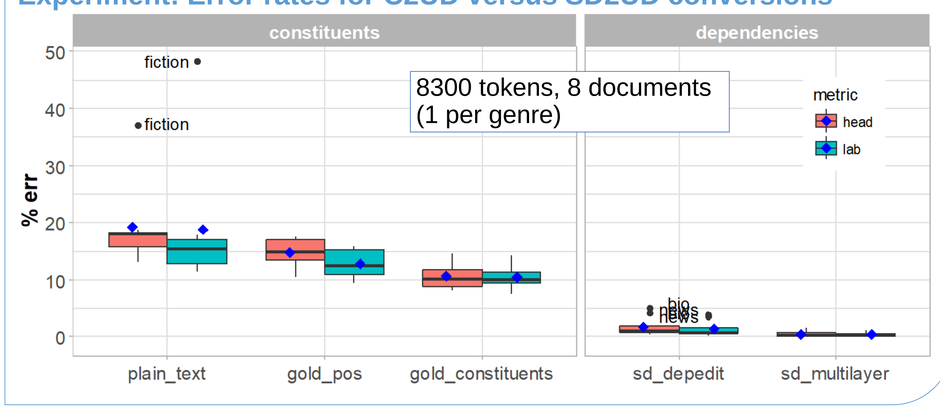
Configurable rule-based conversion from SD to UD

- **Three steps:**
  1. Pull information from other annotation layers (if available, e.g. in GUM)
  2. **Main rule-based conversion** consisting of ~100 rules applied in order
  3. Attaching punctuation using Udapi API (Popel et al. 2017)
- Examples of conversion rules:

attributes	relations	actions
func=/dobj/	none	#1:func=obj
func=/.*/;func=/^cc$/;func=/^conj$/	#1>#2;#1>#3	#3>#2
func=/prep/;pos=/^W.*/;func=/pcomp/	#1>#3;#3>#2	#2:func=pobj;#1>#2;#2>#3;#3:func=rcmod

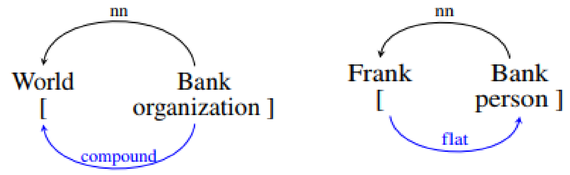


Experiment: Error rates for C2UD versus SD2UD conversions

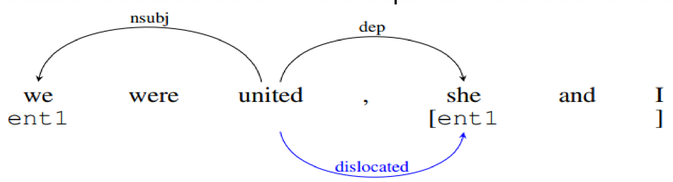


Using Multilayer Annotations

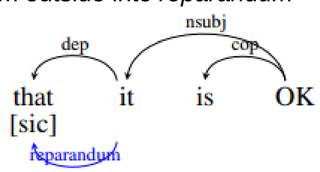
- **Knowing entity type is crucial:**
  - All proper names are annotated as *nn* (noun compound modifier) in SD
  - *Organization* vs. *person* distinguishes *compound* vs. *flat* in UD:
    - *compound*: headed NP with internal structure, e.g. *World Bank*
    - *flat*: headless multi-word expressions, e.g. person names, *Frank Bank*



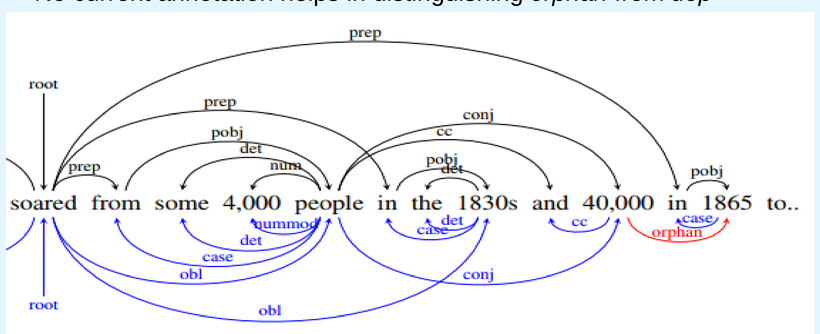
  - Exception: organization names can be headless, e.g. *Wells Fargo*.
- **Coreference is informative for determining dislocation:**
  - *Dislocated* node is coreferent with a dependent of the same verb:



- **TEI annotations provide disfluency information:**
  - *reparandum*: the head of an 'aborted' part that is attached to its repair
  - *TEI XML* tag <*sic*> denotes errors in GUM; convert *dep* inside an error and governed from outside into *reparandum*



- **An unsolved problem: *orphan***
  - Promoted *orphan* dominates the child of missing coordinate parents
  - No current annotation helps in distinguishing *orphan* from *dep*



Results: Top 3 errors by conversion scenario

scenario	head errs		lab errs	
C2UD (gold)	84	nsubj	130	obl
	82	nmod	74	nmod
	71	conj	62	conj
SD (pure)	37	flat	37	flat
	10	nmod	8	obl
	8	appos	7	nsubj
SD (multi)	8	compound	9	compound
	6	nmod	7	obl
	6	flat	6	nmod

- **Other labels are rare but systematically wrong:** *dislocated, reparandum, goeswith* are absent in C2UD

Cross-corpora comparison: non-projectivity

	C2UD	UD V2.2 (corrected)	
EWT	0.34%	0.46%	
	C2UD	**UD V2.2 (from SD multi)**	**original SD**
GUM	0.29%	0.79%	0.63%

- Non-projectivity is more frequent in GUM, 'native dependencies', than in EWT, 'native constituents'
- Low non-projectivity for gold EWT UD may be due to genre differences or reflex of C2UD

Future work

- GUM continues to grow – look for version 5 in winter!
- Plans to use more annotation layers, e.g. using the RST *purpose* annotation to differentiate adverbial clause (*advcl*) from controlled to-infinitives (*xcomp*)
- Figure out what to do with *orphan*… ☹

QR scan: e-poster available

References

- Bies, A. et al. 2012. *English Web Treebank*. Linguistic Data Consortium, Technical Report LDC2012T13, Philadelphia, PA.
- Nivre, J. et al. 2017. *Universal Dependencies 2.0*. Charles University.
- Popel, M. et al. 2017. Udapi: Universal API for universal dependencies. *UDW2017*, 96–101.
- Silveira, N. et al. 2014. A gold standard dependency corpus for English. In *Proc. LREC-2014*. Reykjavik, 2897–2904.
- Zeldes, A. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation* 51(3), 581–612.

- Converter code: https://corpling.uis.georgetown.edu/depedit/
- GUM corpus: http://corpling.uis.georgetown.edu/gum/
- conversions: https://github.com/gucorpling/GUM_UD_LAW2018