ANLP Assignment 2 Report

s2059190
s2041285

1 Investigated question

In this report, how different methods influence the similarities between words will be investigated. Four methods will be evaluated. Two of the methods are PPMI and t-test, which are used to compute context vectors of words; the other two methods are cosine similarity and euclidean distance, which are similarity measures to compare two context vectors of two words. By combining the two methods of calculating the word vector and the method of calculating the similarity, we obtain four methods of calculating the similarity between two words and apply these methods to a given word list.

2 Choices of words

The words used to investigate the methods are chosen by hand. Six groups of words are chosen based on a website[1] about synonyms of words. For each group of words, there is a word as a reference. Based on each reference word, nine words were picked by hand. Three criterion were set in the nine words: words are very similar to reference words, words are moderately similar to reference words and words are not similar to reference words. For each reference word, three words for each criterion were chosen by hand. Besides, some antonyms for the reference words were also chosen. Table 1 shows the chosen reference words and corresponding similar, moderately similar and not similar words.

3 Choices of methods

Except PPMI and cosine similarity, t-test and Euclidean distance are implemented and experiment and analysis are performed on the four methods. Below are the brief introduction of t-test and Euclidean distance.

T-test
The calculation of T-test between two words is as follows[1]. w and f represent two words. In this equation, $P(w, f)$ represents the probability of co-occurrence of word w and word f. $P(w)$ represents the probability of the occurrence of word w and $P(f)$ represents the probability of the occurrence of word f.

$$assoc_{t-test}(w, f) = \frac{P(w,f) - P(w)P(f)}{\sqrt{P(f)P(w)}}$$

Euclidean distance
The calculation of Euclidean distance between two word vectors is as follows[1]. $\overrightarrow{x}$ and $\overrightarrow{y}$ are two word context vectors. x_i represents the i-th element in context vector $\overrightarrow{x}$ and y_i represents the i-th element in context vector $\overrightarrow{y}$.

$$distance_{euclidean}(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

4 Analysis

In this experiment, we firstly use the t-test and PPMI methods to calculate the context vectors of reference words and corresponding selected words, and secondly, use these context vectors to calculate the similarity between words. In order to compare the similarity between the reference word and the corresponding selected word, for each reference word, make pairs of the reference word and each of the corresponding nine selected words. Then calculate the cosine similarity and euclidean distance between each pair of words. Finally, we sort the results (where cosine similarity is sorted according to the result from largest to smallest, and euclidean distance is sorted according to the result from smallest to largest) and analyze them. The ranking results of the four methods are shown in Figure 1.

4.1 Evaluation accuracy of cosine and euclidean distance

In this experiment, the accuracy of the similarity calculation method is calculated according to the formula $acc = \frac{count\ of\ correctly\ classified\ words}{count\ of\ corresponding\ words}$, the final result is shown in the table 2. Here, *correctly classified words* means that the category (ranking 1-3 are similar, ranking 4-6 are moderately similar, and ranking 7-9 are not similar) to which the word belongs after sorting is consistent with the category to which the word originally belongs

[1] https://www.thesaurus.com/browse/synonym

According to the data in the table, the similarity calculated by the combination of t-test and consine similarity has the highest average accuracy. Through comparison, it can be found that the similarity accuracy calculated by the consine method has a higher value for most reference words, while the similarity accuracy calculated by the euclidean distance method has a lower value for most reference words. In summary, it could be seen that the consine method is more suitable for calculation of similarity.

4.2 Calculation method comparison between cosine similarity and Euclidean distance

For each kind of context vectors, the two similarity calculation methods are compared. For some pairs of words, the rankings using cosine similarity have many differences in the rankings using Euclidean distance. For word pair [advice, tip], in the initial word selection, these two words are similar. And the cosine similarity rankings of both their PPMI and t-test rankings are the first, which means that they are similar compared to other words within group. However, for Euclidean distance rankings of the two words, PPMI pair is ranked the last and t-test pair is ranked the second last. The case is similar to word pair [beautiful, lovely], in which the two words are the most similar for cosine similarity and the least similar of Euclidean distance.

One possible reason is that for such word pairs, the two word context vectors either in the representation of PPMI or t-test, they could have vectors with small angle, thus bigger cosine similarity. However, the length of the two word vectors may be very different, thus the euclidean distance is very big although they are similar in cosine similarity. From observation, such word pairs have counts which are very different. For word pair [advice, tip], the count of word 'advic'(stemmed version for advice) is 84389 and the count of word 'tip' is 194211. For word pair [beautiful, lovely], the count for word 'beauti' (stemmed version for 'beautiful') is 629661 and the count for word 'love' (stemmed version for 'lovely') is 5178829. This may cause the same elements in context vectors have big difference and have higher Euclidean distance. For the same group, word pairs with higher similarities calculated in Euclidean distance, they have relatively smaller word count difference and fewer differences in distances, thus could have higher Euclidean similarity.

4.3 The relationship between Antonyms and similarity

When designing the experiment, we evaluated the similarity of the two words by comparing the similarity of the two words in terms of meaning according to daily language habits. This evaluation is subjective. But in this experiment, we discovered an interesting phenomenon: For word a, the similarity between a and its antonym b sometimes even exceeds the similarity between a and the word c which has a similar meaning with a.

When we designed the word_list, some reference words have included their antonyms in "not similar" word and some haven't. Therefore, we re-extract the antonym of reference words (if the reference word has one), use four methods to calculate the similarity and insert it into the similarity ranking list of the original 9 pairs of words. If the position of the antonym is "similar" (that is, the ranking is between 1-3), it means that the antonym is similar to its corresponding reference word. The antonyms we selected are as follows: [equivalent, opposite], [beautiful, ugly], [promise, break]. The results are shown in the following table 3.

From the results in the table, it can be seen that it is unlikely that antonyms and their corresponding reference words have high similarity. It could also be seen that the rankings of the three antonym word pairs are very low using Euclidean distance, and only when the cosine method is used to measure the similarity, the antonyms have high similarity may appear. Cosine similarity of word pair [promise, break] is very high for either PPMI context vectors or t-test context vectors. It may be because that the two words could appear in similar contexts. For example, for sentences with 'break a promise', both break and promise could form similar context vector elements.

5 Conclusion

In this report, three questions are evaluated in section 4. For question 4.1, by observing the cosine similarity and Euclidean distance of the word pairs, we get accuracy for each of the four methods and find that t-test + cosine similarity is more proper than the other three methods in measuring similarity between two words. For question 4.2, by observing the word pairs with high cosine similarity rankings but low Euclidean distance similarity rankings, we analysed that words vectors with small angles does not mean that they have similar vector lengths. For question 4.3, we find that it is not absolutely correct to measure the similarity between two words only by semantic similarity. There may be cases where antonyms are more similar than synonyms by cosine similarity, but this situation does not seem to be universal at present because the amount of antonym pairs are not large enough in this experiment.

References

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

Reference word	similar			moderately similar			not similar		
	w1	w2	w3	w1	w2	w3	w1	w2	w3
information	knowledge	message	data	network	scoop	tidings	ignorance	question	silence
promise	guarantee	assurance	commitment	swear	contract	engagement	marriage	renege	break
advice	recommendation	suggestion	tip	guidance	persuasion	advisement	teaching	forewarning	word
warning	alarm	alert	caution	injunction	tip	forewarning	misinformation	assurance	happy
equivalent	equal	identical	corresponding	like	interchangeable	amounting	near	close	parallel
beautiful	lovely	pretty	stunning	cute	appealing	graceful	fascinating	pleasing	charming

Table 1: word list

	ppmi+cos	ppmi+Euclidean	ttest + cos	ttest + Euclidean
information	0.33	0.11	0.56	0.33
promise	0.33	0.44	0.33	0.44
advice	0.44	0.33	0.44	0.22
warning	0.67	0.44	0.56	0.33
equivalent	0.22	0.44	0.33	0.44
beautiful	0.78	0.22	0.78	0.33
average	0.368	0.423	0.425	0.293

Table 2: The accuracy of 4 similarity calculation methods on different word lists

	ppmi+cos	ppmi+Euclidean	ttest + cos	ttest + Euclidean
equivalent-opposite	7	9	7	8
promise-break	2	8	1	7
beautiful-ugly	6	5	5	5

Table 3: The ranking of antonyms under different calculation methods

information

ppmi + cos		ppmi + Euclidean		ttest + cos		ttest + Euclidean	
data	0.37	scoop	176.45	data	0.25	scoop	0.17
network	0.29	knowledge	179.72	network	0.18	ignorance	0.20
knowledge	0.27	tidings	180.77	knowledge	0.14	knowledge	0.21
question	0.23	question	186.82	question	0.10	tidings	0.24
message	0.13	silence	194.19	message	0.07	silence	0.26
ignorance	0.09	ignorance	194.73	scoop	0.02	message	0.27
silence	0.08	message	210.60	ignorance	0.02	question	0.27
scoop	0.07	network	219.48	tidings	0.01	data	0.30
tidings	0.06	data	226.37	slience	0.01	network	0.37
acc=1/3=0.33		acc=1/9=0.11		acc=5/9=0.56		acc=1/3=0.33	

promise

ppmi + cos		ppmi + Euclidean		ttest + cos		ttest + Euclidean	
commitment	0.17	renege	120.54	break	0.08	renege	0.17
break	0.14	assurance	141.65	swear	0.08	swear	0.19
marriage	0.10	swear	148.08	commitment	0.06	assurance	0.19
assurance	0.09	guarantee	164.94	guarantee	0.04	guarantee	0.21
guarantee	0.09	engagement	168.57	engagement	0.03	engagement	0.23
swear	0.09	commitment	168.92	assurance	0.03	contract	0.24
engagement	0.08	marriage	196.53	marriage	0.02	break	0.26
contract	0.06	break	196.88	contract	0.01	commitment	0.26
renege	0.00	contract	209.69	renege	0.00	marriage	0.32
acc=1/3=0.33		acc=4/9=0.44		acc=1/3=0.33		acc=4/9=0.44	

advice

ppmi + cos		ppmi + Euclidean		ttest + cos		ttest + Euclidean	
tip	0.26	forewarning	114.72	tip	0.24	forewarning	0.11
advisement	0.24	persuasion	126.48	advisement	0.16	persuasion	0.12
suggestion	0.24	guidance	128.19	guidance	0.12	guidance	0.13
recommendatio	0.22	suggestion	135.03	recommendatio	0.11	advisement	0.14
teaching	0.18	advisement	135.32	teaching	0.10	recommendatio	0.17
guidance	0.18	recommendatio	149.92	suggestion	0.08	teaching	0.18
word	0.12	teaching	163.03	word	0.08	suggestion	0.20
persuasion	0.09	word	193.91	persuasion	0.04	tip	0.21
forewarning	0.01	tip	196.87	forewarning	0.00	word	0.28
acc=4/9=0.44		acc=1/3=0.33		acc=4/9=0.44		acc=2/9=0.22	

warning

ppmi + cos		ppmi + Euclidean		ttest + cos		ttest + Euclidean	
alert	0.34	alert	114.72	alert	0.24	forewarning	0.11
assurance	0.12	caution	126.48	caution	0.16	caution	0.12
caution	0.12	misinformatio	128.19	tip	0.12	assurance	0.13
alarm	0.12	assurance	135.03	assurance	0.11	alert	0.14
tip	0.12	alert	135.32	alarm	0.10	misinformatio	0.17
injunction	0.06	injunction	149.92	injunction	0.08	injunction	0.18
misinformatio	0.04	alarm	163.03	misinformatio	0.08	tip	0.20
happy	0.02	tip	193.91	forewarning	0.04	alarm	0.21
forewarning	0.00	happy	196.87	happy	0.00	happy	0.28
acc=2/3=0.67		acc=4/9=0.44		acc=5/9=0.56		acc=1/3=0.33	

equivalent

ppmi + cos		ppmi + Euclidean		ttest + cos		ttest + Euclidean	
amounting	0.34	parallel	156.08	interchangeab	0.01	near	1.14
interchangeab	0.12	near	162.28	parallel	0.00	corresponding	1.14
equal	0.12	corresponding	162.54	identical	0.00	parallel	1.15
identical	0.12	interchangeab	167.46	amounting	0.00	interchangeab	1.15
parallel	0.12	identical	170.85	equal	0.00	amounting	1.15
near	0.06	amounting	178.49	near	0.00	identical	1.15
corresponding	0.04	equal	191.23	corresponding	0.00	equal	1.15
close	0.02	like	200.20	close	0.00	like	1.17
like	0.00	close	222.91	like	0.00	close	1.17
acc=2/9=0.22		acc=4/9=0.44		acc=1/3=0.33		acc=4/9=0.44	

beautiful

ppmi + cos		ppmi + Euclidean		ttest + cos		ttest + Euclidean	
lovely	0.24	pretty	156.08	lovely	0.01	fascinating	1.14
stunning	0.20	pleasing	162.28	pretty	0.00	stunning	1.14
pretty	0.20	fascinating	162.54	stunning	0.00	pleasing	1.15
cute	0.17	stunning	167.46	cute	0.00	pretty	1.15
charming	0.15	charming	170.85	charming	0.00	cute	1.15
graceful	0.10	graceful	178.49	graceful	0.00	appealing	1.15
fascinating	0.08	cute	191.23	fascinating	0.00	graceful	1.15
pleasing	0.06	appealing	200.20	pleasing	0.00	charming	1.17
appealing	0.05	lovely	222.91	appealing	0.00	lovely	1.17
acc=7/9=0.78		acc=2/9=0.22		acc=7/9=0.78		acc=1/3=0.0.33	

Figure 1: The similarity ranking of 6 reference words. The 6 tables reflect the similarity ranking of the 9 corresponding words under the four methods, and The data in the table retain 2 decimal places.