Title:
Subtitle

John Doe
Affiliation
johndoe@gmail.com

John Doe
Affiliation
johndoe@gmail.com

John Doe
Affiliation
johndoe@gmail.com

1 Unknown Token Generation

1.1 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. (1) 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 even 0.06 points above their true mean-squared error value, and therefore reaching the misguided conclusion that FastText embeddings exhibit a high variance in performance between semantic analogies and sentence similarity tasks.

1.2 Improper Candidate Detection

So-called meaningless word embeddings that in fact carry high significance can be spotted through visual inspection of their nearest neighbors. However, they can also be automatically detected through graph theoretic methods.

High clustering coefficient and small characteristic path length are reliable indicators of a small-world subgraph. (**?** , 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 neighboring 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.

< UNFINISHED >

References

[1] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio, “Portuguese word embeddings: Evaluating on word analogies and natural language tasks,” 2017.