CSE712

Topic/Paper Presentation

by

Sajjad Hussain
21166057
Group-19

Context-aware Stand-alone Neural Spelling Correction

- Li *et al.*, (2020)

Problem Statement

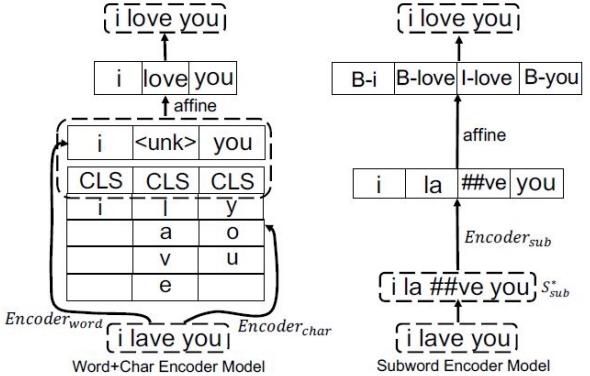
- Spelling Correction
- Man vs. Machine
- Misspelling Types
  - real-word
  - non-word

Objective

- a strong stand-alone spelling corrector
- correct each token
- no addition of tokens
- no deletion of tokens
- preserve maximum origin

Proposition

- Encode spelling & context information
  - Word+Char Encoder
  - Subword Encoder
- Initialized from pretrained language model (LM)
- Dataset
  - Created from *1-Billion-Word-Language-Model-Benchmark*
  - Augment with character-level noise



Performance metrics

	Models	Dev				Test				Real-Word						Non-Word	
										dev			test			dev	test
		Acc	P	R	$F_{0.5}$	Acc	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	P
1	ScRNN (Sakaguchi et al., 2017)	0.958	0.823	0.890	0.836	0.946	0.755	0.865	0.775	0.507	0.592	0.522	0.456	0.523	0.468	0.952	0.873
2	MUDE (Wang et al., 2019)	0.966	0.829	0.952	0.851	0.952	0.751	0.928	0.781	0.595	0.825	0.630	0.533	0.747	0.566	0.945	0.855
3	Char Encoder	0.883	0.517	0.819	0.559	0.870	0.458	0.802	0.501	0.106	0.304	0.122	0.099	0.296	0.113	0.886	0.792
4	Word Encoder	0.932	0.565	0.949	0.615	0.924	0.521	0.903	0.570	**0.916**	0.889	**0.911**	**0.835**	0.792	**0.826**	0.438	0.414
5	Word + Char Encoder	0.988	**0.959**	0.959	**0.959**	0.974	0.882	0.929	0.891	0.900	0.851	0.900	0.819	0.750	0.804	0.979	0.903
6	+ random char	0.986	0.953	0.947	0.951	0.976	**0.898**	0.927	0.904	0.902	0.807	0.881	0.819	0.741	0.802	0.969	0.924
7	Subword Encoder	0.986	0.934	0.972	0.941	0.968	0.831	0.950	0.852	0.804	0.897	0.821	0.715	0.827	0.735	**0.988**	0.877
8	+ Char Encoder	0.980	0.908	0.959	0.917	0.963	0.808	0.939	0.831	0.740	0.848	0.759	0.664	0.786	0.685	0.978	0.867
9	+ random char	0.985	0.931	0.966	0.938	0.973	0.866	0.950	0.881	0.799	0.876	0.813	0.718	0.819	0.736	0.984	0.925
10	+ LM pre-train	**0.990**	0.951	**0.982**	0.957	0.975	0.866	0.962	0.883	0.850	**0.935**	0.866	0.771	0.870	0.789	**0.988**	0.877
11	+ LM pre-train + random char	0.989	0.946	0.979	0.952	**0.980**	0.896	**0.964**	**0.909**	0.845	0.922	0.860	0.787	**0.872**	0.803	0.987	**0.941**

Conclusion

Result Analysis

- Spelling correction requires both spelling and context information
- Pre-trained LM facilitates spelling correction
- Training on additional synthetic character level noise improves model robustness

Possible Improvement

- pre-trained LM initialized version of Word+Char encoder model

Summary

- a strong stand-alone spelling corrector
- combines both spelling and context information
- leverages a pre-trained LM
- uses the synthetic character-level noise

Reference

- X. Li, H. Liu, L. Huang, *"Context-aware Stand-alone Neural Spelling Correction"*, Association for Computational Linguistics, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 407-414, doi: 10.18653/v1/2020.findings-emnlp.37