**Introduction**

Word vectors can be evaluated either Extrinsically (on downstream tasks like sentiment classification, question answering) or Intrinsically (probing tasks like prediction). Most contemporary methods are driven by the need for optimization of model architecture, but should still correlate well with performance on downstream tasks. 

Intrinsic Evaluation so far

Evaluation methods such as Cosine-Similarity and Word-Net are used to extract similarity between two word vectors by computing the similarity between the two word vectors.

The overall performance on these intrinsic tasks is reported as the Pearson / Spearman correlation between predicted and labelled similarity scores. 

Experiments

We set up a series of experiments to compare performances of pretrained word embeddings on (1) Our proposed word association task, (2) other Intrinsic Evaluation tasks, and (3) Downstream tasks.

**Author**

Anil Kumar Singh - IIT BHU

Avijit Thawani - USC

Authors

- Raji V. Prabhu - ConceptNet
- Blake Stroehl - GDFX
- Anil Kumar Singh - IIT BHU

Conclusions

From Figure 3, we see that (albeit 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. 

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 $R_{123}<0.2$, gives us a very narrow confidence interval. We can now produce results on intrinsic evaluation with a decent statistical significance.

**Figure 1:**

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

**Figure 2:**

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!

**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.

**Figure 4:**

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 $R_{123}<0.2$, gives us a very narrow confidence interval. We can now produce results on intrinsic evaluation with a decent statistical significance.