Assignment 1: Evaluating Word Representations

Jack Harding (11623608) Akash Raj Komarlu Narendra Gupta (11617586)

In this report, we experiment with and evaluate various word representation models. The code for the project can be found in our github repository.[1] We do not include the clustering images in the report since they are big (they can be found in the results folder in our repository).

1 Word Similarity Task

The Pearson and Spearman correlation scores for the SimLex task are given in Table 1. As the reader can verify, the word embeddings created using syntactic dependencies as contexts ('Deps' for short) achieved a higher correlation score (on both correlation tests) than the Bag of Words models (referred to as 'BoW2' when the context window $k = 2$ and 'BoW5' when the context window $k = 5$) on the SimLex task. The BoW2 model out-performed the BoW5 model.

In a sense, this is unsurprising. Since it uses syntactic dependencies as its context in Skipgram, Deps tends to capture more of the 'true similarity' that the SimLex data set tests for. For example, BoW models sometimes incorporate pragmatic information (such as the tone of the document in which words appear) into their embeddings. Some potential examples of this are given in Table 8. Note, for example, that 'mushroom' and 'funghi' are synonyms, but the latter tends to be used in more technical contexts; this could explain why Deps assigns a higher similarity to this pair than the BoW models. One problem with BoW models on similarity tasks (and advantage on relatedness tasks), and BoW2 in particular, is that they apportion too much weight to idioms; words that commonly co-occur in idiomatic phrases are counted as similar. We present some examples of this in Table 9.

The Pearson and Spearman correlation scores for the MEN task are given in Table 2. As the reader can verify, BoW5 achieved higher correlation scores than BoW2, which in turn achieved higher scores than Deps. Since the cognitive relatedness of the sort tested by the MEN data set is separable from the sort of semantic similarity discussed above, this is to be expected.

To fine-grain these results, in Table 3, we also present information on the correlations of each model for each range of MEN scores. All three models perform worse when the scores are more extreme (when the relatedness is very high or very low). This could be explained by the high-dimensionality of the vector space in which the word embeddings lie; words are unlikely to be far apart on every dimension, or close on every dimension. In particular, Deps has very low correlation scores when the MEN score of the word pair is less than 10. This could be explained by the fact that Deps tends to capture functional similarity; two words that are intuitively very different could share syntactic features (such as a POS tag), and hence dependencies. Some examples of this sort of drawback of Deps are given in Table 7.

2 Word Analogy Task

In Table 4, we present the accuracy and Mean Reciprocal Rank (MRR) scores of each of the three models on each of the five ontology and nine grammatical tasks in the Google Word Analogy Test Set. Following Mikolov (2012), we have removed query words from consideration. We can see that, although the models have fairly similar scores on the grammatical tasks, the BoW models perform better on the ontology tasks. This could be explained by the fact that functional similarity does not help in these analogy tasks. For example, Deps is likely to return other city names on the analogy $Baghdad \rightarrow Iraq : Athens \rightarrow ?$, since it groups cities close together in the embedding space, prioritising functional similarity over the relatedness needed in the capital city task. So in tasks involving proper names, Deps performs worse. Even within the grammatical tasks, there is considerable variation from category to category. For example, Deps performs very poorly on the Adjective $\rightarrow$ Adverb analogy; this could be because there are very few semantic constraints on which verbs adverbs depend on (such that the dependencies do little to individuate the adverbs). On the other hand, Deps performs well on the Opposite category (relative to the two BoW models). This could be explained by the fact that opposites typically share syntactic features (and hence dependencies), even though they tend to occur in distinct contexts.

[1] https://github.com/Jack-Harding1/ULL

3 Clustering Word Vectors

For the three word representations - {Deps, BoW2, BoW5}, we use K-Means to cluster 2000 nouns with different cluster sizes[2]. Table 5 shows the nearest words using different word representations for the chosen target word.

The first target word, *breakfast*'s cluster contains similar words. We can see that most of these words are unambiguous and give similar meaning in different contexts and therefore these words are grouped together irrespective of the word representations. In the case of *capital*, we see that BoW5 yields words related to an area of a state enjoying primary status. Deps on the other hand yields words that capture the semantic type of the word - wealth in the form of assets. Consider the target word, *bomb*. While Deps considers bomb as a weapon, BoW models talk about the context - a bomb being used to create an *explosion* or *attack* the *enemy*. It is evident that ambiguous words are clustered into different groups in different representations.

Next, we compare Deps for cluster sizes {400, 1000}. In Table 6 we present a few examples. For $k = 400$, consider the cluster [american, christian, german, russian]. When we use more clusters, there is a further division: [american, german, russian] and [christian]. As the number of clusters increases, we observe a further split in the similarity grouping. While increasing the number of cluster has advantages, there are certain shortcomings. For example, the cluster [autumn, summer, fall, spring, winter] which represents the seasons is grouped into 3 clusters for $k = 1000$. This is because while they belong to the same cluster, [spring, fall] are further away from the cluster center due to their ambiguity.

A Tables

Model	Pearson Correlation	Spearman Correlation
Deps	**0.462**	**0.446**
BoW2	0.428	0.414
BoW5	0.376	0.367

Table 1: SimLex Correlation Scores.

Model	Pearson Correlation	Spearman Correlation
Deps	0.597	0.618
BoW2	0.678	0.700
BoW5	**0.708**	**0.723**

Table 2: MEN Relatedness Correlation Scores.

Score Range	Deps		BoW2		BoW5	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
$0 \leq$ MEN Score < 10	-0.026	-0.038	0.096	0.102	0.178	0.169
$10 \leq$ MEN Score < 20	0.240	0.224	0.290	0.280	0.321	0.314
$20 \leq$ MEN Score < 30	0.220	0.218	0.271	0.265	0.288	0.285
$30 \leq$ MEN Score < 40	0.103	0.102	0.122	0.112	0.127	0.123
$40 \leq$ MEN Score < 50	0.112	0.100	0.117	0.105	0.099	0.099

Table 3: MEN Relatedness Correlation Scores for each range.

[2]Cluster sizes = $\{2, 5, 10, 30, 33, 45, 75, 100, 150, 175, 200, 300, 400, 500, 822, 1000\}$

Part of Analogy Task	Deps		BoW2		BoW5	
	Accuracy	MRR	Accuracy	MRR	Accuracy	MRR
Capitals - Common Countries	0.352	0.494	0.836	0.882	**0.941**	**0.964**
Capitals - World	0.112	0.203	0.630	0.719	**0.703**	**0.799**
City - State	0.123	0.221	0.392	0.498	**0.513**	**0.621**
Currency	0.064	0.096	0.113	0.148	**0.122**	**0.169**
Family	0.816	0.855	0.794	0.854	**0.818**	**0.870**
Adjective - Adverb	0.034	0.067	0.159	0.236	**0.169**	**0.272**
Opposite	**0.400**	**0.476**	0.356	0.423	0.363	0.432
Comparative	0.801	0.853	**0.896**	**0.939**	0.830	0.892
Superlative	0.561	0.637	**0.631**	**0.730**	0.571	0.699
Present Participle	0.647	0.740	0.627	0.747	**0.670**	**0.782**
Nationality-Adjective	0.121	0.220	0.742	0.807	**0.824**	**0.865**
Past Tense	**0.659**	**0.732**	0.557	0.663	0.547	0.666
Plural	0.676	0.748	**0.733**	**0.793**	0.668	0.752
Plural Verbs	**0.909**	**0.945**	0.807	0.865	0.736	0.822
Overall (excluding proper names)	0.613	0.676	**0.622**	**0.700**	0.600	0.690
Overall	0.367	0.446	0.593	0.674	**0.623**	**0.712**

Table 4: Scores for the Analogy Task, by category and overall. Query words have been removed from the embedding space for the purposes of prediction.

Target word	Deps	BoW2	BoW5
breakfast	dinner	dinner	dinner
	lunch	host	lunch
	meal	lunch	meal
		meal	
capital	cash	cost	colony
	coin	ease	empire
	currency	expenditure	state
		expense	territory
blow	kick	triumph	tear
	punch	defeat	crack
	shot	face	break
	pass	victory	
independence	isolation	democracy	convention
	separation	freedom	declaration
			peace
			treaty
baby	child	girl	girl
		woman	teenager
			woman
bomb	missile	accident	assault
	mine	boom	attack
	weapon	crash	enemy
		explosion	raid

Table 5: Target words and the words contained in their cluster for different word representations. Cluster size, $k = 400$. Unambiguous words remain in the same cluster in different word representations.

$k = 400$	$k = 1000$		
american	american	christian	
christian	german		
german	russian		
russian			
autumn	autumn	spring	fall
fall	summer		collapse
spring	winter		
summer			
winter			
baby	baby		
child	child		

Table 6: Clustering for `Deps` for cluster sizes 400 and 1000. Each column under $k = 1000$ represents a different cluster.

Word Pair	MEN Score	`Deps` Cosine Score	`BoW2` Cosine Score	`BoW5` Cosine Score
Feline, Pumpkin	8.0	**0.618**	0.310	0.175
Bikini, Pizza	1.0	**0.434**	0.233	0.204

Table 7: Low relatedness word pairs which are nonetheless close together on the `Deps` model.

Word Pair	MEN Score	`Deps` Cosine Score	`BoW2` Cosine Score	`BoW5` Cosine Score
Fungi, Mushrooms	46.0	**0.718**	0.593	0.529
Feline, Kitty	44.0	**0.479**	0.263	0.274
Canine, Poodle	41.0	**0.641**	0.364	0.336
Strawberry, Tomato	34.0	**0.732**	0.512	0.554
Chess, Toys	29.0	**0.357**	0.102	0.08

Table 8: Word pairs with high similarity but which tend to occur in different pragmatic contexts; `Deps` captures the similarity of these words better.

Word Pair	`Deps` Cosine Score	`BoW2` Cosine Score	`BoW5` Cosine Score
Tropical, Storm	0.436	**0.620**	0.588
Ancient, Ruins	0.324	0.425	**0.515**
Swimming, Pool	0.372	0.539	**0.623**
Figure, Skating	0.279	**0.398**	0.310

Table 9: Examples of words with high relatedness and low similarity scores; note the volume of idiomatic phrases here. `BoW` models tend to place embeddings for these words closer together.