Graded Similarity in Context

Tim Lawson

December 13, 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. Historically, however, contextuality 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 is an obstacle to its use in modelling context-dependent phenomena.

Prior to the widespread availability of pre-trained word embeddings (e.g., Mikolov, Chen, et al. 2013; Pennington et al. 2014), 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. 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 continuously-valued human judgment of similarity of the same pair of words in two different contexts (Section 2). I elected to focus on the first subtask, 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 static and contextual representations, and the additive composition of these representations within a fixed-size context window.

2 Task definition

The CoSimLex dataset (Armendariz, Purver, Ulčar, et al. 2020), which served to evaluate the task submissions, extends the SimLex-999 dataset (Hill et al. 2015) to include multiple contexts for each pair of words.

3 Related work

Additive composition (e.g. Kintsch 2001; Mitchell and Lapata 2008; Mikolov, Sutskever, et al. 2013). Analysis of the effect of window size on static embeddings. Analysis of contextual embeddings.

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

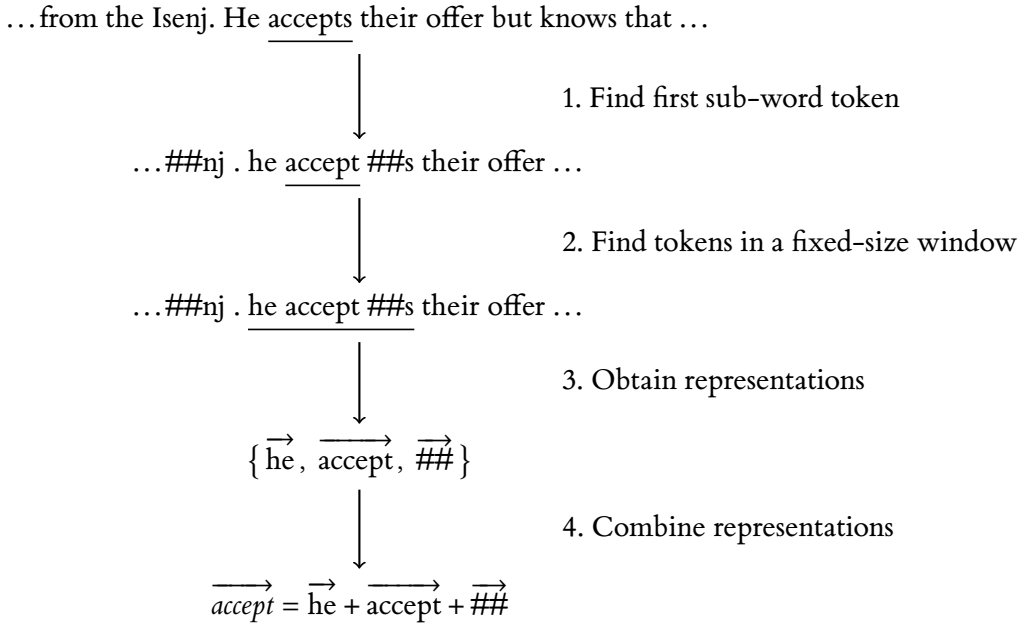

Figure 1: A schematic of the procedure used to obtain a contextualised representation of a target word from either static or contextual embeddings. In this example, the target word is *accept*, the window size is 3, and the composition operation is addition.

4 Methodology

The basic procedure of the methods that I evaluated was as follows. 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 context; finding the tokens within a fixed-size window around the target word's first token; obtaining the representations of the tokens in the window; and combining the the representations to produce a single representation of the target word.

In all cases, the tokenization was performed by and the embeddings were obtained from models available via the HuggingFace *Transformers* library (Wolf et al. 2020). The models that I evaluated for each language are given in Table ??. For the static-embedding variants of the procedure, I used the models' input embeddings; for the contextual-embedding variants, I used the models' outputs. Several of the submissions to SemEval-2020 Task 3 used a combination of the weights of a Transformer model's hidden states; a thorough comparison of the performance of variants of this approach is beyond the scope of this paper.

The use of sub-word tokens means 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 embeddings and the window size is zero. This explains the non-zero scores obtained by models of this kind (Section 7.1), most notably for the Finnish language.

5 Results

5.1 Cost-benefit analysis of contextual embeddings

5.2 Language-specificity of window-size effects

Generally, the scores obtained by both static- and contextual-embedding models are maximised by a non-zero window size. I did not perform an exhaustive search for the optimal window size due to the computational expense. A naïve estimation of the average number of words per context

(segmenting on whitespace) gave a result of 40–60 for the different languages, which motivated the choice of 50 as the maximum window size to evaluate for static-embedding models. The motivation to choose a smaller maximum window size for contextual-embedding models was likewise economical; however, as the window size approaches the length of the sequence, one would expect the sequence-level representation (e.g., BERT's `CLS` token) to be a preferable choice. These heuristics were largely confirmed by the results of the evaluation, which show that the scores decrease as the window size approaches the maximum.

This is intuitive in the case of static embeddings, because the representations of a target word in different contexts only differ otherwise if the target word is represented by different sub-word tokens in each context. In the case of contextual embeddings, a small window size is also beneficial, because a target word may be represented by multiple sub-word tokens. The window size that maximises the score for each language and model is given in Table **??**.

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 evaluation was to compare the performance of static and contextual embeddings and their additive 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 out-performed static embeddings, but at a significant cost in terms of computation and memory usage; and that a small window size improved the performance of both static and contextual embeddings, with strongly language-specific effects.

- The static embeddings of `TurkuNLP/bert-base-finnish-uncased-v1` outperform several of the Finnish submissions with a window size of zero.
- The contextual embeddings of `TurkuNLP/bert-base-finnish-uncased-v1` would have placed fourth and slightly improves on the baseline.

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*.

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.

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.

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

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.

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.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Wolf, Thomas et al. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*.

7 Appendix

7.1 Language-specificity of window-size effects

Model name	Window size	Score
bert-base-cased	18	0.456
bert-base-multilingual-cased	18	0.455
bert-base-multilingual-uncased	19	0.455
bert-base-uncased	14	0.442
bert-large-cased	18	0.441
bert-large-cased-whole-word-masking	18	0.465
bert-large-uncased	16	0.440
bert-large-uncased-whole-word-masking	16	0.471

(a) English

Model name	Window size	Score
TurkuNLP/bert-base-finnish-cased-v1	0	0.452
TurkuNLP/bert-base-finnish-uncased-v1	0	0.564
TurkuNLP/bert-large-finnish-cased-v1	0	0.500
bert-base-multilingual-cased	3	0.394
bert-base-multilingual-uncased	29	0.369

(b) Finnish

Model name	Window size	Score
bert-base-multilingual-cased	32	0.535
bert-base-multilingual-uncased	31	0.558
classla/bcms-bertic	31	0.567

(c) Croatian

Model name	Window size	Score
bert-base-multilingual-cased	8	0.316
bert-base-multilingual-uncased	10	0.291
gerulata/slovakbert	32	0.000

(d) Slovene

Figure 2: The window size that maximises the score for static-embedding models.

Model name	Window size	Score
bert-base-cased	1	0.556
bert-base-multilingual-cased	1	0.574
bert-base-multilingual-uncased	3	0.577
bert-base-uncased	1	0.660
bert-large-cased	3	0.538
bert-large-cased-whole-word-masking	2	0.557
bert-large-uncased	1	0.521
bert-large-uncased-whole-word-masking	1	0.594

(a) English

Model name	Window size	Score
TurkuNLP/bert-base-finnish-cased-v1	1	0.615
TurkuNLP/bert-base-finnish-uncased-v1	1	0.483
TurkuNLP/bert-large-finnish-cased-v1	1	0.679
bert-base-multilingual-cased	3	0.519
bert-base-multilingual-uncased	2	0.423

(b) Finnish

Model name	Window size	Score
bert-base-multilingual-cased	5	0.614
bert-base-multilingual-uncased	6	0.603
classla/bcms-bertic	10	0.423

(c) Croatian

Model name	Window size	Score
bert-base-multilingual-cased	3	0.572
bert-base-multilingual-uncased	9	0.439

(d) Slovene

Figure 3: The window size that maximises the score for contextual-embedding models.

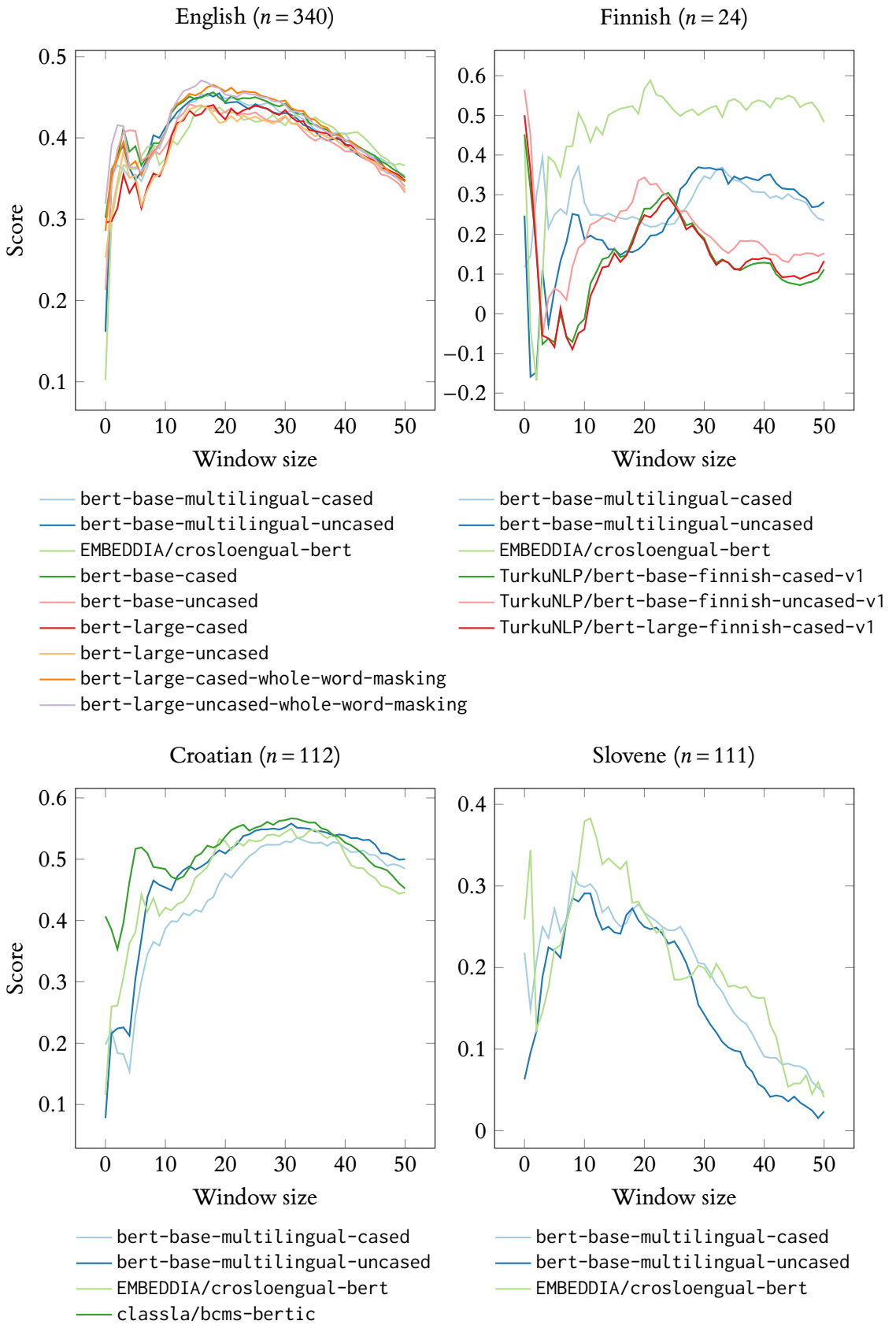

Figure 4: Score by window size for different static-embedding models.

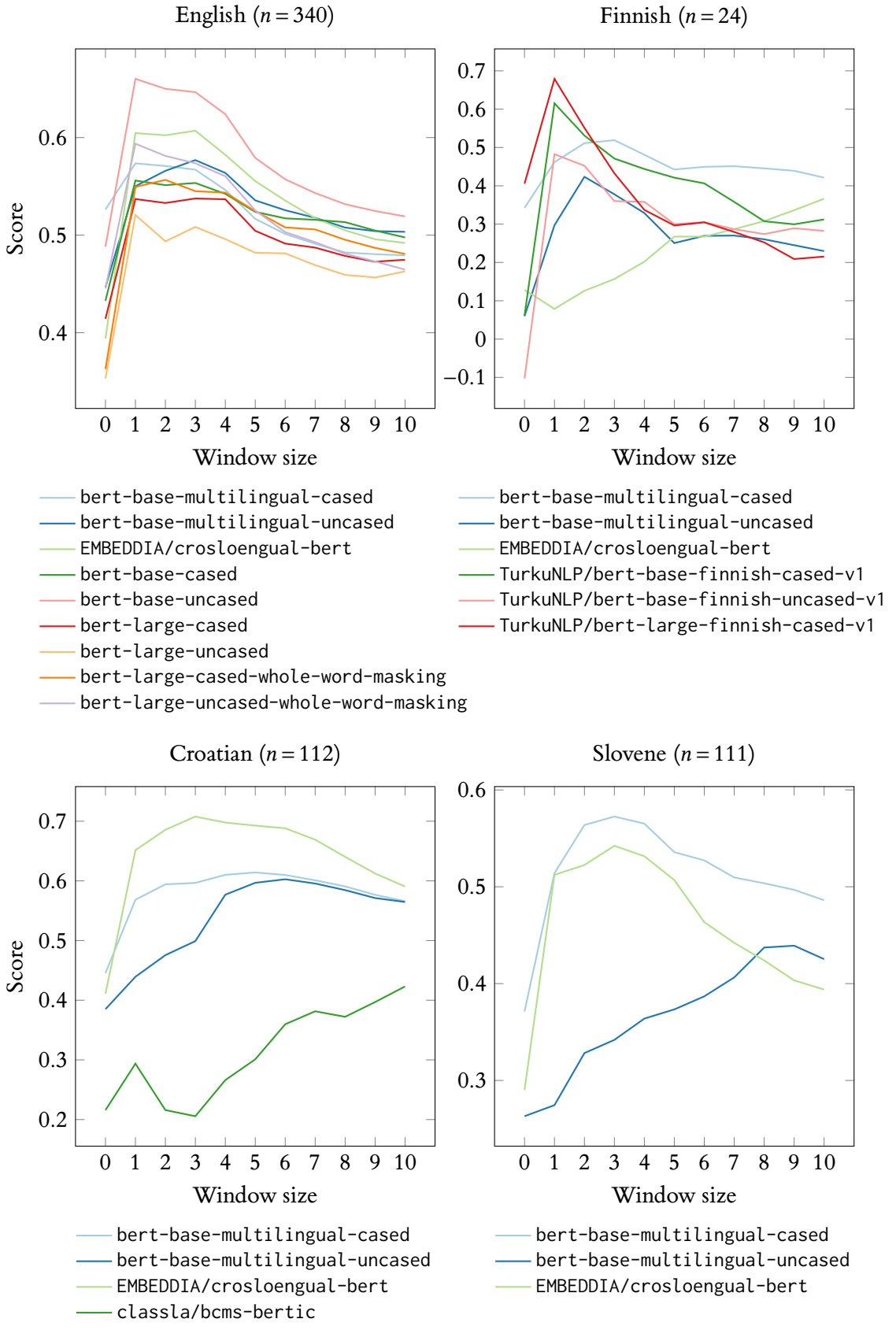

Figure 5: Score by window size for different contextual-embedding models.