Universal Dependency Parsing from Scratch

Peng Qi* Timothy Dozat* Yuhao Zhang* Christopher D. Manning
Stanford University
(* = Equal contribution)

System Highlights

Fully neural pipeline system for all stages required by the Shared Task, including: word/sentence segmentation, POS/UFeats tagging, lemmatization, and dependency parsing.

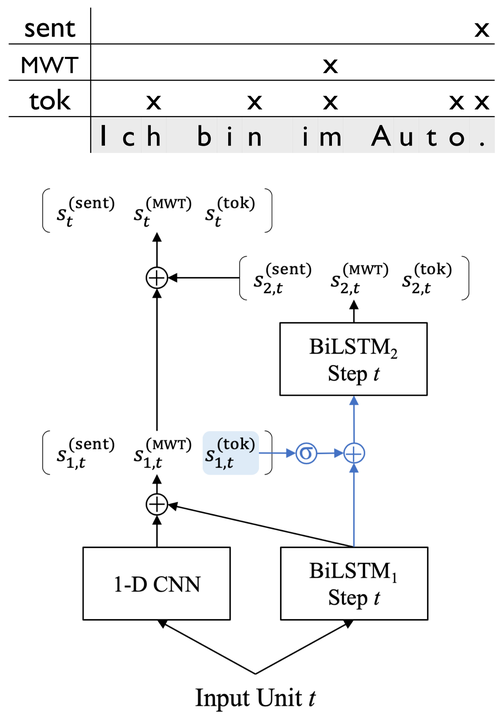
	LAS	MLAS	BLEX
Stanford	72.29 (7th)	60.92 (2nd)	64.04 (5th)
	↓	fixed in the tokenizer	
Stanford+	74.16	62.08	65.28
Top System	75.84	61.25	66.09

Key Contributions

- Symbolic statistics 🔗 Neural nets
- Biaffine POS/UFeats Classifier → Consistency
- Lemmatizer + Edit Classifier → Shortcut for long sequence
- Biaffine parser with relative location features
- Fully open-source implementation/models

Tokenizer / Sentence Segmenter

Joint word and sentence segmentation as tagging.



Multiword-Token (MWT) Expansion

German example: im → in dem

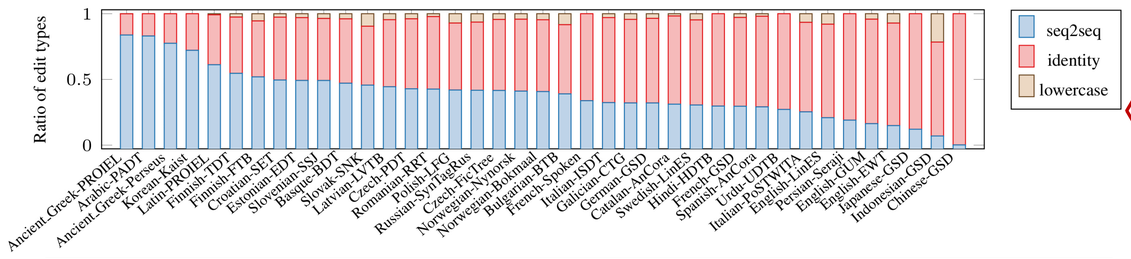
Combining a frequency lexicon (symbolic) with a seq2seq model as fallback (neural).

System	Tokens	Sentences	Words
Stanford+	**99.46**	**91.33**	**99.27**
- conv	99.45	91.03	98.67
- seq2seq	-	-	98.97

Lemmatizer

- 2 Frequency lexicons (word x UPOS, word) + seq2seq
- Edit (e.g., lowercase, identity) classifier to skip decoder

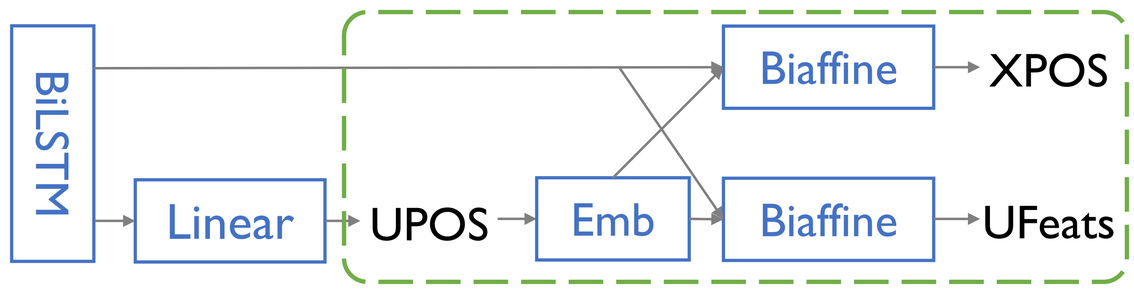
e.g., URL, email, proper nouns, sentence-leading capitalization, etc.



System	Big	Small	Low-Res	All
Stanford	**96.56**	**91.72**	**69.21**	**94.22**
- edit & seq2seq	89.97*	82.68*	63.50*	87.45*
- lexicons	95.37*	90.43*	66.02*	92.89*

POS / UFeats Tagger

Biaffine classifier that conditions on UPOS embeddings to enforce **consistency** between predicted tags/features.



System	UPOS	XPOS	UFeats	AllTags	PMI
Stanford	**96.50**	**95.87**	**95.01**	**92.52**	**.0514**
- biaffine	96.47	95.71*	94.13*	91.32*	.0497*

Dependency Parser

Deep Biaffine parser + terms to model relative positions of heads and dependents:

- Linearization: If there's an edge, are the words in the right order?
- Distance: If there's an edge, are the words close/far enough?

System	LAS	CLAS
Stanford	**87.60**	**84.68**
- linearization	87.55*	84.62*
- distance	87.43*	84.48*

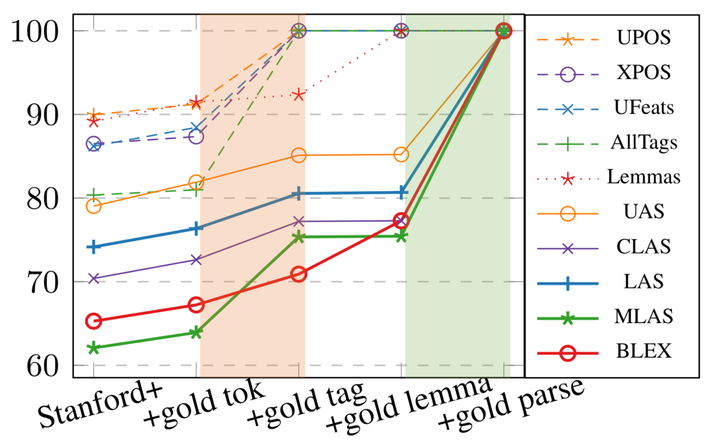
Full Pipeline

Highly accurate system on small and large tree-banks alike.

Treebanks	System	LAS	MLAS
Small	Stanford+	**83.90**	**72.75**
	Top System	69.53	49.24
Low-Res	Stanford+	**63.20**	**51.64**
	Top System	27.89	6.13
PUD	Stanford+	**82.25**	**74.20**
	Top System	74.20	58.75

Easy to make iterative improvements to the performance of the entire pipeline.

Most of the potential for gains still remain with the challenging tasks of **tagging** and **parsing**.



Code and Pretrained Models

Tensorflow tagger/parser code at
https://github.com/tdozat/Parser-v3

PyTorch full pipeline at
https://github.com/stanfordnlp/stanfordnlp

For more details see (also in QR code):
https://stanfordnlp.github.io/stanfordnlp