From Raw Text to Universal Dependencies – Look, No Tags!

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg and Joakim Nivre

UPPSALA UNIVERSITET

Bar-Ilan University

Word and Sentence Segmentation

We model **joint** sentence and word segmentation as a **character-level sequence labeling problem** in a **Bi-RNN-CRF** model.
Example:

```
La sede del condado es Ottawa. En
BEXBIIEXKZJXBIIIIIEXBEXBIIIIETXBE
```

Most multiword tokens are transcribed using dictionary lookups. We use an attention-based LSTM encoder-decoder for the sparse cases.

	word	multiword token
beginning	B	K
inside	I	Z
end	E	J
single-character	S	D
character that does not belong to a token		X
single-character end of sentence boundary		T
other end of sentence boundary		U

Tag set

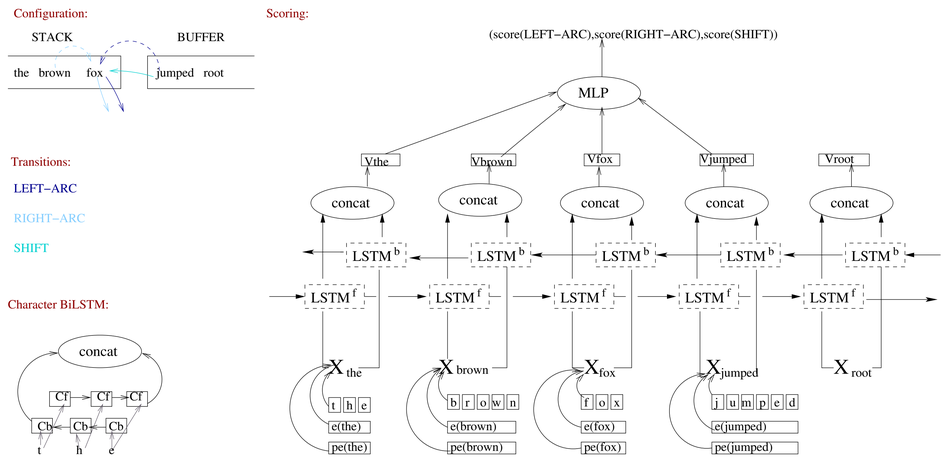
Dependency Parsing

We use transition-based parsing with the **arc-hybrid** system and a **dynamic oracle** for training.

We use **pseudo-projective** parsing to handle non-projective dependencies.

We use feature representations learned by **BiLSTMs** with **word and character embeddings** but **no part-of-speech tags.**

Multilingual models with language embedding used for the low-resource setting.



Pre-Trained Word Embeddings

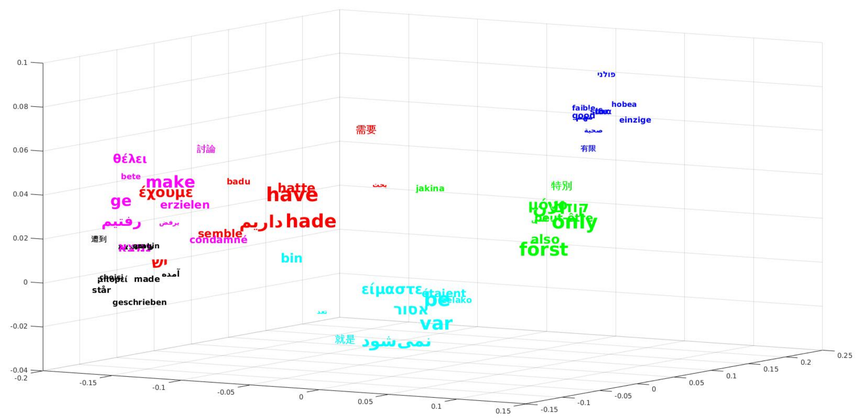
Three step word embedding method:

1. Build a **co-occurrence matrix** over **universal features** of words collected in the UD treebanks.

   Universal features: $\{t_w, t_h, (t_w, d, t_h)\}$
   t_w: POS tag of word
   t_h: POS tag of word's head
   d: dependency relation between word and head

2. **Normalize** the data distribution along each universal feature.

3. Use the **dominant singular vectors** of the normalized data as word vectors.



Results

Language	LAS F1 Test	LAS F1 Corr	LAS F1 Diff	Sentences Test	Sentences Corr	Sentences Diff	Words Test	Words Corr	Words Diff
ar	65.96	68.68	3.38	77.32	78.21	-6.36	94.81	94.99	1.30
ar_pud	47.34	50.70	7.56	97.18	98.66	-1.34	94.32	95.30	4.48
bg	81.25	85.38	1.74	93.36	95.23	2.40	99.70	99.91	0.00
bxr	17.14	18.01	-13.49	86.93	87.37	-4.44	97.77	97.71	-1.64
ca	85.42	87.08	1.69	99.43	99.59	0.64	99.78	99.79	-0.18
cs	85.88	86.83	3.96	93.97	92.79	0.76	99.96	99.98	0.08
cs_cac	83.66	85.75	3.29	99.76	99.68	-0.32	99.97	99.99	0.00
cs_cltt	59.84	75.67	4.03	92.99	96.95	1.89	99.54	99.78	0.43
cs_pud	80.21	82.27	2.47	94.18	95.55	-0.88	98.42	99.25	-0.04
cu	57.88	67.04	4.28	39.71	43.72	7.67	99.73	99.94	-0.02
da	70.63	77.70	4.32	81.12	83.41	4.05	99.93	100.00	0.31
de	72.61	75.27	6.16	80.47	81.47	2.36	99.44	99.67	0.02
de_pud	68.04	70.90	4.37	87.16	86.83	0.34	96.42	96.43	-1.57
el	72.77	80.46	1.20	90.38	91.09	0.30	99.83	99.80	-0.08
en	75.88	79.62	3.78	76.91	80.26	7.04	98.38	99.05	0.38
en_lines	67.52	75.80	2.86	86.84	87.17	1.33	99.82	99.96	0.02
en_partut	63.55	76.11	2.47	98.20	98.10	0.59	99.55	99.54	0.05
en_pud	75.61	80.49	1.54	95.28	96.15	-0.98	99.45	99.59	-0.07
es	82.17	84.26	2.79	95.37	94.16	0.01	99.81	99.84	0.15
es_ancora	84.60	86.79	3.01	98.06	98.46	1.41	99.89	99.92	-0.03
es_pud	78.16	79.01	1.36	93.41	93.39	-0.03	99.39	99.34	-0.13
et	49.01	58.67	-0.12	92.74	93.23	8.03	99.69	99.90	0.13
eu	69.84	73.82	4.67	99.67	100.00	0.42	99.97	100.00	0.04
fa	76.13	81.89	2.65	98.75	99.50	1.50	99.32	99.61	-0.03
fi	74.59	78.41	4.66	90.88	91.48	6.92	99.62	99.71	0.08
fi_ftb	71.85	76.25	2.22	86.98	87.16	3.33	99.91	99.99	0.11
fi_pud	76.22	80.05	1.40	92.02	91.64	-2.03	99.39	99.59	-0.02
fr	80.36	83.66	2.91	93.85	94.32	0.73	99.50	99.53	0.66
fr_partut	69.17	80.84	3.46	99.13	99.50	1.50	99.01	99.50	0.55
fr_pud	73.51	75.25	1.62	93.52	91.33	-0.99	97.38	97.34	-0.83
fr_sequoia	74.96	82.85	2.87	81.89	84.95	1.20	99.31	99.48	0.42
ga	52.81	63.35	1.83	95.70	95.35	-0.46	99.62	99.78	0.49
gl	74.09	79.01	1.70	96.36	96.83	0.68	99.91	99.96	0.04
gl_treegal	56.79	65.85	0.03	82.71	83.79	2.16	98.42	98.23	-0.39
got	56.69	62.62	2.81	29.65	35.01	7.16	100.00	100.00	0.00
grc	50.94	58.83	2.79	98.70	98.93	0.50	96.78	99.98	0.03
grc_proiel	63.86	69.04	3.82	49.31	48.86	5.75	99.99	99.98	-0.02
he	63.72	67.75	10.52	99.29	99.69	0.30	91.18	91.19	6.37
hi	74.34	89.13	2.36	99.29	99.11	-0.09	92.74	99.99	-0.01
hi_pud	45.15	53.31	2.46	94.85	95.00	4.17	92.27	98.65	0.84
hr	75.43	79.51	2.33	97.75	97.25	0.33	99.90	99.91	-0.02
hsb	45.63	47.92	-5.91	91.65	89.88	-0.81	99.28	98.76	-1.08
hu	54.55	65.90	1.60	96.56	97.65	3.80	99.85	99.89	0.07
id	72.11	76.13	1.52	92.66	93.55	2.40	100.00	100.00	0.01
it	84.84	87.33	2.05	99.07	99.38	2.28	99.85	99.86	0.13
it_pud	83.28	85.59	1.89	93.39	93.90	-2.68	99.27	99.28	0.11
ja	65.71	81.54	9.33	94.92	94.92	0.00	84.26	93.59	3.91
ja_pud	71.80	83.26	6.98	97.31	97.31	2.42	86.34	94.30	3.24
kk	18.24	17.14	-7.37	87.52	86.26	4.88	96.56	96.46	1.55
kmr	19.37	20.39	-11.96	94.49	94.08	-2.94	97.15	97.06	-1.79
ko	69.87	74.72	15.63	92.39	93.01	-0.04	99.63	99.99	0.26
la	38.93	46.26	2.49	98.04	97.41	-0.68	100.00	100.00	0.01
la_ittb	80.04	82.34	5.36	94.34	92.93	-0.31	99.97	99.99	0.00
la_proiel	58.74	63.17	5.63	30.24	34.66	8.86	99.99	100.00	0.00
lv	52.36	59.75	-0.20	93.45	93.65	-4.94	99.20	99.13	0.22
nl	69.83	74.41	5.51	75.15	76.16	-0.98	99.73	99.85	-0.03
nl_lassysmall	77.56	83.58	5.43	85.33	87.00	8.38	99.85	99.97	0.04
no_bokmaal	83.22	86.04	2.77	96.44	96.20	0.44	99.84	99.87	0.12
no_nynorsk	81.12	84.41	2.85	94.56	93.67	2.44	99.93	99.92	0.07
pl	77.39	82.33	3.55	98.91	99.46	0.55	99.90	99.93	0.05
pt	80.97	83.25	1.14	90.33	90.43	0.64	99.37	99.45	-0.07
pt_br	86.15	88.19	2.83	96.51	97.04	0.20	99.80	99.87	0.03
pt_pud	72.43	74.48	0.52	93.58	94.50	-1.15	98.39	98.48	-0.94
ro	79.40	81.68	1.80	96.57	96.02	2.60	99.77	99.75	0.11
ru	71.65	77.99	3.96	97.16	96.91	0.49	99.83	99.90	-0.01
ru_pud	65.22	70.78	2.47	98.66	98.80	-0.15	97.31	97.34	0.16
ru_syntagrus	88.04	89.61	2.85	98.64	98.78	0.97	99.51	99.63	0.06
sk	69.35	75.98	3.23	85.32	87.17	3.64	99.97	99.96	-0.04
sl	80.14	84.16	3.01	98.67	98.11	-1.13	99.96	99.97	0.01
sl_sst	36.97	46.76	0.31	19.03	19.52	2.80	97.75	100.00	0.18
sme	11.70	11.72	-18.88	98.27	97.59	-1.20	98.44	96.75	-3.13
sv	73.45	79.86	3.13	97.26	95.96	-0.41	99.86	99.77	-0.07
sv_lines	69.42	76.37	2.08	87.89	88.12	1.68	99.96	99.99	0.01
sv_pud	62.40	69.52	-1.10	84.63	81.14	-9.06	98.56	98.47	0.21
tr	48.29	52.84	-0.35	96.29	96.44	-0.19	96.57	97.51	-0.38
tr_pud	29.79	32.84	-1.69	92.08	90.75	-3.16	96.82	96.93	0.31
ug	28.35	30.98	-3.20	68.76	69.36	5.81	97.82	98.74	0.22
uk	47.00	59.33	-1.43	90.04	92.18	-0.41	99.41	99.52	-0.29
ur	64.96	79.31	2.62	98.60	98.60	0.28	94.55	100.00	0.00
vi	37.99	42.68	5.21	87.30	89.49	-3.10	86.63	86.70	4.23
zh	60.47	65.25	7.85	98.20	98.80	0.61	93.81	93.43	4.52
Average	65.11	70.49	2.14	89.03	89.48	0.99	98.20	98.79	0.30