Handling Noise in Distributional Semantic Models for Large Scale Text Analytics and Media Monitoring - Abstract

Peter Sumbler
Gavagai
peter@gavagai.se

Nina Viereckel
Gavagai
nina@gavagai.se

Nazanin Afsarmanesh
Gavagai
nazanin@gavagai.se

Jussi Karlgren
Gavagai
jussi@gavagai.se

The use of word embeddings in NLP tasks has exploded in the past decade. Distributional Semantic models of this kind have achieved notoriety in the research community, through projects such as GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013), and been successfully deployed as components of commercial systems, such as Gavagai's Living Lexicon[1] (Sahlgren et al., 2016).

While undeniably powerful, the authors have generally found many of the standard pre-trained vector packages to be lacking in their handling of noisy data. This is very important to our use case, which involves large scale analysis of imperfect social media and other online data. Such text is full of variations, misspellings and corruptions.

The popular open-sourced representations tend to be trained on high quality texts, such as Wikipedia dumps or news articles. While this is, of course, comprehensible, the resulting embeddings consequentially lack representations for many noisy terms. Those embeddings that are indeed learned from Social Media usually contain a pre-processing step (spell-checking) to avoid incorporating noise into the model. From our perspective, noise in the model is not only inevitable, but necessary for achieving high coverage.

We capture representations of noisy terms by leaving imperfections in the data used for training. Through capturing the semantic similarity between, for example, words and their misspellings, we can more accurately model topics and entities across corpora of messy text.

In experiments for customers, we have found that the robustness extends to handling OCR errors: ("clear" and "dear" are distributionally similar). Inflectional variants of a lemma, in contrast, are not[2].

We illustrate this phenomenon with lists of semantically similar terms from our Living Lexicon. In the following tables are the five semantically most similar terms for three commonly misspelled words in English. Common misspellings for each word feature in its neighbor list. The noise has been left in the model and adds to its richness.

Word	Cosine Similarity
private business	0.28
core business	0.25
buisness	0.25
online business	0.25
company business	0.24

Table 1: Semantically similar terms to 'business' and their cosine similarities with the target.

Word	Cosine Similarity
seperate	0.63
downstairs	0.23
stand-alone	0.22
en-suite	0.21
backed-up	0.2

Table 2: Semantically similar terms to 'separate' and their cosine similarities with the target.

Word	Cosine Similarity
accomodation	0.56
accommodations	0.45
lodging	0.28
serviced apartment	0.25
lodgings	0.24

Table 3: Semantically similar terms to 'accommodation' and their cosine similarities with the target.

[1] lexicon.gavagai.se

[2] gavagai.se/blog/2017/02/23/morphology/