

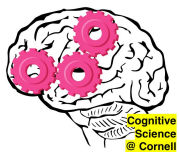
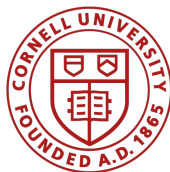
Semantics or Spelling?

Probing contextual word embeddings with orthographic noise

Jacob A. Matthews, John R. Starr and Marten van Schijndel

Cornell University

jam963@cornell.edu



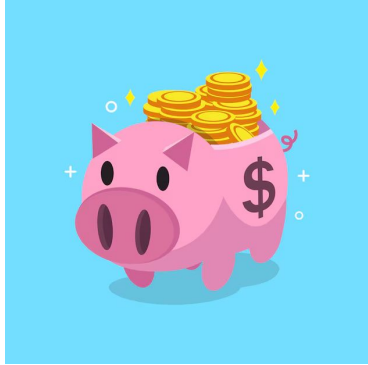
C.Psyd

tl;dr

1. Contextual word embeddings (CWEs) are sensitive to minor spelling changes
2. Changing one character can cause 60% loss in semantic similarity
3. Sensitivity is related to tokenization: fewer tokens = more sensitive

Introduction

piggy **bank**



!=

river **bank**



Introduction



How robust are CWEs to minor orthographic noise?

Methodology

- Character swapping procedure
 - Randomly replace a single character in each word (n=68k) with a random case-matched character

Methodology

- Character swapping procedure
 - Randomly replace a single character in each word (n=68k) with a random case-matched character
- CWE generation
 - Mean-pool hidden states corresponding to tokens in each word.

Methodology

- Character swapping procedure
 - Randomly replace a single character in each word (n=68k) with a random case-matched character
- CWE generation
 - Mean-pool hidden states corresponding to tokens in each word
- Compare CWEs
 - Compute similarity between edited words (with character swaps) and original words.
 - Use both unstandardized similarity (cosine) and standardized similarity (Spearman) metrics.
 - Analysis both with and without context (100 words)

Methodology

Stars with cooler outer atmospheres , including the Sun , can form various diatomic and polyatomic molecules . = = = Diameter = = = Due to their great distance from the Earth , all stars except the Sun appear to the unaided eye as shining points in the night sky that twinkle because of the effect of the Earth 's atmosphere . The Sun is also a star , but it is close enough to the Earth to appear as a disk instead , and to provide daylight . Other than the Sun , the star with the largest

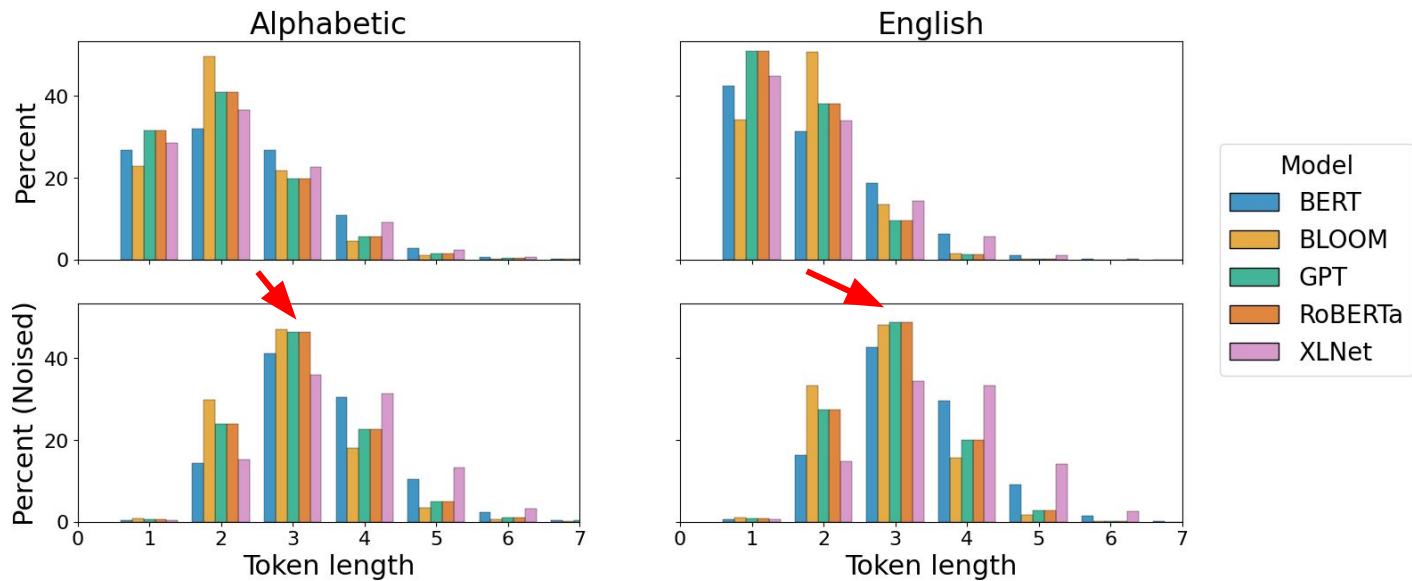
Methodology

Stars with cooler outer atmospheres , including the Sun , can form various diatomic and polyatomic molecules . = = = Diameter = = = Due to their great distance from the Earth , all stars except the Sun appear to the unaided eye as shining points in the night sky that twinkle because of the effect of the Earth 's atmosphere . The Sun is also a star , but it is close enough to the Earth to appear as a disk instead , and to provide daylight . Other than the Sun , the star with the largest

Tokenization

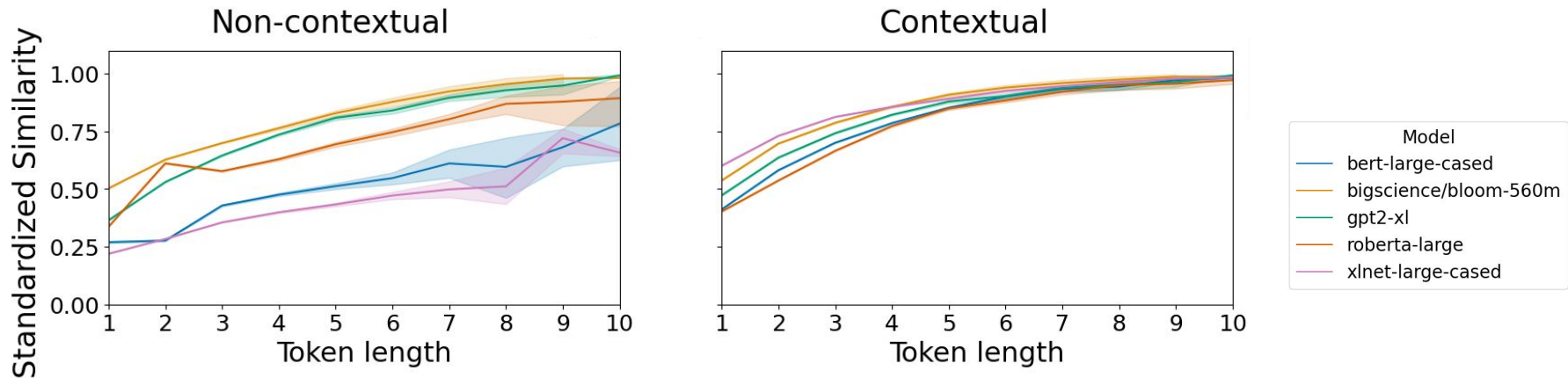
Model	Word	Edited	Word Tokens	Edited Tokens
GPT-2 BERT XLNet	contenders	conteliders	“contenders” “contender”, “s” “contenders”	“cont”, “e”, “ld”, “ers” “con”, “tel”, “ders” “con”, “tel”, “der”, “s””

Tokenization



Most words are represented by only 1-3 tokens, English words by 1 or 2

Representational analysis



Similarity between the edited word and unedited word increases with the number of tokens used to encode the unedited word.

Conclusion

- CWEs are surprisingly sensitive to orthographic noise
- This sensitivity is related to subword tokenization
- Caution is warranted in using CWEs as semantic proxies

Acknowledgements

- Computational Psycholinguistics Discussions (C.Psyd)
- Cornell NLP
- Cognitive Science at Cornell



C.Psyd

