ANLP Assignment 3: Exploring distributional similarity in Twitter

S1866666 S1870915

November 2018

1 Introduction

A limitation of statistical measures of word similarity is that a large corpus is often required to gather enough information, and hence the performance of a model is likely to be affected by the observed frequency of targets. In this work, we report a comparative study of the influence of word frequency on the estimated word similarity in 1% sample of all tweets sent during 2011. Our evaluation is based on four methods for estimating the word similarity, which are are: (1) PPMI with cosine similarity measure, (2) PMI with cosine similarity measure, (3) PPMI with Jaccard measure, and (4) second order co-occurrence PMI. We use the `wordsim353` test collection (Finkelstein et al., 2001) to test our models, as the word pairs in the collection intuitively covers all range of similarity and frequency. Also the set is sufficiently large to support our arguments based on the experiments. We found it is common of the four models that the estimate similarity is positively correlated with the minimum word frequency of a word pair, while negatively correlated with the maximum frequency of the pair. Among the four models we experimented, second order co-occurence PMI is the least influenced by the change of minimum frequency, and PPMI with cosine similarity is the least influenced by the change of maximum frequency.

Section 2 provides explanations of the methods we used in this work. Section 3 briefly discussed the test words we used for experiments. Section 4 interprets the experiments we carried out, and section 5 presents our conclusions and suggestions for future works.

2 Measuring word similarity

Pointwise mutual information (PMI) measures the likelihood of co-occurrence of two words w_1 and w_2 by

$$f^{pmi}(w_1, w_2) = \log_2 \frac{f^c(w_1, w_2)N}{f^t(w_1)f^t(w_2)} \tag{1}$$

where $f^c(w_0, w_1)$ denotes how many times the word w_0 and w_1 co-occur in a same tweet, f^t denotes how many tweets the word w_0 occurs in, and N is the total number of tweets observed. Positive PMI (PPMI) is an improvement of PMI which changes negative PMI values to 0.

Second order co-occurrence PMI (SOCPMI) (Islam & Inkpen, 2006) measures the similarity of two words by incorporating the PMI of one target word and the important context words of the other target word. Specifically, the set of important context words X for W_1 is the top-most β_1 context words sorted in descending order by their PMI value with W_1. Similarly, Y is defined as the top-most $beta_2$ context words sorted by PMI value with W_2. The *beta*'s are defined as

$$\beta_i = (\log(f^t(w_i)))^2 \frac{\log_2(T)}{\delta}, \text{ where } i = 1, 2 \tag{2}$$

where T is the total number of word types, and δ is a constant depending on the size of the corpus. Then for word w_1 and w_2, the $\beta - PMI$ summation of two words are defined as

$$f^{\beta}(w_1) = \sum_{i=1}^{\beta_1} (f^{pmi}(X_i, w_2))^{\gamma}, \text{ and } f^{\beta}(w_2) = \sum_{i=1}^{\beta_2} (f^{pmi}(Y_i, w_1))^{\gamma} \tag{3}$$

Finally, the similarity between the two target words is defined as the sum of average PMI of their top-related context words with the other target words:

$$Sim^{soc}(w_1, w_2) = \frac{f^{\beta}(w_1)}{\beta} + \frac{f^{\beta_2}(w_2)}{\beta_2} \tag{4}$$

Word embedding and cosine similarity. In word embedding, each word is represented by a sparse vector, where each value is a quantified feature of this word, e.g. PMI. The cosine similarity is a way to measure the similarity of two vectors, defined as,

$$Sim^{cos}(v_1, v_2) = \frac{v_1 \cdot v_2}{\mid v_1 \mid\mid v_2 \mid} \tag{5}$$

where $v1$ and $v2$ are vectors.

Jaccard measure is a way of measuring similarity of finite sets. Under this measure, the similarity of two words w_1 and w_2 are defined as

$$Sim^{jac}(w_1, w_2) = \frac{\mid f^s(w_1) \cap f^s(w_2) \mid}{\mid f^s(w_1) \cup f^s(w_2) \mid} \tag{6}$$

where $f^s(w_1)$ is a set containing the context words whose PMI value (or other measures) with w_1 is not 0.

3 Choosing Words

To explore the correlation between word frequency and similarity, we need words that covers a range of similarities and frequencies. Specifically, we use the `wordsim353` test collection (Finkelstein et al., 2001), which includes 353 word pairs along with their relatedness judged by human. An except of it is shown in table 1, in which we can see there are very similar pairs, e.g. money & cash, moderately similar pairs e.g. marriage & morality and not very similar pairs e.g. king & cabbage. We can also see that some are very frequent words e.g. smart, mother, and some are not very frequent e.g. manslaughter, monk. There are 267 words left after we discarded repeated words and filtered those which are not in the corpus. These words together form 35,511 pairs in total and should be sufficient for our experiments.

word1	word2	Human (mean)	word1	word2	Human (mean)
money	cash	9.15	smart	student	4.62
manslaughter	murder	8.53	marriage	morality	3.69
baby	mother	7.85	drink	mother	2.65
computer	internet	7.58	opera	industry	2.63
media	radio	7.42	delay	racism	1.19
drink	mouth	5.96	monk	slave	0.92
governor	man	5.25	king	cabbage	0.23

Table 1: An excerpt of the WordSimilarity-353 Test Collection. Each word pair is followed by an average relatedness rated by human

4 Experiments

We set up four models for the experiments: (1) PPMI with cosine similarity (PPMI-COS), (2) PMI with cosine similarity (PMI-COS), (3) PPMI with Jaccard measure (PPMI-JAC), and (4) second order PMI (SOCPMI). With each model, we computed the similarities of the 35,511 word pairs we choose, and recorded both the larger and smaller frequencies of the word pairs. Shown in the figure 1 are the scatter plots of the word similarity computed by each model against the word frequencies. We would expect an ideal model to be free from the impact of the word frequency and thus to have the same distribution of similarity over all ranges of frequency. From the plots we see that every model shows a skew of different level, meaning each of them suffers the influence of word frequency to a varying extents. In particular, it is common that the distribution of similarity is the widest in the middle and narrower to the sides. This indicates a trend that the word pairs with moderate frequencies (from about 1e4 to 1e6) has a larger chance to get a similarity higher than they does with frequencies too low or too high. To compare the models quantitatively, we evaluated the influence of frequency by Spearman rank correlation and Pearson correlation, as shown in table 2. We see that for all models, the estimate word similarity is positively correlated with the minimum word frequency and negatively correlated with the maximum word frequency. This tells us that words with lower minimum frequency tends to have a lower similarity and words with higher maximum frequency tends to have a lower similarity as well. In either case, the smaller the absolute value of the correlation,

the smaller the impact the word frequency has on the similarity. We can see that second order PMI is least affected by minimum word frequency, and PPMI with cosine similarity is least affected by maximum word frequency.

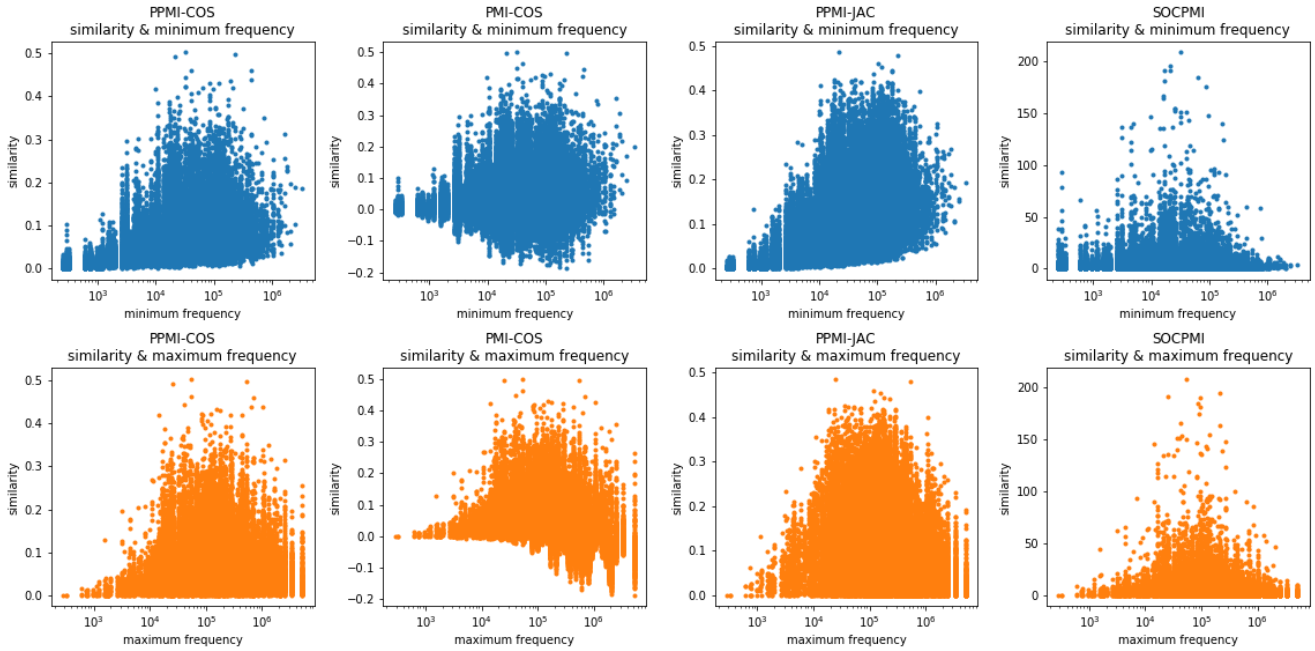

Figure 1: The Similarity distribution under varying word frequencies for each model. The frequencies are plotted in log scale.

method	minimum frequency		maximum frequency	
	pearson	spearman	pearson	spearman
PPMI-COS	0.201	0.499	**-0.124**	**-0.030**
PMI-COS	0.102	0.234	-0.291	-0.288
PPMI-JAC	0.269	0.628	-0.240	-0.171
SOCPMI	**-0.042**	**-0.003**	-0.125	-0.303

Table 2: The correlations between the similarity and word frequencies under two measures for each model

5 Conclusion

Our experiments show that the word frequency does have an impact on the distributional similarity, the extent of which depends on the methods used. In particular, the larger the minimum frequency of the two is, the higher the similarity is likely to be, and the smaller the maximum frequency is, the lower the similarity is likely to be. Therefore, it suggests that when building a statistical model for estimate word similarity, it is preferred to control the frequencies of the words. For example, we can observe same amount of target words and estimate similarity based on the co-occurrence frequency. Or, we can build models that take the effect of frequency into account, for example, scaling the similarity based on the word frequency. Among our four models, we found that second order co-occurence PMI is the most consistent with the change of minimum frequency, and PPMI with cosine similarity is the most consistent with the change of maximum frequency. Future work could aim on the development of a method for estimating similarity which is less affected by both minimum and maximum frequency than the best models evaluated in this work.

similarity	word1	word2	word1 frequency	word2 frequency
0.36	cat	dog	169733	287114
0.17	comput	mous	160828	22265
0.12	cat	mous	169733	22265
0.09	mous	dog	22265	287114
0.07	cat	comput	169733	160828
0.06	comput	dog	160828	287114
0.02	@justinbieber	dog	703307	287114
0.01	cat	@justinbieber	169733	703307
0.01	@justinbieber	comput	703307	160828
0.01	@justinbieber	mous	703307	22265

Table 3: Similarity of each test word pairs using PPMI with cosine similarity

References

Finkelstein, Lev, Gabrilovich, Evgeniy, Matias, Yossi, Rivlin, Ehud, Solan, Zach, Wolfman, Gadi, and Ruppin, Eytan. Placing Search in Context: The Concept Revisited †. Technical report, 2001. URL `http://www.cs.technion.ac.il/ gabr/papers/tois_context.pdf; http://www.cs.technion.ac.il/ gabr/resources/data/wordsim353/wordsim353.html`.

Islam, Md Aminul and Inkpen, Diana. Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. *In Proceedings of LREC 2006, Genoa, Italy*, pp. 1–6, 2006. ISSN 15564681. doi: 10.1145/1376815.1376819. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.612.1098&rep=rep1&type=pdf papers3://publication/uuid/48020432-4DD4-4900-A8A1-BC77595A8662`.