Prague at EPE 2017: The UDPipe System

Milan Straka, Jana Straková, **Jan Hajič**

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
{straka,strakova,**hajic**}@ufal.mff.cuni.cz

EPE 2017, 20th September 2017

Introduction

EPE 2017

- you already know

UDPipe

- a trainable pipeline which performs sentence segmentation, tokenization, POS tagging, lemmatization and dependency parsing
- models for all 50 languages of UD 2.0
- easily trainable using data in CoNLL-U format

Introduction

EPE 2017

- you already know

UDPipe

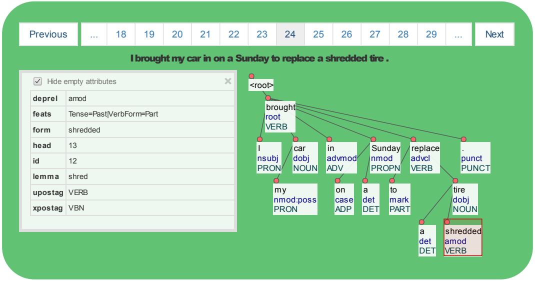
- a trainable pipeline which performs sentence segmentation, tokenization, POS tagging, lemmatization and dependency parsing
- models for all 50 languages of UD 2.0
- easily trainable using data in CoNLL-U format

UDPipe

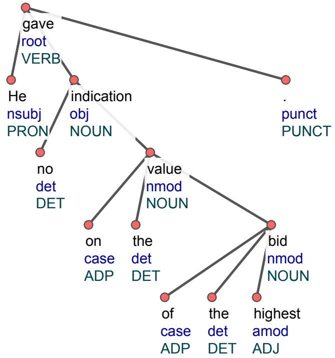
- `http://ufal.mff.cuni.cz/udpipe`
- open-source Mozilla Public License (MPL)
- models under CC BY-SA-NC license
- bindings for C++, Python, Perl, Java, C#

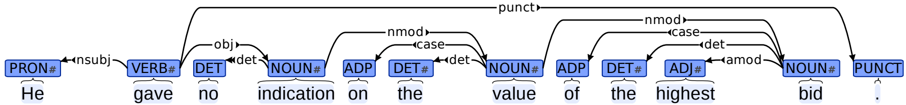
```
pip install ufal.udpipe
cpan UFAL::UDPipe
```

- REST web service
  - `http://lindat.mff.cuni.cz/service/udpipe/`



Universal Dependencies





Tokenizer

- fully trainable from CoNLL-U training data
  - CoNLL-U v1 allows reconstruction of spaces between tokens in one sentence
  - CoNLL-U v1 does not provide markup for paragraph and document boundaries, which are often indicated by visual layout and/or spacing

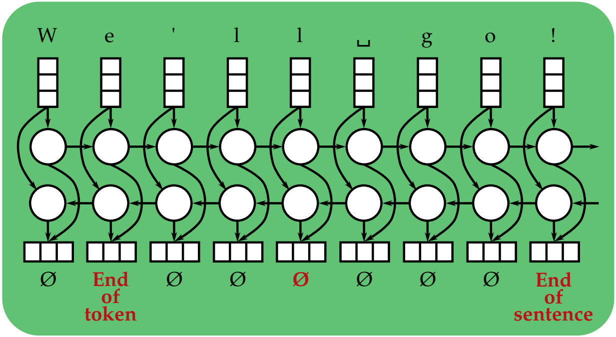
    Keep in touch, / Mike / Michael J. McDermott

    i have two options / using the metro or the air france bus / can anybody tell me if the metro runs directly ...

  - CoNLL-U v2 does include markup for paragraph and document boundaries, but only document boundaries are marked in English UD 2.0 data

Tokenizer Architecture

- bidirectional character-level GRU network
- predicts break type after every character
  - no break
  - token break
  - sentence break



Tagger

- suffix based guesser predicting most frequent *(UPOS, XPOS, FEATS)* candidates from data

SUFFIX	UPOS	XPOS	FEATS	LEMMA RULE
	VERB	VBG	VerbForm=Ger	remove ing
	NOUN	NN	Number=Sing	keep unchanged
	VERB	VBG	VerbForm=Ger	remove ing, append e
-ing	ADJ	JJ	Degree=Pos	keep unchanged
	PROPN	NNP	Number=Sing	keep unchanged
	VERB	VBG	VerbForm=Ger	remove ting
	VERB	VBG	VerbForm=Ger	remove ping

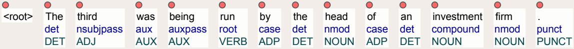
- to generate candidates for a given form
  - all candidates for the same form in the training data
  - several most-frequent candidates using the suffix guesser
- disambiguated by averaged perceptron utilizing a predefined rich set of feature templates and Viterbi decoding of order 3

Lemmatizer

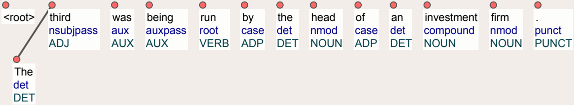
- very similar to tagger, but utilize *(UPOS, lemma rule)* candidates
- *lemma rule* is the shortest formula generating lemma from a given form, using any combination of
  - remove a specific prefix
  - remove a specific suffix
  - append a prefix
  - append a suffix

Transition-Based Dependency Parsing

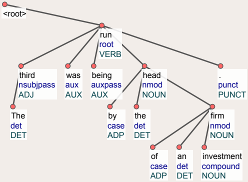
- initial configuration



- transitions modify the configuration

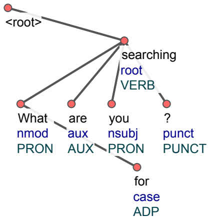


- final configuration



Parser

- fast neural dependency parser
  inspired by Chen and Manning (2014)
- several transition systems
  - projective (`arc standard`)
  - partially non-projective (`arc2`)
  - fully non-projective (`swap`)
- multiple oracles
  - static oracles for all systems
  - dynamic oracle for `arc standard` system
  - search-based oracle for all systems



UDPipe EPE 2017 Runs

Run name	Run	Description	Tokens
UD2.0 En/UDPipe/20	0	UDPipe 1.2, UD 2.0 English data, UDPipe tokenizer, beam size 20	204.5k
UD2.0 En/EPE/20	1	UDPipe 1.2, UD 2.0 English data, EPE provided tokenizer, beam size 20	204.5k
UD2.0 EnMerged/UDPipe/20	2	UDPipe 1.2, UD 2.0 English + English LinES + English ParTUT data, UDPipe tokenizer, beam size 20	292.2k
UD2.0 EnMinus/UDPipe/5	3	UDPipe 1.1, first 95% of UD 2.0 English, UDPipe tokenizer, beam size 5	192.5k
UD1.2 En/UDPipe/5	4	UDPipe 1.0, UD 1.2 English data, UDPipe tokenizer, beam size 5	204.5k
Stanford-Paris	*6*	*UD v1 enhanced dependencies, WSJ+Brown+GENIA data*	*1692.0k*

Extrinsic Results

UDPipe run	Event extraction	Negation resolution	Opinion analysis	Overall score
0-UD2.0 En/UDPipe/20	43.58	58.83	59.79	54.07
1-UD2.0 En/EPE/20	45.54	61.62	61.00	**56.05**
2-UD2.0 EnMerged/UDPipe/20	44.25	59.95	58.71	54.30
3-UD2.0 EnMinus/UDPipe/5	42.70	59.95	58.90	53.85
4-UD1.2 En/UDPipe/5	43.22	50.85	58.53	50.86
Stanford-Paris, run 6	*50.23*	*66.16*	*65.14*	***60.51***

Higher numbers are better.

- overall score of 56.05, lacking behind nearly all other participant systems
- overall scores better systems 56.23, 56.24, 56.65, 56.81, 58.57, 60.51
- best results for EPE-provided tokenizer
- UD 2.0 offers only 200k of English training data
- we emphasize that even though the EPE 2017 shared task focused on English language only, UDPipe is trained in a language agnostic manner for 50 languages without any adaptation for English other than setting up the hyperparameters of the artificial neural networks

Intrinsic Results

Row	Data	Plain text processing						Using gold tokenization			
		Words	Sents	UPOS	XPOS	UAS	LAS	UPOS	XPOS	UAS	LAS
0-UD2.0 En/UDPipe/20	UD 2.0 En	**99.0**	**75.3**	**93.5**	**92.9**	**80.3**	**77.2**	94.4	93.8	**84.6**	**81.3**
	UD 2.0 EnMerged	**98.9**	**79.5**	91.8	—	78.4	73.9	92.7	—	81.4	76.6
	UD 1.2 En	**99.0**	**75.3**	87.9	**92.9**	75.7	63.7	88.8	93.8	79.1	66.8
1-UD2.0 En/EPE/20	UD 2.0 En	96.2	59.9	90.7	90.0	74.6	71.8	94.4	93.8	**84.6**	**81.3**
	UD 2.0 EnMerged	97.8	71.0	90.6	—	75.8	71.4	92.7	—	81.4	76.6
	UD 1.2 En	96.2	59.9	85.1	90.0	70.3	58.7	88.8	93.8	79.1	66.8
2-UD2.0 EnMerged/UDPipe/20	UD 2.0 En	**99.0**	**75.3**	93.4	92.6	79.8	76.7	94.4	93.6	84.0	80.6
	UD 2.0 EnMerged	**98.9**	**79.5**	**92.0**	—	**79.1**	**74.9**	**92.9**	—	**82.2**	**77.7**
	UD 1.2 En	**99.0**	**75.3**	87.8	92.6	75.6	63.4	88.7	93.6	78.9	66.3
3-UD2.0 EnMinus/UDPipe/5	UD 2.0 En	98.7	73.2	93.1	92.4	78.9	75.8	**94.5**	**93.9**	83.8	80.7
	UD 2.0 EnMerged	98.8	78.6	91.6	—	77.7	73.1	92.8	—	81.1	76.3
	UD 1.2 En	98.7	73.2	87.5	92.4	74.6	62.6	88.9	**93.9**	78.6	66.3
4-UD1.2 En/UDPipe/5	UD 2.0 En	98.4	72.3	87.3	92.2	73.9	62.0	88.8	93.8	78.8	66.3
	UD 2.0 EnMerged	98.7	77.8	86.5	—	73.9	60.1	87.6	—	77.2	63.0
	UD 1.2 En	98.4	72.3	**92.9**	92.2	**78.3**	**75.1**	**94.5**	93.8	**84.2**	**80.7**

Higher numbers are better.

Observations

Tokenization Issues

- UDPipe tokenizer of lower quality – EPE-provided tokenizer improves score by 2 points in EPE 2017
- opposite results in intrinsic evaluation on UD 2.0 test set

Merged English Treebanks

- merged UD 2.0 English treebanks show improvements in extrinsic evaluation, even if they have inconsistent XPOS tags and show performance drop in intrinsic evaluation

Negation Resolution Drop of Run 4

- poor Unicode character handling by UDPipe 1.0 tokenizer as no Unicode in UD 1.2 English training data

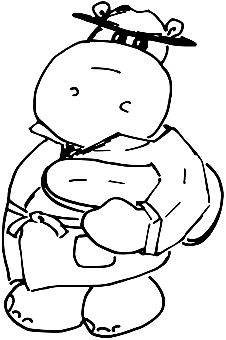
Conclusions and Future Work

Conclusions

- evaluation of language-agnostic UDPipe pipeline in the EPE 2017 shared task

Immediate Future Work

- when the paragraph boundaries are annotated in the UD data, does the trained sentence segmenter achieve better performance
- can a rule-based English tokenizer also improve the results
- what effect would larger training data (like WSJ) have
- what performance would a state-of-the-art dependency parser attain using the UD 2.0 data only



Questions?