Character-level approach in sentence segmentation

Mikhail Florinsky

Considered tasks

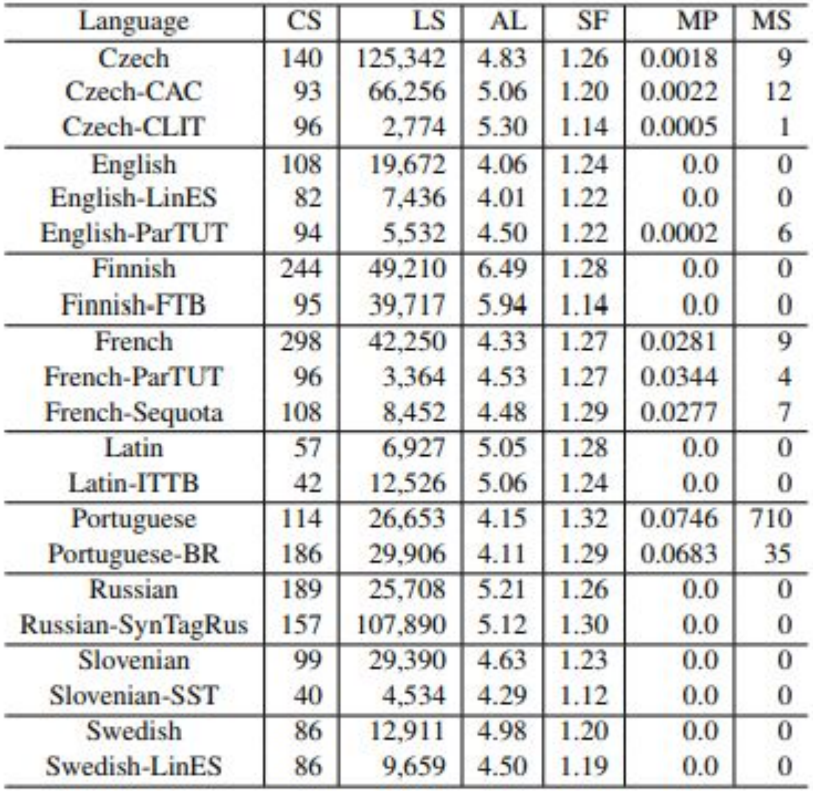
Every task is seq2seq labeling task:

- Word segmentation https://arxiv.org/abs/1807.02974
- Cross-lingual morphological tagging https://arxiv.org/abs/1708.09157
- Text segmentation using sentence parsing

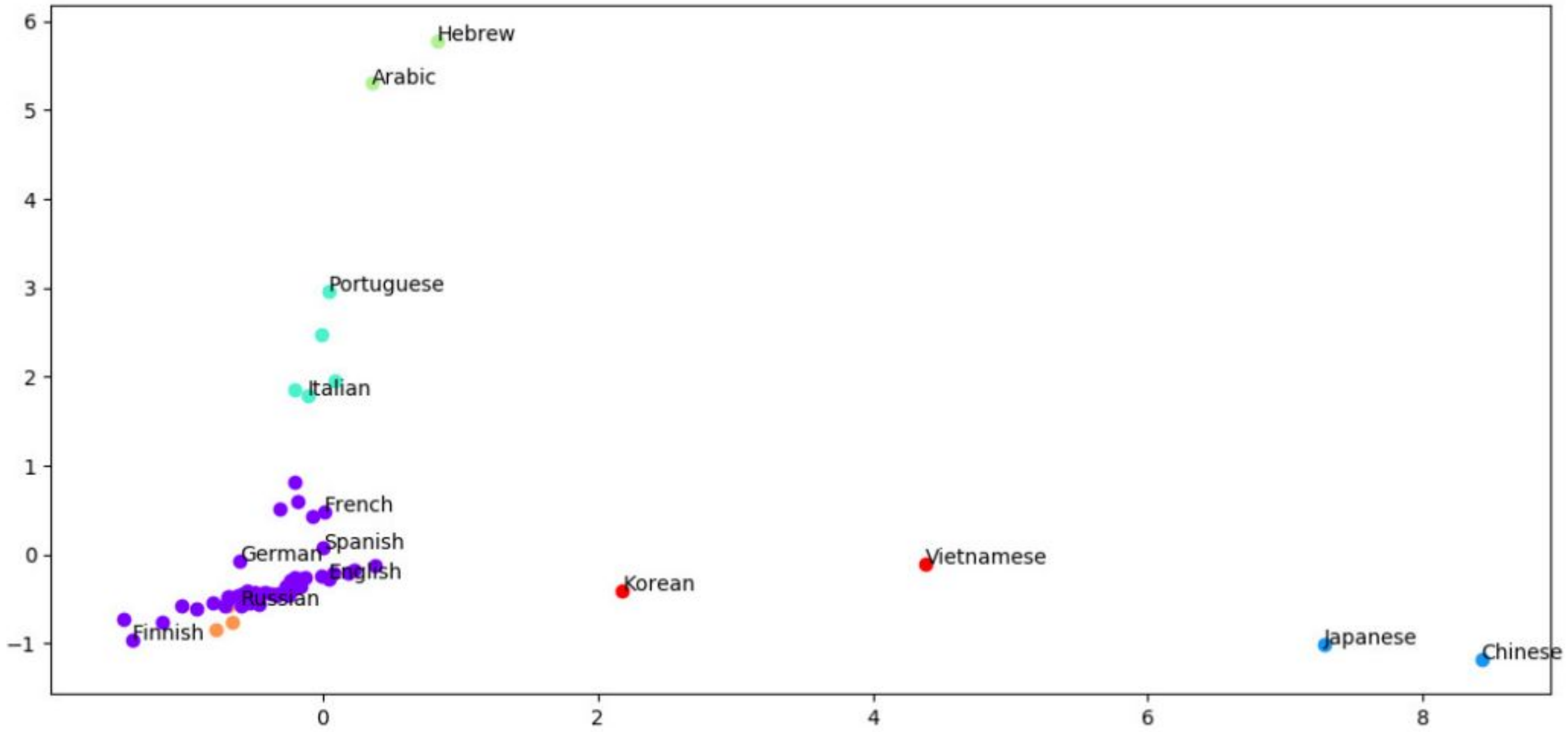
Dataset: UD treebank

Language typological factors

- Character Set Size
- Lexicon Size
- Average Word Length
- Segmentation Frequency
- Multiword Token Portion
- Multiword Token Set Size



Language	CS	LS	AL	SF	MP	MS
Czech	140	125,342	4.83	1.26	0.0018	9
Czech-CAC	93	66,256	5.06	1.20	0.0022	12
Czech-CLIT	96	2,774	5.30	1.14	0.0005	1
English	108	19,672	4.06	1.24	0.0	0
English-LinES	82	7,436	4.01	1.22	0.0	0
English-ParTUT	94	5,532	4.50	1.22	0.0002	6
Finnish	244	49,210	6.49	1.28	0.0	0
Finnish-FTB	95	39,717	5.94	1.14	0.0	0
French	298	42,250	4.33	1.27	0.0281	9
French-ParTUT	96	3,364	4.53	1.27	0.0344	4
French-Sequota	108	8,452	4.48	1.29	0.0277	7
Latin	57	6,927	5.05	1.28	0.0	0
Latin-ITTB	42	12,526	5.06	1.24	0.0	0
Portuguese	114	26,653	4.15	1.32	0.0746	710
Portuguese-BR	186	29,906	4.11	1.29	0.0683	35
Russian	189	25,708	5.21	1.26	0.0	0
Russian-SynTagRus	157	107,890	5.12	1.30	0.0	0
Slovenian	99	29,390	4.63	1.23	0.0	0
Slovenian-SST	40	4,534	4.29	1.12	0.0	0
Swedish	86	12,911	4.98	1.20	0.0	0
Swedish-LinES	86	9,659	4.50	1.19	0.0	0

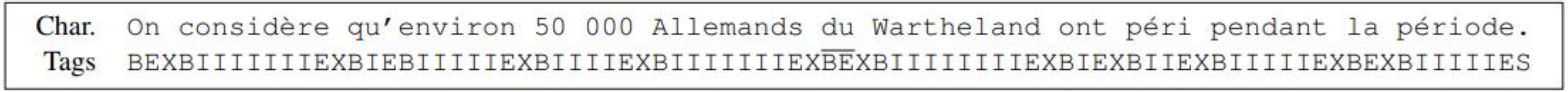


K-Means clustering (K = 6) of the UD languages

Char-based vs Word-based approach

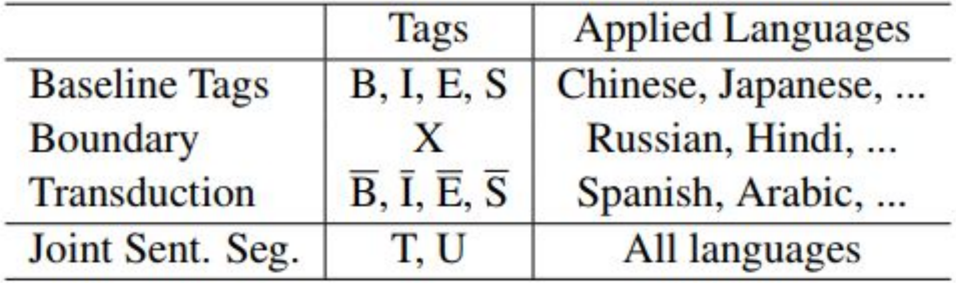
- Dictionary size
- Morphemes similarity
- Cross-lingual

Word segmentation



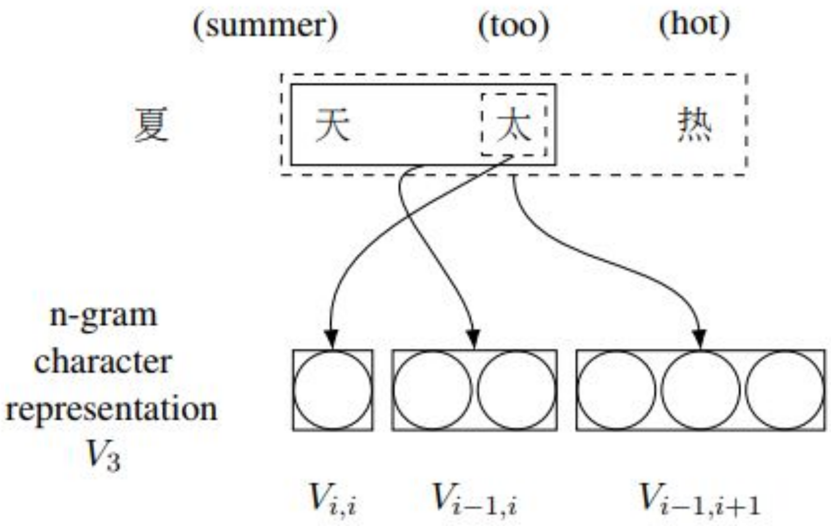
Char.	`On considère qu'environ 50 000 Allemands du Wartheland ont péri pendant la période.`
Tags	BEXBIIIIIIIEXBIEBIIIIIEXBIIIIEXBIIIIIIIEX$\overline{\mathrm{BE}}$XBIIIIIIIIEXBIEXBIIEXBIIIIIEXBEXBIIIIIES

- B - begin of token
- I - inner symbol
- E - end of token
- S - single symbol
- X - bound (Space)
- T - single end of sentence
- U - end of sentence
- Upperscored - for multiword tokens

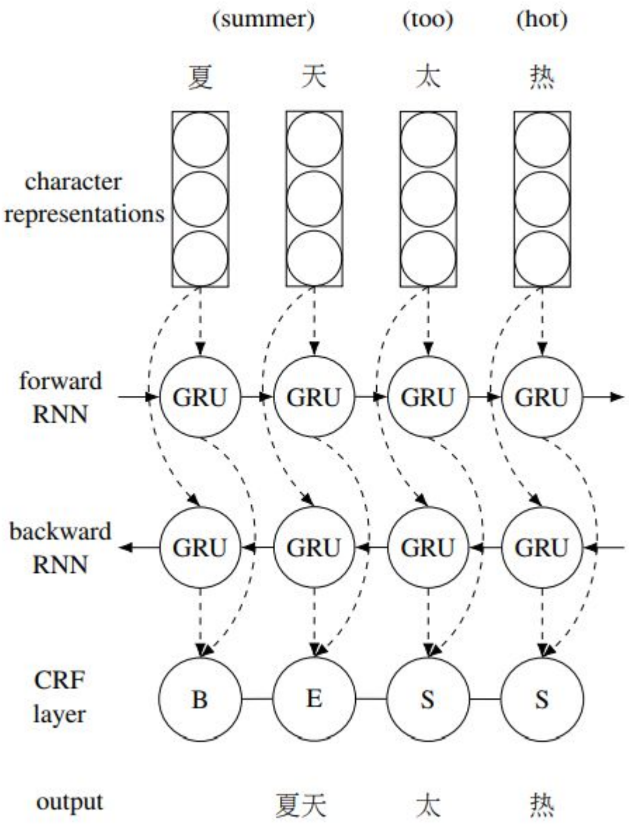


	Tags	Applied Languages
Baseline Tags	B, I, E, S	Chinese, Japanese, ...
Boundary	X	Russian, Hindi, ...
Transduction	$\overline{\mathrm{B}}, \overline{\mathrm{I}}, \overline{\mathrm{E}}, \overline{\mathrm{S}}$	Spanish, Arabic, ...
Joint Sent. Seg.	T, U	All languages

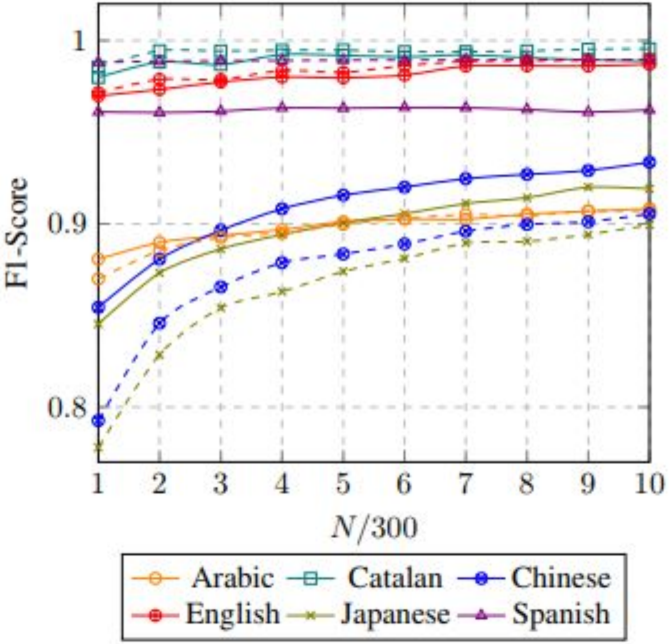
Architecture: embedding



Architecture: BiLSTM + CRF

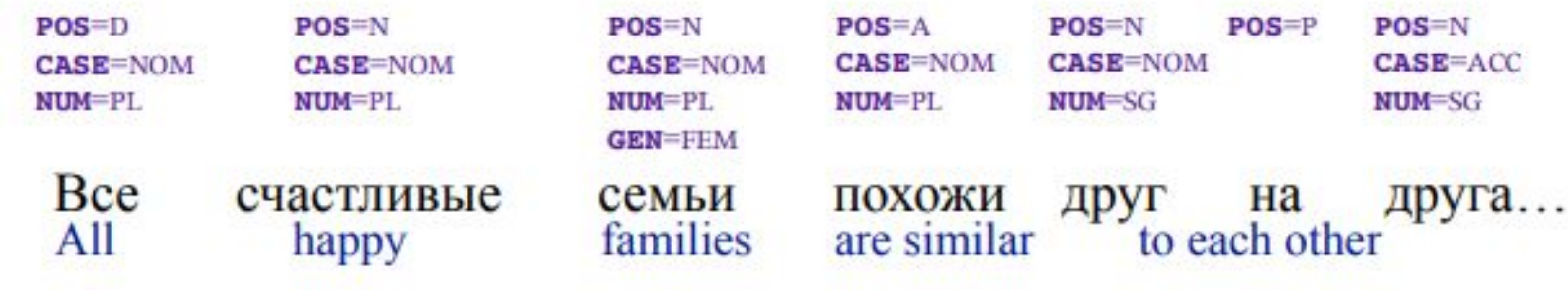


Results



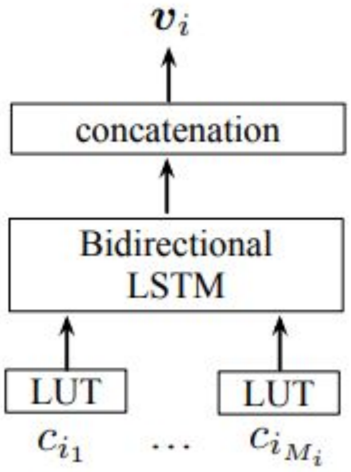
F1-Score
1
0.9
0.8
1
2
3
4
5
6
7
8
9
10
N/300
Arabic
Catalan
Chinese
English
Japanese
Spanish

Cross-Lingual Morphological Transfer

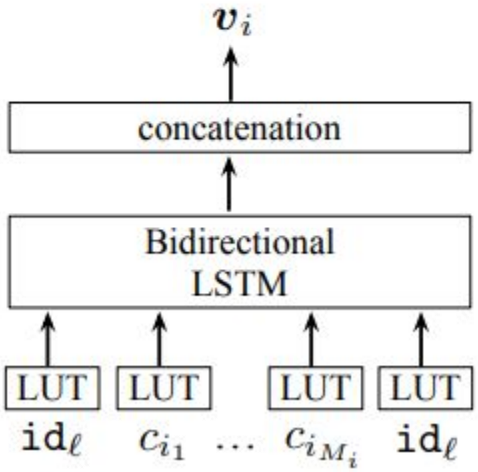


POS=D CASE=NOM NUM=PL	POS=N CASE=NOM NUM=PL	POS=N CASE=NOM NUM=PL GEN=FEM	POS=A CASE=NOM NUM=PL	POS=N CASE=NOM NUM=SG	POS=P	POS=N CASE=ACC NUM=SG
Все	счастливые	семьи	похожи	друг	на	друга…
All	happy	families	are similar	to each other		

Architecture: embedding

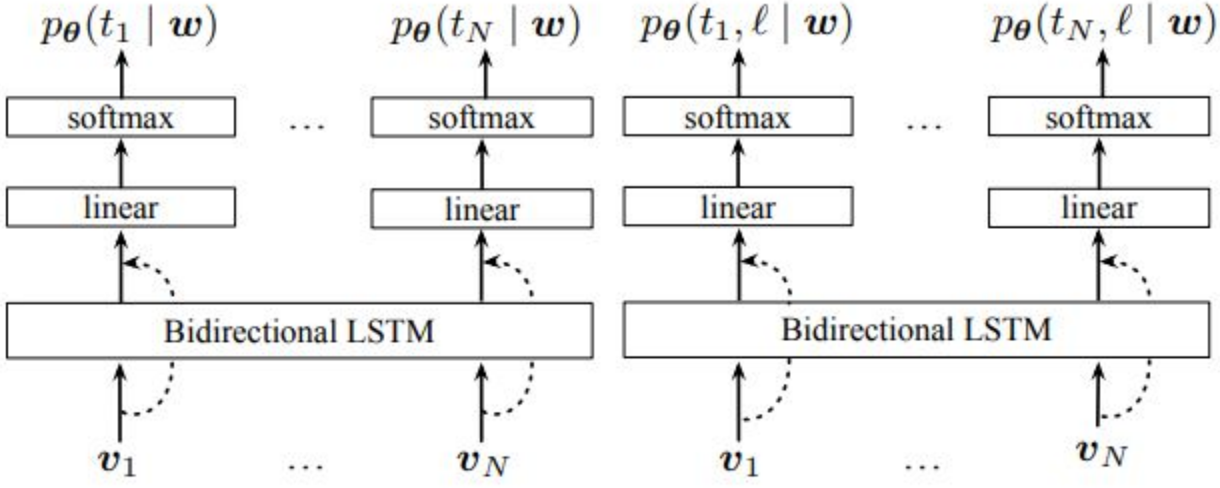


default



language-specific

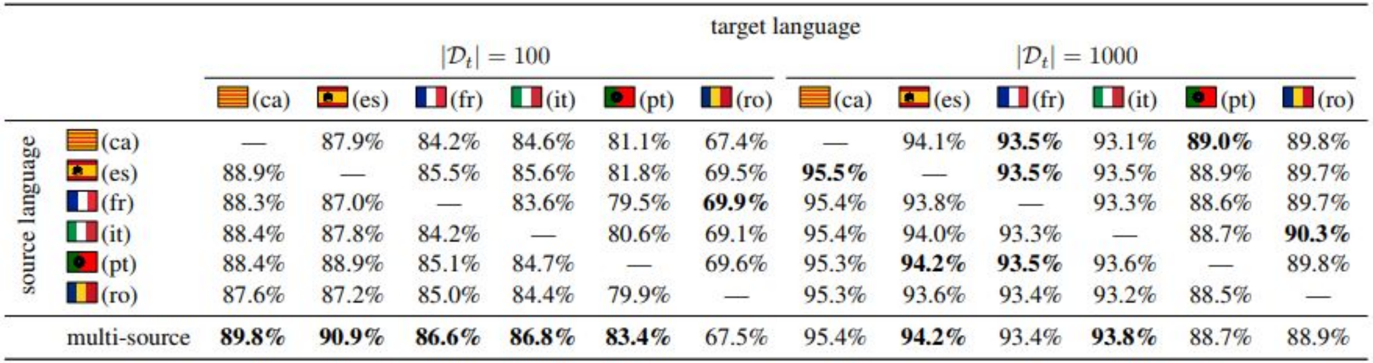
Architecture: BiLSTM + softmax


$p_{\boldsymbol{\theta}}(t_1 \mid \boldsymbol{w})$
$p_{\boldsymbol{\theta}}(t_N \mid \boldsymbol{w})$
$p_{\boldsymbol{\theta}}(t_1, \ell \mid \boldsymbol{w})$
$p_{\boldsymbol{\theta}}(t_N, \ell \mid \boldsymbol{w})$
softmax
...
softmax
softmax
...
softmax
linear
linear
linear
linear
Bidirectional LSTM
Bidirectional LSTM
$\boldsymbol{v}_1$
...
$\boldsymbol{v}_N$
$\boldsymbol{v}_1$
...
$\boldsymbol{v}_N$

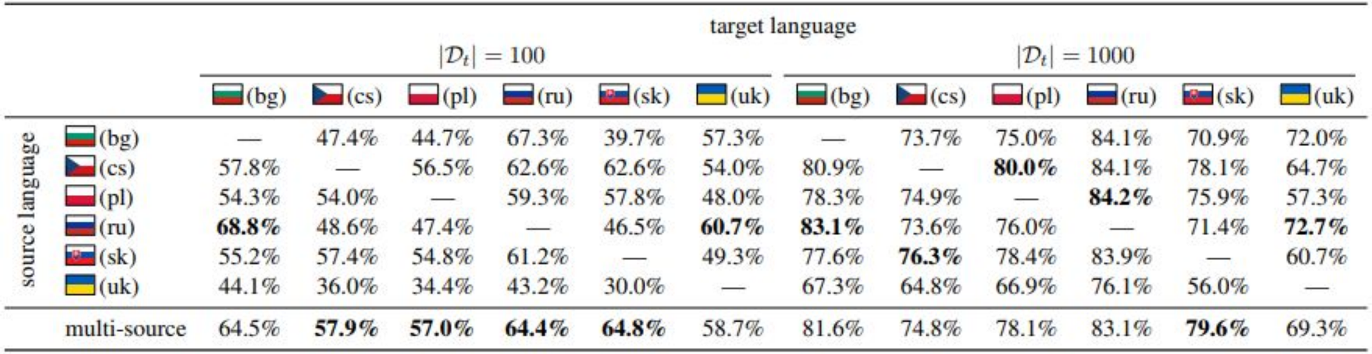
default

language-specific

Results



		target language											
		$\lvert\mathcal{D}_t\rvert = 100$						$\lvert\mathcal{D}_t\rvert = 1000$					
		(ca)	(es)	(fr)	(it)	(pt)	(ro)	(ca)	(es)	(fr)	(it)	(pt)	(ro)
source language	(ca)	—	87.9%	84.2%	84.6%	81.1%	67.4%	—	94.1%	**93.5%**	93.1%	**89.0%**	89.8%
	(es)	88.9%	—	85.5%	85.6%	81.8%	69.5%	**95.5%**	—	**93.5%**	93.5%	88.9%	89.7%
	(fr)	88.3%	87.0%	—	83.6%	79.5%	**69.9%**	95.4%	93.8%	—	93.3%	88.6%	89.7%
	(it)	88.4%	87.8%	84.2%	—	80.6%	69.1%	95.4%	94.0%	93.3%	—	88.7%	**90.3%**
	(pt)	88.4%	88.9%	85.1%	84.7%	—	69.6%	95.3%	**94.2%**	**93.5%**	93.6%	—	89.8%
	(ro)	87.6%	87.2%	85.0%	84.4%	79.9%	—	95.3%	93.6%	93.4%	93.2%	88.5%	—
	multi-source	**89.8%**	**90.9%**	**86.6%**	**86.8%**	**83.4%**	67.5%	95.4%	**94.2%**	93.4%	**93.8%**	88.7%	88.9%



		target language											
		$\lvert\mathcal{D}_t\rvert = 100$						$\lvert\mathcal{D}_t\rvert = 1000$					
		(bg)	(cs)	(pl)	(ru)	(sk)	(uk)	(bg)	(cs)	(pl)	(ru)	(sk)	(uk)
source language	(bg)	—	47.4%	44.7%	67.3%	39.7%	57.3%	—	73.7%	75.0%	84.1%	70.9%	72.0%
	(cs)	57.8%	—	56.5%	62.6%	62.6%	54.0%	80.9%	—	**80.0%**	84.1%	78.1%	64.7%
	(pl)	54.3%	54.0%	—	59.3%	57.8%	48.0%	78.3%	74.9%	—	**84.2%**	75.9%	57.3%
	(ru)	**68.8%**	48.6%	47.4%	—	46.5%	**60.7%**	**83.1%**	73.6%	76.0%	—	71.4%	**72.7%**
	(sk)	55.2%	57.4%	54.8%	61.2%	—	49.3%	77.6%	**76.3%**	78.4%	83.9%	—	60.7%
	(uk)	44.1%	36.0%	34.4%	43.2%	30.0%	—	67.3%	64.8%	66.9%	76.1%	56.0%	—
	multi-source	64.5%	**57.9%**	**57.0%**	**64.4%**	**64.8%**	58.7%	81.6%	74.8%	78.1%	83.1%	**79.6%**	69.3%

Text segmentation and boundaries detection

how are you i am ok thanks

[How are you?] [I am OK, thanks.]
[How are you?] [I am OK.] [Thanks.]

Sentence parsing

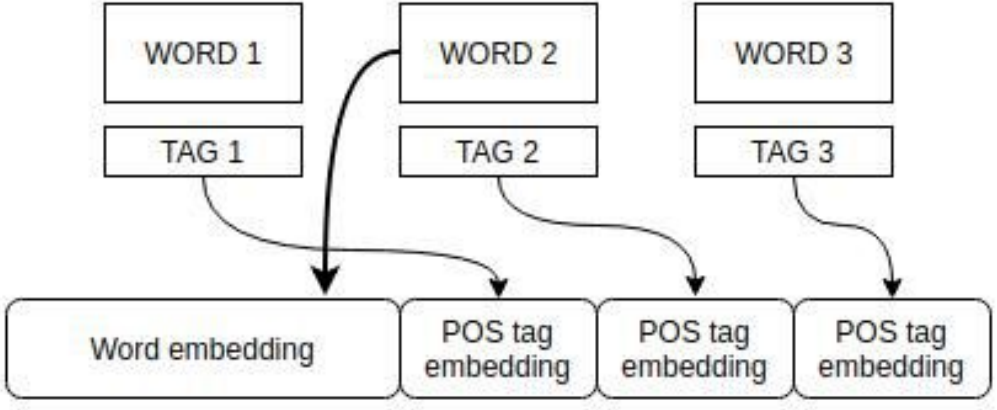
Let’s try to predict distance between word and its parent:

$$\begin{cases} tag(word) = word.parent\_id - word.id \\ root.parent\_id = root.id \end{cases}$$

Data preparing

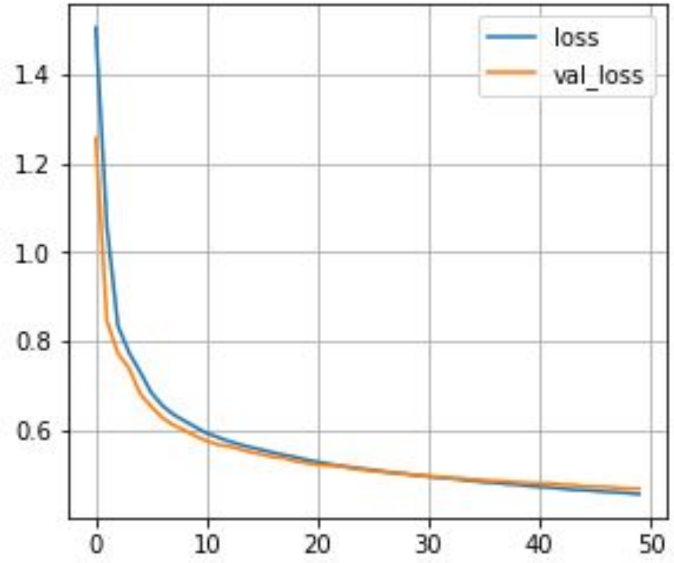
- Each sample is two joint sentences
- Concat word and pos
- Lowercase every word and remove punctuation

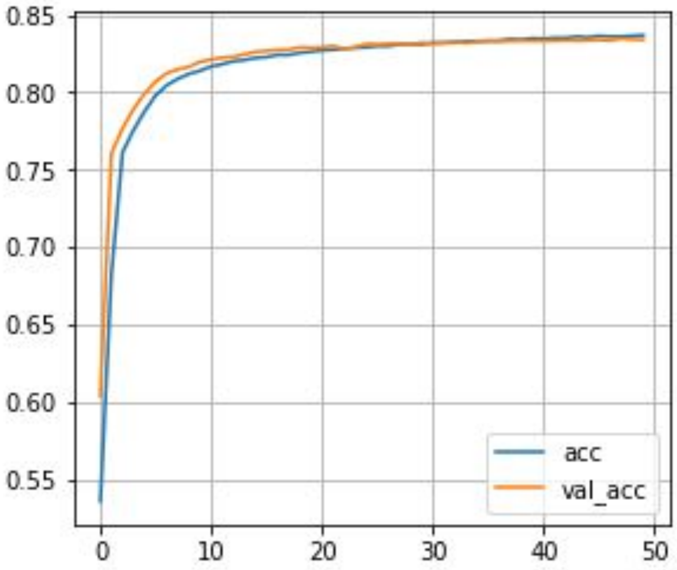
POS-based architecture



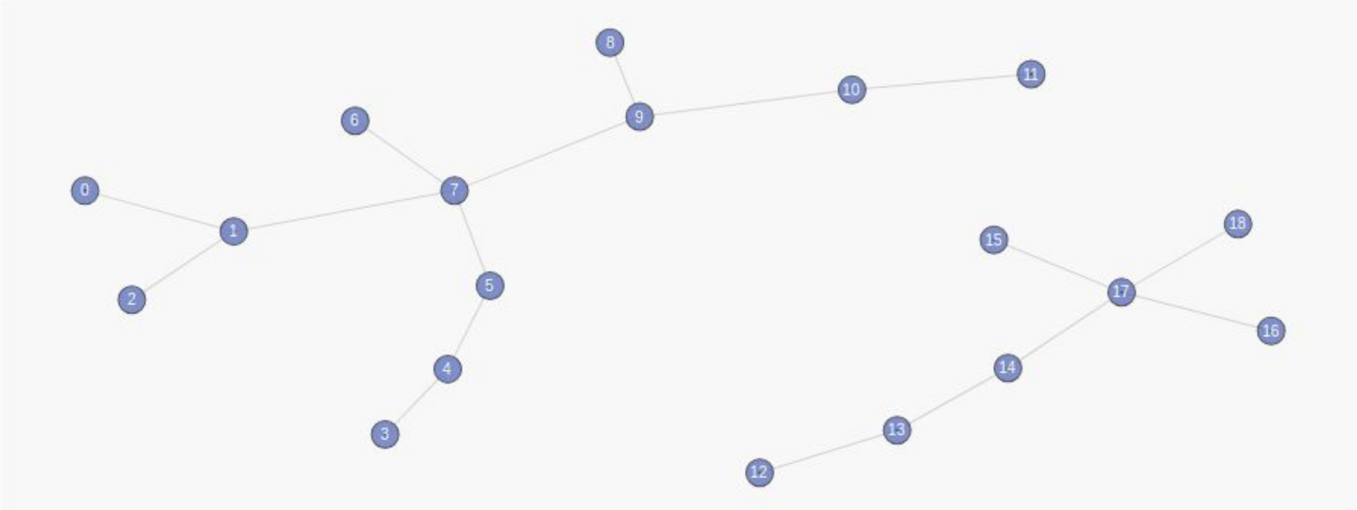
Main model is similar to model in first task

Results: not so good

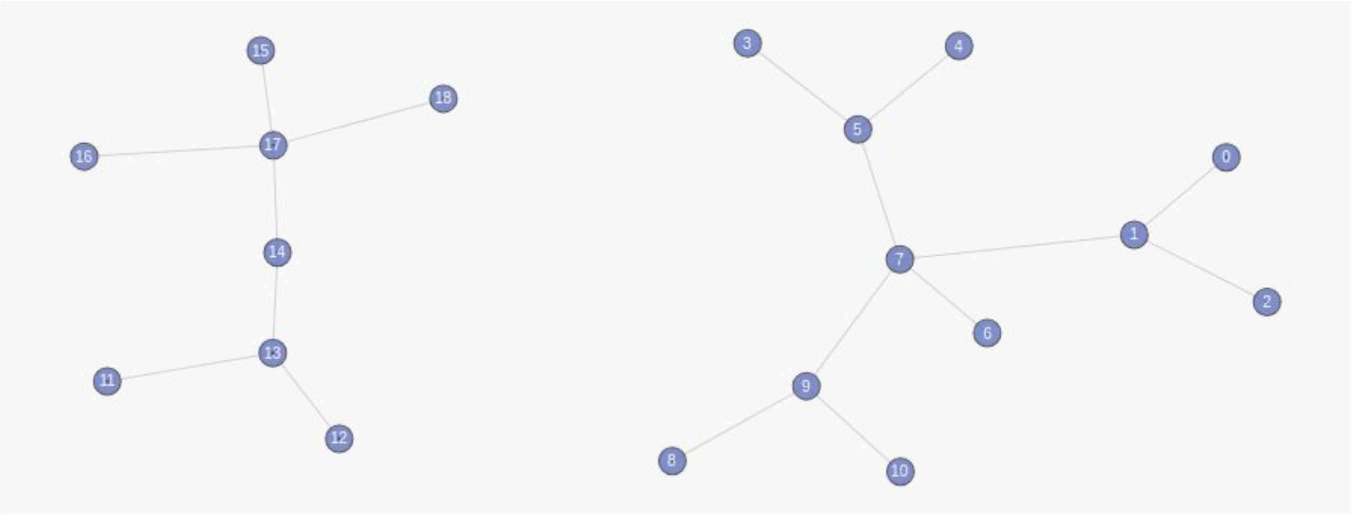




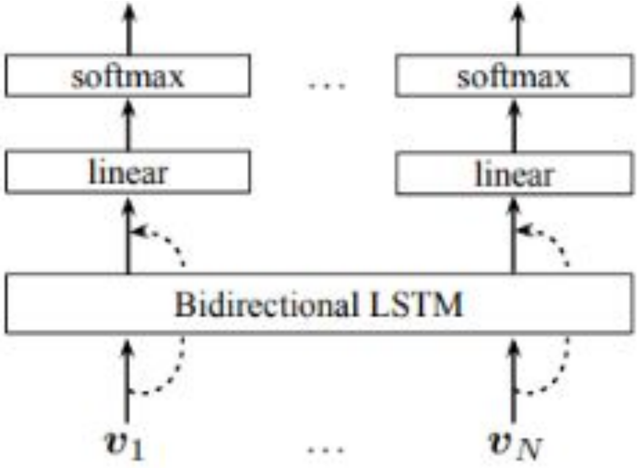
True

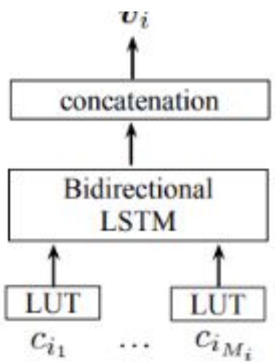


Predicted

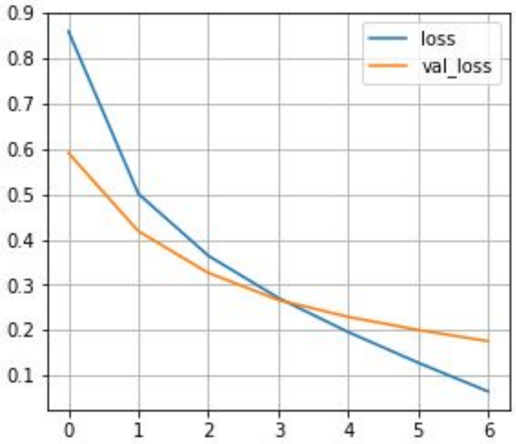


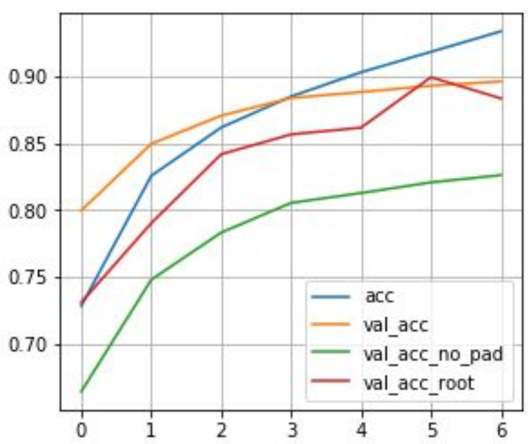
In progress





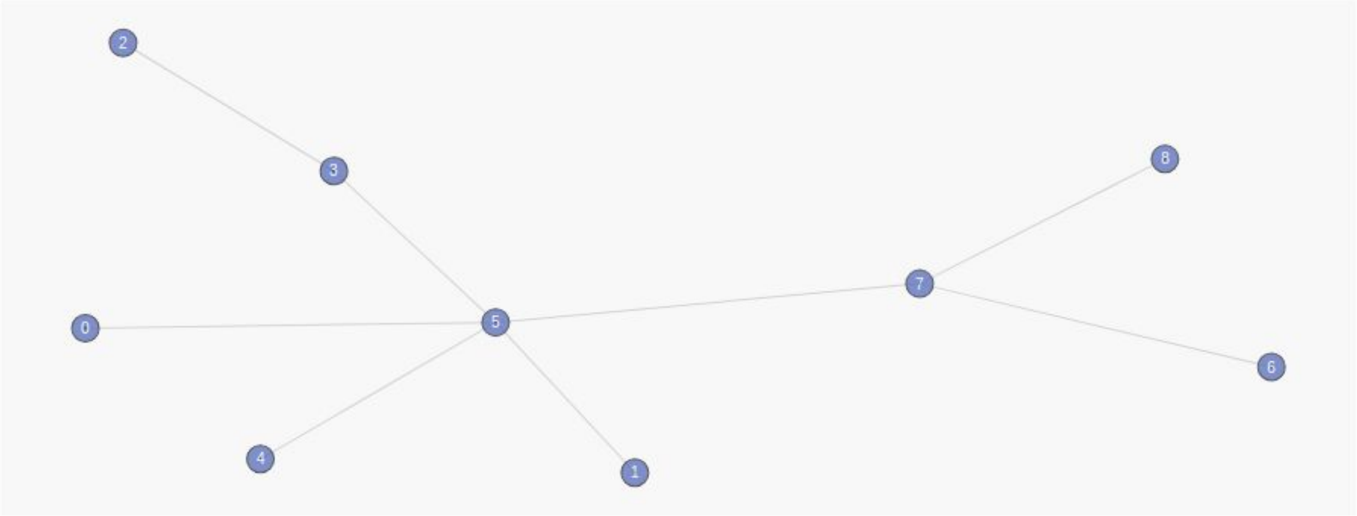
+ POS tag





>90% accuracy expected

True



Predicted