ELMo Evaluation
Subtitle

Autor 1[1], **Autor 2**[2]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)

[2]Instituto 2

Abstract. *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibendum enim facilisis gravida. Nisl nunc mi ipsum faucibus vitae aliquet nec ullamcorper. Amet luctus venenatis lectus magna fringilla.*

Resumo. *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibendum enim facilisis gravida. Nisl nunc mi ipsum faucibus vitae aliquet nec ullamcorper. Amet luctus venenatis lectus magna fringilla.*

Sentence Similarity

Avoiding Error-Prone Methods

ELMo can be useful for tasks in which it may be necessary to generate word embeddings for unknown words. As it is a purely character based representation, it will always generate a valid word embedding for any sequence of characters. Therefore, it can be useful in tasks where simply ignoring unknown words is not an option, as we may be using models that take word position into account.

Although in some cases unknown words may simply be replaced by a symbol such as a UNKNOWN, it can lead to highly distorted results if positional details about the word embedding for this symbol are taken for granted, as it may happen to be placed inside a densely connected cluster of correlated words within the network.

[Hartmann et al. 2017] achieved distorted benchmarks for the ASSIN sentence similarity task with FastText by inadvertently using the string *unk* as a token for unknown words. Due to the employment of character n-grams, FastText mapped *unk* to a cluster of words commonly associated with musical subjects, such as *g-funk, g-punk and punk-funk*. This has led the authors of the paper to reach benchmarks highly divergent from what could be obtained by simply ignoring unknown tokens.

Table 1 shows the results reported by [Hartmann et al. 2017] side by side with our results on the embeddings that were most strongly affected by using the *unk* token. We repeated their experiments exactly as described, except for the removal of the *unk* tokens from the pre-processed data. Although we reproduced all their tests with exactly the same pre-processed data and word embeddings, other word embeddings did not display such a large difference in test results after the word *unk* was removed (Table 2).

Table 1. Mean squared error and Pearson correlation coefficient

	MSE	Pearson	MSE	Pearson
FastText, CBOW (600)	0.68	0.33	0.63	0.4
FastText, skip-gram (100)	0.58	0.49	0.52	0.55
FastText, skip-gram (1000)	0.56	0.52	0.49	0.59
FastText, skip-gram (300)	0.53	0.55	0.5	0.58
FastText, skip-gram (50)	0.61	0.45	0.55	0.52
FastText, skip-gram (600)	0.64	0.40	0.49	0.59
Wang2Vec, CBOW (300)	0.55	0.53	0.5	0.57

Table 2. Mean squared error and Pearson correlation coefficient

	MSE	Pearson	MSE	Pearson
FastText, CBOW (600)	0.68	0.33	0.63	0.4
FastText, skip-gram (100)	0.58	0.49	0.52	0.55
FastText, skip-gram (1000)	0.56	0.52	0.49	0.59
FastText, skip-gram (300)	0.53	0.55	0.5	0.58
FastText, skip-gram (50)	0.61	0.45	0.55	0.52
FastText, skip-gram (600)	0.64	0.40	0.49	0.59
Wang2Vec, CBOW (300)	0.55	0.53	0.5	0.57

Improper Candidate Detection

Visual inspection of the nearest neighbors for a token within a word embedding is not always a reliable technique to determine if it's fit to become a symbol for unknown words. However, this can be reliably determined through graph theoretic methods.

High clustering coefficient and small characteristic path length are reliable indicators of a small-world subgraph. [Cecchini 2017, p.35]. If our unknown token unk belongs to such a graph, it can significantly impact our results, specially if it replaces a large percentage of the tokens.

Let G be the graph formed by taking all words within our model as nodes. There will be an edge between two nodes a and b if their cosine similarity $s(a, b)$ is above a certain threshold t, and it will be attributed to this edge a weight of $1 - s(a, b)$.

We wish to know whether a candidate for unknown token u belongs to a small-world subgraph S' within G. Therefore, we perform a random walk starting from u, proceeding to a neighbor node c on each step, which may or may not already have been visited. If no sudden variations in the local clustering coefficient of c and the characteristic path length of the neighborhood $N_G[c]$ occur within the first few steps, then this is a strong indicator that we started our random walk inside a small-world subgraph, and therefore it may be presumed that our unknown token candidate u has the potential of distorting our results.

References

Cecchini, F. M. (2017). Graph-based clustering algorithms for word sense induction.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks.