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 et al. 2020b, 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 et al. 2020a), presents an opportunity to make such a comparison. Briefly, the task is to predict the human judgment of similarity of a 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 et al. 2020b, pp. 39–42) but only a minimal 'practice kit' of fewer than ten instances was provided in advance.

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

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

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 et al. 2020b, 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 composition and contextualization 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 et al. 2013b, p. 9; Mikolov et al. 2013a, p. 7), though its generality as an evaluation methodology has been questioned (Lenci et al. 2022, p. 1300). The relations between distributional semantics and compositionality have been surveyed by Erk (2012), Clark (2015), and Boleda and Herbelot (2016).

3.2 Costs and benefits of contextual models

Contextual language models have achieved widespread success on benchmark tasks (Bommasani 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 necessary computational resources may be prohibitive to a smaller organization or in a resource-constrained environment (Bommasani et al. 2022, pp. 142–145, 154). For instance, Lenci et al. (2022) found that static embeddings generally outperform BERT (Devlin et al. 2019) on word-similarity and -association tasks, provided optimal hyperparameters. Relatedly, Arora et al. (2020) have shown that static and even *random* embeddings can perform comparably to contextual embeddings, which Gupta et al. (2019, pp. 5244–5246) and Bommasani et al. (2020, pp. 4760–4762) have demonstrated for word-similarity tasks. Surveys of contextual and static embeddings are given by Liu et al. (2020) and Torregrossa et al. (2021), cross-lingual embeddings by Ruder et al. (2019), and further analyses of contextual language models by Reif et al. (2019) and Brunner et al. (2019), for example. The costs and benefits of contextual models for the task at hand are discussed 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]			✓	

Table 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 (Batchkarov et al. 2016, pp. 8–9). In this case, Armendariz et al. (2020b, p. 8) and Armendariz et al. (2020a, 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; Westera and Boleda 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 et al. 2020a, pp. 36, 42–45), so I chose to evaluate a variety of pre-trained Transformer models. 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 table 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 …

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

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

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

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

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

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

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

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

Figure 1: 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 the static input and contextual output representations of these models. 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 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 pooling hidden-states. However, a thorough comparison of its variants is beyond the scope of this paper. Hereafter, I refer to the 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. 1. 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 primarily investigated element-wise addition and multiplication as composition operations to contextualize embeddings. Preliminary experiments indicated that multiplication performed poorly across all languages, models, and window sizes, so it was discarded before the final analysis on the evaluation dataset. Initially, I also investigated the concatenation ('stacking') of embeddings. In the case that the number of embeddings was fewer than that expected from the window size, i.e., the target word was too close to the beginning or end of its context, I right-padded the concatenated embeddings with zeros to obtain representations of equal length. This approach was also generally inferior to addition, as discussed in section 5.1. The cosine similarity between two vectors is invariant with respect to the multiplication of the vectors by scalars, so 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.

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 of the evaluation dataset, i.e., segmenting on whitespace, gave 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 economical, due to their greater computational expense (section 5.2). However, as the window size approaches the length of the sequence, I expected 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 on the evaluation dataset, which showed that the scores decrease as the window size approaches the maximum.

5 Results

5.1 Hyperparameter search

In sections 5.2 and 5.3, I present the results for different models on the *evaluation* dataset. However, it would not have been possible to optimize hyperparameters on the evaluation dataset. Therefore, I also optimized them on the 'practice kit' dataset (section 2) to select a candidate model for each language and kind of embedding. This data was not provided for Finnish, so it was excluded from the analysis. In comparison to addition, the scores for the other composition operations varied more widely with respect to the window size (figs. 2 to 4). Hence, I excluded them before selecting candidate models. The models and their scores on the two datasets are listed in table 2. As expected, the scores on the 'practice kit' dataset of fewer instances are generally higher than those on the evaluation dataset. In some cases, the scores on the evaluation dataset are close to the maxima in table 3. Generally, the benefit of additional 'training' data is evident.

5.2 Cost-benefit analysis of contextual embeddings

In the main, greater scores were achieved with contextual and pooled embeddings than with static embeddings (tables 2 and 3). However, static embeddings make up a small fraction of the size of a contextual language model. For example, BERT's vocabulary size is approximately 30000, the dimensions of the `bert-base` and `bert-large` variants' hidden-states are 768 and 1024, and their total parameters are 110M and 340M respectively (Devlin et al. 2019, pp. 4173–4174). Static embeddings thus make up approximately 21% and 9% of the total parameters. It is also much faster to compute a contextualized representation from static embeddings than to run inference on a language model. 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 is shown in fig. 6. It is notably greater for contextual embeddings. The right-most cluster is due to the `large` model variants.

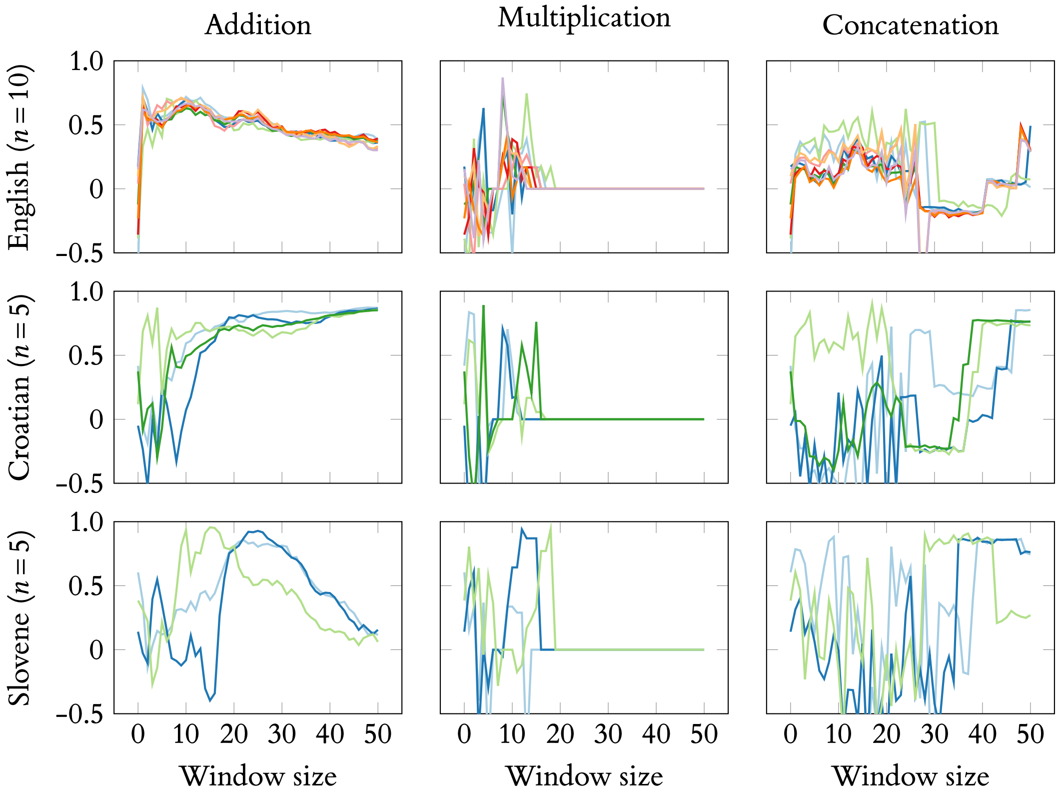

Figure 2: The score on the 'practice kit' dataset against window size for *static* embedding models. The model-name legends are omitted for brevity but match fig. 7.

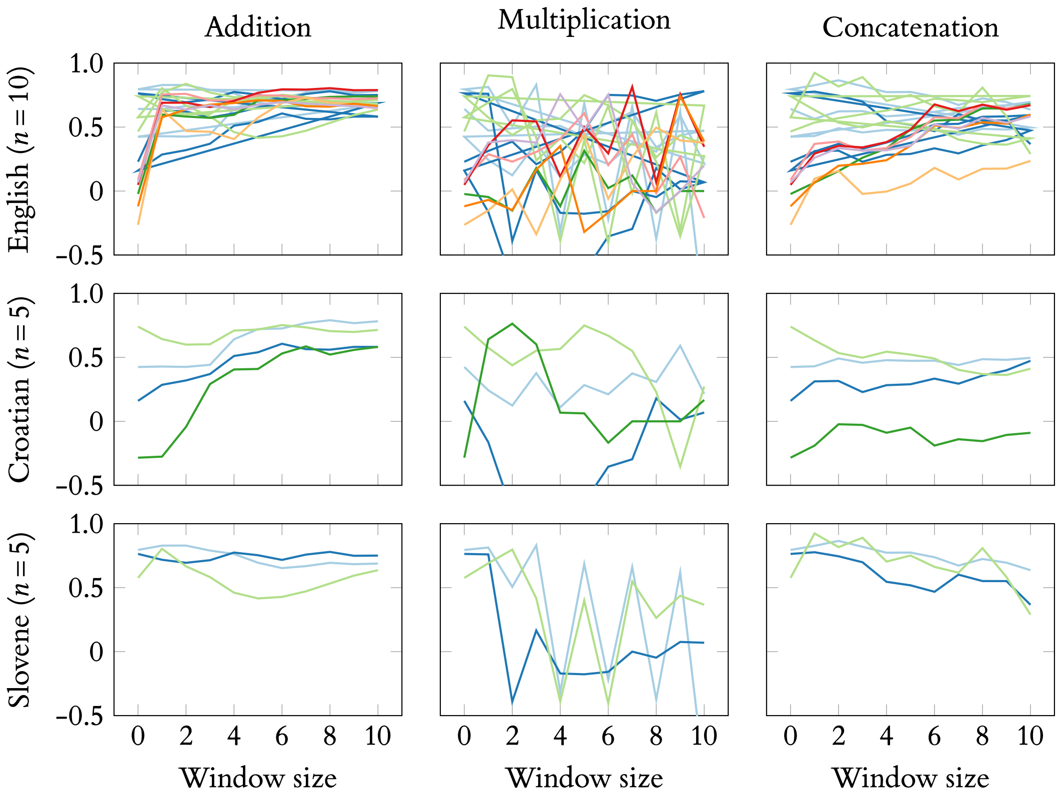

Figure 3: The score on the 'practice kit' dataset against window size for *contextual* embedding models. The model-name legends are omitted for brevity but match fig. 8.

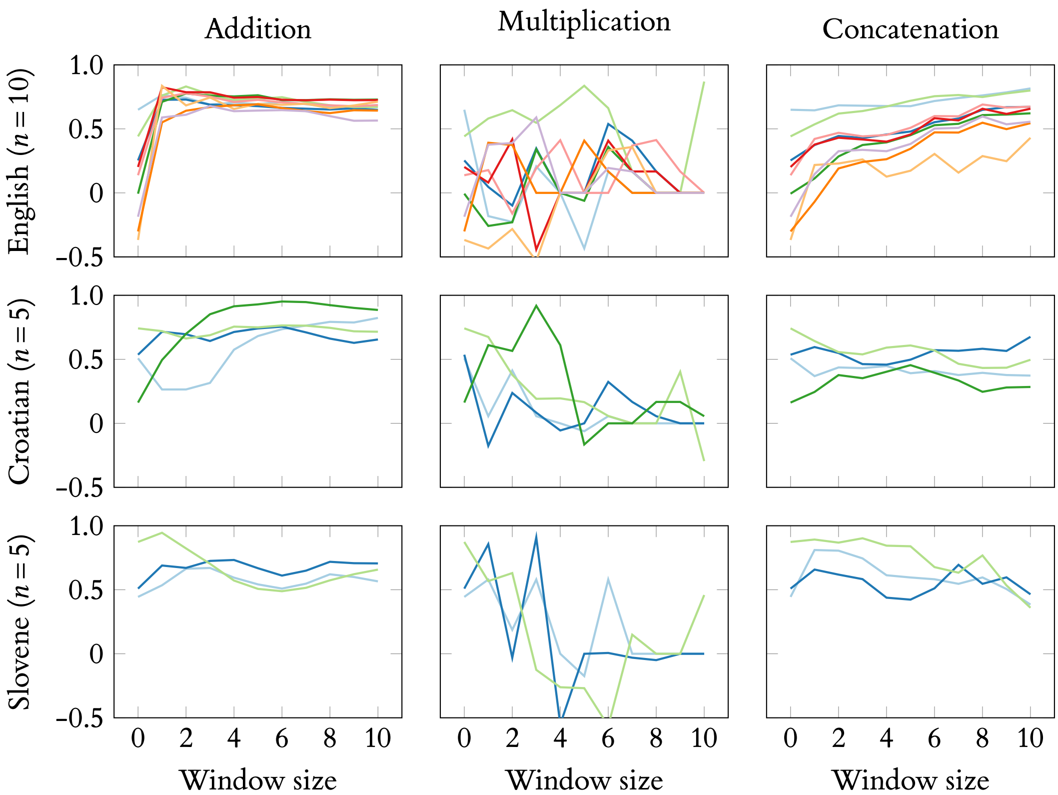

Figure 4: The score on the 'practice kit' dataset against window size for *pooled* embedding models. The model-name legends are omitted for brevity but match fig. 9.

Language	Model name	Window size	Practice	Evaluation
`en`	`bert-base-multilingual-cased`	1	0.785	0.352
`hr`	`bert-base-multilingual-cased`	48	0.873	0.492
`sl`	`EMBEDDIA/crosloengual-bert`	15	0.956	0.327

(a) Static

Language	Model name	Window size	Practice	Evaluation
`en`	`EMBEDDIA/crosloengual-bert`	2	0.838	0.602
`hr`	`bert-base-multilingual-cased`	8	0.790	0.591
`sl`	`bert-base-multilingual-cased`	2	0.829	0.564

(b) Contextual

Language	Model name	Window size	Practice	Evaluation
`en`	`bert-large-uncased`	1	0.836	0.619
`hr`	`classla/bcms-bertic`	6	0.952	0.604
`sl`	`EMBEDDIA/crosloengual-bert`	1	0.945	0.539

(c) Pooled

Table 2: The best scores on the 'practice kit' dataset for each kind of embedding, and the corresponding scores on the evaluation dataset. The results were limited to the composition operation of addition due to the variability of the scores with multiplication and concatenation (figs. 2 to 4).

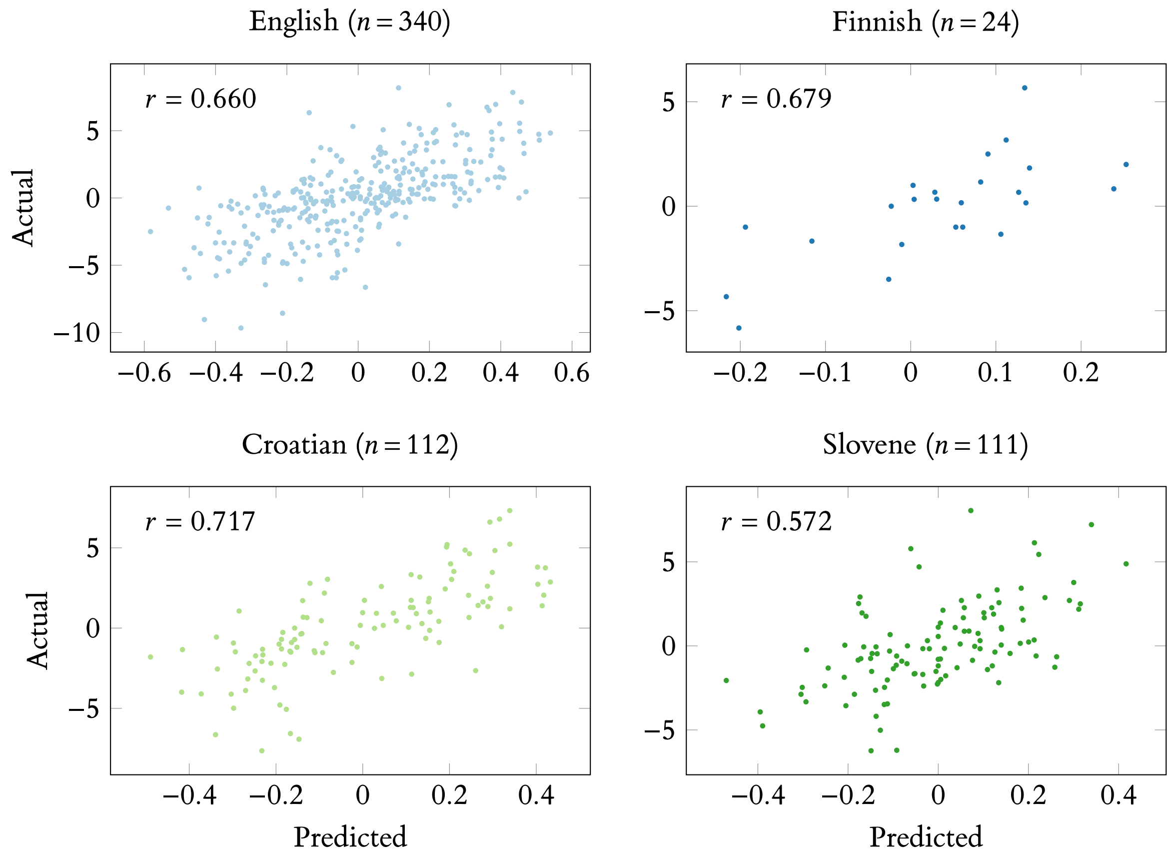

Figure 5: The predicted and actual human judgments of the change in similarity of the best models for each language on the evaluation dataset. The best models are highlighted in table 3. The zero-mean Pearson correlation coefficient (section 2), i.e., the score, is given in the top-left corner of each plot.

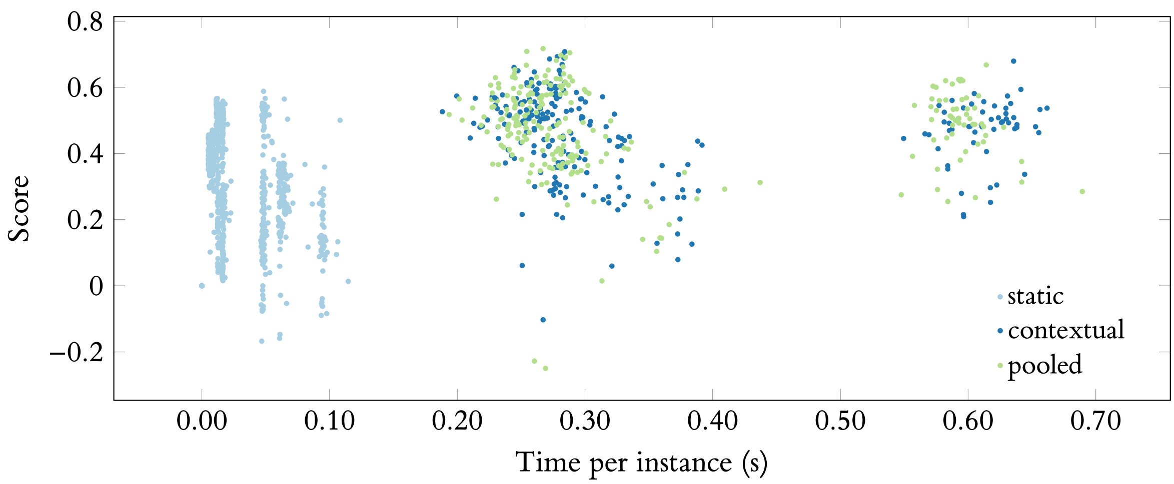

Figure 6: The scores on the evaluation dataset against the approximate time per instance, i.e., the total time divided by the number of instances, with the composition operation of addition.

Language	Model name	Window size	Score
en	`bert-large-uncased-whole-word-masking`	16	0.471
fi	`EMBEDDIA/crosloengual-bert`	21	0.588
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

Table 3: The best scores on the evaluation dataset for each kind of embedding – in all cases, it was obtained with the composition operation of addition. The best overall score for each language is underlined. The predicted and actual human judgments of the change in similarity for the models with the best overall scores are shown in fig. 5.

I quantified the significance of the differences between the scores obtained with different kinds of embeddings by paired t-tests and the Nemenyi test (Demšar 2006) on the scores of the best models over ten random samples of 90% of the evaluation dataset (table 3). For each pair of models, the null hypothesis was that the differences between the mean scores were due to chance. For each language, either contextual or pooled embeddings significantly outperformed static embeddings, but did not differ significantly from each other (table 4). Hence, the results support the conclusion that contextual embeddings are more effective than static embeddings for this task.

Model 1		Model 2				
Embeddings	Score	Embeddings	Score	t-statistic	p-value	$p < 0.05$
Contextual	0.665	Static	0.477	58.6	$1.00 \cdot 10^{-3}$	✓
Contextual	0.665	Pooled	0.658	4.7	$6.53 \cdot 10^{-2}$	
Pooled	0.658	Static	0.477	58.7	$6.53 \cdot 10^{-2}$	

(a) English ($n = 340$)

Model 1		Model 2				
Embeddings	Score	Embeddings	Score	t-statistic	p-value	$p < 0.05$
Contextual	0.679	Static	0.606	4.3	$2.30 \cdot 10^{-3}$	✓
Contextual	0.679	Pooled	0.668	2.4	0.37	
Pooled	0.668	Static	0.606	3.7	0.11	

(b) Finnish ($n = 24$)

Model 1		Model 2				
Embeddings	Score	Embeddings	Score	t-statistic	p-value	$p < 0.05$
Contextual	0.707	Static	0.576	22.3	$6.53 \cdot 10^{-2}$	
Contextual	0.707	Pooled	0.715	−4.7	$6.53 \cdot 10^{-2}$	
Pooled	0.715	Static	0.576	22.7	$1.00 \cdot 10^{-3}$	✓

(c) Croatian ($n = 112$)

Model 1		Model 2				
Embeddings	Score	Embeddings	Score	t-statistic	p-value	$p < 0.05$
Contextual	0.589	Static	0.391	16.8	$1.00 \cdot 10^{-3}$	✓
Contextual	0.589	Pooled	0.547	5.5	$6.53 \cdot 10^{-2}$	
Pooled	0.547	Static	0.391	15	$6.53 \cdot 10^{-2}$	

(d) Slovene ($n = 111$)

Table 4: The t-statistics from paired t-tests, and p-values from the Nemenyi test, on the scores obtained by the best models for each language and kind of embedding over ten random samples of 90% of the evaluation dataset. The best models are highlighted in table 3. A positive t-statistic indicates that the mean score of 'Model 1' is greater than that of 'Model 2'.

5.3 Language-specificity of window-size dependence

Generally, I found that the scores obtained by all three types of embeddings were maximized by a non-zero context-window size (tables 2 and 3). 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 contexts if the word is represented by different sub-word tokens in the different contexts. A similar argument applies to contextual embeddings, in that a target word may be represented by multiple sub-word tokens that differ between contexts. For the composition operation of addition, the scores against window size for each language, kind of embedding, and model are shown in figs. 7 to 9.

Virtanen et al. (2019, p. 3) have noted that, for a random sample of 1% of the relevant Wikipedia dataset, the number of sub-word tokens that represent a word is greater for Finnish (1.97) than for English (1.16) with the multilingual BERT model. This is attributed to the morphological complexity of Finnish and its comparatively small fraction of the model's vocabulary. Accordingly, I found that Finnish-specific models generally outperformed multilingual ones and that the scores varied more widely with window size for Finnish than the other languages.

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 significantly outperformed static embeddings but at a computational cost (section 5.2). Composition benefited both static and contextual embeddings of sub-word tokens, with a language-dependent optimal window size (section 5.3).

These results 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, so had limited opportunity to optimize hyperparameters like the window size (section 5.1). The models that I used were also not necessarily available to the authors (section 4.1). With these caveats, I achieved several notable results (table 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 one would have placed fourth among the Finnish submissions, with a score of 0.679.

References

Armendariz, Carlos Santos et al. (2020a). "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.

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

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.

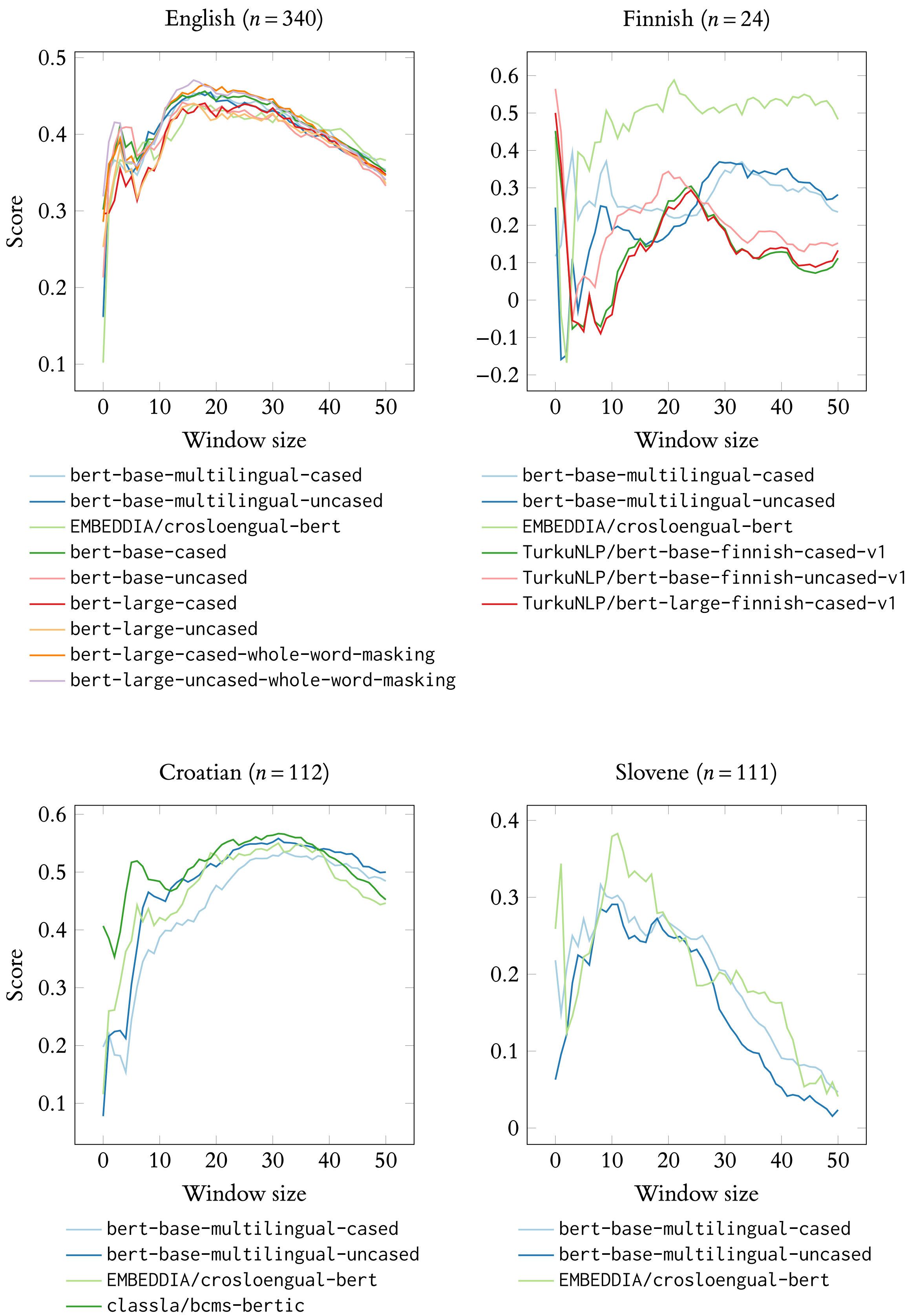

Figure 7: The scores on the evaluation dataset against window size for *static* embedding models with the composition operation of addition.

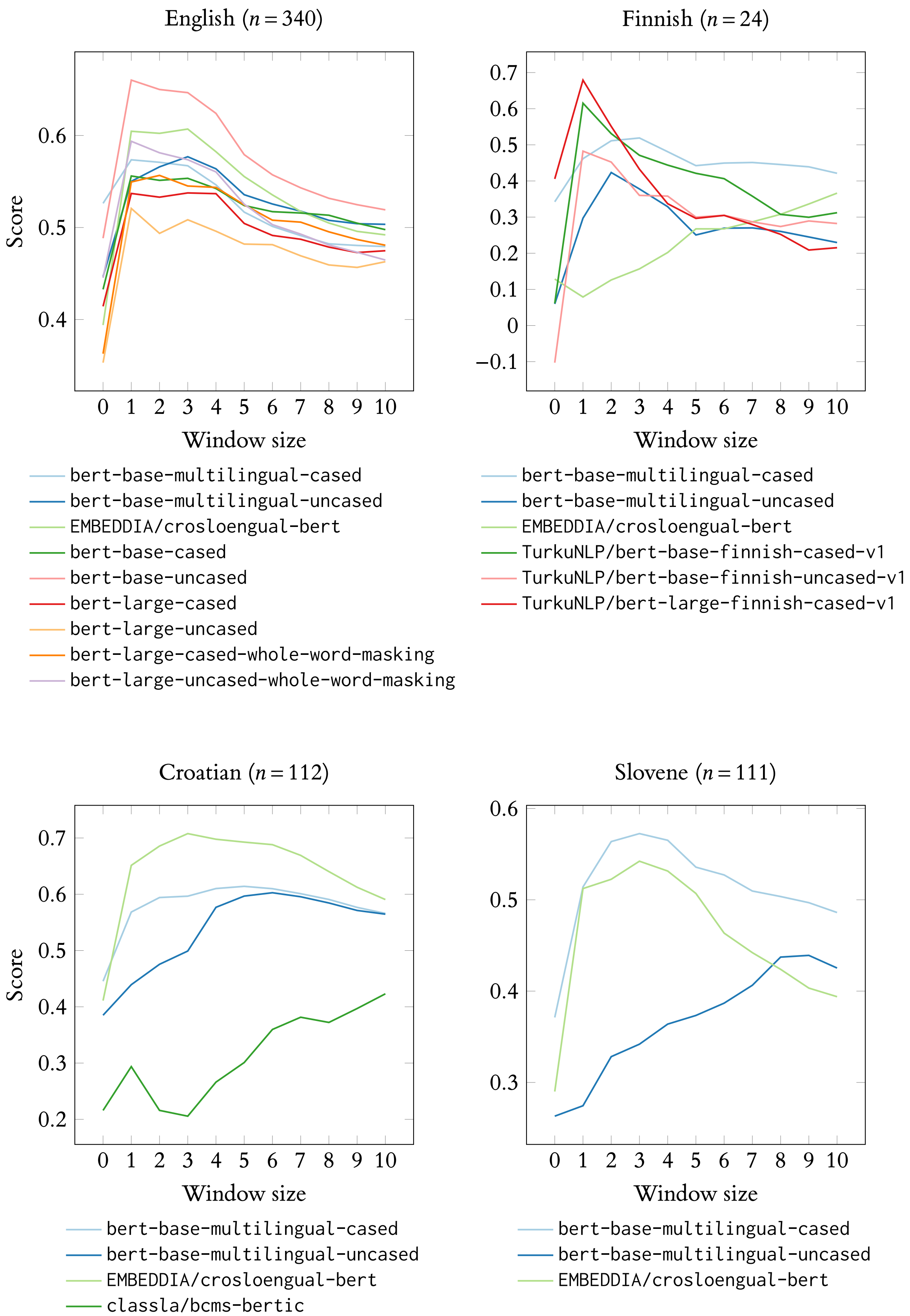

Figure 8: The scores on the evaluation dataset against window size for *contextual* embedding models with the composition operation of addition.

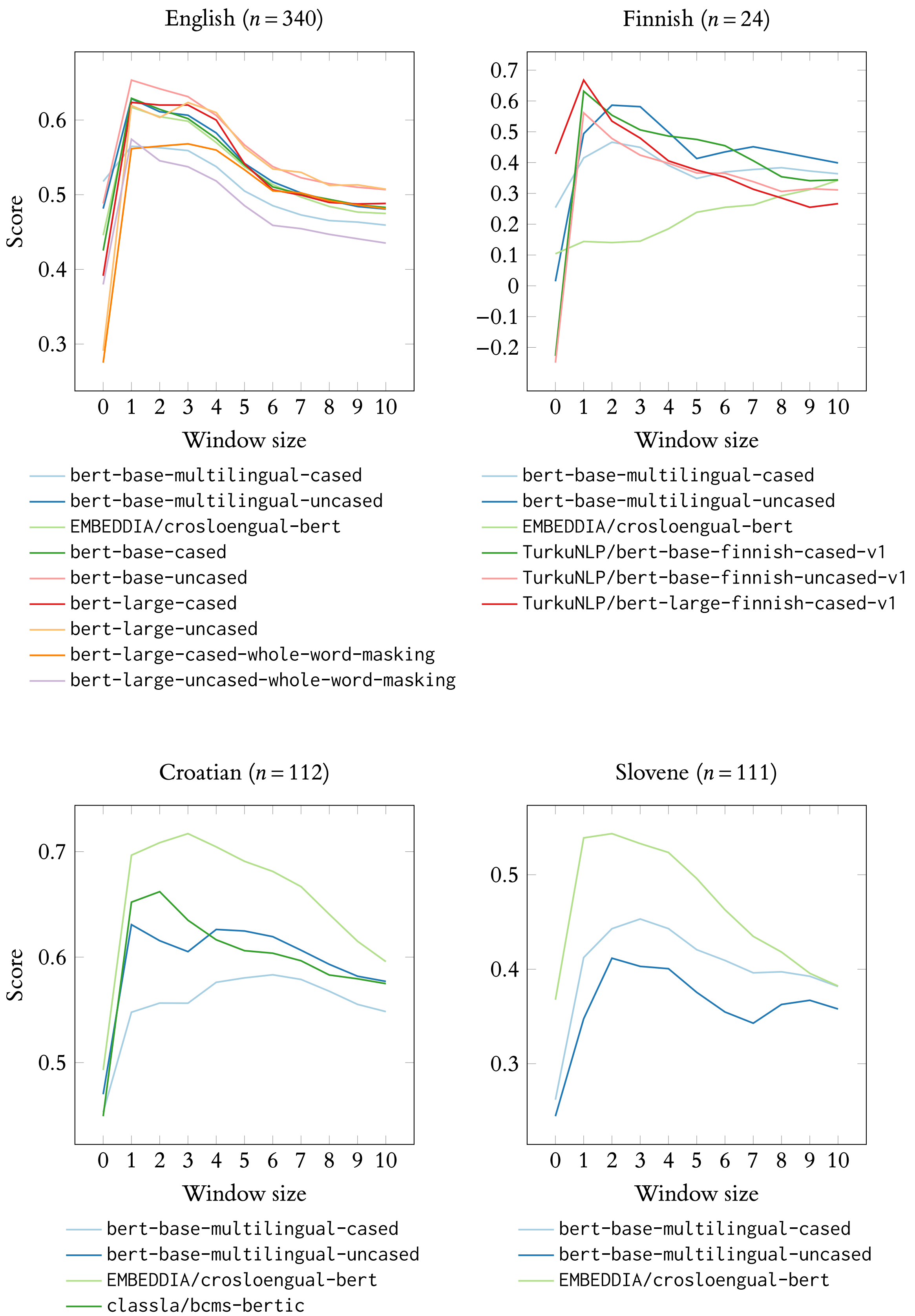

Figure 9: The scores on the evaluation dataset against window size for *pooled* embedding models with the composition operation of addition.

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.

Boleda, Gemma and Aurélie Herbelot (2016). "Formal Distributional Semantics: Introduction to the Special Issue". In: *Computational Linguistics* 42.4, pp. 619–635.

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 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. URL: `https://openreview.net/pdf?id=BJg1f6EFDB`.

Clark, Stephen (2015). "Vector Space Models of Lexical Meaning". In: *The Handbook of Contemporary Semantic Theory*. John Wiley & Sons. Chap. 16, pp. 493–522.

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 Aurélie 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.

Demšar, Janez (2006). "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *The Journal of Machine Learning Research* 7, pp. 1–30.

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.

Erk, Katrin (2012). "Vector Space Models of Word Meaning and Phrase Meaning: A Survey". In: *Language and Linguistics Compass* 6.10, pp. 635–653.

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 Aurélie 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 Aurélie 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 et al. (2022). "A comparative evaluation and analysis of three generations of Distributional Semantic Models". In: *Language Resources and Evaluation* 56.4, pp. 1269–1313.

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 et al. (2013a). "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.

Mikolov, Tomáš et al. (2013b). "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.

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.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2019). "A Survey of Cross-lingual Word Embedding Models". In: *Journal of Artificial Intelligence Research* 65, pp. 569–631.

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`.

Westera, Matthijs and Gemma Boleda (2019). "Don't Blame Distributional Semantics if it can't do Entailment". In: *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*. Ed. by Simon Dobnik, Stergios Chatzikyriakidis, and Vera Demberg. Gothenburg, Sweden: Association for Computational Linguistics, pp. 120–133.

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.