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.*

Dealing with OOV (Out of Vocabulary) Words

Naive Approaches

For some tasks, acceptable results can be obtained by simply ignoring OOV words. However, this may not be a suitable choice, specially if OOV words constitute a significant portion of the dataset, or if factors such as word position are relevant to the task at hand.

It may seem tempting to use a single generic OOV (out of vocabulary) embedding for all words not seen during the training phase, but this approach can lead to highly undesirable results. [] has noticed that such a method fails to tell known words that have been deliberately obfuscated from non-obfuscated and rare words. [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 the subject of pop music, 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 sentence similarity 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 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

Combined ELMo Embeddings

An alternative approach

[**?**] has shown that

Smooth Inverse Frequency

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.