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

Tim Lawson

University of Bristol

tim.lawson@bristol.ac.uk

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" (1960, p. xvii). This 'context principle' is intuitive: words are frequently polysemous, or assume different connotations and emphasis within different expressions (Armendariz, Purver, Pollak, et al. 2020, pp. 2–3). Historically, however, context-dependence has been a problem for distributional meaning representations. Founded on the distributional hypothesis (e.g., Turney and Pantel 2010, pp. 142–143), 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* embeddings 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 (contextualization). 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 embeddings with the application of prior methods of contextualization to static embeddings.

SemEval-2020 Task 3, "Graded Word Similarity in Context" (Armendariz, Purver, Ulčar, 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. Specifically, I evaluated the results obtained by computing the cosine similarity between the different kinds of embeddings for a variety of pre-trained language models, and their composition with the embeddings within a fixed-size context window.[2]

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, Pollak, et al. 2020, pp. 39–42) but only a minimal 'practice kit' of fewer than ten instances was provided in

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

[2]The code that produced these results is available at tslwn/graded-similarity.

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). 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, Pollak, et al. 2020, p. 42). This metric is equivalent to the cosine similarity between the two vectors of results:

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

3 Related work

3.1 Composition and contextualization

Many context-based approaches to word-sense disambiguation have been proposed since the advent of count-based models of word meaning. With reference to Latent Semantic Analysis (Deerwester et al. 1990), for example, Landauer and Dumais argued that taking the average of the high-dimensional representation of a word and the representations of the words in its context may suffice to determine the word's contextual meaning (1997, pp. 229–230). Thus, the contextualization 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 procedure to contextualize 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 (Mikolov, Chen, et al. 2013, p. 9; Mikolov, Sutskever, et al. 2013, p. 7). Recent reviews of distributional semantic models are given by Lenci (2018) and Boleda (2020).

3.2 Costs and benefits of contextual models

Contextual language models have achieved widespread success on benchmark tasks (Bommasani, Hudson, et al. 2022, pp. 22–27). There is, however, cause to criticize the suitability of typical benchmarks for characterizing the capabilities of language models (Srivastava et al. 2023, pp. 5–6). Furthermore, the social and environmental costs of deploying a large model may not be justifiable; and the computational resources it requires may be prohibitive to an organization or in a resource-constrained environment (Bommasani, Hudson, et al. 2022, pp. 142–145, 154). In the case of systems based on contextual embeddings, for instance, Arora et al. (2020) have shown 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) have compared different kinds of embeddings for a variety of word-similarity tasks and found that static embeddings can be obtained from contextual models that outperform their contextual counterparts while reducing the cost of inference. Surveys of contextual and static embeddings are given by Liu et al. (2020) and Torregrossa et al. (2021); further analyses of contextual language models are provided by Reif et al. (2019) and Brunner et al. (2019), for example. I briefly discuss the costs and benefits of contextual models in section 5.2.

3.3 Word similarity

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 is ambiguous (Elekes et al. 2020) and encompasses a broad range of semantic relations (Padó and Lapata 2003, p. 2),

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.

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 in comparison to more specific tasks (ibid., pp. 8–9). In this case, Armendariz, Purver, Pollak, et al. (2020, p. 8) and Armendariz, Purver, Ulčar, et al. (2020, p. 42) reported similar inter-annotator correlations between the different languages and to those of the SimLex-999 dataset (Hill et al. 2015, pp. 678–680). In the present context, we are explicitly concerned with the ability of pre-trained embeddings to capture context-dependent similarity judgments. However, the interpretation of distributional semantic models as explanatory theories of human linguistic processing is subject to debate (Günther et al. 2019), and it may be that less data-intensive models are more appropriate for a specific task of this kind (De Deyne et al. 2016).

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. 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),[3] and the vast majority of the task submissions were based on Transformers (Armendariz, Purver, Ulčar, et al. 2020, pp. 36, 42–45). Therefore, I chose to evaluate a variety of pre-trained Transformers. Because both static and contextual embeddings can be obtained from a Transformer model, this approach facilitated a direct comparison between them. The models that I evaluated were accessed via the HuggingFace *Transformers* library (Wolf et al. 2020) and are listed in fig. 1.

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

…from the Isenj. He <u>accepts</u> their offer but knows that …

$$\downarrow \quad \text{1. Find first sub-word token}$$

$$\left[\ldots,\ \text{\#\#nj},\ .,\ \text{he},\ \underline{\text{accept}},\ \text{\#\#s},\ \text{their},\ \text{offer},\ \ldots\right]$$

$$\downarrow \quad \text{2. Find tokens in a fixed-size window}$$

$$\left[\,\text{he},\ \text{accept},\ \text{\#\#s}\,\right]$$

$$\downarrow \quad \text{3. Obtain embeddings}$$

$$\left[\,\overrightarrow{\text{he}},\ \overrightarrow{\text{accept}},\ \overrightarrow{\text{\#\#s}}\,\right]$$

$$\downarrow \quad \text{4. Compose embeddings}$$

$$\overrightarrow{\textit{accept}} = \overrightarrow{\text{he}} + \overrightarrow{\text{accept}} + \overrightarrow{\text{\#\#s}}$$

Figure 2: A schematic of the procedure used to obtain a contextualized 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.

The primary comparison that I made was between these models' static input and contextual output representations. Several of the task submissions used a combination of a Transformer's hidden-states (e.g. Gamallo 2020, p. 276; Costella Pessutto et al. 2020, p. 61; Hettiarachchi and Ranasinghe 2020, p. 145). This method 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 method – a thorough comparison of its variants is, however, beyond the scope of this paper. Hereafter, I refer to the three kinds 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.

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 contextualized representation of a target word by:

1. finding the index of the target word's first sub-word token within the tokens of its context;
2. finding the tokens within a fixed-size window around its first token;
3. obtaining the embeddings of the tokens in the window; and
4. composing the embeddings to produce a single representation.

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 similarity between the representations of a pair of target words may be different in each context, even if the representations are individual static embeddings. This is the cause of the non-zero scores obtained by models of this kind, particularly for the Finnish language (section 5.3).

Inspired by Landauer and Dumais (1997), Kintsch (2001), and Mitchell and Lapata (2008), I predominantly investigated element-wise addition[4] and multiplication as composition operations to contextualize embeddings. However, preliminary experiments indicated that multiplication performed poorly across all languages, models, and window sizes; hence, it was discarded before the final analysis. Initially, I also investigated 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 too close to the beginning or 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 5.2); 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 models (Devlin et al. 2019, p. 4174). These heuristics were largely vindicated by the results of the evaluation, which demonstrated that the scores decrease as the window size approaches the maximum.

5 Results

5.1 Hyperparameter search

In the following sections, I present the results achieved on the evaluation dataset. However, by comparing the scores achieved with different kinds of embeddings, window sizes, etc., I have effectively performed hyperparameter search on the evaluation dataset, which would not have been possible or legitimate as a task submission. Therefore, I also explicitly implemented a hyperparameter search procedure on the 'practice kit' to select a candidate model for each language and computed the scores of these models on the evaluation dataset.

1: Add best 'practice kit' candidates and results on evaluation data.

5.2 Cost-benefit analysis of contextual embeddings

In the main, greater scores were achieved with contextual and pooled embeddings than with static embeddings. However, static embeddings make up a small fraction of the size of a contextual language model, and it is much faster to compute a contextualized representation from static embeddings than to run inference on a language model. The second of these advantages is shown in fig. 4. For a naïve implementation of the procedure described in section 4, the approximate time taken to compute the change in similarity between two words in context, i.e., the total time divided by the number of instances, is notably greater for contextual embeddings than for static embeddings. The right-most cluster is due to the `large` model variants.

2: Plot the gold-standard values against the predicted values for the best model for each language to visualize the Pearson correlation coefficient.

[4]The cosine similarity between two vectors is invariant with respect to the multiplication of the vectors by scalars. Therefore, the results of composing the embeddings within a fixed-size context window by addition or the arithmetic mean are equal. Hence, I did not also investigate the arithmetic mean.

Language	Model name	Window size	Score

(a) static

Language	Model name	Window size	Score

(b) contextual

Language	Model name	Window size	Score

(c) pooled

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

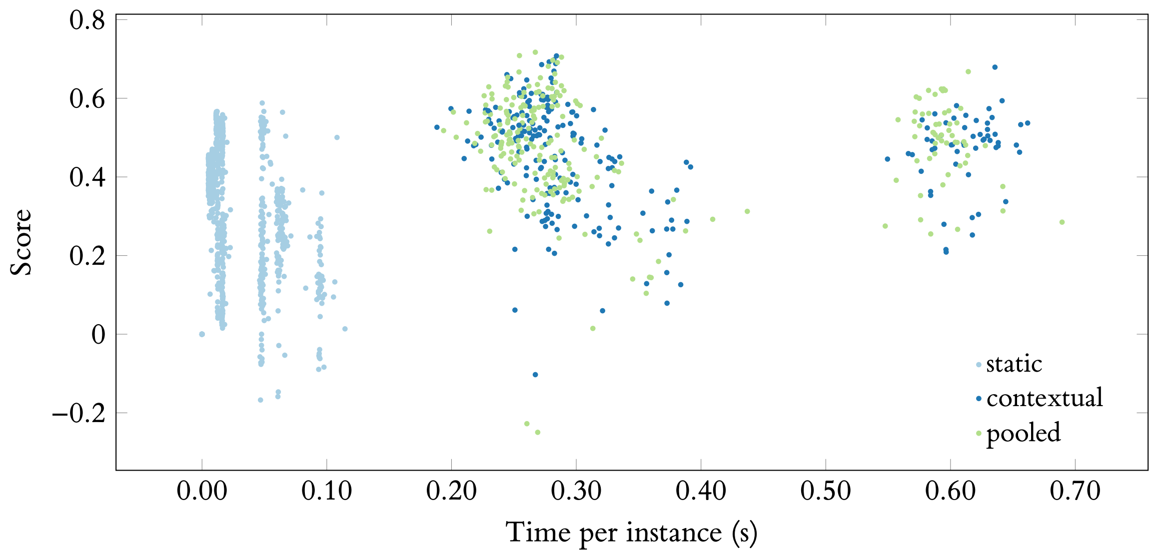

Figure 4: The score obtained against the approximate time per instance, i.e., the total time divided by the number of instances, of the different models and languages with the additive composition operation.

5.3 Language-specificity of window-size effects

Generally, I found that the scores obtained by all three types of embeddings were maximized 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. Accordingly, I found that Finnish-specific models generally outperformed multilingual ones and that the score varied more widely with window size for Finnish than the other languages.

5.4 Significance tests

Finally, to quantify the significance of the differences between the scores obtained with different kinds of embeddings, I computed paired t-tests of the scores of the best models in each class over the same ten random samples of 90% of the evaluation dataset. For each pair of model classes, the null hypothesis is that the two sets of scores have the same mean, i.e., there is no significant difference between the mean score obtained by the two classes.

3: Add significance test results.

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 optimize 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 embeddings 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 embeddings 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 embeddings 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.

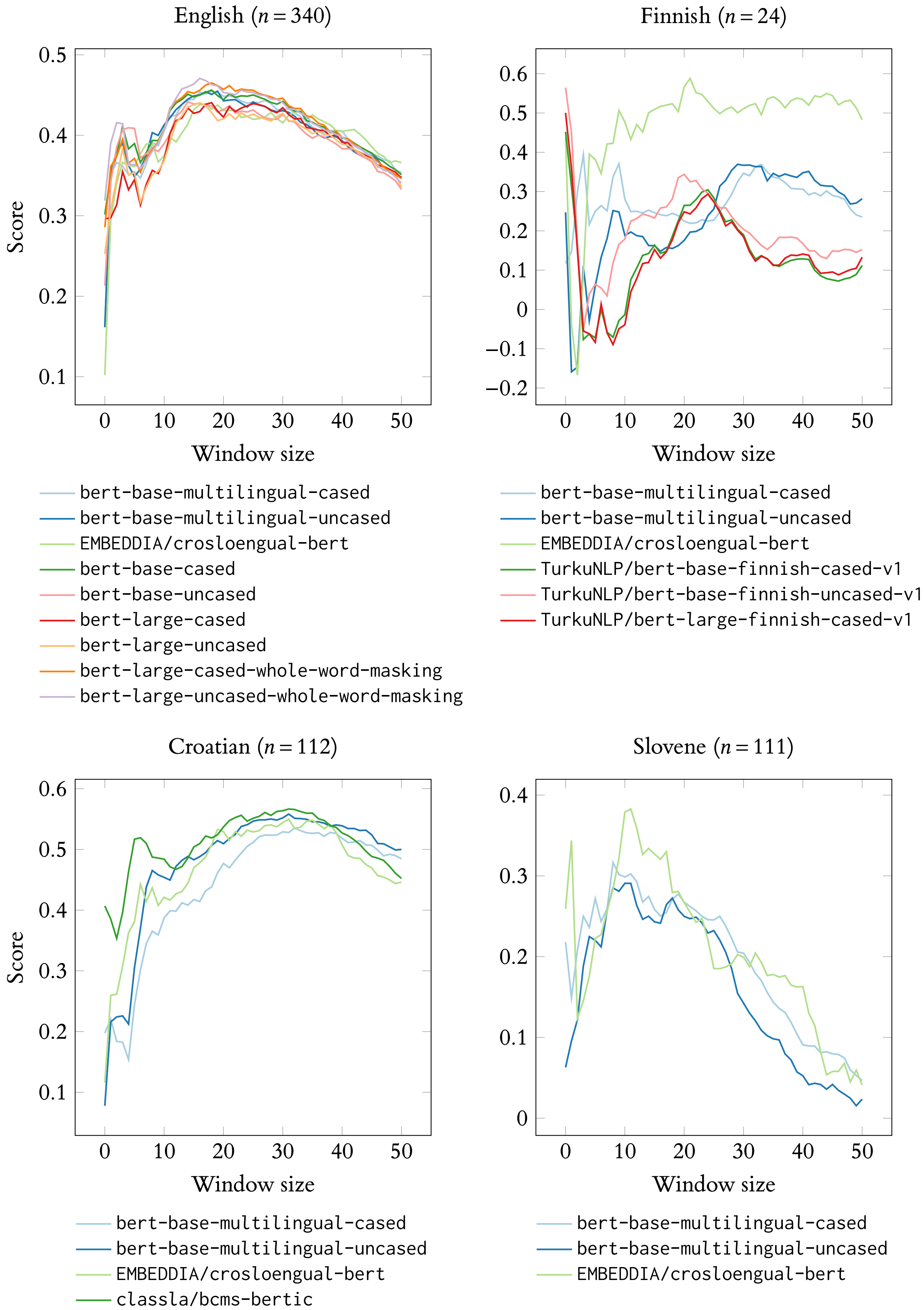

Figure 5: The score against window size for static embedding models and additive composition.

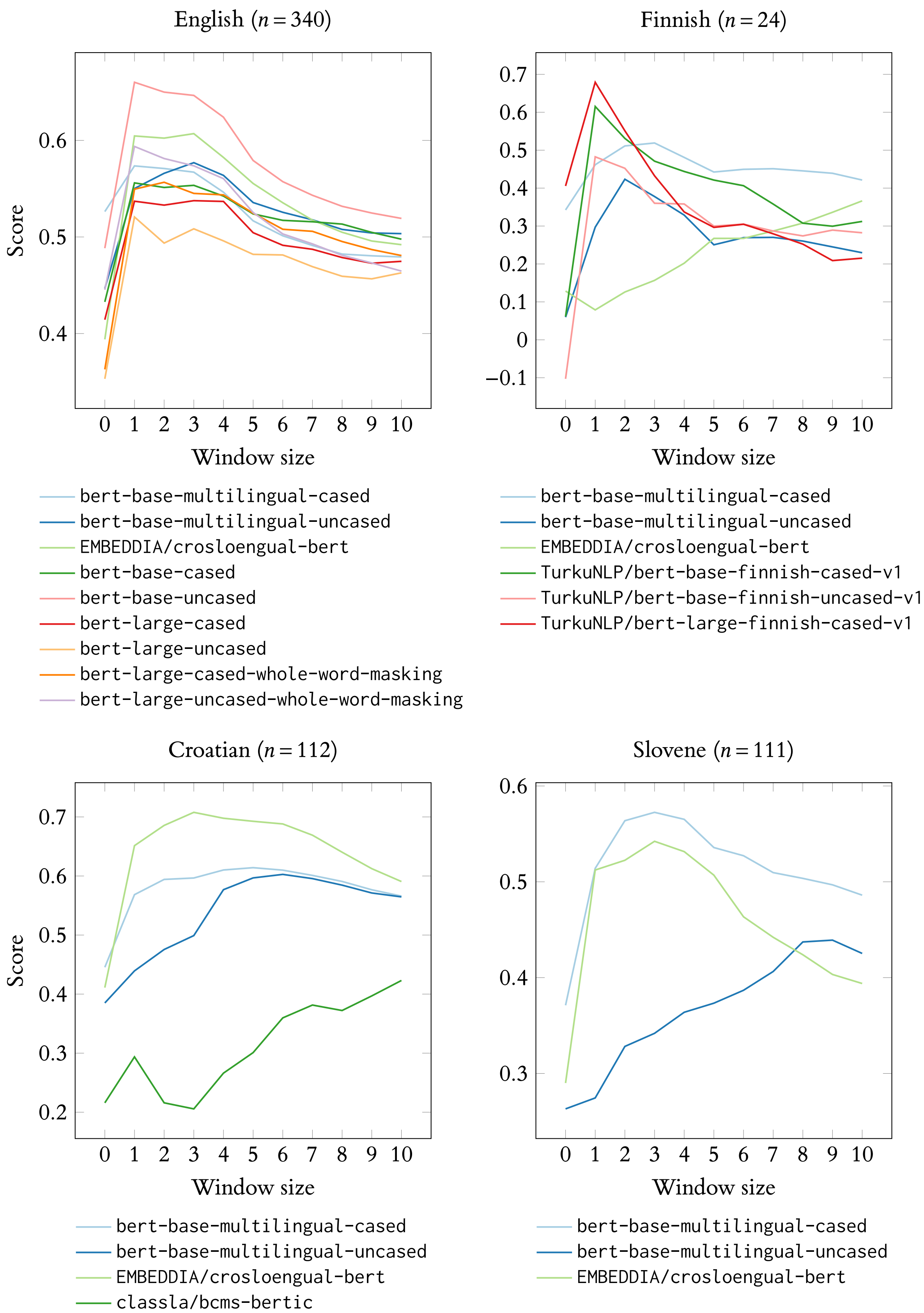

Figure 6: The score against window size for contextual embedding models and additive composition.

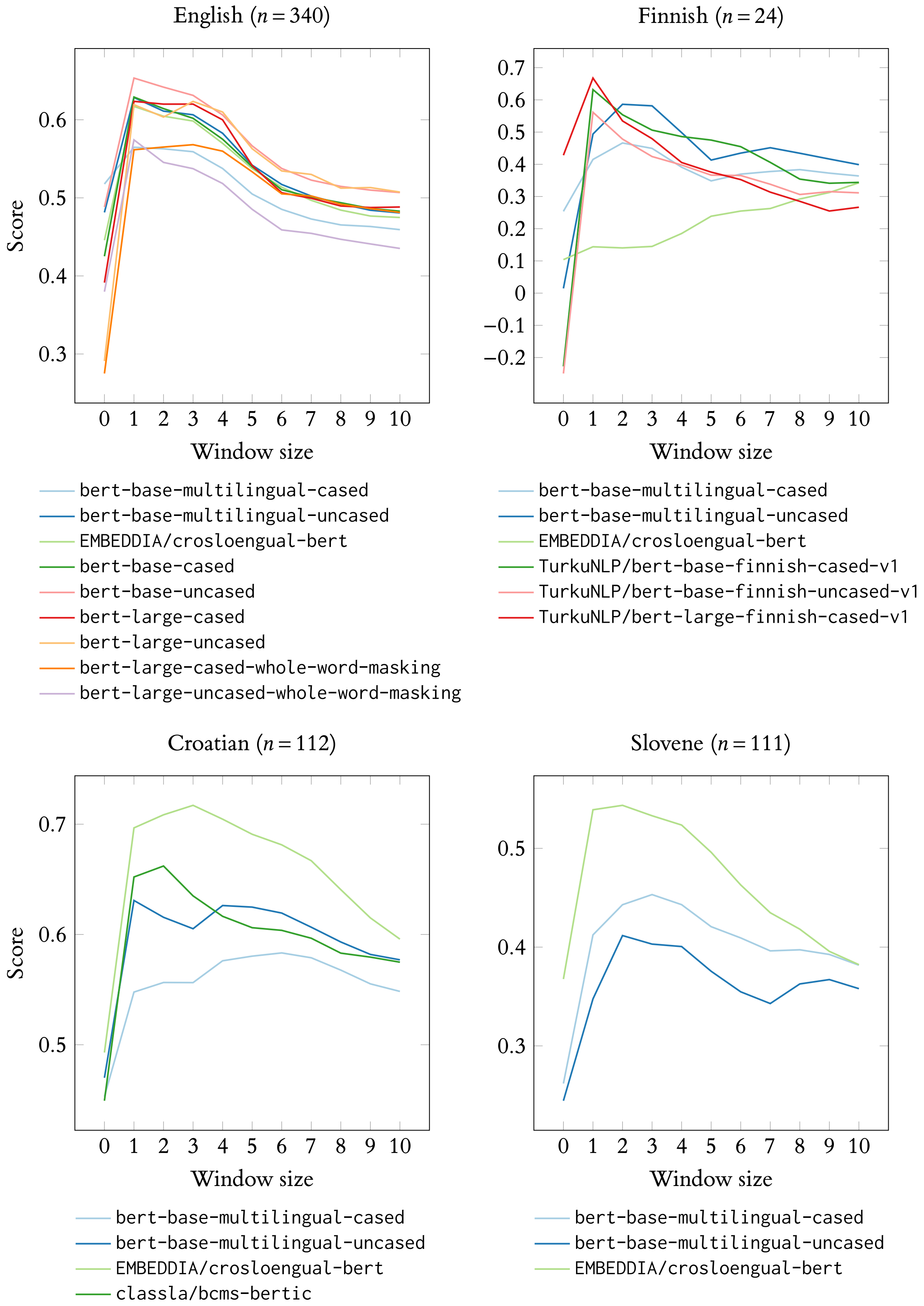

Figure 7: The score against window size for pooled embedding models and additive composition.

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". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 5878–5886.

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. (2019). "On Identifiability in Transformers". In: *The Eighth International Conference on Learning Representations*. Addis Ababa, Ethiopia.

Costella Pessutto, Lucas Rafael et al. (2020). "BabelEnconding at SemEval-2020 Task 3: Contextual Similarity as a Combination of Multilingualism and Language Models". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by Aurelie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, pp. 59–66.

De Deyne, Simon, Amy Perfors, and Daniel J Navarro (2016). "Predicting human similarity judgments with distributional models: The value of word associations." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Ed. by Yuji Matsumoto and Rashmi Prasad. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1861–1870.

Deerwester, Scott et al. (1990). "Indexing by Latent Semantic Analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

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.

Elekes, Ábel et al. (2020). "Toward meaningful notions of similarity in NLP embedding models". In: *International Journal on Digital Libraries* 21.2, pp. 109–128.

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.

Frege, Gottlob (1960). *The Foundations of Arithmetic: A Logico-Mathematical Enquiry Into the Concept of Number*. Trans. by J. L. Austin. Second Revised Edition. New York, NY: Harper & Brothers.

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.

Günther, Fritz, Luca Rinaldi, and Marco Marelli (2019). "Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions". In: *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 14.6, pp. 1006–1033.

Gupta, Prakhar, Matteo Pagliardini, and Martin Jaggi (2019). "Better Word Embeddings by Disentangling Contextual n-Gram Information". 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. 933–939.

Hettiarachchi, Hansi and Tharindu Ranasinghe (2020). "BRUMS at SemEval-2020 Task 3: Contextualised Embeddings for Predicting the (Graded) Effect of Context in Word Similarity". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by Aurelie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, pp. 142–149.

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`.

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. Kyiv, Ukraine: Association for Computational Linguistics, pp. 37–42.

Mikolov, Tomáš, Kai Chen, et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. Scottsdale, AZ.

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

Mitchell, Jeff and Mirella Lapata (2008). "Vector-based Models of Semantic Composition". In: *Proceedings of ACL-08: HLT*. Ed. by Johanna D. Moore et al. Columbus, OH: 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.

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, LA: 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 32 (NeurIPS 2019)*. Ed. by H. Wallach et al. Vol. 32. Curran Associates.

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.

Turney, Peter D. and Patrick Pantel (2010). "From Frequency to Meaning: Vector Space Models of Semantics". In: *Journal of Artificial Intelligence Research* 37.1, pp. 141–188.

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. Brno, Czech Republic: Springer Nature Switzerland, 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 30 (NIPS 2017)*. Ed. by I. Guyon et al. Vol. 30. Curran Associates.

Virtanen, Antti et al. (2019). *Multilingual is not enough: BERT for Finnish*. URL: `http://arxiv.org/abs/1912.07076`.

Wolf, Thomas et al. (2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, pp. 38–45.