UNIVERSITY OF CAMBRIDGE

Code and models available at:
https://samueljamesbell.com/context

Context is Key: Grammatical Error Detection with Contextual Word Representations

Samuel Bell
Helen Yannakoudakis
Marek Rei

Motivation

Grammatical error detection (GED) can be considered a low-resource task, making purely-supervised approaches challenging due to limitations in dataset size and class label imbalance. Contextualized word representations (e.g. BERT, ELMo and Flair Embeddings) optimized on large amounts of unsupervised data, capture compositional information in language and present a useful tool in improving the state-of-the-art in GED.

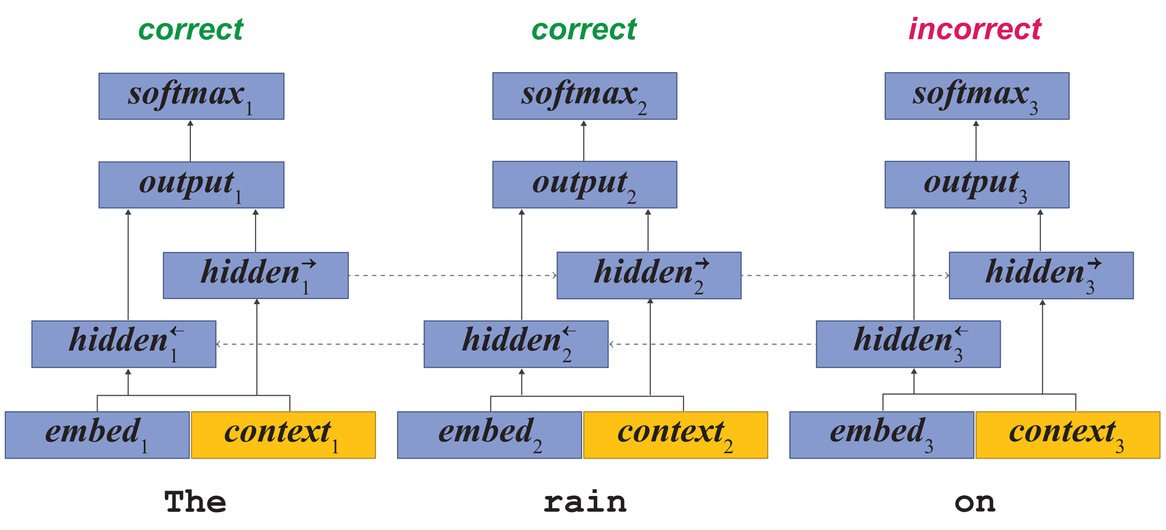
The	***correct***
rain	***correct***
on	***incorrect***
Spain	***correct***
falls	***correct***
manely	***incorrect***
…	…

Method

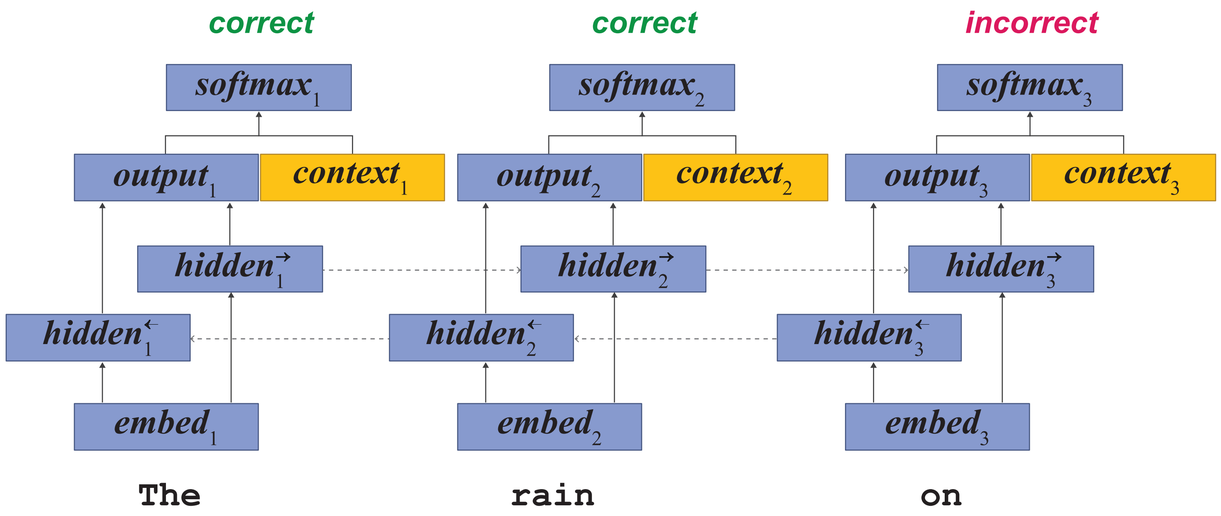
- We extend a **neural sequence labeller**, using the existing SOTA in GED [1].
- Using contextual embeddings from **BERT, ELMo, Flair**.
- **Trained only on the FCE training set** [2] to evaluate out-of-domain transfer.
- Models are evaluated on the FCE test set, the CoNLL-14 test set, the JFLEG test set, and the BEA '19 shared task dev and test sets.
- Automatically-induced error types across all datasets using ERRANT [3].

Model

- **Word-level + char-level bi-LSTM** sequence labeller.
- Softmax layer to output probability distribution over ***correct***/***incorrect***.
- Secondary language modelling objective.
- Contextual embeddings concatenated with **input** word+char embeddings:



- We also evaluate concatenating with LSTM **outputs**:



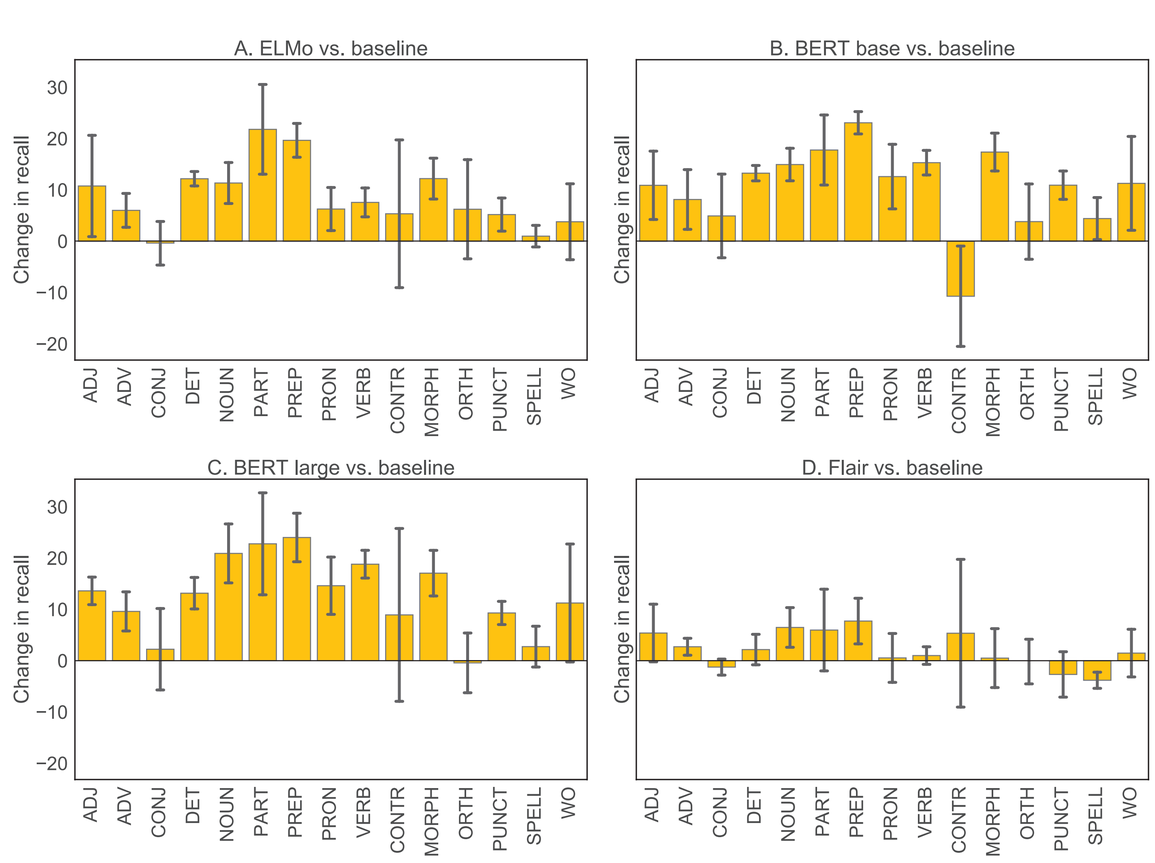
Contributions

- A new state-of-the-art in GED with publicly-released code and models.
- A neural sequence labeler enriched with contextual embeddings.
- A systematic comparison of BERT, ELMo and Flair on GED datasets.
- Particular benefits for transfer to out-of-domain datasets.
- A detailed analysis of error-type specific performance changes.

Results

- **A new SOTA on the FCE test set (+*35.9% rel.*).**
- Contextual embeddings always improve performance across datasets.
- Strong improvements in transfer performance to out-of-domain data.
- **BERT base yields the strongest performance across most datasets.**
- **Earlier contextual embeddings lower in the model performs best.**
- Relatively little improvement when using Flair Embeddings.
- BERT improves performance on nouns, prepositions, particles and morphological errors..
- BERT offers limited improvement around conjugation, orthography and spelling.

	$F_{0.5}$					
	CoNLL 1	CoNLL 2	FCE	JFLEG	BEA Dev	BEA Test
Rei (2017) [1]	17.86	25.88	48.48	-	-	-
Rei et al. (2017) [4]	21.87	30.13	49.11	-	-	-
Kasewa et al. (2018) [5]	28.3	35.5	55.6	-	-	-
Baseline	19.73	27.55	42.15	50.65	28.58	38.93
Flair	25.79	34.35	49.97	54.08	36.45	50.00
ELMo	29.14	40.15	52.81	58.54	42.96	56.15
BERT base	35.70	**46.29**	**57.28**	**61.98**	**48.50**	**63.57**
BERT large	**36.94**	45.80	56.96	61.52	47.75	61.26



Conclusions

- Contextual embeddings of all forms improve GED performance.
- Using BERT (base) yields the strongest improvement.
- Integrating contextual embeddings lower/earlier in a model is best.
- Contextual embedding performance varies with error type.

References

[1] Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting ofthe Association for Computational Linguistics*, pages 2121–2130.
[2] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.
[3] Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In P*roceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 793–805.
[4] Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. 2017. Artificial error generation with machine translation and syntactic patterns. *arXiv:1707.05236*.
[5] Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. *arXiv:1810.00668*.