MEASURING SENTENCE SIMILARITY FROM DIFFERENT ASPECTS

LIN LI, XIA HU, BI-YUN HU, JUN WANG, YI-MING ZHOU

School of Computer Science and Engineering, Beihang University, Beijing 100191, China
E-MAIL: lilinbuaa@cse.buaa.edu.cn, huxia@cse.buaa.edu.cn, hubiyun@cse.buaa.edu.cn, junwang8151@163.com, zhouyiming@buaa.edu.cn

Abstract:

The paper proposes to determine sentence similarities from different aspects. Based on the information people get from a sentence, *Objects-Specified Similarity*, *Objects-Property Similarity*, *Objects-Behavior Similarity* and *Overall Similarity* are defined to determine sentence similarities from four aspects. Experiments show that the proposed method makes the sentence similarity comparison more exactly and give out a more reasonable result, which is similar to the people's comprehension to the meanings of the sentences.

Keywords:

Sentence similarity; Chunking; Semantic vector

1. Introduction

At present a lot of methods to compute sentence similarities have been proposed in the research and application community of text knowledge representation and discovery, such as text mining, information extraction [1], automatic question-answering [2, 3], text summarization [4], text classification [5] and machine translation [6]. To deliver the sentence meaning more exactly, nowadays more and more applications require not only comparing the overall similarity between sentences but also the similarity between parts of these sentences. In daily life, people can evaluate sentence meaning from different aspects. For two sentences, “Young people like running.” “Old people like walking.” From the overall meaning, both sentences say that people like exercises, which expresses a strong similarity. But considering subjects and objects, there exists a significant difference that different people prefer different exercises.

To simulate human’s comprehension to sentence meaning and make sentence similarity comparison more meaningful, we propose to measure sentence similarities from different aspects. In this paper, based on information people draw from a sentence, we define sentence similarities by sentence chunking to represent the above differences.

The paper is organized as follows. Section 2 reviews some related work. The detailed new definitions of sentence similarities and their implementations are described in section 3. Section 4 presents the experiments to evaluate our work. Finally, Section 5 summarizes our work, draws some conclusions, and proposes future works.

2. Related work

Both semantic and syntactic information make contributions to the meaning of a sentence. When comparing sentence similarity, previous methods usually consider semantic, POS, syntactic (word order) information or their combinations, and give out an overall similarity of two compared sentences. Mandreoli et al. [7] proposed a method adopting the Edit Distance as similarity measure between (parts of) sentences, and the method mainly pays attention to the similarity of syntax structure. Hatzivassiloglou et al. [8] presented a composite similarity metric over short passages, which only utilize semantic information. Mihalcea et al. [9] developed a method to score the semantic similarity of sentences by exploiting the information that can be drawn from the similarity of the component words, but the syntax structure is ignored. Li et al. [10] presented a method that combines semantic and word order information. Liu et al. [11] also gave out another method that combines semantic and word order information. In [12], Liu et al. measured semantic similarity between sentences by Dynamic Time Warping (DTW) technique based on the analysis of parts of speech (POS). Palakorn et al. [3] tested different combinations of sentence vector similarity, word order similarity, POS similarity and also took into account question category similarity to measure the question similarities.

We can see that all of the above methods take sentences as a whole to compare their similarities, which insufficiently exploit the sentence information and not suitable for some applications.

978-1-4244-3703-0/09/$25.00 ©2009 IEEE

3. Methodologies

Due to the complexity of natural languages, only a few types of sentences in text have all the three components of subject, predicate verb and object with normal orders, a lot of compound and short sentences exist with absent or complemental components, or reversed order. In natural language processing, people usually use parsing to find out the detailed information in sentences. At present, the cost of parsing is expensive in time and resources, and the accuracy always proves unsatisfied. So except those applications that really need to compare similarities between the subjects, predicate verbs, objects or other components in sentences, it is much inefficient and even unrealistic to compare sentence similarities based on their fully parsed trees.

In order to compare sentence similarity in an economical way, we make a compromise between the overall sentence similarity comparison and the component similarity comparison of parsed sentences. Instead of parsing sentences, we chunk them and with this chunking information, we propose our sentence similarity definitions, which make the computing process more resemble the human's comprehension to sentence meanings and provide a more reasonable result in sentence similarity comparison.

3.1. Four similarity definitions

Usually, people obtain information from a sentence on three aspects, or some of them: *objects* the sentence describes, *properties* of these *objects* and *behaviors* of these *objects*. We try to measure sentence similarities from those three aspects. We define *Objects-Specified Similarity* to demonstrate the similarity between the *objects* which the two sentences describe; *Objects-Property Similarity* to show the similarity between the *objects properties* of the two sentences; and *Objects-Behavior Similarity* to express the similarity between the *objects behaviors*. Then, we define *Overall Similarity* to denote the overall similarity of the two sentences, which combines the above three. After that, we can measure sentence similarities from those four aspects respectively as well.

3.2. Similarities calculation

3.2.1. Sentence chunking

Chunking, which is also known as shallow parsing [13], is a natural language processing technique that attempts to provide a sentence structure which machine can interpret. It is a middle step between identifying the part of speech of individual words and a full parsed tree of a sentence. A chunker divides a sentence into series of words that compose a grammatical unit (mostly noun, verb, or preposition phrase). It is an easier natural language processing task than parsing.

In order to find out the information in sentences that we need to compute the above four similarities, we chunk each sentence and extract all noun phrases and verb phrases. Then we choose all nouns in noun phrases as the *objects specified* in the sentence, all adjectives and adverbs in noun phrases as the *objects properties* and all verb phrases as the *objects behaviors*. Then we calculate the four similarities.

3.2.2. Semantic similarity between words

Word similarity measurement performance directly impacts on the result of the sentence similarity comparison. Currently, there are many approaches to compute word similarities [14], which perform comparatively well. Here, Liu's [14] model is adopted, which performs almost the best in the experiments.

In this edge-counting based technique, the model considers the relationship between common and different features of two compared words to compute their semantic similarity. Based on WordNet [15], it uses the depth of least common subsumer as their common features and the shortest path length as their different features. Given two words w_1 and w_2, the semantic similarity $S_w(w_1, w_2)$ is calculated as:

$$S_w(w_1, w_2) = \frac{f(d)}{f(d) + f(l)} \tag{1}$$

Where l is the shortest path length between w_1 and w_2, d the depth of the least common subsumer in the hierarchical semantic net, and $f(x)$ the transfer function for d and l. For $S_w(w_1, w_2)$, the interval of similarity is [0, 1], 1 for the same and 0 for no similarity. We choose $f(x)$ as a nonlinear function, $f(x) = e^x-1$, according to the good experimental results in [14], then Formula (1) becomes:

$$S_w(w_1, w_2) = \frac{e^{\alpha d} - 1}{e^{\alpha d} + e^{\beta l} - 2} (0 < \alpha, \beta \le 1) \tag{2}$$

Where α and β are smoothing factors. The optimal values for α and β depend on the lexical knowledge base and can be determined by a set of word pairs with human similarity ratings. For using with WordNet, the optimal values are: α =0.25 and β = 0.25, as reported in [13].

3.2.3. Semantic similarity in one aspect

Due to the good sentence representation [3, 10], we improve Li's basic semantic vector method to compute the similarity in each aspect. The process to form the vectors and compute their similarities is as follows. Take *Objects-Specified Similarity* computation as an example.

First, we map all nouns (*objects specified*) derived from noun phrases of a sentence into an *objects specified* vector, which is conceptually similar to a typical vector space representation used in a standard IR approach, but it only takes the nouns from noun phrases of the two compared sentences as the feature set instead of employing all indexed terms in the corpus. Each entry in the vector is derived from computing the word similarity, as illustrated in 3.2.2, between a word feature w_i and each noun from noun phrases of a sentence. After that, the maximum score from the matching words that exceeds certain similarity threshold θ will be chosen. Here we take $\theta = 0.2$ as reported in [10].

Here is an example that illustrates the formation of *objects specified* vectors. Suppose sentence s_1 and s_2 are the two sentences to be compared and there are two noun phrases in s_1, which are $np_{11}= \{w_1, w_2, w_3, w_4\}$ and $np_{12}= \{w_1, w_5, w_6, w_4\}$, and the nouns in np_{11} and np_{12} are $noun_1= \{w_3, w_4, w_6\}$. There is one noun phrase in s_2, which is $np_{21}= \{w_1, w_7, w_3\}$, the nouns in np_{21} are $noun_2= \{w_3, w_7\}$. Words in $noun_1$ form s_1's *objects specified* set. Words in $noun_2$ form s_2's *objects specified* set. So the feature set of *objects specified* vector is $vf_{os}= noun_1 \cup noun_2 = \{w_3, w_4, w_6, w_7\}$. For each word w_i in the vector entries, the formations of *objects specified* vector v_{os1} and v_{os2} of s_1 and s_2 are shown below in Table 1:

Table 1. Formation of *Objects Specified* Vectors

	w_3	w_4	w_6	w_7
v_{os1}	1	1	1	$S_w(w_7, noun_1)$
v_{os2}	1	$S_w(w_4, noun_2)$	$S_w(w_6, noun_2)$	1

Where $S_w(w_i, noun_j)$ is a maximum word similarity score of w_i and the matching word in $noun_j$. If the two words are lexically identical, then $S_w(w_i, noun_j)$ is equal to 1.

Secondly, the similarity between *objects specified* of two sentences is derived from the cosine coefficient between the two vectors by:

$$Sim_{os} = \frac{v_{os1} \bullet v_{os2}}{\| v_{os1} \| \bullet \| v_{os2} \|} \qquad (3)$$

Then we get the *Objects-Specified Similarity*, which is denoted as Sim_{os}.

And *objects properties* vectors and *objects behaviors* vectors are formed in the similar way as *objects specified* vectors. Then the *Objects-Property Similarity* and the *Objects-Behavior Similarity* can be obtained using cosine coefficient of vectors, which are denoted as Sim_{op} and Sim_{ob}.

3.2.4. Overall sentence similarity

After we have the *Objects-Specified Similarity*, the *Objects-Property Similarity* and the *Objects-Behavior Similarity*, we calculate the *Overall Similarity* of two compared sentences based on the three. We combine them as follows:

$$Sim_{sent} = aSim_{os} + bSim_{op} + cSim_{ob} \qquad (4)$$

Where a, b and c are the coefficients which denote the contribution of each part to the overall sentence similarity, $a+b+c = 1$; $a, b, c \in (0, 1)$.

4. Experiments

With the increasing research in the sentence similarity comparison, many data sets have been built to test the performance of the methods, such as Text REtrieval Conference (TREC) sources and MS paraphrase corpus [16]. But at present there isn't a data set that can be used to test the similarity comparison between parts of sentences, that is, comparing sentence similarity from four aspects proposed in this paper.

We choose MS paraphrase corpus to do our experiments. There are 1,725 test pairs and 4,076 training pairs in the corpus, and the pairs were automatically collected from thousands of news sources, then subsequently labeled by two human annotators who determined whether the two sentences in a pair were semantically equivalent or not.

4.1. Parameter training

When computing the *Overall Similarity*, our method requires three parameters, a, b and c to be determined. To get the optimal a, b and c, we train them based on the MS paraphrase training corpus. For each sentence pair in the corpus, we compute its *Overall Similarity*, if the value exceeds a certain threshold η, then we label the candidate pair as a correctly identified paraphrase. Based on the conditions that of $a, b, c, \eta \in (0, 1)$ and $a+b+c = 1$, we set the increasing step of a, b, c and η to 0.05 to do the training and take the maximal *F-measure* as the criterion to get the optimal a, b, c and η.

Table 2 shows the first 5 maximal values of *F-measure* and the corresponding a, b, c and η. We take the optimal a, b, c as 0.45, 0.35, 0.20 respectively and the corresponding η as 0.40. From Table 2, we can see that $a > b$ and $a > c$, which means that different information in a sentence contributes differently to its overall meaning, and *objects specified* are more significant than *objects properties* and *objects behaviors* in determining a sentence's meaning. This is up to our intuition that there are more nouns in sentences and a sentence meaning is usually delivered by its nouns. We also can see that there is no obvious difference in *b* and *c*, which means *objects properties* and *objects behaviors* are relatively equivalent in determining a sentence's meaning.

Table 2. Training results for *a*, *b*, *c* and *η*

a	*b*	*c*	*η*	F-measure
0.45	0.35	0.20	0.40	81.35%
0.45	0.40	0.15	0.40	81.33%
0.50	0.30	0.20	0.45	81.32%
0.45	0.25	0.30	0.40	81.27%
0.55	0.30	0.15	0.50	81.27%

4.2. Overall similarities

With the trained *a*, *b*, *c* and *η* in section 4.1, we calculate the four similarities of each news pair in MS paraphrase test corpus. Then we test the paraphrase identification performance of our method and present the results in terms of *Accuracy*, representing the number of correctly identified positive and negative samples, as well as *Precision*, *Recall* and *F-measure* with respect to the true or false values of positive sample.

To prove the effectiveness of our method, we compare our results Sim_{sent} with the *PMI-IR* measure [9], Li's measure Sim_{li} [10] and Liu's measure Sim_{liu} [11]. Previous experiment results are taken from literatures of [9, 11]. The results are shown in Table 3.

Table 3. Experimental results

Methods	Precision	Recall	F-measure	Accuracy
PMI-IR	70.4%	94.4%	80.7%	69.9%
Sim_{li}	69.3%	98.8%	81.5%	70.1%
Sim_{liu}	74.5%	91.6%	82.2%	73.6%
Sim_{sent}	71.3%	97.3%	82.3%	72.1%

From Table 3 we can see that our method performs best in *F-measure*, which reaches 82.3%. This comes as no surprise that we take *F-measure* as the criterion in training. Our method also achieves good performance in other criterions. In *Precision* and *Accuracy*, our result is higher than that of *PMI-IR* method and Li's. And our *Recall* is much higher than that of *PMI-IR* measure and Liu's.

4.3. Similarities in four aspects

From the calculated similarities in MS paraphrase test corpus, we randomly select four positive and four negative samples, and their corresponding similarities from four aspects to demonstrate the feasibility and validity of our method. Table 4 shows the results.

The selected eight samples are as follows.

Sample1:

"Taha is married to former Iraqi oil minister Amir Muhammed Rasheed, who surrendered to U.S. forces on April 28." "Taha's husband, former oil minister Amer Mohammed Rashid, surrendered to U.S. forces on April 28."

Sample2:

"On July 22, Moore announced he would appeal the case directly to the U.S. Supreme Court." "Moore of Alabama says he will appeal his case to the nation's highest court."

Sample3:

"Six Democrats are vying to succeed Jacques and have qualified for the Feb. 3 primary ballot." "Six Democrats and two Republicans are running for her seat and have qualified for the Feb. 3 primary ballot."

Sample4:

"Agriculture Secretary Luis Lorenzo told Reuters there was no damage to the vital rice crop as harvesting had just finished." "Agriculture Secretary Luis Lorenzo said there was no damage to the vital rice crop as the harvest had ended."

Sample5:

"A soldier was killed Monday and another wounded when their convoy was ambushed in northern Iraq." "On Sunday, a U.S. soldier was killed and another injured when a munitions dump they were guarding exploded in southern Iraq."

Sample6:

"Perkins will travel to Lawrence today and meet with Kansas Chancellor Robert Hemenway." "Perkins and Kansas Chancellor Robert Hemenway declined comment Sunday night."

Sample7:

"'I am proud that I stood against Richard Nixon, not with him,' Kerry said." "'I marched in the streets against Richard Nixon and the Vietnam War,' she said."

Sample8:

"The report by the independent expert committee aims to dissipate any suspicion about the Hong Kong government's handling of the SARS crisis." "A long awaited report on the Hong Kong government's handling of the SARS outbreak has been released."

Table 4: Similarities from four aspects

	Original Decision	OS Similarity	OP Similarity	OB Similarity	Overall Similarity
Sample1	1	0.869	1.0	0.577	0.857
Sample2	1	0.853	0.0	0.996	0.583
Sample3	1	0.972	1.0	0.707	0.929
Sample4	1	1.0	1.0	0.873	0.975
Sample5	0	0.576	0.0	0.577	0.375
Sample6	0	0.721	0.0	0.283	0.381
Sample7	0	0.500	0.0	0.807	0.386
Sample8	0	0.884	0.0	0.0	0.398

For positive Sample1, according to our definition, the *objects specified* in the first sentence include 'Taha', 'Iraqi', 'oil', 'minister', 'Amir Muhammed Rasheed', 'U.S. forces', 'April', and in the second sentence it includes 'Taha', 'husband', 'oil', 'minister', 'Amir Muhammed Rasheed', 'U.S. forces', 'April'. We calculate its *Objects-Specified Similarity* and get a similarity of 0.869, which demonstrates a high similarity. The *objects properties* in the first sentence only include 'former' and in the second sentence it also only include 'former'. So we get an *Objects-Property Similarity* of 1.0. For the *objects behaviors*, in the first sentence it includes 'married', 'surrendered', and in the second sentence, it only includes 'surrendered'. We get an *Objects-Behavior Similarity* of 0.577, which is lower than *Objects-Specified Similarity* and *Objects-Property Similarity*. At last, we calculate the *Overall Similarity* of the two sentences, which is 0.857. We can see that for the two compared sentences, they have a high *Objects-Specified Similarity*, *Objects-Property Similarity*, *Overall Similarity* but a low *Objects-Behavior Similarity*, which more exactly demonstrates the similarity between the two sentences and is identical to the people's comprehension to the meanings of the sentences.

For negative Sample5, according to our definition, the *objects specified* in the first sentence include 'soldier', 'Monday', 'convoy', 'northern', 'Iraq', and in the second sentence, it includes 'Sunday', 'U.S.', 'soldier', 'munitions', 'dump', 'southern', 'Iraq'. We calculate the *Objects-Specified Similarity* and get a similarity of 0.576. Because there are no adjectives and adverbs in the two sentences for *the Objects-Property Similarity* calculation, we get an *Objects-Property Similarity* of 0.0. For the *objects behaviors*, in the first sentence it includes 'killed', 'wounded', 'ambushed', and in the second sentence, it includes 'killed', 'injured', 'guarding', 'exploded', and we get an *Objects-Behavior Similarity* of 0.577. At last, we calculate the *Overall Similarity* of the two sentences, which is 0.375. We can see that for the two compared sentences, although they show a relatively higher *Objects-Specified Similarity* and *Objects-Behavior Similarity*, their *Overall Similarity* and *Objects-Property Similarity* is low, which reasonably demonstrates the similarity between the two sentences and is identical to the people's comprehension to the meanings of the sentences.

Similar results can be gained by analyzing the rest samples, and can also be gained from most of the pairs in the test corpus.

5. Conclusions

The paper proposes a new way to determine sentence similarities from different aspects. Probably based on information people can obtain from a sentence, which is *objects* the sentence describes, *properties* of these *objects* and *behaviors* of these *objects*. Four aspects, *Objects-Specified Similarity*, *Objects-Property Similarity*, *Objects-Behavior Similarity* and *Overall Similarity* are defined to determine sentence similarities. First, two compared sentences are respectively chunked with noun phrases and verb phrases. Secondly, for each sentence, all nouns in noun phrases are chosen as the *objects specified* in the sentence, all adjectives and adverbs in noun phrases as the *objects properties* and all verb phrases as the *objects behaviors*. Then, the four similarities are calculated based on a semantic vector method.

Experiments show that the proposed method makes the sentence similarity comparison more intuitive and render a more reasonable result, which imitates the people's comprehension to the meanings of the sentences.

In the future, we would like to further investigate the people's comprehension to sentences and simulate it with sentence processing technique to make the sentence similarity comparison more robust, accurate and suitable to some applications.

Acknowledgements

The work is supported by the National Basic Research Program of China under Grant No.2007CB310803.

References

[1] Poon, H., and Domingos, P., "Joint inference in information extraction", Proceeding of the Twenty-Second AAAI Conference on Artificial Intelligence, pp. 913–918, 2007.

[2] Lin, D., and Pantel, P., "Discovery of inference rules for question answering", Natural Language Engineering, Vol. 7, No. 2, 2001.

[3] Achananuparp P., Hu X., Xiaohua Zhou, Xiaodan Zhang, "Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community", WWW 2008 Workshop on Question Answering on the Web, April 22, Beijing, China 2008.

[4] Erkan, G., and Radev, D., "Lexrank: Graph-based lexical centrality as salience in text summarization", Journal of Artificial Intelligence Research Vol. 22, pp. 457–479, 2004.

[5] Ko, Y., Park, J., and Seo, J., "Improving text categorization using the importance of sentences", Information Processingand Management Vol. 40, No.1, pp. 65–79, 2004.

[6] Papineni, K., Roukos, S., Ward, T., and Zhu, W., "Bleu: a method for automatic evaluation of machine translation", Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, 2002.

[7] Mandreoli, F., Martoglia, R., and Tiberio, P., "A syntactic approach for searching similarities within sentences", Proceeding of International Conference on Information and Knowledge Management, pp. 656–637, 2002.

[8] Hatzivassiloglou, V., Klavans, J., and Eskin, E., "Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning", Proceeding of Empirical Methods in natural language processing and Very Large Corpora, 1999.

[9] Mihalcea, R., Corley, C., and Strapparava, C., "Corpus-based and knowledge-based measures of text semantic similarity" Proceeding of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, 2006.

[10] Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K., "Sentence similarity based on semantic nets and corpus statistics", IEEE Transactions on Knowledge and Data Engineering Vol. 18, No. 8, pp. 1138–1150, 2006.

[11] Liu, X., Zhou, Y. and Zheng, R., "Measuring Semantic Similarity within Sentences", Proceeding of ICMLC2008 Conference, Kunming, 2008.

[12] Liu, X., Zhou, Y. and Zheng, R., "Sentence Similarity based on Dynamic Time Warping", International Conference on Semantic Computing 2007, ICSC2007, pp. 250-256, 2007.

[13] James H., Miles O., Susan A., Walter D., "Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing", Journal of Machine Learning Research, Vol.2, pp. 551-558, 2002

[14] Liu, X., Zhou, Y. and Zheng, R., "Measuring Semantic Similarity in Wordnet", Proceeding of ICMLC2007 Conference, Hongkong, 2007.

[15] Miller, G., "Wordnet: a lexical database for English", Communications of the ACM, Vol. 38, No. 11, pp. 39–41, 1995.

[16] Dolan, W., Quirk,C., and Brockett, C., "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources", Proceeding of the 20th International Conference on Computational Linguistics, 2004.