Evaluating Word Representations

Tarun Krishna
tarun.krishna@student.uva.nl

Dhruba Pujary
dhruba.pujary@student.uva.nl

April 20, 2018

Word Similarity Task Given evaluation datasets capture their own notion of similarity. MEN which emphasizes more on association and relatedness, while on the other hand SimLex-999[1] quantifies similarity rather than association or relatedness so that pairs of entities that are associated but not actually similar have a low rating. As a consequence, MEN effectively penalize models for learning the evident truth. Evaluations on these datasets can be summarizes as: **1** BoW based model captures topical similarity while dependency based model captures *functional similarities*[2]. As a result DEPS has a better Pearson-Spearman Correlation as compared to BoW2 and BoW5 for SimLex-999 as depicted in Figure1. Although the gains are marginals but DEPs based model gains from the fact that it captures syntactic context. **2** Although BoW2 perform slightly better than BoW5 that's because a window size of 5 captures broad topical content, whereas smaller windows contain more focused information about the target word.

3 For MEN dataset1, BoW5 supersede all other models(broader topical similarity). Dependency based models suffers because it emphasizes on functional similarity i.e it yields more focused embeddings. This is also attributed to the fact that MEN dataset captures association and relatedness rather then similarity which is in contrast with SimLex-999.

[1]**Qualitative Analysis** In Figure2and3 we manually inspect 5 common word in both dataset. For simLex-999 in word *cloud*, DEPs is more consistent with ranking as compared to other two but in MEN BoW5 is more consistent. For *band* DEPs remains consistent but BoW5 fails to even retrieve it. Interestingly in MEN, DEPs ranks musicians higher than BoW(2,5) this could be attributed to the notion of syntactic context it captures i.e something like "*band of musicians*"(generally as it is said) is captured by DEPs rather than topical similarity. For *happy* and *car* story remains the same for SimLex-999 i.e is DEPs is fairly better than BoW, but for *beach* BoW5 and BoW2 are more consistent than DEPs this is because the similarity scores for SimLex for this word is more of a topical. In MEN for *happy* and *car* BoW are better than DEPs. The overall conclusion which can be drawn is that these model capture different notions of similarity. And their performances varies with different evaluation dataset. Discussed models are not really consistent in performance for given evaluation datasets but these evaluations gives us some measure of lackness in these models. Also, we cannot ignore the fact that association and similarity are neither mutually exclusive nor independent. This could lead to a conflict as in the case of word *band* and *beach* for MEN and for word *beach* in SimLex-999.

Analogy The word analogy task is to predict the word given a group of word pairs that supposedly have the same relationship. In figure 4 , we see that overall accuracy and MRR of BoW5 representation has a higher score than BoW2 followed by DEPS. Category wise, i.e. gram7-past-tense, gram8-plural and gram9-plural verbs relations are captured better by the DEPs embeddings as compared to BoW5 and BoW2. Whereas for all other category BoW5 performs better (Figure5, 6, 7). From this analysis, we can say that the relationship such as capital-common-countries that is predicted by BoW model is better because these words appear more within the context window of a sentence. The difference between BoW(k=2) and (k=5) gives different results because of different context window size. The DEPs on the other hand captures relationships like past tense and plural verbs better semantically(syntactical context), for example think:thinks::talks talks is predicted as talk,brags,talks by DEPs whereas by BoW2 as talk, talkpage, vandalizes, thinks etc. DEPs predicts words *brags* which is also plural of similar meaning as *talks* whereas the BoW2 only predicts the words which are not so much related to *talks* directly but are plural. There are counter examples as well where DEPs predicts more semantically similar plural words but BoW2 and BoW5 predicts the correct words. In conclusion, we can say that BoW models can predict words based on direct occurrence in word embedding space to a given relationship better than DEPs whereas the DEPs model predict words that behave similar to a given relationship.

Clustering In noun word clustering, for the three models, we can see words form distinct clusters(Figure8). For BoW2 model, the words are very localized and have similar contextual meaning such as titles, names of body parts, names of location etc. In case of BoW5 model, the clusters have more contextual meaning, for example foods, negative actions, cause

[1] In this discussion first dataset is SimLex-999 and other is MEN.

like drugs, murders etc, or clusters related to scientific measurements and physics. In DEPs model, we see the words are related to each other in more contextual or semantic way are close together. For example,the occupation of people are in one cluster, nature related things like mist, sunlight etc, or things related to money. We can say that BoW2 keeps words together with small contextual similarity. The BoW5 brings together such small contextual similar words together binding them giving a bigger contextual meaning i.e a larger topical similarity. DEPs on the other hand keeps words together that are more syntactically related then their meaning.

References

[1] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2015.

[2] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[3] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations, 2013.

Code:
https://github.com/druv022/ULL/tree/master/Lab1

A Word Similarity

For Figure2 and 3 we took 5 common words in SimLex-999 and MEN with their top-5 scores as indicated in Figure2and3. We tried to compute cosine similarity scores with the given embeddings BoW(2,5) and DEPs within in each of the data set. We display again top-5 scores for each of them.

B Analogy

Using the vector offset method[3], the computed word may not exist. So we search for the best matching word by ranking the word's cosine similarity to every other word embeddings. The accuracy is measured by counting the number of best matching word which is also the correct word and divide it by the total number of example. The mean reciprocal rank(MRR) is evaluated by summing all the reciprocal rank of all the examples and divide it by the total number of examples. We observed that the predicted rank of any analogy task mostly contains the same example words followed by the correct word. This could happen if the vector offset is very small and or the nearest words are the example words only near that offset. We calculated the accuracy(type 2) and MRR(type 2) by excluding the example words in the rank and obtain better scores.(figure 4)

C Clustering

We performed clustering on the three model that minimizes total distance of cluster assignment to the data points. We obtain 50 clusters works best for BoW2, 30 for BoW5 and 40 and DEPs respectively.

SimLex-999			
Correlation Type	BoW2	BoW5	DEPS
Spearman	0.414	0.367	0.446
Pearson	0.428	0.376	0.469
MEN Natural Form			
Spearman	0.70	0.723	0.617
Pearson	0.677	0.708	0.5974

Figure 1: **Comparison of Correlation coefficients**

Target Word	SimLex-999	BoW 5	BoW2	DEPS
cloud	haze 0.732 mist 0.667 fog 0.6 storm 0.56 weather 0.487	**mist** **haze** drizzle **fog** snow	**fog** **mist** rain dense **haze**	**mist** **haze** **fog** drizzle bubble
band	orchestra 0.708 parade 0.392	metal song singer soul ballad	song choir **orchestra** singer soul	choir **orchestra** singer company orthodontist
happy	cheerful 0.955 glad 0.917 young 0.2 angry 0.128 mad 0.095	**glad** proud unhappy confident afraid	**glad** unhappy proud **cheerful** sad	**glad** anxious **cheerful** unhappy stupid
car	cab 0.742 carriage.513 bicycle 0.47 highway 0.34 factory 0.275	vehicle driver motor bicycle taxi	vehicle **bicycle** **carriage** boat wagon	vehicle wagon **carriage** boat **cab**
beach	seashore 0.833 island 0.56 sea 0.468 reef 0.377	palm shore **seashore** **island** **reef**	**seashore** shore **island** lake **reef**	shore **seashore** **reef** ledge **island**

Figure 2: **SimLex-999**

Target Word	MEN	BoW 5	BoW2	DEPS
cloud	sky 0.84	**mist**	fog	**mist**
	rain 0.8	fog	shadow	fog
	mist 0.74	snow	**mist**	shadow
	droplets 0.72	**rain**	**rain**	puddle
	misty 0.7	**droplets**	flame	**droplets**
band	musicians 0.84	punk	punk	punk
	music 0.82	rock	outfit	**musicians**
	guitar 0.82	metal	**musicians**	outfit
	played 0.74	jazz	cheerleader	rock
	concert 0.74	**musicians**	metal	cheerleaders
happy	smile 0.8	**love**	fun	cute
	love 0.78	**smile**	sweet	**sexy**
	kids 0.58	fun	**love**	sunny
	sexy 0.56	kiss	**smile**	colorful
	tears 0.54	**sexy**	cute	dirty
car	vehicle 0.92	truck	**vehicle**	truck
	motor 0.82	automobile	truck	**vehicle**
	garage 0.82	**vehicle**	motorcycle	automobile
	parking 0.72	motorcycle	automobile	motorcycle
	park 0.665	bike	bike	bike
beach	sand 0.96	boardwalk	pier	boardwalk
	sea 0.88	surf	boardwalk	**shore**
	shore 0.84	palm	**shore**	pier
	coast 0.84	**shore**	pond	bay
	harbour 0.82	pier	surf	sidewalk

Figure 3: **MEN Natural form full**

Embedding	MMR Type 1	MMR Type 2	Accuracy Type 1	Accuracy Type 2
BOW2	0.367	0.575	0.0873	0.5838
BOW5	0.407	0.597	0.102	0.61325
Dependency	0.237	0.411	0.0264	0.36062

Figure 4: MMR and Accuracy on google dataset

Category	MMR Type 1	MMR Type 2	Accuracy Type 1	Accuracy Type 2
capital-common-countries	0.546	0.651	0.229	0.82597
capital-world	0.376	0.659	0.0597	0.63019
currency	0.0639	0.0996	0.00924	0.084296
city-in-state	0.198	0.497	0.00081	0.39238
family	0.492	0.729	0.123	0.79447
gram1-adjective-to-adverb	0.114	0.234	0.00202	0.15927
gram2-opposite	0.212	0.414	0.00985	0.35591
gram3-comparative	0.538	0.808	0.313	0.89565
gram4-superlative	0.254	0.65	0.0365	0.59358
gram5-present-participle	0.399	0.708	0.0388	0.62689
gram6-nationality-adjctive	0.678	0.265	0.527	0.74171
gram7-past-tense	0.349	0.643	0.0192	0.55705
gram8-plural	0.403	0.78	0.0128	0.73273
gram9-plural-verbs	0.492	0.746	0.109	0.8069

Figure 5: MMR and Accuracy on different catagory of google dataset for BOW2 representation

Category	MMR Type 1	MMR Type 2	Accuracy Type 1	Accuracy Type 2
capital-common-countries	0.569	0.741	0.223	0.94071
capital-world	0.44	0.74	0.0588	0.70292
currency	0.0775	0.101	0.0231	0.091224
city-in-state	0.309	0.61	0.113	0.51277
family	0.523	0.697	0.172	0.81818
gram1-adjective-to-adverb	0.14	0.267	0.00504	0.16935
gram2-opposite	0.219	0.415	0.0172	0.3633
gram3-comparative	0.549	0.69	0.198	0.83033
gram4-superlative	0.337	0.615	0.0419	0.53743
gram5-present-participle	0.417	0.755	0.0265	0.67045
gram6-nationality-adjctive	0.742	0.255	0.595	0.82364
gram7-past-tense	0.362	0.645	0.0212	0.54679
gram8-plural	0.398	0.727	0.0248	0.66817
gram9-plural-verbs	0.481	0.705	0.116	0.73563

Figure 6: MMR and Accuracy on different catagory of google dataset for BOW5 representation

Category	MMR Type 1	MMR Type 2	Accuracy Type 1	Accuracy Type 2
capital-common-countries	0.215	0.492	0.00198	0.35178
capital-world	0.108	0.202	0.00177	0.11207
currency	0.0421	0.0585	0.00693	0.04388
city-in-state	0.11	0.221	0.0	0.12282
family	0.478	0.763	0.090	0.81621
gram1-adjective-to-adverb	0.0426	0.0599	0.00504	0.034274
gram2-opposite	0.246	0.476	0.0	0.40025
gram3-comparative	0.461	0.775	0.0788	0.80105
gram4-superlative	0.311	0.522	0.0472	0.52763
gram5-present-participle	0.403	0.705	0.035	0.64678
gram6-nationality-adjctive	0.137	0.214	0.00563	0.12133
gram7-past-tense	0.402	0.683	0.0487	0.65897
gram8-plural	0.4	0.716	0.0315	0.67568
gram9-plural-verbs	0.553	0.798	0.147	0.9092

Figure 7: MMR and Accuracy on different catagory of google dataset for Dependency representation

Embedding	Clusters
BoW2	**cluster 1**’duke’, ’lord’, ’count’, ’king’, ’emperor’, ’queen’, ’knight’, ’earl’, ’princess’, ’lady’, ’prince’ **cluster 2** ’wound’, ’cancer’, ’gene’, ’stroke’, ’hair’, ’heart’, ’corpse’, ’body’, ’tongue’, ’brain’, ’skin’, ’surgery’, ’vein’, ’limb’, ’tooth’, ’blood’, ’spine’, ’leaf’, ’bone’, ’stomach’, ’muscle’, ’tissue’, ’throat’, ’cell’, ’root’, ’breast’, ’stem’, ’nerve’ **cluster 3** ’team’, ’football’, ’coach’, ’player’, ’captain’, ’league’, ’manager’, ’defender’, ’receiver’, ’champion” **cluster 4** ’incidence’, ’length’, ’quantity’, ’accuracy’, ’rating’, ’size’, ’speed’, ’average’, ’profile’, ’height’, ’extent’,’cost’, ’register’, ’sensitivity’, ’value’, ’volume’, ’limit’, ’proportion’, ’availability’, ’minimum’, ’amount’, ’level’, ’share’ **cluster 5** ’australia’, ’america’, ’italy’, ’leeds’, ’india’, ’germany’, ’paris’, ’britain’, ’israel’, ’england’, ’york’
BoW5	**cluster 1**’beer’, ’sugar’, ’salt’, ’cream’, ’bottle’, ’juice’, ’coffee’, ’potato’, ’apple’, ’drink’, ’grain’, ’fruit’, ’alcohol’, ’ingredient’, ’champagne’, ’wine’ **cluster 2** ’drug’, ’death’, ’violence’, ’illness’, ’crime’, ’incident’, ’killer’, ’offender’, ’arrest’, ’abuse’, ’punishment’, ’accident’, ’suicide’, ’prison’, ’killing’, ’gang’, ’offence’, ’terror’, ’affair’, ’conviction’ **cluster 3** ’momentum’, ’flow’, ’signal’, ’experiment’, ’beam’, ’spectrum’, ’impulse’, ’particle’, ’wave’, ’envelope’, ’error’, ’array’, ’emission’, ’output’, ’measurement’, ’sample’, ’angle’, ’medium’, ’frequency’ **cluster 4** ’estate’, ’temple’, ’ground’, ’grounds’, ’chapel’, ’shelter’, ’accommodation’, ’housing’, ’site’, ’castle’, ’land’, ’facility’, ’villa’, ’tenant’, ’mine’, ’hotel’, ’house’, ’terrace’
DEPs	**cluster 1**’photographer’, ’scientist’, ’researcher’, ’author’, ’writer’, ’practitioner’ ’consultant’, ’designer’, ’lawyer’, ’architect’, ’editor’, ’politician’, ’historian’, ’teacher’ **cluster 2** ’mist’, ’wind’, ’darkness’, ’water’, ’shadow’, ’breeze’, ’rain’, ’tide’, ’sunlight’, ’snow’, ’weather’, ’flame’, ’fire’, ’dust’, ’storm’, ’dawn’, ’terror’, ’light’, ’smoke’, ’heat **cluster 3** ’pound’, ’exchange’, ’fortune’, ’coin’, ’deposit’, ’import’, ’fund’, ’penny’, ’bond’, ’rent’, ’lease’, ’market’, ’transaction’ ’share’, ’credit’, ’dollar’, ’loan’

Figure 8: **Noun clusters of different word embeddings** (only few samples)