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 precludes its use in modelling context-dependent phenomena.

Prior to the widespread availability of pre-trained word embeddings (e.g., Mikolov 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 similarity in each context. Specifically, I compared the results of computing the similarity between both kinds of representation of the target words and their composition with the representations of the words in a fixed-size context window, inspired by Kintsch (2001) and Mitchell and Lapata (2008).

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

4 Methodology

5 Results

6 Conclusion

- Language-specific models perform better on their target language(s).
- A small context window improves the outcomes for both static and contextualised embeddings.

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

Model	Model name	Score
contextual	`bert-base-multilingual-uncased`	0.060
contextual	`EMBEDDIA/crosloengual-bert`	0.129
contextual	`TurkuNLP/bert-base-finnish-cased-v1`	0.061
contextual	`TurkuNLP/bert-base-finnish-uncased-v1`	−0.103
contextual	`TurkuNLP/bert-large-finnish-cased-v1`	0.406
static	`bert-base-multilingual-cased`	0.117
static	`bert-base-multilingual-uncased`	0.247
static	`EMBEDDIA/crosloengual-bert`	0.437
static	`TurkuNLP/bert-base-finnish-cased-v1`	0.452
static	`TurkuNLP/bert-base-finnish-uncased-v1`	0.564[†]
static	`TurkuNLP/bert-large-finnish-cased-v1`	0.500

Figure 1: The results in Finnish with a zero-size context window.

- The contextualised embeddings outperform the static embeddings.
- However, the static embeddings are much quicker to compute.

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 et al. (2013). *Efficient Estimation of Word Representations in Vector Space*.

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.

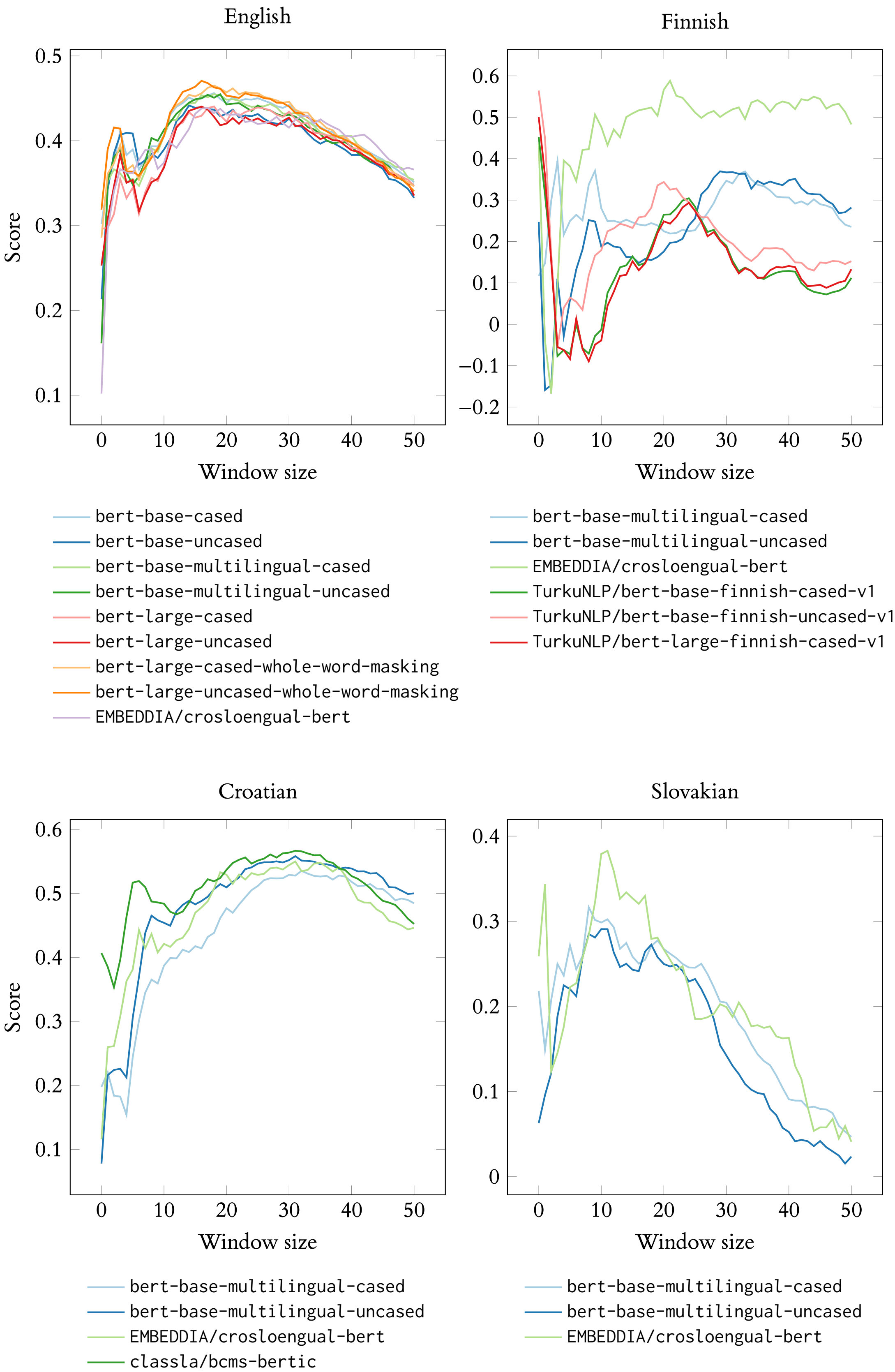

Figure 2: Score by window size for different static-embedding models.