SemEval-2012 Task 4: Evaluating Chinese Word Similarity

Peng Jin
School of Computer Science
Leshan Normal University
Leshan, 614000, China
jandp@pku.edu.cn

Yunfang Wu
Institute of Computational Linguistics
Peking University
Beijing, 100871, China
wuyf@pku.edu.cn

Abstract

This task focuses on evaluating word similarity computation in Chinese. We follow the way of Finkelstein et al. (2002) to select word pairs. Then we organize twenty undergraduates who are major in Chinese linguistics to annotate the data. Each pair is assigned a similarity score by each annotator. We rank the word pairs by the average value of similar scores among the twenty annotators. This data is used as gold standard. Four systems participating in this task return their results. We evaluate their results on gold standard data in term of Kendall's tau value, and the results show three of them have a positive correlation with the rank manually created while the taus' value is very small.

1 Introduction

The goal of word similarity is to compute the similarity degree between words. It is widely used in natural language processing to alleviate data sparseness which is an open problem in this field. Many research have focus on English language (Lin, 1998; Curran and Moens, 2003; Dinu and Lapata, 2010), some of which rely on the manual created thesaurus such as WordNet (Budanitsky and Hirst, 2006), some of which obtain the similarity of the words via large scale corpus (Lee, 1999), and some research integrate both thesaurus and corpus (Fujii et al., 1997). This task tries to evaluate the approach on word similarity for Chinese language. To the best of our knowledge, this is first release of benchmark data for this study.

In English language, there are two data sets: Rubenstein and Goodenough (1965) and Finkelstein et al. (2002) created a ranking of word pairs as the benchmark data. Both of them are manually annotated. In this task, we follow the way to create the data and annotate the similarity score between word pairs by twenty Chinese native speakers. Finkelstein et al. (2002) carried out a psycholinguistic experiment: they selected out 353 word pairs, then ask the annotators assign a numerical similarity score between 0 and 10 (0 denotes that words are totally unrelated, 10 denotes that words are VERY closely related) to each pair. By definition, the similarity of the word to itself should be 10. A fractional score is allowed.

It should be noted that besides the rank of word pairs, the thesaurus such as Roget's thesaurus are often used for word similarity study (Gorman and Curran, 2006).

The paper is organized as follows. In section 2 we describe in detail the process of the data preparation. Section 3 introduces the four participating systems. Section 4 reports their results and gives a brief discussion.. And finally in section 5 we bring forward some suggestions for the next campaign and conclude the paper.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 374–377,
Montréal, Canada, June 7-8, 2012.

2 Data Preparation

2.1 Data Set

We use wordsim 353 (Finkelstein et al., 2002) as the original data set. First, each word pair is translated into Chinese by two undergraduates who are fluent in English. 169 word pairs are the same in their translation results. To the rest 184 word pairs, the third undergraduate student check them following the rules:

(i) Single character vs. two characters. If one translator translate one English word into the Chinese word which consists only one Chinese character and the other use two characters to convey the translation, we will prefer to the later provided that these two translations are semantically same. For example, "tiger" is translated into "虎" and "老虎", we will treat them as same and use "老虎" as the final translation. This was the same case in "drug" ("药" and "药物" are same translations).

(ii) Alias. The typical instance is "potato", both "土豆" and "马铃薯" are the correct translations. So we will treat them as same and prefer "土豆" as the final translation because it is more general used than the latter one.

(iii) There are five distinct word pairs in the translations and are removed.

At last, 348 word pairs are used in this task. Among these 348 word pairs, 50 ones are used as the trial data and the rest ones are used as the test data[1].

2.2 Manual Annotation

Each word pair is assigned the similarity score by twenty Chinese native speakers. The score ranges from 0 to 5 and 0 means two words have nothing to do with each other and 5 means they are identically in semantic meaning. The higher score means the more similar between two words. Not only integer but also real is acceptable as the annotated score. We get the average of all the scores given by the annotators for each word pair and then sort them according to the similarity scores. The distribution of word pairs on the similar score is illustrated as table 1.

Score	0.0-1.0	1.0-2.0	2.0-3.0	3.0-4.0	4.0-5.0
# Word pairs	39	90	132	72	13

Table1: The distribution of similarity score

Rank	Word in Chinese/English	Word 2 in Chinese/ English	Similarity score	Std. dev	RSD (%)
1	足球/football	足球/soccer	4.98	0.1	2.0
2	老虎/tiger	老虎/tiger	4.89	0.320	6.55
3	恒星/planet	恒星/star	4.72	0.984	20.8
4	入场券 /admission	门票/ticket	4.60	0.516	11.2
5	钱/money	现金/cash	4.58	0.584	12.7
6	银行/bank	钱/cash	4.29	0.708	16.5
7	手机/cell	电话/phone	4.28	0.751	17.5
8	宝石/gem	珠宝/jewel	4.24	0.767	18.1
9	类型/type	种类/kind	4.24	1.000	23.6
10	运算 / calculation	计算 / computation	4.14	0.780	19.0
Avg	-	-	4.496	0.651	14.80

Table 2: Top ten similar word pairs

Table 2 and table 3 list top ten similar word pairs and top ten un-similar word pairs individually. Standard deviation (Std. dev) and relative standard deviation (RSD) are also computed. Obviously, the relative standard deviation of top ten similar word pairs is far less than the un-similar pairs.

2.3 Annotation Analysis

Figure 1 illustrates the relationship between the similarity score and relative standard deviation. The digits in "x" axes are the average similarity score of every integer interval, for an instance, 1.506 is the average of all word pairs' similarity score between 1.0 and 2.0.

3 Participating Systems

Four systems coming from two teams participated in this task.

[1] In fact there are 297 word pairs are evaluated because one pair is missed during the annotation.

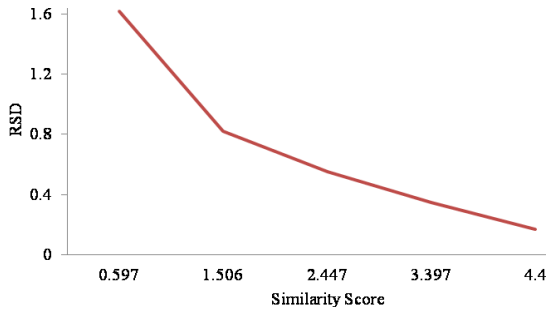

Figure 1. The relationship between RSD and similar score

Rank	Word1 in Chinese/in English	Word2 in Chinese/in English	Similarity score	Std. dev	RSD(%)
1	中午/noon	线绳/string	0.06	.213	338.7
2	国王/king	卷心菜/cabbage	0.16	.382	245.3
3	产品/production	徒步/hike	0.17	.432	247.5
4	延迟/delay	种族主义/racism	0.26	.502	191.1
5	教授/professor	黄瓜/cucumber	0.30	.62	211.1
6	股票/stock	美洲虎/jaguar	0.30	.815	268.2
7	签名/sign	暂停/recess	0.30	.655	215.4
8	股票/stock	CD/CD	0.31	.540	173.6
9	喝/drink	耳朵/ear	0.31	.833	264.8
10	公鸡/rooster	航程/voyage	0.33	.771	236.7
Avg	-	-	0.25	.576	239.2

Table 3: Top ten un-similar word pairs

MIXCC: This system used two machine readable dictionary (MRD), HIT IR-Lab Tongyici Cilin (Extended) (Cilin) and the other is Chinese Concept Dictionary (CCD). The extended CiLin consists of 12 large classes, 97 medium classes, 1,400 small classes (topics), and 17,817 small synonym sets which cover 77,343 head terms. All the items are constructed as a tree with five levels. With the increasing of levels, word senses are more fine-grained. The Chinese Concept Dictionary is a Chinese WordNet produced by Peking University. Word concepts are presented as synsets corresponding to WordNet 1.6. Besides synonym, antonym, hypernym/hyponym, holonym/meronym, there is another semantic relation type named as *attribute* which happens between two words with different part-of-speeches.

They first divide all word pairs into five parts and rank them according to their levels in Cilin in descending order. For each part, they computed word similarity by Jiang and Conrath (1997) method[2].

MIXCD: Different form MIXCC, this system used the trial data to learn a multiple linear regression functions. The CCD was considered as a directed graph. The nodes were synsets and edges were the semantic relations between two synsets. The features for this system were derived from CCD and a corpus and listed as follows:

- the shortest path between two synsets which contain the words
- the rates of 5 semantic relation types
- mutual information of a word pair in the corpus

They used the result of multiple linear regressions to forecast the similarity of other word pairs and get the rank.

GUO-ngram: This system used the method proposed by (Gabrilovich and Markovitch, 2007). They downloaded the Wikipedia on 25th November, 2011 as the knowledge source. In order to bypass the Chinese segmentation, they extract one character (uni-gram) and two sequential characters (bi-gram) as the features.

GUO-words: This system is very similar to GUO-ngram except that the features consist of words rather than n-grams. They implemented a simple index method which searches all continuous character strings appearing in a dictionary. For example, given a text string ABCDEFG in which ABC, BC, and EF appear in the dictionary. The output of the tokenization algorithm is the three words ABC, BC, EF and the two characters E and G.

[2] Because there is no sense-tagged corpus for CCD, the frequency of each concept was set to 1 in this system.

4 Results

Each system is required to rank these 500 word pairs according to their similarity scores. Table 4 gives the overall results obtained by each of the systems.

Rank	Team ID	System ID	Tau's value
1	lib	MIXCC	0.050
2		MIXCD	0.040
3	Gfp1987	Guo-ngram	0.007
4		Guo-words	-0.011

Table 4: The results of four systmes

The ranks returned by these four systems will be compared with the rank from human annotation by the Kendall Rank Correlation Coefficient:

$$\tau = 1 - \frac{2S(\pi,\sigma)}{N(N-1)/2}$$

Where N is the number of objects. π and σ are two distinct orderings of a object in two ranks. $S(\pi,\sigma)$ is the minimum number of adjacent transpositions needing to bring π and σ (Lapata, 2006). In this metric, tau's value ranges from -1 to +1 and -1 means that the two ranks are inverse to each other and +1 means the identical rank.

From table 4, we can see that except the final system, three of them got the positive tau's value. It is regret that the tau's is very small even if the MIXCC system is the best one.

5 Conclusion

We organize an evaluation task focuses on word similarity in Chinese language. Totally 347 word pairs are annotated similarity scores by twenty native speakers. These word pairs are ordered by the similarity scores and this rank is used as benchmark data for evaluation.

Four systems participated in this task. Except the system MIXCD, three ones got their own rank only via the corpus. Kendall's tau is used as the evaluation metric. Three of them got the positive correlation rank compared with the gold standard data

Generally the tau's value is very small, it indicates that obtaining a *good* rank is still difficult. We will provide more word pairs and distinct them relatedness from similar, and attract more teams to participate in the interesting task.

Acknowledgments

This research is supported by National Natural Science Foundation of China (NSFC) under Grant No. 61003206, 60703063.

References

A. Budanitsky and G. Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 2006, 32(1):13-47.

J. Curran and M. Moens. Scaling Context Space. *Proceedings of ACL*, 2002, pp. 231-238.

G. Dinu and M. Lapata. Measuring Distributional Similarity in Context. *Proceedings of EMNLP*, 2010, pp. 1162-1172.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116-131.

A. Fujii, T. Hasegawa, T. Tokunaga and H. Tanaka. Integration of Hand-Crafted and Statistical Resources in Measuring Word Similarity. 1997. *Proceedings of Workshop of Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. pp. 45-51.

E. Gabrilovich and S. Markovitch, Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, *Proceedings of IJCAI,* Hyderabad, 2007, pp. 1606—1611.

J. Gorman and J. Curran. Scaling Distributional Similarity to Large Corpora. *Proceedings of ACL*, 2006, pp. 361-368.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.

M. Lapata. Automatic Evaluation of Information Ordering: Kendall's Tau. *Computational Linguistics*, 2006, 32(4):471-484.

D. Lin. Automatic Retrieval and Clustering of Similar Words. *Proceedings of ACL / COLING*, 1998, pp. 768-774.

L. Lee. Measures of Distributional Similarity. *Proceedings of ACL*, 1999, pp. 25-32.

H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627-633.