Unsupervised Statistical Text Simplification Using Pre-trained Language Modeling for Initialization

Jipeng QIANG, Feng ZHANG, Yun LI, Yunhao YUAN, Yi ZHU, Xindong WU

Frontiers of Computer Science, DOI: 10.1007/s11704-022-1244-0

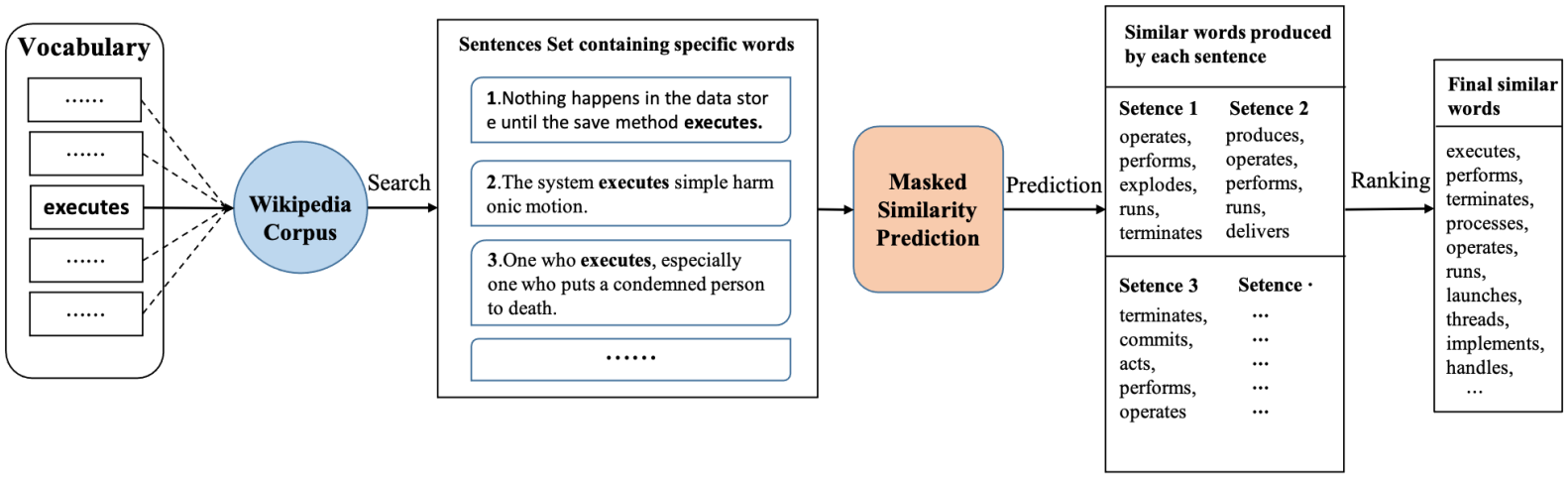
Problems & Ideas

Problems of UnsupPBMT:

- UnsupPBMT aligns a large number of non-similar words to initialize the phrase tables, which will bring noise to the simplification system.
- Many simple high-similar words are hard to find using a word embedding modeling, which is not following the TS task.

Ideas:

- a novel unsupervised statistical text simplification wherein a pre-trained LM is used as a general knowledge base for initialization.



The overview of the process of obtaining similar words using BERT.

Experimental results

Table 1. Performance of baselines and our method on three corpora.

	WikiLarge		WikiSmall		NewSela	
Method	SARI ↑	FKGL ↓	SARI ↑	FKGL ↓	SARI ↑	FKGL ↓
Baselines						
Complex	28.70	8.11	4.34	12.40	2.74	8.65
Reference	49.89	8.26	63.62	8.96	70.25	3.48
Supervised Methods						
PBMT-R (2012)	38.56	8.30	15.97	11.52	15.77	7.95
Hybrid (2014)	31.40	**4.70**	30.46	9.55	30.00	4.15
EncDecA (2017)	35.66	8.67	13.61	11.41	24.12	5.49
Dress (2017)	37.08	6.79	27.48	7.62	27.37	4.19
Dress-LS (2017)	37.27	6.62	27.24	7.55	26.63	4.21
EntPar (2018)	37.45	7.41	28.24	**6.93**	**32.98**	**1.38**
EditNTS (2019)	38.22	7.30	**32.35**	5.47	31.41	3.40
ACCESS (2020)	**41.87**	7.22	-	-	-	-
Unsupervised Methods						
UNMT (2019)	33.72	8.23	-	-	-	-
UNTS (2019)	35.29	7.84	-	-	-	-
UnsupPBMT (2021)	39.08	8.26	25.12	10.66	23.75	7.36
UnsupPBMT-BERT	**40.10**	**7.52**	**29.08**	**7.04**	**27.36**	**5.55**

Conclusions: (1) A good strategy to find synonyms or high-similar words using pre-trained language modeling BERT, compared with synonym dictionary and word embedding modeling; (2) Our method outperforms significantly all unsupervised text simplification methods and has comparable performance to strong supervised methods.