# Unsupervised Language Learning: Lab 1

Aron Hammond & Joris Mollinga

## Word similarity

For this task we compared the cosine similarity of word vectors to two sets of similarity scores: SimLex and MEN. Looking at both Pearson and Spearman correlations, the similarity scores for `bow5` seem to correlate best with the MEN evaluations and worst with SimLex. This is most likely due to the fact that SimLex scores measure similarity rather than relatedness and a larger window captures relatively more topical (ie. related) meaning. As expected, the dependency based embeddings have the highest correlation on SimLex. What is curious, is that correlations with MEN are stictly larger than those on SimLex. This suggests that all three models are still better at capturing relatedness rather than similarity. However, it could also be the case that the noise is more dominant in the SimLex correlations because of the smaller sample size.

## Word analogy

Another task used to evaluate embeddings is completing analogies. The following section will briefly discuss the differences of the three models performance on this task. Due to computational constraints, the MRR and accuracy were computed for a random sample of 10.000 analogies (approximately half of the available data). The highest MRR@5 is scored by `bow5` (0.399) followed by `bow2` (0.356) and `deps` (0.216). The accuracies for these models are ranked the same and are 0.103, 0.087 and 0.024 respectively. These scores are lower than could be due to the fact that often the top-ranked candidate is one of the input words. Excluding these from the candidates could significantly boost the results [Mikolov et al.2013], as demonstrated by a smaller sample of 100 analogies where the accuracies are 0.68, 0.7 and 0.29. A much smaller source of error comes from the absence of the target word in the vocabulary of the models. We also visually inspected some analogies along with the cadidates proposed by each model. This was done for candidates that don't consider the input words to declutter the list. All models perform noticably better on the grammatical categories. For the geographic analogies the dependency based vectors seem to perform poorly. An example of a failure in this category is $bankok : thailand :: hanoi :?$ for which the following candidates were proposed: ethiopia, uzbekistan, iran, tajikistan and turkmenistan. These are all countries, but they have very little in common with Hanoi. The other models both suggested candidates that were in the same region with the correct anwser ranked relatively high. The lower performance of deps hints at the conclusion that relatedness is more usefull in analogies than similarity. Another interesting failure was caused by the ambiguous word 'real'. It is both a word to describe the existence of someting and the Brazilian currency. So for a land-currency analogy the models suggested words to do with realism rather than currencies.

## Clustering word vectors

We implemented the PCA and TSNE algorithm to visualize the word vectors in two dimensions, but because these are very low representations of high dimensional data points, it is difficult to draw conclusions based on these plots. When clustering the 2000 nouns, results start to make sense from 25 clusters. At that point the algorithm realizes that human body parts belong in one cluster, while there is another cluster related to money, albeit clusters are still noisy. Results continue to improve for increasing number of clusters, but remain a bit noisy. In general the BOW model with $k = 2$ is the least noisy while capturing the most appropriate similar words, although in absence of a ground truth it is impossible to compare quantatively.[1]

## References

[Mikolov et al.2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[1] Code available at: https://github.com/jorism1993/ULL-Practicals