SWOW-8500: Word Association Task for Intrinsic Evaluation of Word Embeddings

Introduction

Word vectors can be evaluated either Extrinsically (downstream tasks like sentiment classification, question answering) or Intrinsically (smaller probing tasks like predicting similarity score between two given words). Intrinsic evaluation should be simple, independent of model architecture, but should still correlate well with performance on downstream tasks.

Word 1	Word 2	Similarity
tea	coffee	4
dog	jeans	0
war	battle	5
353 more such rows		

Figure 1a: WordSim-353 sample data

Cue	Response	R123	out of N
would	should	63	288
would	could	63	288
would	will	24	288
would	can	11	288
978908 more such rows			

Figure 1b: Small World of Words sample data

Intrinsic Evaluation so far:
WordSim and SimLex type resources are a collection of a few hundred word pairs manually labelled with a score of their similarity (Figure 1a). For each word pair, the set of word embeddings to be evaluated (referred as E henceforth) predicts a similarity score, usually by simply computing the cosine similarity between the two words' vectors. The overall performance of E on this intrinsic task is reported as the Pearson / Spearman correlation between predicted and labelled similarity scores. Faruqui et al. 2016 state that these intrinsic evaluations suffer from a lack of statistical significance.

Word Association datasets:
Word Association games are those wherein participants are asked to utter the first (or first few) words that occur to them when given a trigger / cue / stimulus word. For example, given the cue word King, one could respond with words like Rule, Queen, Kingdom, or even Kong (from the movie King Kong). Word associations have long intrigued psychologists including Carl Jung and hence large studies have been conducted in this direction. The Small World of Words project (SWOW) is based on one such study which involved 90000+ participants. The dataset is in the format of the number of participants (called R123 score) which responded to a given Cue word with a particular Response word (Figure 1b).

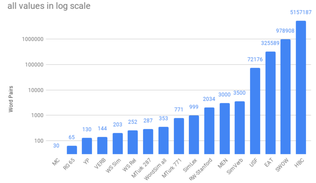

Figure 2: Sizes of similarity-based intrinsic evaluation datasets and 4 word association datasets (USF - Univ of Southern Florida Free Association Norms, EAT - Edinburgh Association Thesaurus, SWOW - Small World of Words, HBC - Human Brain Cloud) in terms of number of word pairs each.

What's the first word that comes to your mind when I say 'Minneapolis'?

The **Word Association** task can act as a proxy for Intrinsic Evaluation of Word Embeddings, with:

- similar results, but
- better confidence intervals
- and for FREE!

Task

One way to use word association datasets for intrinsic evaluation is to simply list all word pairs and proceed exactly like WordSim or SimLex. But that doesn't work very well since several cue-response pairs in word association datasets are meaningless and have a low frequency / R123 score associated with them. Even when filtering only the top occurring cue-response pairs, we are left with so many positive examples but no negative ones (see the dog-jeans example in Figure 1a).

Instead we propose the following task: Given a set of pretrained word embeddings E and a subset of the word association dataset's cue-response pairs CR (chosen by a threshold over the R123 score), we ask E to predict which are the top k closest responses to each cue in CR. Let the top 3 responses, according to SWOW, for the cue Would be Should, Could, and Will. Let a certain E guess Will, Might, and Should as the closest words to Would, based on cosine similarity. Then the True Positives are Will and Should. Might is a True Negative. Could is a False Positive. Precision = 2/3. Recall = 2/3.

Experiments

Figure 3 shows results from our experiments where we compare performances of pretrained word embeddings on (1) Our proposed task, (2) other Intrinsic Evaluation tasks, and (3) Downstream Tasks. We settle upon the following candidates for E (all with 300 dimensions; vocabulary of 7779 words):

1. Word2vec skip gram
2. GloVe
3. FastText
4. ConceptNet Numberbatch

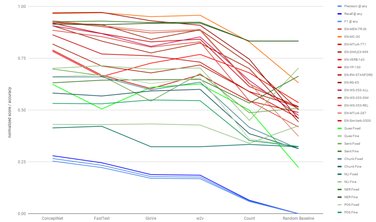

Figure 3: Performance Scores of the 6 candidate embeddings on (1) Our proposed word association task - shades of BLUE, (2) 13 other Intrinsic Evaluation tasks - shades of RED, and (3) 6 Downstream tasks - shades of GREEN.

Conclusions

From Figure 3, we see that (unlike prior reports) intrinsic methods of evaluation seem to correlate well with performance on downstream tasks. Also, our word association task SWOW-8500 correlates well with both types of tasks! This begs the question: if our task captures the same properties already captured, then why introduce a new task?

Figure 4 shows the confidence intervals of scores reported by all intrinsic evaluation methods, including our proposed word association task. We see how SWOW, even with a very modest threshold of R123 / N > 0.2, gives us a very narrow confidence interval. We can now pronounce our results on intrinsic evaluation with a decent statistical significance.

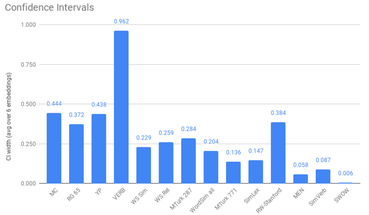

Figure 4: Confidence Intervals of accuracy / error scores for 13 existing Intrinsic Evaluation tasks, as well as our proposed SWOW-8500 task (last stack).

Bonus

- According to Aristotle, there exist 3 types of word associations: Similarity, Contiguity, and Contrast. For the word King, three associated words could be Emperor (Similarity), Kong (Contiguity), and Queen (Contrast). We could study Distributional Semantics even further by classifying the "company of a word" based on what kind of association it keeps. WordNet may come in handy.
- Check out this word game: https://research.google.com/semantris/
- Check out our code & data at https://github.com/avi-jit/SWOW-eval/

Authors

- Avijit Thawani (thawani@usc.edu)
- Biplav Srivastava - IBM New York
- Anil Kumar Singh - IIT BHU

Take a picture to view demo code and data!