Bootstrapping the Annotation of UD Learner Treebanks

Arianna Masciolini
Språkbanken Text, Department of Swedish, Multilingualism, Language Technology, University of Gothenburg

Abstract

Learner data comes in a variety of formats, making corpora difficult to compare with each other. Universal Dependencies (UD) has therefore been proposed as a replacement for the various *ad-hoc* annotation schemes. Nowadays, the time-consuming task of building a UD treebank often starts with a round of automatic annotation. The performance of the currently available tools trained on standard language, however, tends to decline substantially upon application to learner text. Grammatical errors play a major role, but a significant performance gap has been observed even between standard test sets and normalized learner essays. In this paper, we investigate how to best bootstrap the annotation of UD learner corpora. In particular, we want to establish whether grammar-corrected learner sentences are suitable training data for fine-tuning a parser aimed for original (ungrammatical) L2 material. We perform experiments using English and Italian data from two of the already available UD learner corpora. Our results show manually annotated corrections to be highly beneficial and suggest that even automatically parsed sentences of this kind might be helpful, if available in sufficiently large amounts.

Keywords: **Second Language Acquisition, Learner Corpora, Dependency Parsing, Universal Dependencies**

The Case for UD Learner Treebanks

UD can benefit learner corpus research since:

- it provides a **uniform morphosyntactic annotation layer**, making it easier to compare different corpora
- it allows for **cross-lingual comparisons** between standard and learner language, as well as between different L2s
- if learner sentences are paired with corrections, UD annotation can replace explicit error tagging. This is format is known as ***L1-L2 treebank***
- the **existing parsers** can semi-automate the annotation process

L2 Parsing is Challenging

- **grammatical errors** are widely known to negatively affect performance
- parsers are usually trained on other text types (news, social media, Wikipedia articles, fiction...), → **learner productions are out of domain** (regardless of their grammaticality)

Research Questions

How to best bootstrap the annotation of new L1-L2 treebanks?

And more specifically:

- **does fine-tuning on corrections improve parser performance on learner originals?**
- if this is the case, **are automatically annotated corrections sufficient?**

Approach

1. **select** suitable **BERT models**
2. **train baselines** by fine-tuning for dependency parsing on standard language, using large reference treebanks, until performance is comparable to off-the-shell tools*
3. further **fine-tune on**:
   - **gold-annotated corrections** (when available)
   - **silver-annotated corrections****
4. **evaluate** the resulting parsers **on**:
   - **standard test sets** (expecting a potential performance decline)
   - **corrections** (expecting substantial performance improvements)
   - **original L2 sentences** (expecting smaller, but still significant improvements)

* fine-tuning utilizes the MaChAmp toolkit, the reference parser is UDPipe 2
** gold = manually annotated/validated; silver = automatically annotated

Data

treebank	language	# sentences		
		train	dev	test
EWT	standard en	12544	2001	2077
ESL	learner en	2×5124	2×100	2×5024
ISDT	standard it	13121	564	482
VALICO	learner it	2×1613	2×233	2×398

Learner treebanks:

- **ESL** (English as a Second Language treebank)
  - based on the First Certificate in English corpus
  - short essays for upper intermediate (B1) English test
  - wide variety of language backgrounds
  - fully manually annotated
- **VALICO** (*Varietà Apprendimento Lingua Italiana Corpus Online****)
  - narrative texts elicited by comic strips
  - native speakers of 4 western European languages
  - 1 to 4 years of study
  - only the test set it manually validated

*** "online corpus of learner varieties of the Italian language"

Conclusions

- the experiments on English data strongly suggest that **fine-tuning on gold-annotated corrections produces performance improvements on original learner productions**
- **whether automatically annotated corrections are helpful** is less clear and **might depend on proficiency level**, as well as on training set size

Recommendation

When annotating new learner treebanks, start with manually annotating/validating corrections and use them to train a domain-specific parser to bootstrap the annotation of learner original.

Other observations

- UDPipe 2 models seem to have better cross-domain generalization capabilities
- MaChAmp is extremely effective for training domain-specific parsers

Acknowledgements

This work is supported by the Swedish national research infrastructure Nationella Språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

Results

	EWT		ESL L1		ESL L2	
	LAS	UAS	LAS	UAS	LAS	UAS
baseline	**91.79**	**93.64**	86.43	90.18	85.21	89.38
ft-gold	84.32	88.67	**98.92**	**99.65**	**95.28**	**97.05**
ft-silver	86.61	90.55	90.70	93.44	89.32	92.46
UDPipe 2	90.56	92.62	90.70	93.44	89.42	92.51

Table 1: LAS and UAS score for the full-scale English experiment

	EWT		ESL L1		ESL L2	
	LAS	UAS	LAS	UAS	LAS	UAS
baseline	**91.79**	**93.64**	86.43	90.18	85.21	89.38
ft-gold-s	84.11	88.70	**94.53**	**96.50**	**92.21**	**94.82**
ft-silver-s	86.27	90.32	90.46	93.24	88.95	92.18
UDPipe 2	90.56	92.62	90.70	93.44	89.42	92.51

Table 2: LAS and UAS scores for a smaller-scale English experiment

	ISDT		VALICO L1		VALICO L2	
	LAS	UAS	LAS	UAS	LAS	UAS
baseline	**93.64**	**95.21**	89.25	91.86	85.99	89.94
-	-	-	-	-	-	-	-
ft-silver	89.96	93.15	88.49	91.46	85.59	89.77
UDPipe 2	93.34	94.96	**90.22**	**92.86**	**87.69**	**91.61**

Table 3: LAS and UAS scores for the Italian experiment

Results reported in Table 2, obtained by fine-tuning a 1613-sentence ESL sample, are to be compared with those in Table 1 (same language, different training set size) and 3 (different language, same training set size).