Towards Better UD Parsing:
Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, Ting Liu

Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology

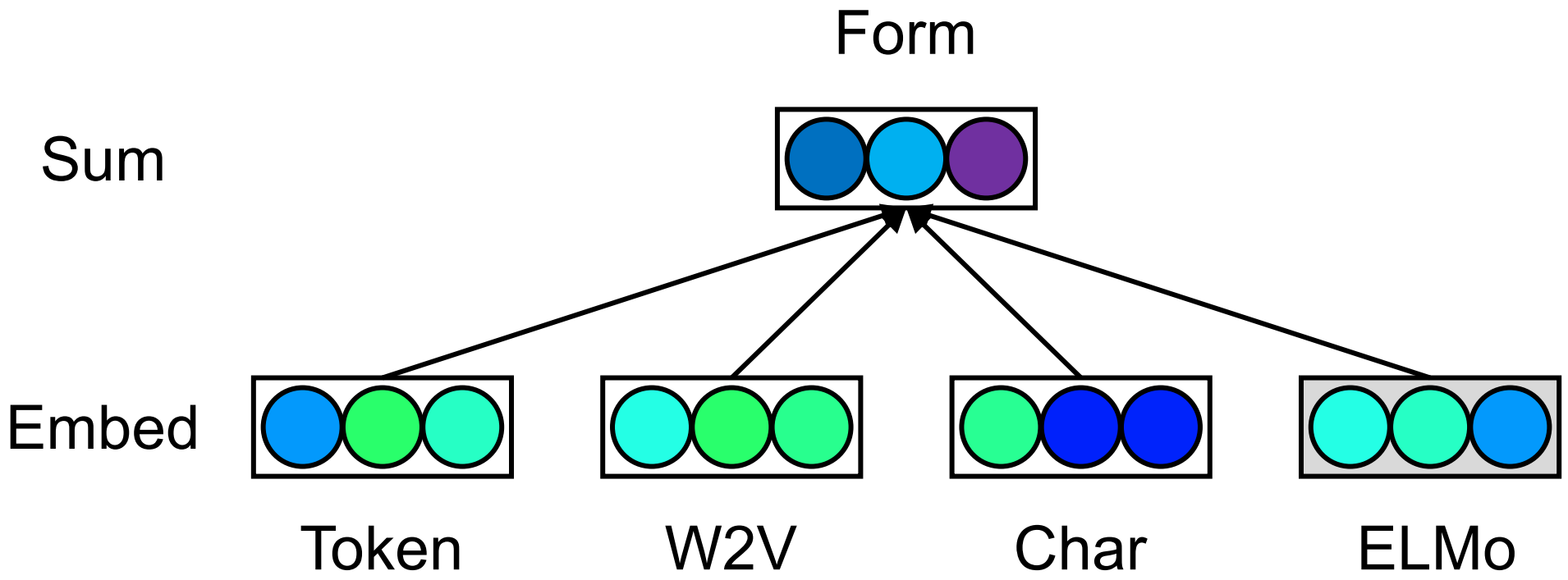
Overview of Our Techniques

- Rank 1st according to LAS
- Baseline model: Dozat et al. (2017)
- Winning strategies:
  - ELMo: +0.84
  - Ensemble: +0.55
  - Treebank Concat.: +0.42 (estimated on Dev set.)

LAS Ranking

1.	HIT-SCIR (Harbin)	75.84 ± 0.14 [OK]	(p<0.001)
2.	TurkuNLP (Turku)	73.28 ± 0.14 [OK]	(p=0.039)
3-5.	UDPipe Future (Praha)	73.11 ± 0.13 [OK]	(p=0.221)
3-5.	LATTICE (Paris)	73.02 ± 0.14 [OK]	(p=0.461)
3-5.	ICS PAS (Warszawa)	73.02 ± 0.14 [OK]	(p<0.001)
6.	CEA LIST (Paris)	72.56 ± 0.14 [OK]	(p=0.036)
7-8.	Uppsala (Uppsala)	72.37 ± 0.15 [OK]	(p=0.191)
7-8.	Stanford (Stanford)	72.29 ± 0.14 [OK]	(p<0.001)
9-10.	AntNLP (Shanghai)	70.90 ± 0.15 [OK]	(p=0.242)
9-10.	NLP-Cube (București)	70.82 ± 0.14 [OK]	(p=0.032)
11.	ParisNLP (Paris)	70.64 ± 0.14 [OK]	(p<0.001)
12.	SLT-Interactions (Bengaluru)	69.98 ± 0.14 [OK]	(p<0.001)
13.	IBM NY (Yorktown Heights)	69.11 ± 0.16 [OK]	(p<0.001)
14.	UniMelb (Melbourne)	68.66 ± 0.15 [OK]	(p=0.002)
15.	LeisureX (Shanghai)	68.31 ± 0.16 [OK]	(p<0.001)
16.	KParse (İstanbul)	66.58 ± 0.16 [OK]	(p=0.015)
17.	Fudan (Shanghai)	66.34 ± 0.15 [OK]	(p<0.001)
18.	BASELINE UDPipe 1.2 (Praha)	65.80 ± 0.15 [OK]	(p=0.048)
19.	Phoenix (Shanghai)	65.61 ± 0.16 [OK]	(p<0.001)
20.	CUNI x-ling (Praha)	64.87 ± 0.16 [OK]	(p<0.001)
21.	BOUN (İstanbul)	63.54 ± 0.15 [OK]	(p<0.001)
22.	ONLP lab (Ra'anana)	58.35 ± 0.15 [81]	(p<0.001)
23.	iParse (Pittsburgh)	55.83 ± 0.11 [65]	(p<0.001)
24.	HUJI (Yerushalayim)	53.69 ± 0.15 [80]	(p<0.001)
25.	ArmParser (Yerevan)	47.02 ± 0.11 [66]	(p<0.001)
26.	SParse (İstanbul)	1.95 ± 0.00 [2]	

Our Extension to Dozat et al. (2017)



$$ELMo_i = \sum_{j=0}^{2} h_{i,j}^{(LM)}$$

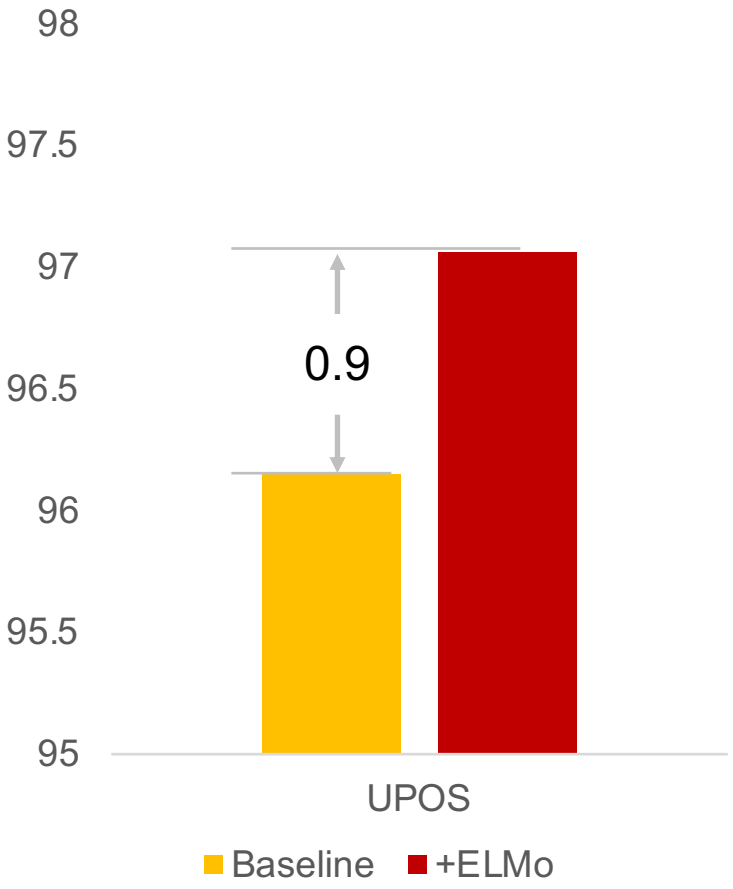
Two Extensions on AllenNLP ELMo

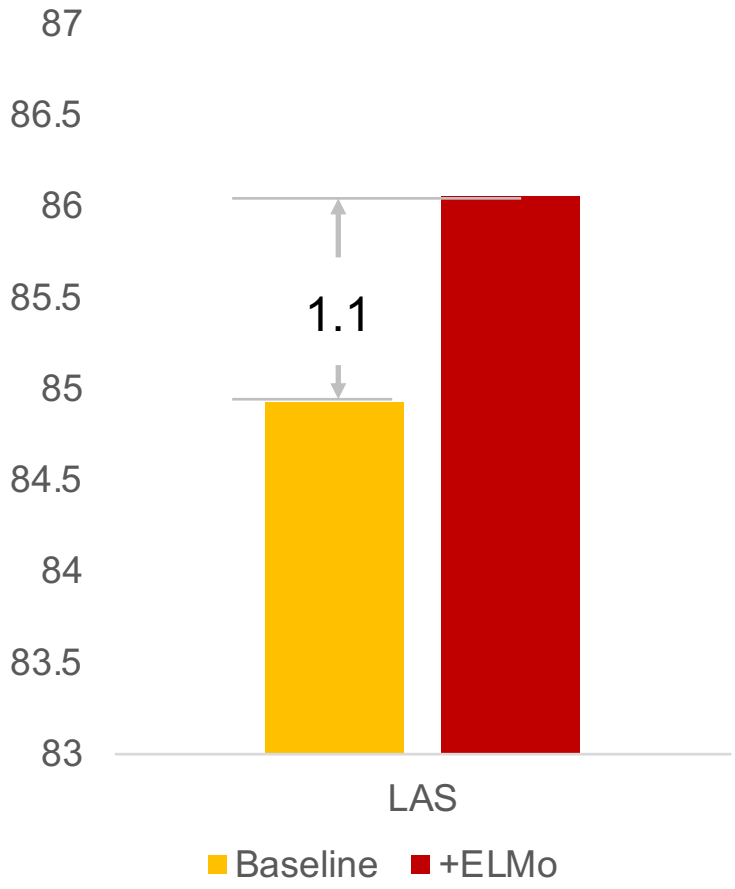
- Supporting Unicode range
- Training with *sample softmax*
  - use a window of 8192 surrounding words as negative samples
  - More stable training and better performance
- One language takes 3 days on Nvida P100

Other Techniques that Contributes

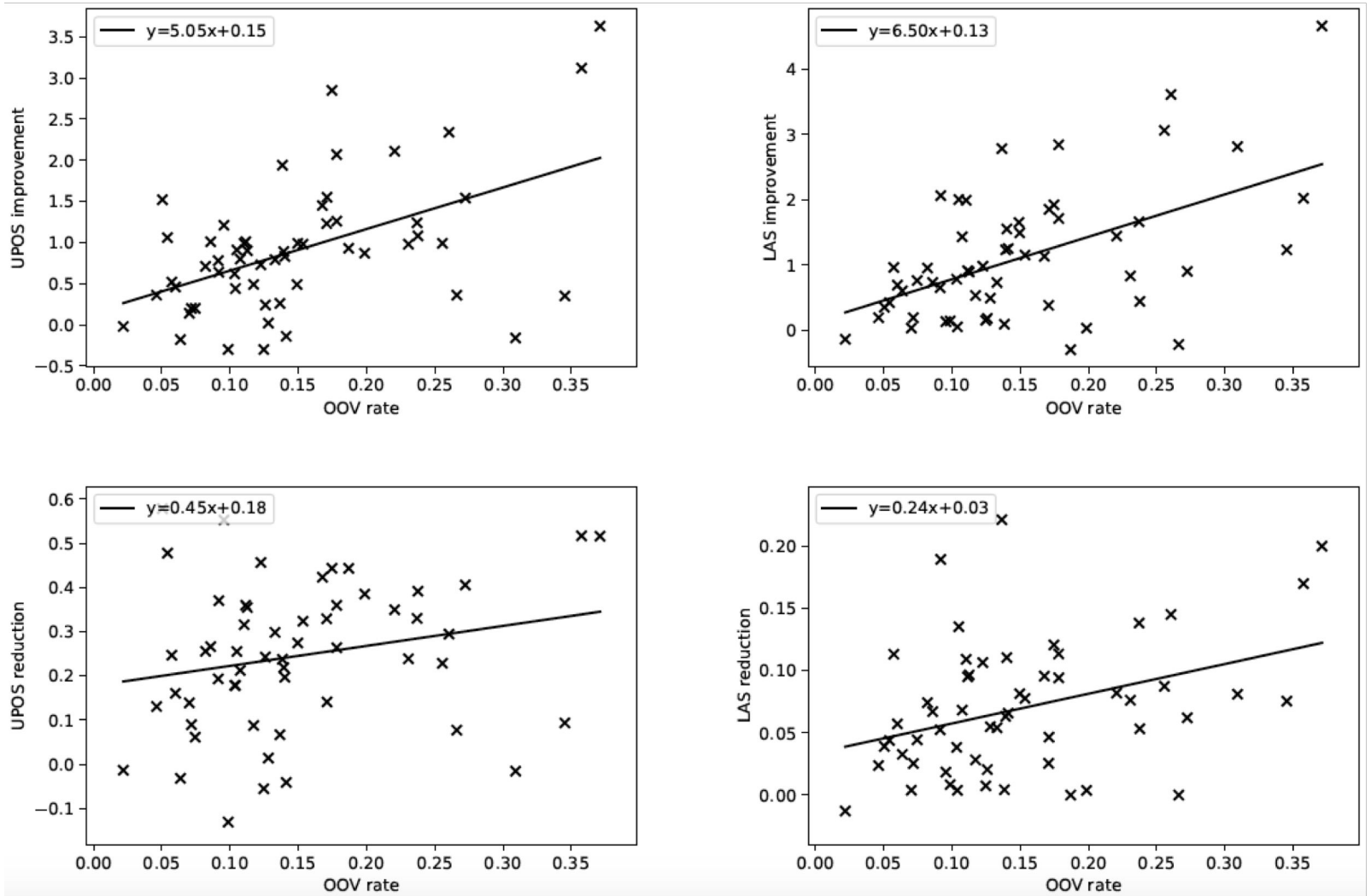
- Improved POS tagging:
  - Ranked 2nd in the UPOS evaluation (1st on the big treebanks)
  - Biaffine tagger + ELMo
- Improved tokenization:
  - Ranked 2nd in the Tokenization-F1 evaluation
  - BiLSTM sequence labeling + unigram character ELMo
  - Wins in zh_gsd (+6.6), ja_gsd (+4.1), vi_vtb (+7.2), ja_morden (+9.7)

Effects of ELMo: Improvements on Gold Segmentation

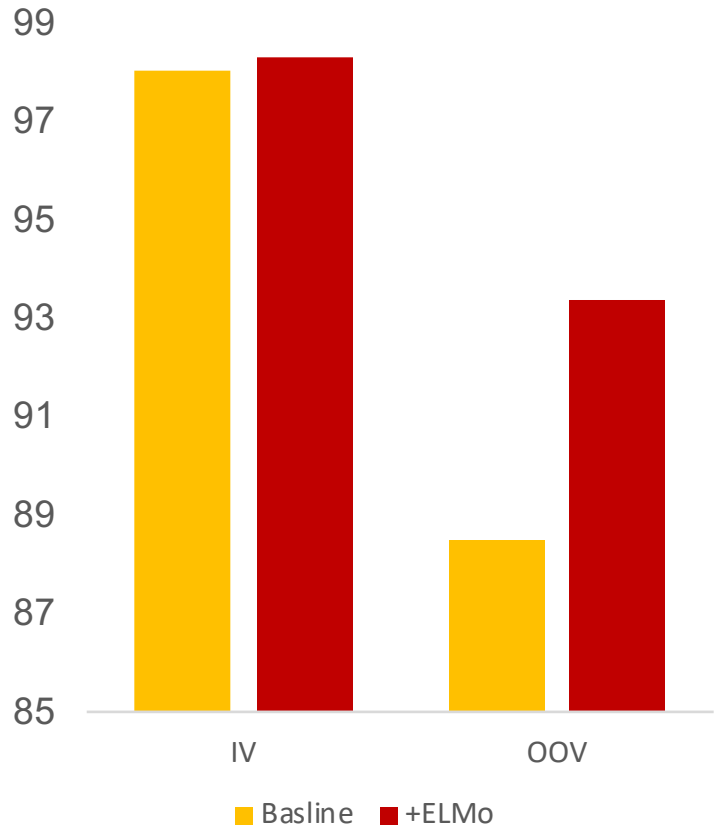


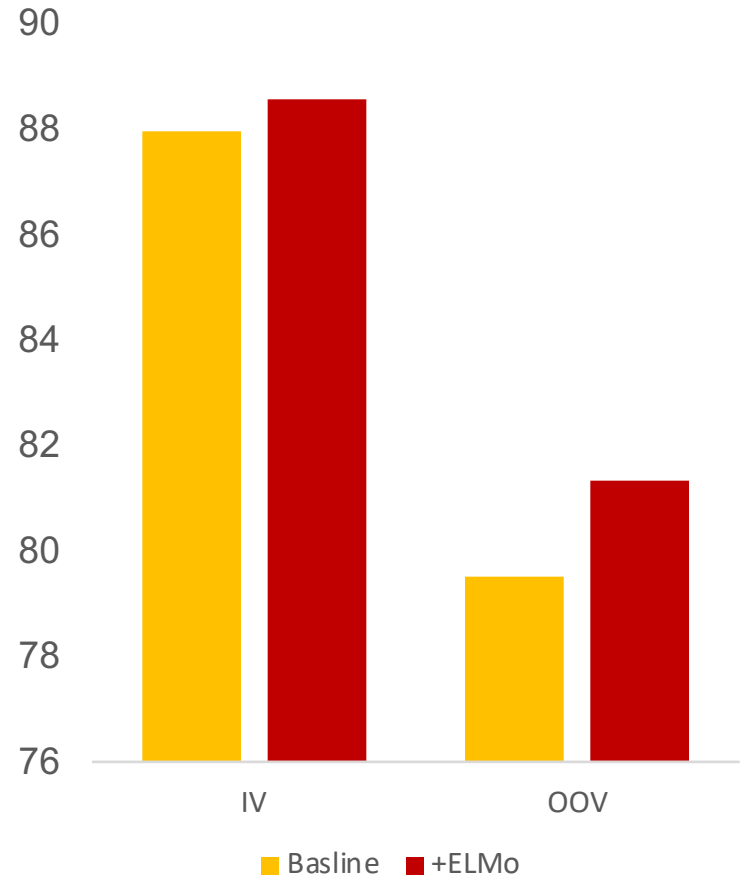


Effects of ELMo: OOV Rate against Improvements

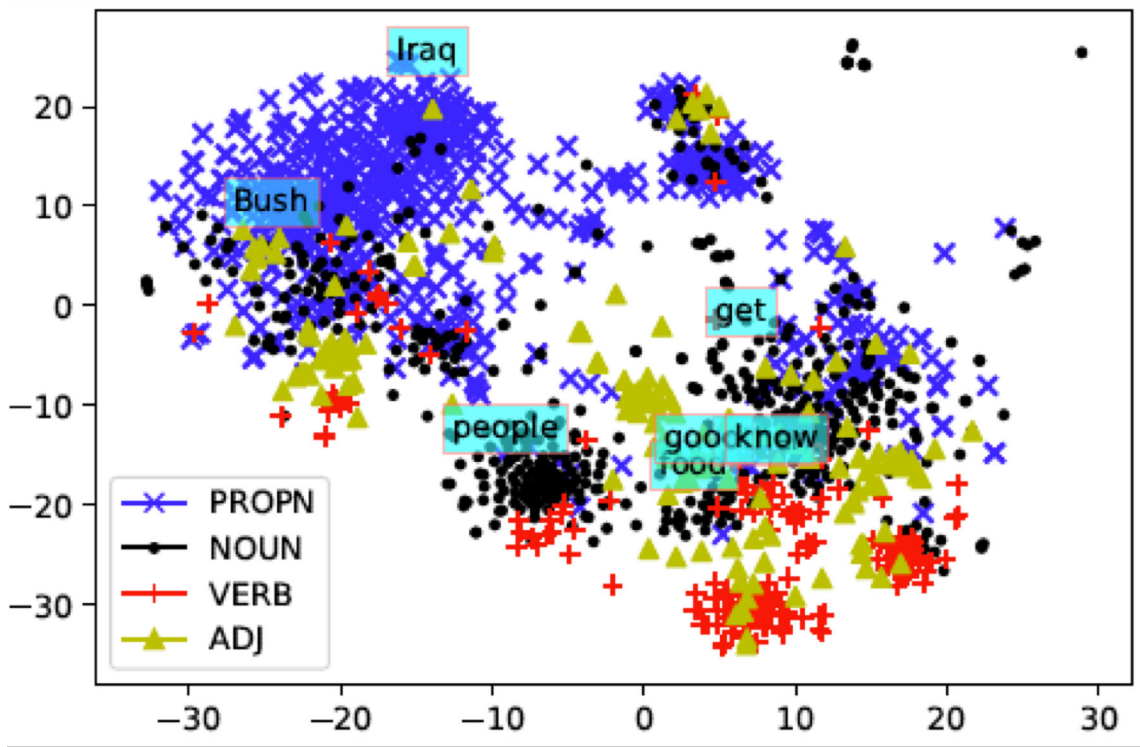


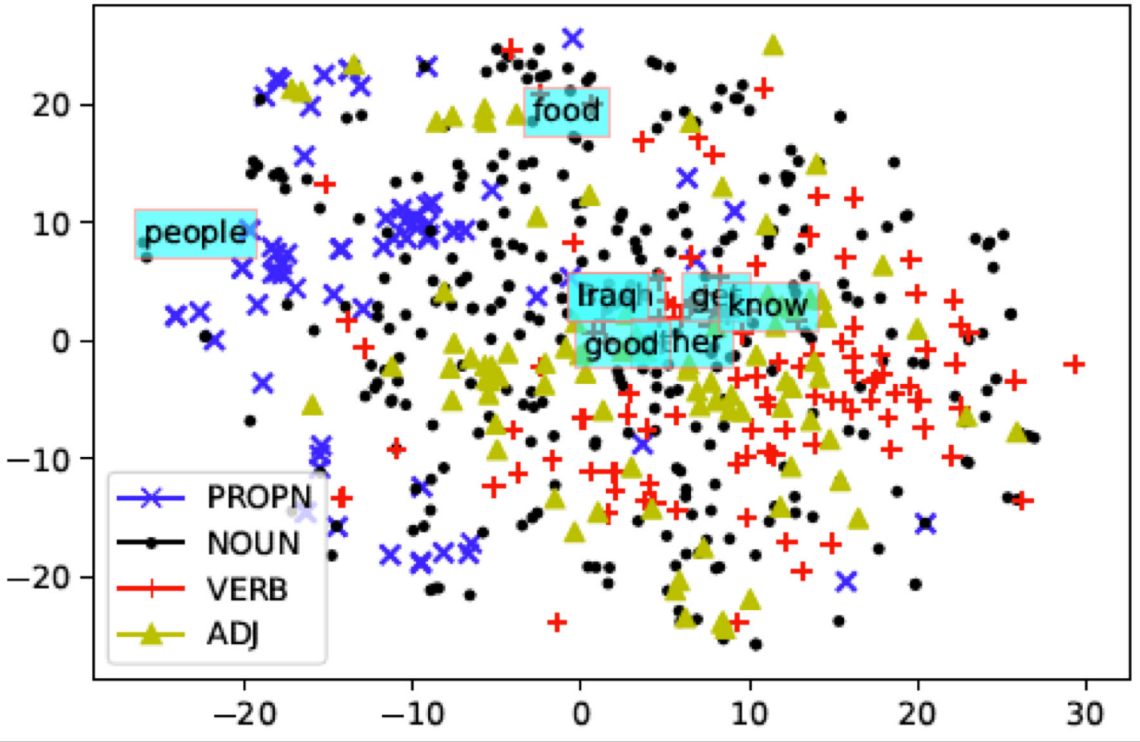
Effects of ELMo:
ELMo's Effects on IV and OOV





Effects of ELMo: Better OOV Abstraction





Conclusion

- We made several improvements including incorporating deep contextualized word embeddings, parser ensemble, and treebank concatenation.
- Analysis shows that ELMo mainly improves the OOV performance via learning better abstraction.
- We release the pre-trained ELMo at: https://github.com/HIT-SCIR/ELMoForManyLangs