Semantics or spelling?
Probing contextual word embeddings with orthographic noise

Jacob A. Matthews, John R. Starr, and Marten van Schijndel
Cornell University
jam963@cornell.edu

Research Question

★ Contextual word embeddings are assumed to capture semantic info.
★ How sensitive are these embeddings to non-semantic features in text input?
★ Our approach: a simple character swapping procedure to introduce minor orthographic noise.

Background

➔ Prior work has investigated the effect of noise on downstream task performance [1, 2, 3].
➔ Known problems with CWEs generated with PLMs (anisotropy, rogue dimensions) [4, 5].
➔ No work on the effect of textual noise on contextual embeddings.

Methods

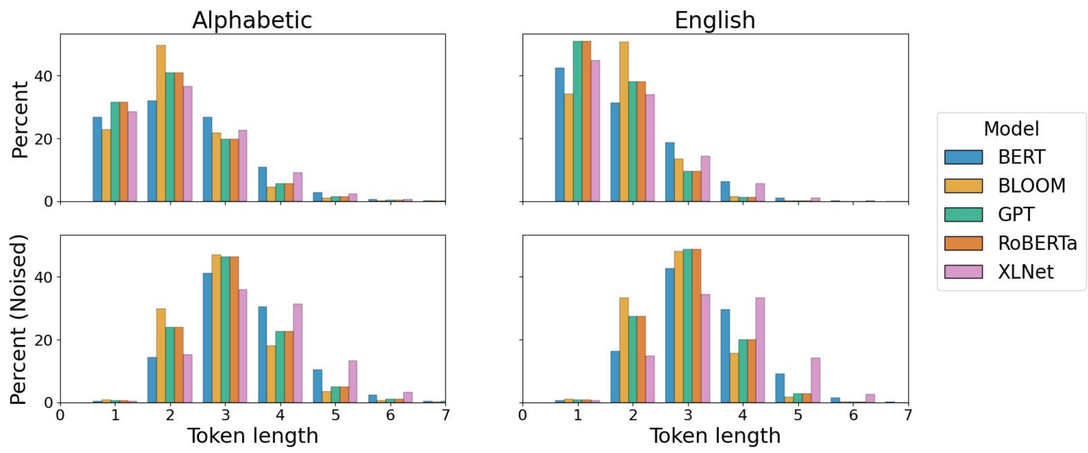
➔ Character swapping procedure: **swap a character in each word** with a random case-matched character (n=68k).
➔ Compare **similarity of edited word embedding and unedited word embedding**, both **with and without 100 words of context** from Wikitext.
➔ Similarity metrics: cosine similarity, Spearman correlation (to mitigate effects of anisotropy)

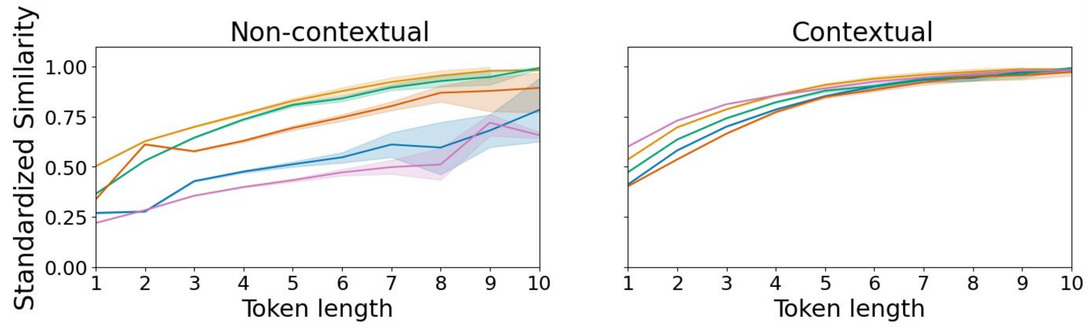
Effect on Tokenization

Model	Word	Edited	Word Tokens	Edited Tokens
GPT-2 BERT XLNet	contenders	contelders	“contenders” “contender” ,“s” “contenders”	“cont”, “e”, “ld”, “ers “con”, “tel”, “ders” “con”, “tel”, “der”, “s””

Tokenization

➔ Methods like BPE [6] result in the majority of words being represented by 1-3 tokens.
➔ English words are often only 1-2 tokens.
➔ Minor orthographic **noise causes complex and unpredictable “splitting” in token-level representation.**





Results

★ We find that CWEs are **highly sensitive** to minor orthographic noise.
  ○ Sensitivity is **related to subword tokenization: fewer tokens, higher sensitivity.**
★ Single character swaps (particles vs partfcles) result in up to **60% loss** of a word’s semantic identity.
★ Most English words are represented by 1-2 tokens, making them vulnerable.
★ Context **does not significantly mitigate** this effect for any model.

References

[1] Xue et al. 2022. “ByT5: Towards a token-free future with pre-trained byte-to-byte models”. *TACL*. **[2]** Niu et al. 2020. “Evaluating robustness to input perturbations for neural machine translation”. *ACL*. **[3]** Karpukhin et al. 2019. “Training on synthetic noise improves robustness to natural noise in machine translation”. *W-NUT 2019*. **[4]** Kawin Ethayarajh. 2019. “How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings”. *EMNLP-IJCNLP* **[5]** Timkey and van Schijndel. 2021. “All bark and no bite: Rogue dimensions in transformer language models obscure representational quality”. *EMNLP* **[6]** Sennrich et al. 2016. “Neural machine translation of rare words with subword units”. *ACL*