SemEval-2020 Task 3: Graded Word Similarity in Context by Composing Pre-trained Embeddings

Tim Lawson

December 18, 2023

1 Introduction

In his *Foundations of Arithmetic*, Frege promises "never to ask for the meaning of a word in isolation, but only in the context of a proposition" (1980, p. xvii). This 'context principle' is intuitive: words are frequently polysemous, or assume different connotations and emphasis within different expressions (Armendariz, Purver, Ulčar, et al. 2020, pp. 2–3). Historically, however, context-dependence has been a problem for distributional meaning representations. Founded on the distributional hypothesis (Harris 1954; Firth 1957), both count-based and predictive models of word meaning[1] originally produced a single representation for each word in the model's vocabulary. One of these *static* representations must, therefore, encode all of a word's senses and connotations, which may obstruct its use in modelling context-dependent phenomena.

Prior to the widespread availability of pre-trained language models, this problem was generally addressed by one of two approaches: firstly, by producing a representation for each sense of a target word and disambiguating between them in the given context (word-sense disambiguation); or secondly, by composing the representation of the target word with the representations of the words in its context (contextualisation). These approaches have been largely overshadowed by the advent of model architectures that take sequences as inputs and naturally produce *contextual* representations of the items in the sequence, such as Transformers (Vaswani et al. 2017).

To my knowledge, however, there has been scant direct comparison of the performance of these contextual representations with the application of prior methods of contextualisation to static representations (section 3). SemEval-2020 Task 3, "Graded Word Similarity in Context" (Armendariz, Purver, Pollak, et al. 2020), presents an opportunity to make such a comparison. Briefly, the task is to predict the human judgment of similarity of the same pair of words in two different contexts. I elected to focus on the first sub-task, which is to predict the *change* in similarity, rather than the absolute similarity in each context (section 2). Specifically, I evaluated the results obtained by computing the cosine similarity between static and contextual embeddings and the composition of these embeddings within a fixed-size context window (section 4).

2 Task definition

The first sub-task of SemEval-2020 Task 3 is to predict the direction and magnitude of the change in the human judgment of similarity of the same pair of target words in two different contexts. The task is unsupervised: the submissions were evaluated on the CoSimLex dataset (Armendariz, Purver, Ulčar, et al. 2020, pp. 39–42) but only a minimal 'practice kit' of fewer than ten instances was provided in advance. CoSimLex is an extension of SimLex-999 (Hill et al. 2015) that consists of pairs of target words and their contexts in four languages: English (n = 340), Finnish (n = 24), Croatian (n = 112), and Slovene (n = 111).

[1] This terminological distinction is due to Baroni et al. (2014).

The score for the first sub-task was computed by the uncentered (zero-mean) Pearson correlation coefficient between the predicted changes in similarity and the human judgments represented in the CoSimLex dataset (Armendariz, Purver, Ulčar, et al. 2020, p. 42). This metric is equivalent to the cosine similarity between the two vectors of results:

$$\text{score}(\vec{\hat{y}}, \vec{y}) = \frac{\sum_{i=1}^{n} \hat{y}_i y_i}{\left(\sum_{i=1}^{n} \hat{y}_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)} = \frac{\vec{\hat{y}} \cdot \vec{y}}{\|\vec{\hat{y}}\| \|\vec{y}\|} \tag{1}$$

Notably, this metric is invariant with respect to multiplication by a scalar quantity, so the results of composing an equal number of representations by addition or the arithmetic mean are equal. Hence, I only investigated addition among the two (section 4).

3 Related work

Many context-based approaches to word-sense disambiguation have been proposed since the advent of count-based models of word meaning. In the paper that introduced Latent Semantic Analysis (LSA), for example, the authors argued that taking the average of the high-dimensional representation of a word with those of its immediate context may be sufficient to determine the word's contextual meaning (Landauer and Dumais 1997, pp. 229–230). Thus, the contextualisation of representations of word meanings is intimately related to their composition to form representations of more complex expressions. This relationship is evident, for example, in the work of Kintsch (2001), who proposed a a procedure to contextualise the representation of a predicate according to its argument, and in the adaptation of this demonstration by Mitchell and Lapata (2008) to evaluate alternative composition operations. Vector addition and averaging continue to be 'surprisingly effective' means to compose word embeddings (Boleda 2020, p. 10), and addition produces plausible results for the word-analogy task (e.g., Mikolov, Chen, et al. 2013, p. 9; Mikolov, Sutskever, et al. 2013, p. 7). A review of distributional semantic models is given by Lenci (2018).

Nevertheless, models that produce contextual embeddings have achieved widespread success on benchmark tasks that involve language understanding (Bommasani, Hudson, et al. 2022, pp. 22–27). There is, however, cause to criticise the suitability of typical benchmarks for characterising the capabilities of language models (Srivastava et al. 2023, pp. 5–6). The computational cost of inference with contextual embeddings may also be prohibitive or unjustified (section 5.1). For instance, Arora et al. (2020) demonstrated that static and even *random* embeddings can achieve similar performance given sufficient data and linguistically simple tasks. Furthermore, Gupta et al. (2019, pp. 5244–5246) and Bommasani, Davis, et al. (2020, pp. 4760–4762) compared the performance of different embedding methods on a variety of word-similarity tasks, and demonstrated that static embeddings can be obtained from contextual models that outperform their contextual counterparts while reducing the computational cost of inference. Surveys of contextual and static embeddings are given by Q. Liu et al. (2020) and Torregrossa et al. (2021); further analyses of their nature are provided by Hewitt and Manning (2019), Y. Liu et al. (2019), Reif et al. (2019), and Brunner et al. (2020).

Batchkarov et al. (2016) critically analyse word similarity as an evaluation methodology for distributional semantic models. In particular, the notion of 'similarity' manifested by these models encompasses a broad range of semantic relations (e.g., Padó and Lapata 2003, p. 2), with the consequence that performance on an intrinsic word-similarity task does not necessarily translate to extrinsic downstream tasks (Batchkarov et al. 2016, pp. 7–8). Moreover, inter-annotator agreement is generally poor for word-similarity tasks in comparison to more specific downstream tasks (ibid., pp. 8–9). In this case, Armendariz, Purver, Ulčar, et al. (2020, p. 8) and Armendariz, Purver, Pollak, et al. (2020, p. 42) reported similar correlation scores between the different languages and in comparison to SimLex-999 (Hill et al. 2015, pp. 678–680).

Model name	English	Finnish	Croatian	Slovene
`EMBEDDIA/crosloengual-bert`[1]	✓	✓	✓	✓
`TurkuNLP/bert-base-finnish-cased-v1`[2]		✓		
`TurkuNLP/bert-base-finnish-uncased-v1`[2]		✓		
`TurkuNLP/bert-large-finnish-cased-v1`[2]		✓		
`bert-base-cased`	✓			
`bert-base-multilingual-cased`	✓	✓	✓	✓
`bert-base-multilingual-uncased`	✓	✓	✓	✓
`bert-base-uncased`	✓			
`bert-large-cased`	✓			
`bert-large-cased-whole-word-masking`	✓			
`bert-large-uncased`	✓			
`bert-large-uncased-whole-word-masking`	✓			
`classla-bcms-bertic`[3]			✓	

Figure 1: The pre-trained models from the HuggingFace *Transformers* library (Wolf et al. 2020) that I evaluated for each language. The corresponding references are [1]Ulčar and Robnik-Šikonja (2020a), [2]Virtanen et al. (2019), [3]Ljubešić and Lauc (2021), and Devlin et al. (2019) otherwise.

4 Methodology

4.1 Embedding models

I undertook this task to investigate the relative performance of pre-trained static and contextual embeddings for a context-dependent word-similarity task. To this end, Transformer models (Vaswani et al. 2017) were a natural choice. The baseline models for the task were the multilingual BERT model (Devlin et al. 2019) and ELMo models (Peters et al. 2018) trained on Finnish, Croatian, and Slovene datasets (Ulčar and Robnik-Šikonja 2020b); and the vast majority of the task submissions used Transformers (Armendariz, Purver, Pollak, et al. 2020, pp. 36, 42–45). The models that I evaluated were accessed via the HuggingFace *Transformers* library (Wolf et al. 2020) and are listed in fig. 1.

The primary comparison that I made was between these models' static input and contextual output representations. Several of the submissions used a combination of a Transformer's hidden-states (e.g., Gamallo 2020, p. 276; Pessutto et al. 2020, p. 3; Hettiarachchi and Ranasinghe 2021, p. 4). This choice is supported by the analysis of Ethayarajh (2019), who found that the upper layers of Transformer models produce more context-dependent representations. Hence, I also evaluated an example of this approach—a thorough comparison of the performance of its variants is, however, beyond the scope of this paper. Hereafter, I refer to the three types of embeddings that I evaluated as:

- `static`, the model's input embeddings;
- `contextual`, the model's output embeddings; and
- `pooled`, the sum of the model's last four hidden-states.

I note that I did not directly reproduce the baseline model because it requires the `bert-embedding` Python package, which is incompatible with Apple's ARM-based processors and has been deprecated since 2020 (Lai 2023). However, the baseline is notionally equivalent to the `contextual` embeddings of the `bert-base-multilingual-cased` model with a window size of zero.

4.2 Composition operations

The basic procedure that I employed is described in fig. 2. For each pair of target words and each of the two contexts in which they appear, I obtained a contextualised representation of a target word by: finding the index of the target word's first sub-word token within the tokens of the target word's

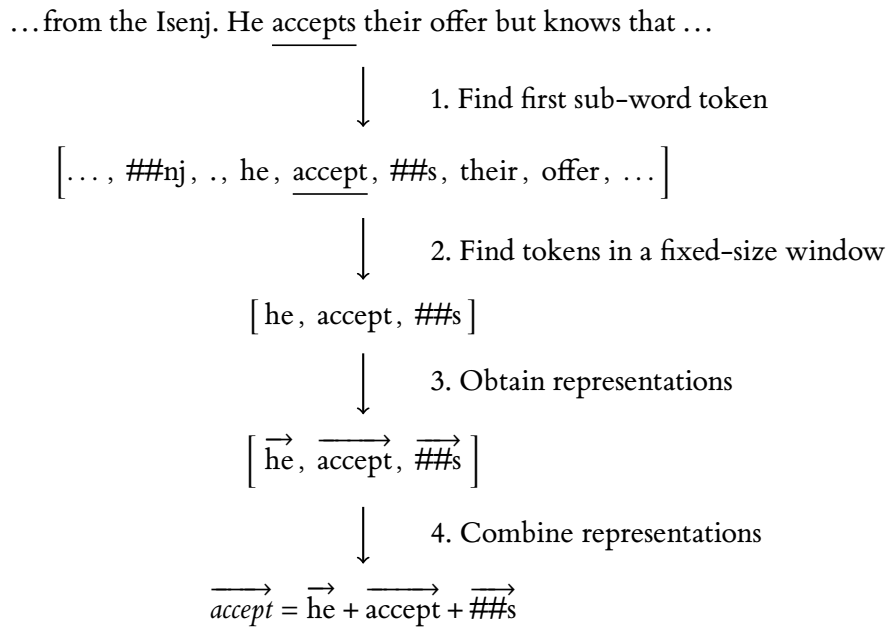

Figure 2: A schematic of the procedure used to obtain a contextualised representation of a target word from pre-trained embeddings. In this example, the target word is "accept", the window size is one (either side of the target word), and the composition operation is addition.

context; finding the tokens within a fixed-size window around the target word's first token; obtaining the embeddings of the tokens in the window; and combining the the embeddings to produce a single representation of the target word. Notably, the use of a sub-word vocabulary by the models in question (e.g., Devlin et al. 2019, p. 4174) dictates that a target word may be represented by a different number of tokens in each context. As a result, the representations of a pair of target words may be different in each context, even if they are static and the window size is zero. This is the cause of the non-zero scores obtained by models of this kind, particularly for the Finnish language (fig. 5).

Inspired by Landauer and Dumais (1997), Kintsch (2001), and Mitchell and Lapata (2008), I predominantly investigated the application of element-wise addition and multiplication as composition operations. However, preliminary experiments indicated that multiplication performed poorly across all languages, models, and window sizes; hence, it was discarded before the final analysis. Additionally, I chose to evaluate the concatenation ('stacking') of embeddings. In the case that the number of embeddings was fewer than that expected for the window size, i.e., the target word was close to the beginning or the end of its context, I right-padded the concatenated embedding with zeros to obtain contextual embeddings of equal length. This approach was also inferior to addition for practically all combinations of parameters.

4.3 Window size

Due to the computational expense of exhaustively searching the possible window sizes, I applied heuristics to constrain the search space. A naïve estimation of the average number of words in each context, i.e., segmenting on whitespace, gave a result of between 40 and 60 for the different languages. Therefore, for the static-embedding models, I chose 50 as an upper bound on the window size on either side of the target word. The motivation to choose a smaller maximum window size for contextual-embedding models was similarly economical (section 5.1); however, as the window size approaches the length of the sequence, one would expect a combination of token representations to be superseded by the sequence-level representation of the model, e.g., the special `CLS` token of BERT variants (Devlin et al. 2019, p. 4174). These heuristics were largely vindicated by the results of the evaluation, which

Language	Model name	Window size	Score
en	bert-large-uncased-whole-word-masking	16	0.471
fi	EMBEDDIA/crosloengual-bert	10	0.633
hr	classla/bcms-bertic	31	0.567
sl	EMBEDDIA/crosloengual-bert	11	0.383

(a) static

Language	Model name	Window size	Score
en	bert-base-uncased	1	0.660
fi	TurkuNLP/bert-large-finnish-cased-v1	1	0.679
hr	EMBEDDIA/crosloengual-bert	3	0.708
sl	bert-base-multilingual-cased	3	0.572

(b) contextual

Language	Model name	Window size	Score
en	bert-base-uncased	1	0.653
fi	TurkuNLP/bert-large-finnish-cased-v1	1	0.668
hr	EMBEDDIA/crosloengual-bert	3	0.717
sl	EMBEDDIA/crosloengual-bert	2	0.544

(c) pooled

Figure 3: The best scores obtained by each embedding model for each language. In all cases, the best score was obtained with the additive composition operation.

demonstrated that the scores decrease as the window size approaches the maximum.

5 Results

5.1 Cost-benefit analysis of contextual embeddings

1: Write this section.

5.2 Language-specificity of window-size effects

Generally, I found that the scores obtained by all three types of embeddings were maximised by a non-zero context-window size. The influence of the window size is intuitive in the case of static embeddings. Without a context window, the representations of a target word only differ between expressions if the word is represented by different sub-word tokens in the different expressions. A similar argument applies to contextual embeddings, in that a target word may be represented by multiple sub-word tokens. For the additive composition operation, the scores against window size for each language and model are given in figs. 5 to 7.

Virtanen et al. (2019, p. 3) noted that, for a random sample of 1% of the relevant Wikipedia dataset, the number of sub-word tokens used to represent a word by a multilingual BERT model is greater for Finnish (1.97) than for English (1.16). This is attributed to the morphological complexity of Finnish and its comparatively small fraction of the multilingual model's vocabulary. My results are broadly consistent with those of Virtanen et al., i.e., the Finnish-specific models generally outperform the multilingual ones.

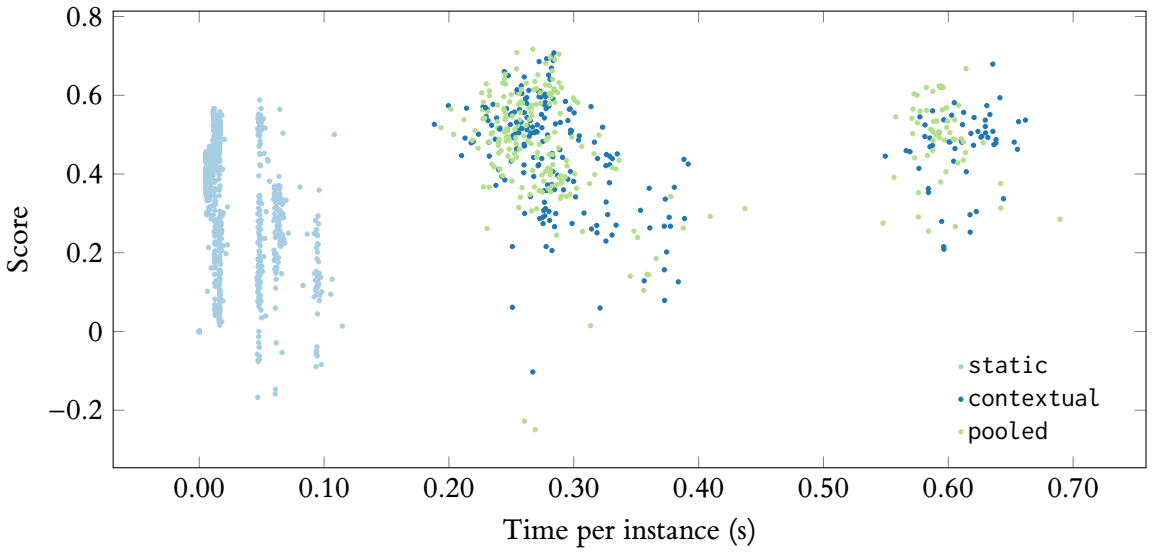

Figure 4: The score obtained against the time per instance of different models and languages with the additive composition operation. The `static` variants take less time per instance and show less variability. The `large` models among the `contextual` and `pooled` variants take the most time per instance.

6 Conclusion

In this paper, I have presented the results of a hypothetical submission to SemEval-2020 Task 3, "Graded Word Similarity in Context". The purpose of this investigation was to compare the performance of static and contextual embeddings and their composition within a fixed-size context window on the task of predicting the change in the human judgment of similarity of a pair of words in two different contexts. I found that contextual embeddings generally outperformed static embeddings, but at a significant computational cost, and that composition benefited both static and contextual embeddings of sub-word tokens, with highly language-specific dependence on the window size.

The results that I have given must be interpreted in context: the original submission authors did not have access to the evaluation dataset prior to submitting their results, only the 'practice kit' of very few instances, and were therefore unable to optimise a parameter such as the window size prior to submission. Additionally, the models that I used were not necessarily available to the authors. With these caveats, I achieved several notable results (fig. 3):

- The `pooled` variant of `EMBEDDIA/crosloengual-bert` with a window size of three would have placed second among the Croatian submissions, with a score of 0.717.
- The `contextual` variant of `TurkuNLP/bert-base-finnish-uncased-v1` with a window size of zero would have placed fourth among the Finnish submissions, with a score of 0.679.
- The `static` variant of `TurkuNLP/bert-base-finnish-uncased-v1` with a window size of zero outperform several of the Finnish submissions, including the baseline, with a score of 0.564.

Given the significant expense of applying contextual-embedding models, these results highlight the importance of analysing the complexity of the task at hand and considering the possibility that a simpler model produces adequate results.

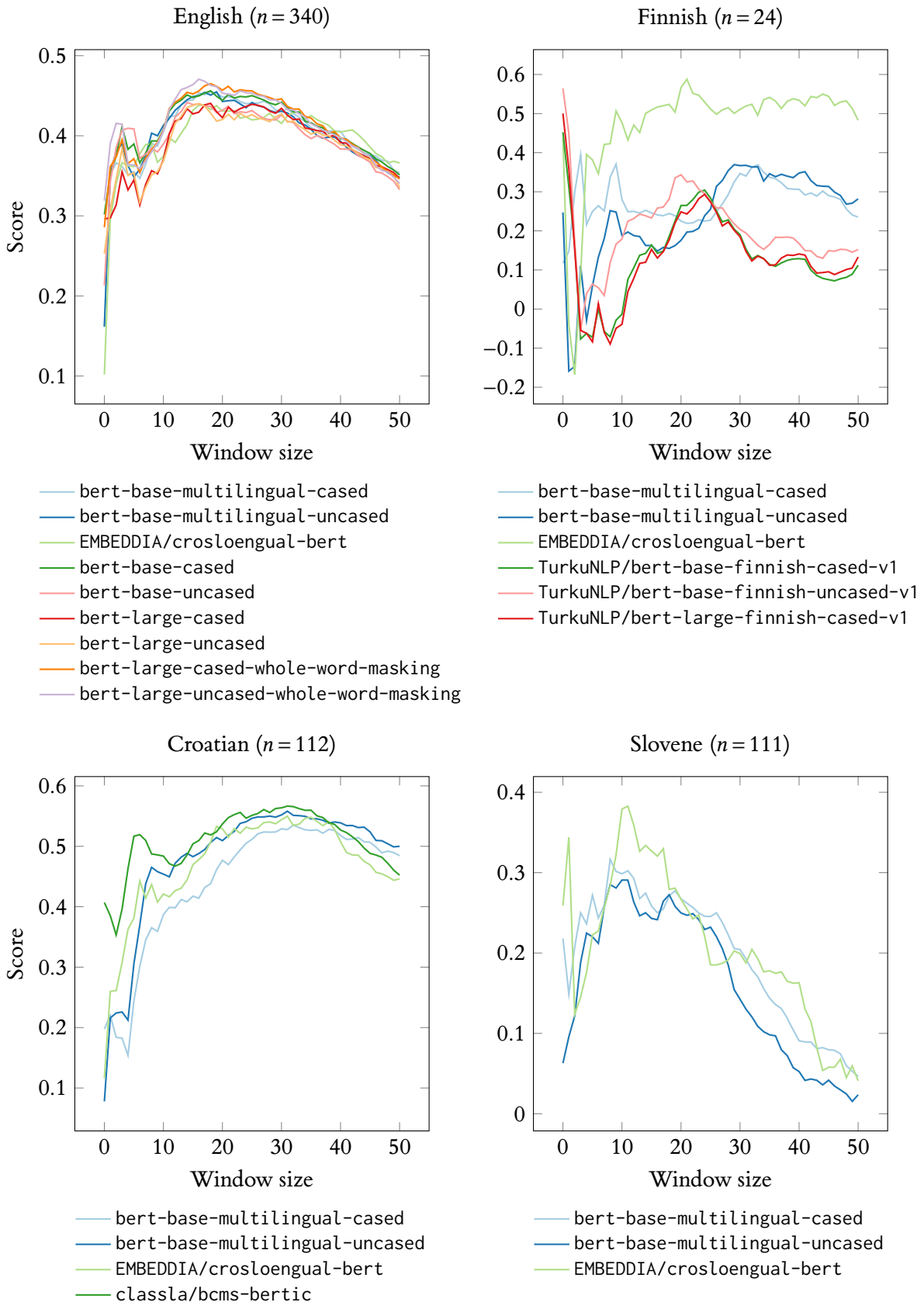

Figure 5: The score against window size for the `static` model variants with the additive composition operation.

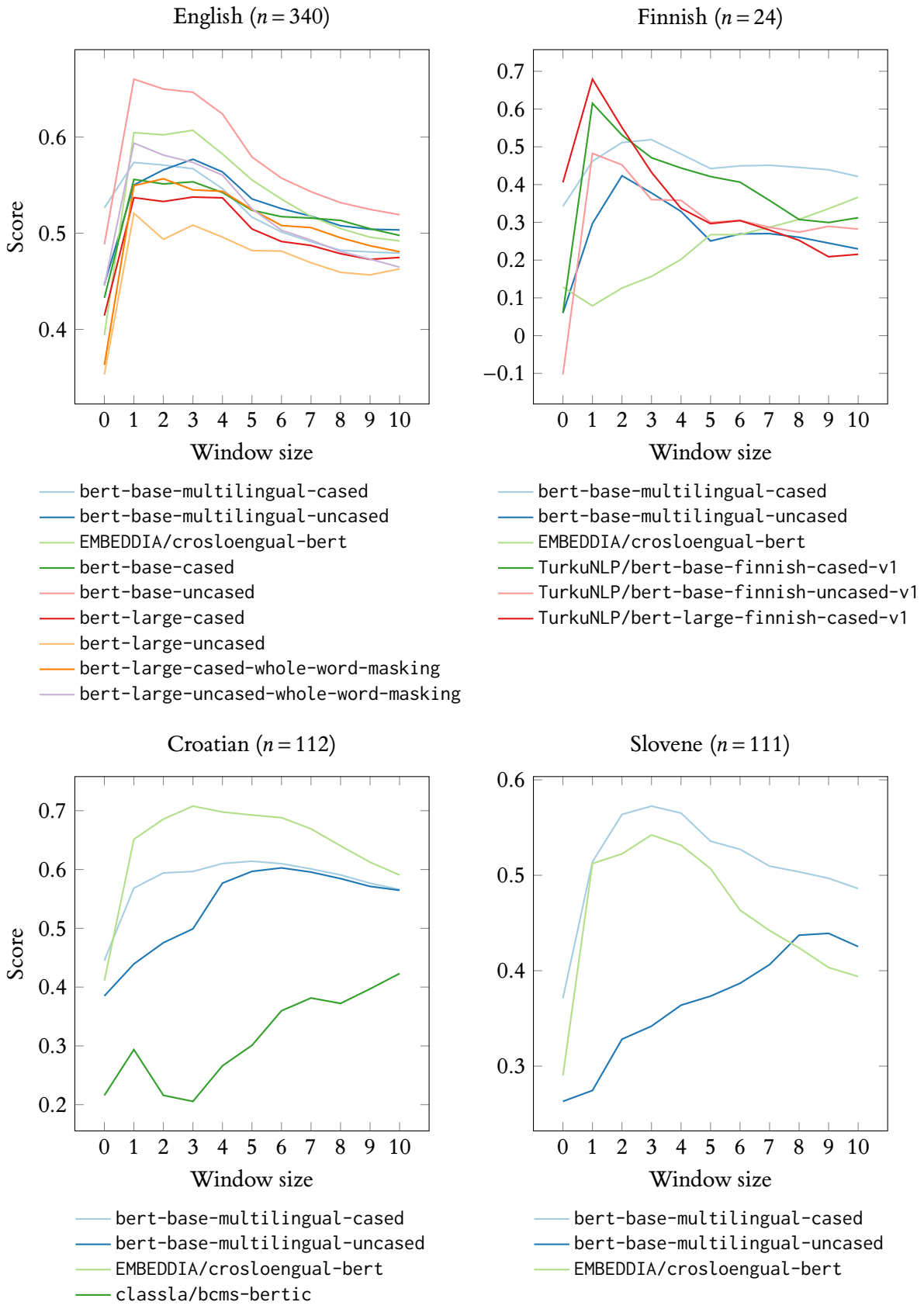

Figure 6: The score against window size for the contextual model variants with the additive composition operation.

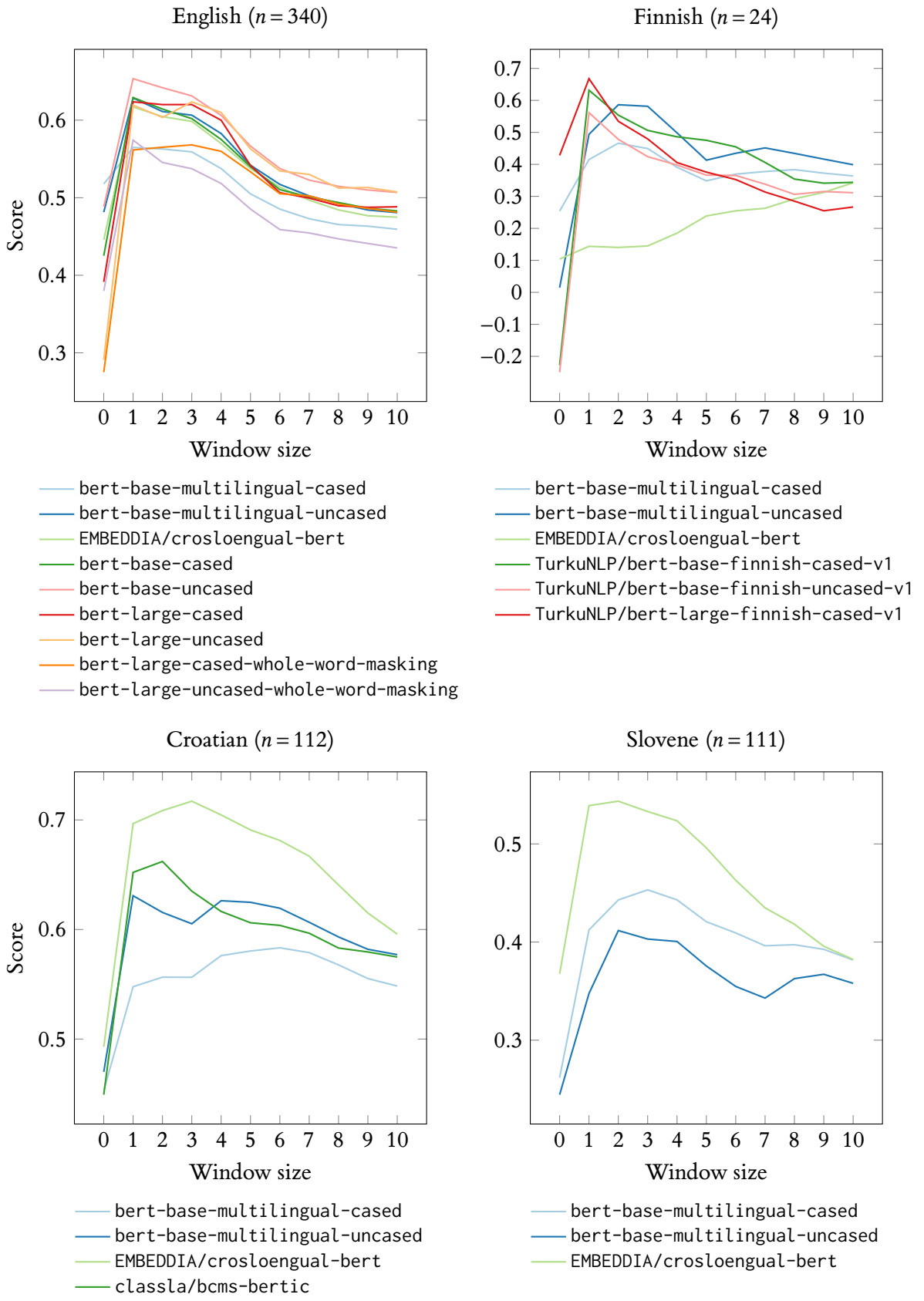

Figure 7: The score against window size for the pooled model variants with the additive composition operation.

References

Armendariz, Carlos Santos, Matthew Purver, Senja Pollak, et al. (2020). "SemEval-2020 Task 3: Graded Word Similarity in Context". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, pp. 36–49.

Armendariz, Carlos Santos, Matthew Purver, Matej Ulčar, et al. (2020). *CoSimLex: A Resource for Evaluating Graded Word Similarity in Context*.

Arora, Simran et al. (2020). "Contextual Embeddings: When Are They Worth It?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 2650–2663.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, MD: Association for Computational Linguistics, pp. 238–247.

Batchkarov, Miroslav et al. (2016). "A Critique of Word Similarity as a Method for Evaluating Distributional Semantic Models". In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 7–12.

Boleda, Gemma (2020). "Distributional Semantics and Linguistic Theory". In: *Annual Review of Linguistics* 6.1, pp. 213–234.

Bommasani, Rishi, Kelly Davis, and Claire Cardie (2020). "Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4758–4781.

Bommasani, Rishi, Drew A. Hudson, et al. (2022). *On the Opportunities and Risks of Foundation Models*. URL: http://arxiv.org/abs/2108.07258.

Brunner, Gino et al. (2020). *On Identifiability in Transformers*.

Devlin, Jacob et al. (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, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, MN: Association for Computational Linguistics, pp. 4171–4186.

Ethayarajh, Kawin (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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 55–65.

Firth, J. R. (1957). "A Synopsis of Linguistic Theory, 1930–1955". In: *Studies in Linguistic Analysis*. Oxford, UK: Basil Blackwell, pp. 1–32.

Frege, Gottlob (1980). *The Foundations of Arithmetic: A Logico-Mathematical Enquiry Into the Concept of Number*. 2nd rev. ed. Evanston, Ill: Northwestern University Press.

Gamallo, Pablo (2020). "CitiusNLP at SemEval-2020 Task 3: Comparing Two Approaches for Word Vector Contextualization". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by Aurelie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, pp. 275–280.

Gupta, Prakhar, Matteo Pagliardini, and Martin Jaggi (2019). *Better Word Embeddings by Disentangling Contextual N-Gram Information*.

Harris, Zellig S. (1954). "Distributional Structure". In: *WORD* 10.2-3, pp. 146–162.

Hettiarachchi, Hansi and Tharindu Ranasinghe (2021). *BRUMS at SemEval-2020 Task 3: Contextualised Embeddings for Predicting the (Graded) Effect of Context in Word Similarity*. URL: http://arxiv.org/abs/2010.06269.

Hewitt, John and Christopher D. Manning (2019). "A Structural Probe for Finding Syntax in Word Representations". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, MN: Association for Computational Linguistics, pp. 4129–4138.

Hill, Felix, Roi Reichart, and Anna Korhonen (2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation". In: *Computational Linguistics* 41.4, pp. 665–695.

Kintsch, Walter (2001). "Predication". In: *Cognitive Science* 25.2, pp. 173–202.

Lai, Gary (2023). *Imgarylai/Bert-Embedding*. URL: https://github.com/imgarylai/bert-embedding.

Landauer, Thomas K. and Susan T. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge". In: *Psychological Review* 104.2, pp. 211–240.

Lenci, Alessandro (2018). "Distributional Models of Word Meaning". In: *Annual Review of Linguistics* 4.1, pp. 151–171.

Liu, Qi, Matt J. Kusner, and Phil Blunsom (2020). *A Survey on Contextual Embeddings*. URL: http://arxiv.org/abs/2003.07278.

Liu, Yijia et al. (2019). "Deep Contextualized Word Embeddings for Universal Dependency Parsing". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 19.1, 9:1–9:17.

Ljubešić, Nikola and Davor Lauc (2021). "BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian". In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Ed. by Bogdan Babych et al. Kiyv, Ukraine: Association for Computational Linguistics, pp. 37–42.

Mikolov, Tomáš, Kai Chen, et al. (2013). *Efficient Estimation of Word Representations in Vector Space*.

Mikolov, Tomáš, Ilya Sutskever, et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. Burges et al. Vol. 26. Curran Associates, Inc.

Mitchell, Jeff and Mirella Lapata (2008). "Vector-Based Models of Semantic Composition". In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, pp. 236–244.

Padó, Sebastian and Mirella Lapata (2003). "Constructing Semantic Space Models from Parsed Corpora". In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, pp. 128–135.

Pessutto, Lucas R. C. et al. (2020). *BabelEnconding at SemEval-2020 Task 3: Contextual Similarity as a Combination of Multilingualism and Language Models*. URL: http://arxiv.org/abs/2008.08439.

Peters, Matthew E. et al. (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.

Reif, Emily et al. (2019). "Visualizing and Measuring the Geometry of BERT". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Srivastava, Aarohi et al. (2023). "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models". In: *Transactions on Machine Learning Research*.

Torregrossa, François et al. (2021). "A Survey on Training and Evaluation of Word Embeddings". In: *International Journal of Data Science and Analytics* 11.2, pp. 85–103.

Ulčar, Matej and Marko Robnik-Šikonja (2020a). "FinEst BERT and CroSloEngual BERT". In: *Text, Speech, and Dialogue*. Ed. by Petr Sojka et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 104–111.

— (2020b). "High Quality ELMo Embeddings for Seven Less-Resourced Languages". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 4731–4738.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.
Virtanen, Antti et al. (2019). *Multilingual Is Not Enough: BERT for Finnish*.
Wolf, Thomas et al. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*.