MIXCD: System Description for Evaluating Chinese Word Similarity at SemEval-2012

Yingjie Zhang
Nanjing University
22 Hankou Road
Jiangsu P. R. China
jillzhyj@139.com

Bin Li
Nanjing University
Nanjing Normal University
122 Ninghai Road
Jiangsu P. R. China
gothere@126.com

Xinyu Dai
Nanjing University
22 Hankou Road
Jiangsu P. R. China
dxy@nju.edu.cn

Jiajun Chen
Nanjing University
22 Hankou Road
Jiangsu P. R. China
cjj@nju.eud.cn

Abstract

This document describes three systems calculating semantic similarity between two Chinese words. One is based on Machine Readable Dictionaries and the others utilize both MRDs and Corpus. These systems are performed on SemEval-2012 Task 4: Evaluating Chinese Word Similarity.

1 Introduction

The characteristics of polysemy and synonymy that exist in words of natural language have always been a challenge in the fields of Natural Language Processing (NLP) and Information Retrieval (IR). In many cases, humans have little difficulty in determining the intended meaning of an ambiguous word, while it is extremely difficult to replicate this process computationally. For many tasks in psycholinguistics and NLP, a job is often decomposed to the requirement of resolving the semantic similarity between words or concepts.

There are two ways to get the similarity between two words. One is to utilize the machine readable dictionary (MRD). The other is to use the corpus.

For the 4th task in SemEval-2012 we are required to evaluate the semantic similarity of Chinese word pairs. We consider 3 methods in this study. One uses MRDs only and the other two use both MRD and corpus. A post processing will be done on the results of these methods to treat synonyms.

In chapter 2 we introduce the previous works on the evaluation of Semantic Similarity. Chapter 3 shows three methods used in this task. Chapter 4 reveals the results of these methods. And conclusion is stated in chapter 5.

2 Related Work

For words may have more than one sense, similarity between two words can be determined by the best score among all the concept pairs which their various senses belong to.

Before constructed dictionary is built, Lesk similarity (Lesk, 1986) which is proposed as a solution for word sense disambiguation is often used to evaluating the similarity between two concepts. This method calculates the overlap between the corresponding definitions as provided by a dictionary.

$$sim_{Lesk}(c_1, c_2) = |gloss(c_1) \cap gloss(c_2)|$$

Since the availability of computational lexicons such as WordNet, the taxonomy can be represented as a hierarchical structure. Then we use the structure information to evaluate the semantic similarity. In these methods, the hierarchical structure is often seen as a tree and concepts as the nodes of the tree while relations between two concepts as the edges.

(Resnik, 1995) determines the conceptual similarity of two concepts by calculating the information content (IC) of the least common subsumer (LCS) of them.

$$sim_{res}(c_1, c_2) = IC(LCS(c_1, c_2))$$

where the IC of a concept can be quantified as follow:

$$IC(\mathrm{c}) = log^{-1}P(c)$$

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 425–429,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

This method do not consider the distance of two concepts. Any two concepts have the same LCS will have the same similarity even if the distances between them are different. It is called node-based method.

(Leacock and Chodorow, 1998) develops a similarity measure based on the distance of two senses c_1 and c_2. They focus on hypernymy links and scaled the path length by the overall depth D of the tree.

$$sim_{lch}(c_1, c_2) = -log\frac{length(c_1, c_2)}{2 \times D}$$

(Wu and Palmer, 1994) combines the depth of the LCS of two concepts into a similarity score.

$$sim_{wup}(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

These approaches are regarded as edge-based methods. They are more natural and direct to evaluating semantic similarity in taxonomy. But they treat all nodes as the same and do not consider the different information of different nodes.

(Jiang and Conrath, 1998) uses the information content of concept instead of its depth. So both node and edge information can be considered to evaluate the similarity. It performs well in evaluating semantic similarity between two texts (Zhang et al., 2008; Corley and Mihalcea, 2005; Pedersen, 2010).

$$sim_{jnc}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2))}$$

SemCor is used in Jiang's work to get the frequency of a word with a specific sense treated by the Lagrange Smoothing.

3 Approaches

For SemEval-2012 task 4, we use two MRDs and one corpus as our knowledge resources. One MRD is HIT IR-Lab Tongyici Cilin (Extended) (Cilin) and the other is Chinese Concept Dictionary (CCD). The corpus we used in our system is People's Daily. Three systems are proposed to evaluate the semantic similarity between two Chinese words. The first one utilizes both the MRDs called MIXCC (Mixture of Cilin and CCD) and other two named MIXCD1 (Mixture of Corpus and Dictionary) and MIXCD2 respectively combine the information derived from both corpus and dictionary into the similarity score. A post processing is done to trim the similarity of words with the same meaning.

3.1 Knowledge Resources

HIT IR-Lab Tongyici Cilin (Extended) is built by Harbin Institute of Technology which contained 77343 word items. Cilin is constructed as a tree with five levels. With the increasing of the level, word senses are more fine-grained. All word items in Cilin are located at the fifth level. The larger level the LCS of an item pair has, the closer their concepts are.

Chinese Concept Dictionary (CCD) is a Chinese WordNet produced by Peking University. Word concepts in it are represented as Synsets and one-one corresponding to WordNet 1.6. There are 4 types of hierarchical semantic relations in CCD as follows:

- Synonym: the meanings of two words are equivalence
- Antonym: two synsets contain the words with opposite meaning
- Hypernym and Hyponym: two synsets with the IS-A relation
- Holonym and Meronym: two synsets with the IS-PART-OF relation

Additionally there is another type of semantic relation such as Attribute in CCD This relation type often happens between two words with different part-of-speech. Even though it is not the hierarchical relation, this relation type can make two words with different POS have a path between them. In WordNet it is often shown as a Morphological transform between two words, while it may happen on two different words with closed meaning in CCD.

The corpus we use in our system is People's Daily 2000 from January to June which has been manually segmented.

3.2 MIXCC

MIXCC utilizes both Cilin and CCD to evaluate the semantic similarity of word pair. In this method we get the rank in three steps.

First, we use Cilin to separate the list of word pairs into five parts and sort them in descending order of LCS's level. The word pairs having the same level of LCS will be put in the same part.

Second, for each part we compute the similarity almost by Jiang and Conrath's method mentioned in Section 2 above. Only Synonym and Hypernym-Hyponym relations of CCD concepts are considered in this method. So CCD could be constructed as a forest. We add a root node which combined the forest into a tree to make sure that there is a path between any two concepts.

$$\mathrm{sim}_{jnc}(\mathrm{w}_1, \mathrm{w}_2) = \max_{\substack{\mathrm{c}_1 \in \mathrm{concept}(\mathrm{w}_1) \\ \wedge \mathrm{c}_2 \in \mathrm{concept}(\mathrm{w}_2)}} \mathrm{sim}_{\mathrm{jnc}}(\mathrm{c}_1, \mathrm{c}_2)$$

w_1 and w_2 compose a word pair needed to calculate semantic similarity between them. c_1 (c_2) is the Synset in CCD which contains w_1 (w_2).

Because there is no sense-tagged corpus for CCD, the frequency of every word in each concept is always 1.

After $\mathrm{sim}_{\mathrm{jnc}}(\mathrm{w}_1, \mathrm{w}_2)$ of all word pairs in the same part are calculated, we sort the scores in a decreasing order again. Then we get five groups of ranked word pairs.

At last the five groups are combined together as the result shown in table 1.

3.3 MIXCD

MIXCD combines the information of corpus and MRDs to evaluate semantic similarity.

In this system we use trial data to learn a multiple linear regression function. There are two classes of features for this study which are derived from CCD and People's Daily respectively. One class of feature is the mutual information of a word pair and the other is the shortest path between two concepts containing the words of which the similarity needed to be evaluated.

We consider CCD as a large directed graph. The nodes of the graph are Synsets and edges are the semantic relations between two Synsets. All five types of semantic relation showed in Section 3.1 will be used to build the graph.

For each word pair, the shortest path between two Synsets which contain the words respectively is found. Then the path is represented in two forms.

In one form we record the vector consisting of the counts of every relation type in the path. The system using this path's form is called MIXCD0.

For example the path between "心理学 (psychology)" and "精神病学 (psychiatry)" is represented as (0, 0, 3, 2, 0). It means that "心理学" and "精神病学" are not synonym and the shortest path between them contained 3 IS-A relations and 2 IS-PART-OF relations.

We suppose that the path's length is a significant feature to measure the semantic similarity of a word pair. So in the other form the length is added into the vector as the first component. And the counts of each relation are recorded in proportion to the length. This form of path representation is used in the submitted system called MIXCD. Then the path between "心理学" and "精神病学" is represented as (5, 0, 0, 0.6, 0.4, 0).

In both forms, the Synonym feature will be 1 if the length of the path is 0.

The mutual information of all word pairs is calculated via the segmented People's Daily.

Last we use the result of multiple linear regression to forecast the similarity of other word pairs and get the rank.

3.4 Post Processing

The word pair with the same meaning may be consisted of two same words or two different words belong to the same concept. It is difficult for both systems to separate one from the other. Therefore we display a post processing on our systems to make sure that the similarity between the same words has a larger rank than two different words of the same meaning.

4 Experiments and Results

We perform our systems on trial data and then use Kendall tau Rank Correlation (Kendall, 1995; Wessa, 2012) to evaluate the results shown in Table 1. The trial data contains 50 word pairs. The similarity of each pair is scored by several experts and the mean value is regarded as the standard answer to get the manual ranking.

Method	Kendall tau	2-sided p value
MIXCC	**0.273469**	0.005208
MIXCD0	0.152653	0.119741
MIXCD	0.260408	0.007813
Manual(upper)	0.441633	6.27E-06

Table 1: Kendall tau Rank Correlation of systems on trial

From Table 1, we can see the tau value of MIXCD0 is 0.1526 and MIXCD is 0.2604. MIXCD performed notably better than MIXCD0. It shows that path's length between two words is on an important position of measuring semantic similarity. This feature does improve the similarity result. The 2-sided p value of MIXCD0 is 0.1197. It is much larger than the value of MIXCD which is 0.0078. So the ranking result of MIXCD0 is much more occasional than result of MIXCD.

The tau value of MIXCC is 0.2735 and it is much smaller than the manual ranking result which is 0.4416 seen as the upper bound. It shows that the similarity between two words in human's minds dose not only depend on their hierarchical relation represented in Dictionary. But the value is larger than that of MIXCD. It seems that the mutual information derived from corpus which is expected to improve the result reduces the correction of rank result contrarily. There may be two reasons on it.

First, because of the use of trial data in MIXCD, the result of similarity ranking strongly depended on this data. The reliability of trial data's ranking may influent the performance of our system. We calculate the tau value between every manual and the correct ranking. The least tau value is 0.4416 and the largest one is 0.8220 with a large disparity. We use the Fleiss' kappa value (Fleiss, 1971) to evaluate the agreement of manual ranking and the result is 0.1526 which showed the significant disagreement. This disagreement may make the regression result cannot show the relation between features and score correctly. To reduce the disagreement's influence we calculate the mean of manual similarity score omitting the maximum and minimum ones and get a new standard rank (trial2). Then we perform MIXCD on trail2 and show the new result as MIXCD-2 in Form 2. MIXCC's result is also compared with trail2 shown as MIXCC-2.

	MIXCC-2	MIXCD-2	MIXCC	MIXCD
Kendall tau	**0.297959**	0.265306	0.273469	0.260408

Table 2: tau value on new standard (omit max/min manual scores)

From Table 2 we can see the tau values of MIXCC rose to 0.2980 and MIXCD to 0.2653. It shows that omitting the maximum and minimum manual scores can reduce some influence of the disagreement of artificial scoring.

Second, the combination method of mutual information and semantic path in MRD may also influent the performance of our system. The ranks between MIXCD and MIXCC are also compared and the tau value is 0.2065. It shows a low agreement of semantic similarity measurements between MRD and Corpus. The mutual information exerts a large influence on the measure of similarity and sometimes may bring the noise to the result making it worse.

We also perform our systems on test data containing 297 words pairs in the same form of trial data and got the follow result:

Method	Kendall tau
MIXCC	0.050
MIXCD0	-0.064
MIXCD	0.040

Table 3 tau values of the result of test data

The ranking on test data of our systems shows an even worse result. Because of the low confidence of trial data ranking, multiple linear regression function learning from the trial data performs bad on other word pairs.

5 Conclusion

In this paper we propose three methods to evaluate the semantic similarity of Chinese word pairs. The first one uses MRDs and the second one adds the information derived from corpus. The third one uses the same knowledge resources as the second one but highlights the path length of the word pair. The results of the systems show a large difference and all have a low score. From the results we can see the similarity showed in corpus is much different from the one expressed in MRD. One reason of the low score is that the manual rank given by the task has a low agreement among them. We get a new manual rank which reduces some influence of disagreement by calculating the mean value of scores omitting the maximum and minimum ones. Comparing the result of our systems with the new ranking, all of them get a higher tau value.

Acknowledgement

This paper is supported in part by National Natural Science Fund of China under contract 61170181, National Social Science Fund of China under contract 10CYY021, State Key Lab. for Novel Software Technology under contract KFKT2011B03, Jiangsu PostDoc Fund under contract 1101065C.

References

Mike E. Lesk, 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference* 1986, Toronto, June.

Philip Resnik, 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada.

Claudia Leacock and Martin Chodorow, 1998. Combining local context and WordNet sense similiarity for word sense disambiguation. In *WordNet, An Electronic Lexical Database*. The MIT Press.

Zhibiao Wu and Martha Palmer, 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico.

Jay J. Jiang and David W. Conrath, 1998. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*.

Ce Zhang , Yu-Jing Wang , Bin Cui , Gao Cong, 2008. Semantic similarity based on compact concept ontology. In *Proceeding of the 17th international conference on World Wide Web*, April 21-25, 2008, Beijing, China

Courtney Corley , Rada Mihalcea, 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, p.13-18, June 30-30, 2005, Ann Arbor, Michigan.

Ted Pedersen, 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p.329-332, June 02-04, 2010, Los Angeles, California.

M. G. Kendall, 1955. *Rank Correlation Methods*. New York: Hafner Publishing Co.

P. Wessa, 2012. *Free Statistics Software, Office for Research Development and Education*, version 1.1.23-r7, URL http://www.wessa.net/

Jordan L. Fleiss, 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382.