Semantics or spelling? Probing contextual word embeddings with orthographic noise

Jacob A. Matthews   John R. Starr   Marten van Schijndel
Cornell University
{jam963, jrs673, mv443}@cornell.edu

Abstract

Pretrained language model (PLM) hidden states are frequently employed as contextual word embeddings (CWE): high-dimensional representations that encode semantic information given linguistic context. Across many areas of computational linguistics research, similarity between CWEs is interpreted as semantic similarity. However, it remains unclear exactly what information is encoded in PLM hidden states. We investigate this practice by probing PLM representations using minimal orthographic noise. We expect that if CWEs primarily encode semantic information, a single character swap in the input word will not drastically affect the resulting representation, given sufficient linguistic context. Surprisingly, we find that CWEs generated by popular PLMs are highly sensitive to noise in input data, and that this sensitivity is related to subword tokenization: the fewer tokens used to represent a word at input, the more sensitive its corresponding CWE. This suggests that CWEs capture information unrelated to word-level meaning and can be manipulated through trivial modifications of input data. We conclude that these PLM-derived CWEs may not be reliable semantic proxies, and that caution is warranted when interpreting representational similarity.

1 Introduction

Contextual word embeddings (CWE) have become commonplace, enabling a wide range of research and industry applications. These representations, typically generated using transformer language models like BERT (Devlin et al., 2019), differ from their non-contextual predecessors like word2vec (Mikolov et al., 2013) in that these embeddings vary based on the linguistic context of the words they represent. However, the extent to which these representations capture contextual (Ethayarajh, 2019), linguistic (Miaschi and Dell'Orletta, 2020), and social (Kurita et al., 2019; Guo and Caliskan, 2020) information remains an area of significant interest. In emerging subfields like semantic shift detection (SSD), CWEs are frequently employed as semantic proxies (Montanelli and Periti, 2023). BERT or similar pretrained language models (PLMs) are commonly used to generate CWEs which function as continuous representations of semantic information. This use of CWEs assumes that PLMs encode semantic information in their hidden states, and that the similarity between CWEs can be directly interpreted as a measure of semantic similarity.

We investigate this assumption using a simple probing task.[1] We generate CWEs of alphabetic words in authentic contexts, then we compare them pairwise to CWEs of the same word with a single character swapped for another of the same case. Since the rest of each word and context sequence is otherwise kept identical, we would expect that the unedited characters and linguistic context should be sufficient to generate extremely similar CWEs to their unedited counterparts. Put differently, if CWEs primarily capture semantic information, they should be robust to minor orthographic noise.

In fact, this is not the case. We observe that CWEs are quite sensitive to minor orthographic noise, regardless of the PLM used to generate them. Moreover, this effect is strongest for single token words, which account for over a third of English words in our test data. We find that these representations are highly dependent on tokenization: words represented by fewer tokens are more sensitive to orthographic noise. Changes at the character level result in drastic differences in token-level representation, which in turn causes the model to produce distinct CWEs.

---

[1] All code and data are available at https://github.com/jam963/semantics-or-spelling/.

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

Table 1: Example of the effect of noise on model tokenization. The second "n" in "contenders" has been replaced with an "l". Though GPT-2 and XLNet tokenize the original word identically, a single character replacement results in distinct splitting behavior in the edited tokenization.

2 Background

While some prior work has addressed the robustness of PLMs to noisy input sequences (Xue et al., 2022; Niu et al., 2020; Karpukhin et al., 2019), these studies rely on downstream task performance to quantify noise robustness. However, to our knowledge, there is no thorough account of *representational* noise robustness in PLMs. Additionally, there are known problems with representations generated by PLMs, such as anisotropy (Mimno and Thompson, 2017; Ethayarajh, 2019), the influence of rogue dimensions (Timkey and van Schijndel, 2021), and similarity underestimation due to frequency (Zhou et al., 2022).

One notable difference between non-contextual word-based embedding models such as word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) and PLMs is tokenization. Whereas the former utilize a fixed model vocabulary, contemporary transformer-based PLMs generally rely on *subword tokenization* strategies for efficiency and flexibility. Most models use a small number of very similar methods. Byte Pair Encoding (BPE; Sennrich et al., 2016), specifically byte-level BPE, is used in a wide range of popular PLMs, including GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), RoBERTa (Liu et al., 2019a), and BLOOM (Scao et al., 2022), among others. BPE employs a set of merge rules to convert a fixed vocabulary into a segmented subword vocabulary. This approach allows out-of-vocabulary words to be tokenized into sequences of in-vocabulary subword tokens. Besides BPE, there are a number of other subword tokenization methods, including SentencePiece (Kudo and Richardson, 2018) and WordPiece (Schuster and Nakajima, 2012).

In addition, character-based (Tay et al., 2022) and tokenizer-free methods (Clark et al., 2022) have been investigated with transformer models, as well as regularization approaches to address known shortcomings with existing tokenizers (Provilkov et al., 2020). For a comprehensive overview and taxonomy of contemporary tokenizers, see Mielke et al. (2021).

3 Data and model selection

For our analyses, we establish a test vocabulary by isolating every unique, alphabetic, whitespace-delimited word of more than 3 characters in the `Wikitext-2-raw-v1` training set (Merity et al., 2016). This results in a total vocabulary size of 68,725 alphabetic words. To generate edited words, we randomly select a character in each word and replace it with another random alphabetic character of the same case (Table 1). We will refer to these as *edited words*.

The PLMs we analyze represent a range of families, sizes, and types, all of which are freely available via the HuggingFace `transformers` package (Wolf et al., 2020), which we also use here for our analyses. We generate CWEs using versions of GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), XL-Net (Yang et al., 2019) and BLOOM (Scao et al., 2022). While GPT-2 and BLOOM are causal language models which are typically not used for generating CWEs, they represent a class of extremely popular and widely-available PLMs. Additional model details are described in the Appendix (Table 2).

4 Measuring similarity of CWEs

We use two measures of representational similarity between CWEs: cosine similarity and Spearman correlation. We first define a mean-pooled word embedding $\bar{\mathbf{w}}$ as

$$\bar{\mathbf{w}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{h_i} \qquad (1)$$

where $\mathbf{h_i}$ is a hidden state corresponding to token i in a word of n tokens. The cosine similarity cos between a word embedding $\bar{\mathbf{w}}$ and its noised coun-
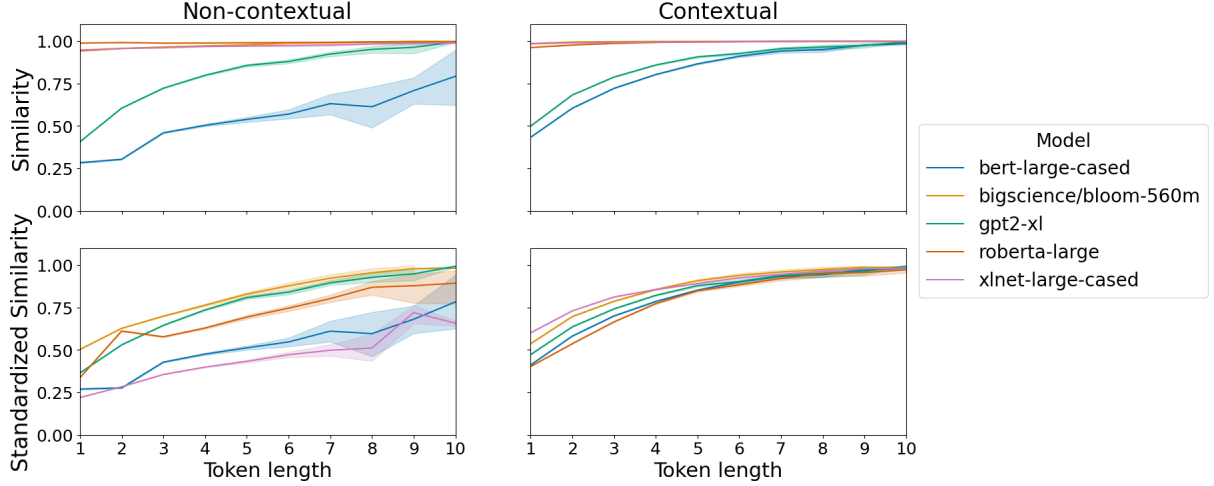
Figure 1: Similarities between CWEs and edited-CWEs, using cosine (top) and Spearman (bottom) for individual words without contexts (left) and with contexts (right). The shaded area around each line represents the 99% confidence interval.

terpart $\bar{\mathbf{w}}_{edit}$ is given by

$$cos(\bar{\mathbf{w}}, \bar{\mathbf{w}}_{edit}) = \frac{\bar{\mathbf{w}} \cdot \bar{\mathbf{w}}_{edit}}{\|\bar{\mathbf{w}}\| \|\bar{\mathbf{w}}_{edit}\|} \qquad (2)$$

where values closer to 1.0 are interpreted as more similar and therefore more noise robust.

Though commonly used, cosine similarity is *unstandardized*. As such, small numbers of outlier or "rogue" dimensions often dominate these similarity scores (Timkey and van Schijndel, 2021; Kovaleva et al., 2021). Moreover, PLMs have been shown to exhibit *anisotropy*, where all representations occupy a narrow region of the embedding space (Mimno and Thompson, 2017; Ethayarajh, 2019), which distorts similarity measures like cosine. Thus we also calculate the Spearman rank correlation coefficient, ρ, for all embedding pairs,

$$\rho_{\bar{\mathbf{w}}, \bar{\mathbf{w}}_{edit}} = \frac{cov(\mathbf{R}(\bar{\mathbf{w}}), \mathbf{R}(\bar{\mathbf{w}}_{edit}))}{\sigma_{\mathbf{R}(\bar{\mathbf{w}})} \sigma_{\mathbf{R}(\bar{\mathbf{w}}_{edit})}} \qquad (3)$$

where $cov(\mathbf{R}(\bar{\mathbf{w}}), \mathbf{R}(\bar{\mathbf{w}}_{edit}))$ is the covariance of the rank-transformed CWEs $\bar{\mathbf{w}}$ and $\bar{\mathbf{w}}_{edit}$. Because the rank transformation mitigates the effects of rogue dimensions and anisotropy, Spearman is a more robust similarity measure than cosine (Zhelezniak et al., 2019).

5 Representational similarity analyses

5.1 Non-contextual similarity

To determine the baseline influence of our editing procedure on word embeddings without linguistic context, we first generate non-contextual word

embeddings for every word and edited word. We pass each word and corresponding edited word to all models independently, without any additional context, then mean pool the final layer hidden states corresponding to the tokens in each word. This gives us a pair of word embeddings for all vocabulary items $(\mathbf{w}, \mathbf{w}_{edit})$, which we compare using cosine similarity and Spearman correlation. For both similarity measures, values closer to 1.0 suggest that a model has similar representations for both the edited and unedited words.

5.1.1 Results

We report results from the largest models tested in Figure 1 (Left), with similarity scores binned by the number of tokens used to represent the unedited word, or the *token length*. While some models (BLOOM, GPT2-small, and the RoBERTa models) have average cosine similarities close to 1.0 for all token lengths, standardized similarities (Spearman) are universally low for all but the highest number of tokens, falling between 0.2 and 0.5 for single token words.[2]

5.2 Contextual similarity

Since CWEs are designed to modulate their representations according to linguistic context, we may expect that context would correct these representational discrepancies, allowing the model to better ignore minor orthographic noise.

To simulate the conditions under which these models are typically employed, we investigate

---

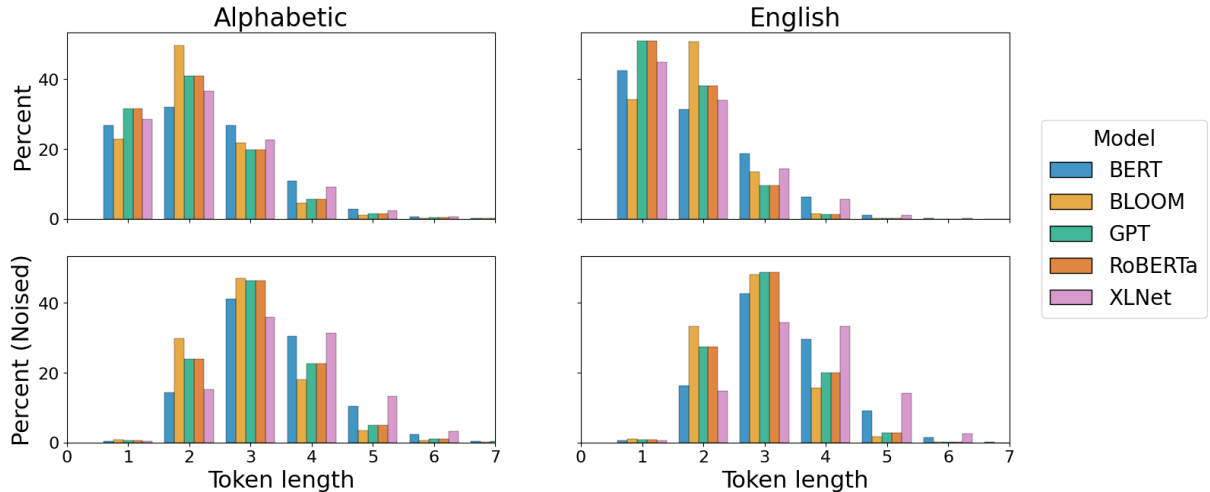[2]See Figures 3 and 4 in Section A.2 for all model results.

Figure 2: Distribution of token lengths for all alphabetic words (left) and English words (right), tokenized without (top) and with noise (bottom). For most models, English words tend to be tokenized with fewer tokens (top right). Noise tends to increase the token length, with most noisy words requiring 3 or more tokens instead of 1 or 2.

how noise affects CWE representations given authentic linguistic contexts. We repeat our analysis from Section 5.1, but provide contexts from Wikitext for each word when generating hidden layer representations. We find naturally occurring 100-word contexts that contain each unedited test word, and we pool hidden layer representations in that context to generate CWEs corresponding to the test word and their edited twin. Thus, between a given word and edited word, their respective 100-word input sequence varies only by a single character. As in section 5.1, we compare the resulting CWEs for the original and edited items using cosine similarity and Spearman correlation.

5.2.1 Results

Though context notably increases similarity and standardized similarity for some models like BERT and XLNet, it has almost no effect on BLOOM, RoBERTa, and larger GPT-2 variants (Figure 1, Right; see also Figures 3 and 4 in Appendix). Just as in the non-contextual case, words originally represented by fewer tokens are the least noise robust (Spearman averaging 0.5 for single token words). The fact that similarity scores approach 1.0 at high numbers of tokens indicates that PLMs are in principle able to use context to mitigate input noise; however, the dependence of subword tokenization schemes on occurrence frequencies seems to inhibit this behavior and make models more susceptible to orthographic noise.

6 PLM token length distributions

To study the impact of frequency on the behaviors we've found, we analyze the distribution of word lengths (in tokens) in our test vocabulary (Figure 2, Top). Because the Wikitext dataset contains non-English words, we also use the MorphoLex database (Sánchez-Gutiérrez et al., 2018) to identify authentic English words (40,013 words). Though specific tokenizer implementations vary somewhat between models, we find that the token length distributions are generally similar (Figure 2): as many as 90% of alphabetic words are less than 3 tokens in length, while English words are mostly composed of even fewer tokens, usually only one or two. Coupled with our results from section 5.2.1, this suggests that most English words will exhibit very low robustness to minor orthographic noise, even when observed in long linguistic contexts.

Based on how subword tokenizers work, edited words tend to be represented by more tokens than unedited words (Figure 2, Bottom). Though subword tokenizers can accommodate orthographic noise, they can only do so by combining multiple tokens that are not present in the unedited word. For all but very rare words composed of many tokens, the mismatch in token-level representation is too much for context to representationally "repair," even when the noise in question only consists of a single character.

7 Discussion

We show that contextual word embeddings (CWEs) are highly sensitive to orthographic noise and that this effect is related to the way text data is processed by subword tokenizers. Our results indicate that CWE representations can be manipulated through trivial orthographic changes to the input – given 100 words of context, 40%-60% of a word's semantic identity is lost when a single character is changed – challenging the assumption that CWEs capture primarily word-level semantic information. In addition, this study suggests that CWEs and representational similarity measures are surprisingly fragile, and common noise sources such as misspellings and optical character recognition errors may result in unpredictable model behavior downstream. As such, we advise caution when relying on PLM-generated CWEs as semantic proxies in computational research, especially when dealing with noisy data.

8 Limitations

This study relies on third-party packages and open-source models which may differ from the implementation described in cited literature. Our procedure for generating CWEs, designed to be compatible with a range of models, may not be optimal, as we only combine the last hidden states through mean pooling. Though we believe the number of words in our test data to be sufficient, the vocabulary in Wikitext is still limited and may not be representative of English words in all contexts.

We note two limitations of our noising procedure, which 1) operates on random characters and 2) applies to all classes of words. Regarding the first limitation, our procedure may not reflect the actual distribution of orthographic noise in human-generated training data, which may already have some systematic character-level noise present (i.e. "teh" for "the", swapping "j" for "h"). However, we believe that our approach provides a generalizable view of PLM behavior rather than a focused study of identifiable noise sources. Because these models are deployed across a vast range of linguistic contexts, we do not take a strong view of what constitutes "authentic" orthographic noise.

With respect to the second limitation, our study does not address the influence of part-of-speech or other word-level characteristics on CWE noise robustness. By limiting our investigation to words with over 3 characters, our study does not address the noise robustness of many function words in English.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Wenqian Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Find-

ings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentence-Piece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Workshop on Representation Learning for NLP*.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *ArXiv*, abs/2304.01666.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Claudia H. Sánchez-Gutiérrez, Hugo Mailhot, S. Hélène Deacon, and Maximiliano A. Wilson. 2018. Morpholex: A derivational morphological database for 70,000 english words. *Behavior Research Methods*, 50:1568–1580.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre

Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma Sharma, S. Longpre, So maieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation coefficients and semantic textual similarity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 951–962, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Model details

Family	Model	Parameters	Tokenizer	Training Objective	Released
BERT	bert-base-cased	109M	WordPiece	MLM, NSP	2018
	bert-large-cased	335M			
GPT	gpt2	137M	BPE	CLM	2019
	gpt2-medium	380M			
	gpt2-large	812M			
	gpt2-xl	1.61B			
RoBERTa	roberta-base	125M	BPE	MLM	2019
	roberta-large	335M			
XLNet	xlnet-base-cased	150M	SentencePiece	PermLM	2019
	xlnet-large-cased	340M			
BLOOM	bloom-560m	560M	BPE	CLM	2022

Table 2: Summary of PLMs analyzed. Training objectives include masked language modeling (MLM), next-sentence prediction (NSP), causal language modeling (CLM), and permutation language modeling (PermLM).

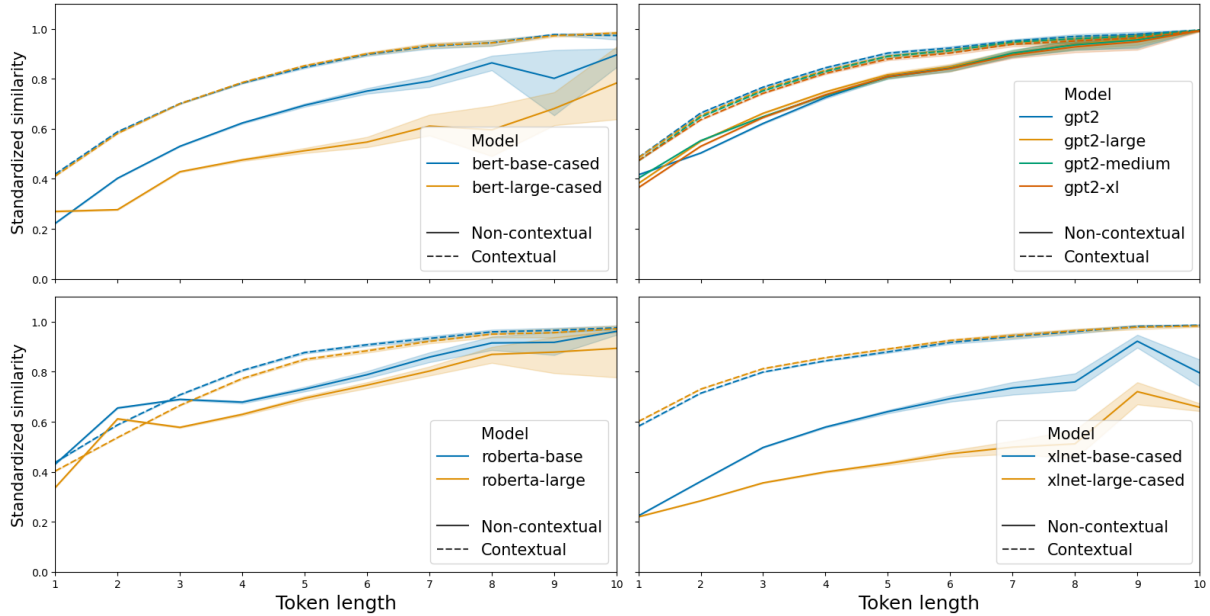## A.2 Full representational analysis results

Figure 3: Standardized similarities (Spearman), grouped by model family. Token length is measured as the number of tokens used to encode the original, unedited word. We show similarities both for CWEs generated with context (dashed line) and without (solid line), and the shaded area around each line represents the 99% confidence interval.
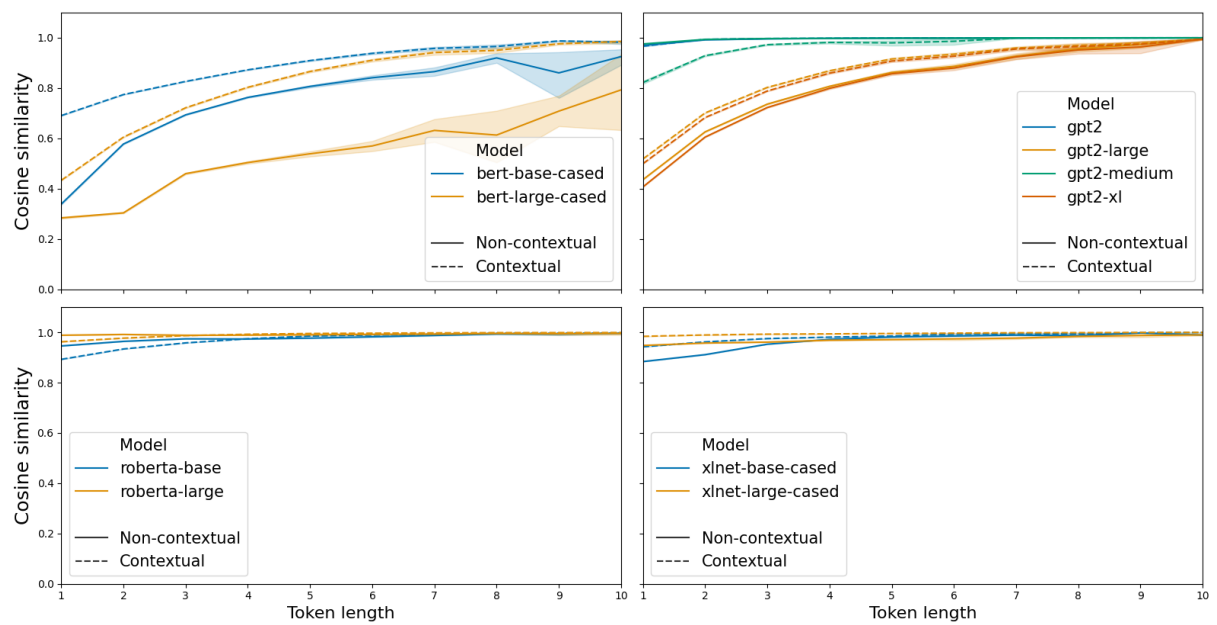
Figure 4: Cosine similarities between CWEs and edited-CWEs, grouped by model family. Formatting is otherwise identical to Figure 3.