The Evaluation of Sentence Similarity Measures

Palakorn Achananuparp[1], Xiaohua Hu[1,2], and Xiajiong Shen[2]

[1] College of Information Science and Technology
Drexel University, Philadelphia, PA 19104
[2] College of Computer and Information Engineering, Hehan University, Henan, China
pkorn@drexel.edu, thu@cis.drexel.edu, shenxj@henu.edu.cn

Abstract. The ability to accurately judge the similarity between natural language sentences is critical to the performance of several applications such as text mining, question answering, and text summarization. Given two sentences, an effective similarity measure should be able to determine whether the sentences are semantically equivalent or not, taking into account the variability of natural language expression. That is, the correct similarity judgment should be made even if the sentences do not share similar surface form. In this work, we evaluate fourteen existing text similarity measures which have been used to calculate similarity score between sentences in many text applications. The evaluation is conducted on three different data sets, TREC9 question variants, Microsoft Research paraphrase corpus, and the third recognizing textual entailment data set.

Keywords: Sentence similarity, Paraphrase Recognition, Textual Entailment Recognition.

1 Introduction

Determining the similarity between sentences is one of the crucial tasks which have a wide impact in many text applications. In information retrieval, similarity measure is used to assign a ranking score between a query and texts in a corpus. Question answering application requires similarity identification between a question-answer or question-question pair [1]. Furthermore, graph-based summarization also relies on similarity measures in its edge weighting mechanism. Yet, computing sentence similarity is not a trivial task. The variability of natural language expression makes it difficult to determine semantically equivalent sentences. While many applications have employed certain similarity functions to evaluate sentence similarity, most approaches only compare sentences based on their surface form. As a result, they fail to recognize equivalent sentences at the semantic level. Another issue pertains to the notions of similarity underlying sentence judgment. Since sentences convey more specific information than documents, a general notion of topicality employed in document similarity might not be appropriate for this task. As Murdock [16] and Metzler et al. [14] point out, there are multiple categories of sentence similarity based on topical specificity. Furthermore, specific notions such as paraphrase or entailment might be needed for certain applications. In this work, we investigate the performance of three classes of measures: word overlap, TF-IDF, and linguistic measures. Each

I.-Y. Song, J. Eder, and T.M. Nguyen (Eds.): DaWaK 2008, LNCS 5182, pp. 305–316, 2008.

sentence pair is judged based on the notion that they have identical meaning. For example, two sentences are considered to be similar if they are a paraphrase of each other, that is, they talk about the same event or idea judging from the common principal actors and actions. Next, two sentences are similar if one sentence is a superset of the other. Note that this is also a notion used in textual entailment judgment where directional inference between two sentences is made.

The paper is organized as follows. First, we review the work related to our study. Next, we briefly describe fourteen similarity measures used in the evaluation. In section 4, we explain the experimental evaluation, including evaluation metrics and data sets, used in this study. We discuss about the result and conclude the paper in section 5 and 6, respectively.

2 Related Work

Previous works have been done to evaluate different approaches to measure similarity between short text segments [15]. Specifically, many studies have focused on a comparison between probabilistic approaches and the existing text similarity measures in a sentence retrieval experiment [14][16][3]. For example, Metzler et al. [14] evaluate the performance of statistical translation models in identifying topically related sentences compared to several simplistic approaches such as word overlap, document fingerprinting, and TF-IDF measures. In [15], the effectiveness of lexical matching, language model, and hybrid measures, in computing the similarity between two short queries are investigated. Next, Balasubramanian et al. [3] compare the performance of nine language modeling techniques in sentence retrieval task. Despite their superiority in coping with vocabulary mismatch problem, most probabilistic measures do not significantly outperform existing measures in sentence retrieval task [17]. Although we share the same goal of comparing the performance of sentence similarity measures, there are a few key differences in this study. First, our focus is to evaluate the effectiveness of measures in identifying the similarity between two arbitrary sentences. That is, we perform a pair-wise comparison on a set of sentence pairs. In contrast, sentence retrieval evaluation concentrates on estimating the similarity between the reference query or sentence and the top-N retrieved sentences. Second, the text unit in the previous research is a short text segment such as a short query while we are interested in a syntactically well-formed sentence. Lastly, we conduct the comparative evaluation on public data sets which contain different notions of text similarity, e.g. paraphrase and textual entailment, whereas the prior studies evaluate the effectiveness of measures based on the notion of topical relevance.

3 Sentence Similarity Measures

We describe three classes of measures that can be used for identifying the similarity between sentences. The similarity score produces by these measures has a normalized real-number value from 0 to 1.

3.1 Word Overlap Measures

Word overlap measures is a family of combinatorial similarity measure that compute similarity score based on a number of words shared by two sentences. In this work, we consider four word overlap measures: Jaccard similarity coefficient, simple word overlap, IDF overlap, and phrasal overlap.

3.1.1 Jaccard Similarity Coefficient

Jaccard similarity coefficient is a similarity measure that compares the similarity between two feature sets. When applying to sentence similarity task, it is defined as the size of the intersection of the words in the two sentences compared to the size of the union of the words in the two sentences.

3.1.2 Simple Word Overlap and IDF Overlap Measures

Metzler et al. [14] defined two baseline word overlap measures to compute the similarity between sentence pairs. Simple word overlap fraction ($sim_{overlap}$) is defined as the proportion of words that appear in both sentences normalized by the sentence's length, while IDF overlap ($sim_{overlap,IDF}$) is defined as the proportion of words that appear in both sentences weighted by their inverse document frequency.

3.1.3 Phrasal Overlap Measure

Banerjee and Pedersen [4] introduced the overlap measure based on the Zipfian relationship between the length of phrases and their frequencies in a text collection. Their motivation stems from the fact that a traditional word overlap measure simply treats sentences as a bag of words and does not take into account the differences between single words and multi-word phrases. Since a phrasal n-word overlap is much rarer to find than a single word overlap, thus a phrasal overlap calculation for m phrasal n-word overlaps is defined as a non-linear function displayed in equation 1 below.

$$overlap_{phrase}(s_1, s_2) = \sum_{i=1}^{n} \sum\nolimits_m i^2 \tag{1}$$

where m is a number of i-word phrases that appear in sentence pairs. Ponzetto and Strube [19] normalized equation 1 by the sum of sentences' length and apply the hyperbolic tangent function to minimize the effect of the outliers. The normalized phrasal overlap similarity measure is defined in equation 2.

$$sim_{overlap,phrase}(s_1, s_2) = \tanh\left(\frac{overlap_{phrase}(s_1, s_2)}{|s_1| + |s_2|} \right) \tag{2}$$

3.2 TF-IDF Measures

Three variations of measures that compute sentence similarity based on term frequency-inverse document frequency (TF-IDF) are considered in this study.

3.2.1 TF-IDF Vector Similarity

Standard vector-space model represents a document as a vector whose feature set consists of indexing words. Term weights are computed from TF-IDF score. For sentence similarity task, we adopt the standard vector-space approach to compare the similarity between sentence pairs by computing a cosine similarity between the vector representations of the two sentences. A slight modification is made for sentence representation. Instead of using indexing words from a text collection, a set of words that appear in the sentence pair is used as a feature set. This is done to reduce the degree of data sparseness in sentence representation. The standard TF-IDF similarity ($sim_{TFIDF,vector}$) is defined as cosine similarity between vector representation of two sentences. For a baseline comparison, we also include $sim_{TF,vector}$ which utilizes term frequencies as the basic term weights.

3.2.2 Novelty Detection and Identity Measure

Allan et al. [2] proposed TF-IDF measure ($sim_{TFIDF,nov}$) for detecting topically similar sentences in TREC novelty track experiment. The formulation is based on the sum of the product of term frequency and inverse document frequency of words that appear in both sentences. Identity measure ($sim_{identity}$) [9] is another variation of TF-IDF similarity measure originally proposed as a measure for identifying plagiarized documents or co-derivation. It has been shown to perform effectively for such application. Essentially, the identity score is derived from the sum of inverse document frequency of the words that appear in both sentences normalized by the overall lengths of the sentences and the relative frequency of a word between the two sentences. The formulation of the two measures can be found in [14].

3.3 Linguistic Measures

Linguistic measures utilize linguistic knowledge such as semantic relations between words and their syntactic composition, to determine the similarity of sentences. Three major linguistic approaches are evaluated in this work. Note that there are several approaches that utilize word semantic similarity scores to determine similarity between sentences. For a comprehensive comparison of word similarity measures, we recommend the readers to the work done by Budanitsky and Hirst [5]. In this work, we use Lin' universal similarity [12] to compute word similarity scores.

3.3.1 Sentence Semantic Similarity Measures

Li et al. [10] suggest a semantic-vector approach to compute sentence similarity. Sentences are transformed into feature vectors having words from sentence pair as a feature set. Term weights are derived from the maximum semantic similarity score between words in the feature vector and words in a corresponding sentence. In addition, we simplify Li et al.'s measure by only using word similarity scores as term weights. Moreover, we only compute semantic similarity of words within the same part-of-speech class. Then, semantic similarity between sentence pair (sim_{ssv}) is defined as a cosine similarity between semantic vectors of the two sentences.

Another semantic measure, proposed by Mihalcea et al. [16], also combines word semantic similarity scores with word specificity scores. Given two sentences s_1 and s_2, the sentence similarity calculation begins by finding the maximum word similarity score for each word in s_1 with words in the same part of speech class in s_2. Then, apply the same procedure for each word in s_2 with words in the same part of speech class in s_1. The derived word similarity scores are weighted with *idf* scores that belong to the corresponding word. Finally, the sentence similarity formulation is defined in equation 3.

$$sim_{sem,IDF}(s_1,s_2)=\frac{1}{2}\left(\frac{\sum_{w\in\{s_1\}}(\max Sim(w,s_2)\times idf(w))}{\sum_{w\in\{s_1\}} idf(w)}+\frac{\sum_{w\in\{s_2\}}(\max Sim(w,s_1)\times idf(w))}{\sum_{w\in\{s_2\}} idf(w)}\right) \quad (3)$$

where *maxSim*(*w*,s_i) is the maximum semantic similarity score of *w* and words in s_i that belong to the same part-of-speech as *w* while *idf*(*w*) is an inverse document frequency of *w*. The reason for computing the semantic similarity scores only between words in the same part of speech class is that most WordNet-based measures are unable to compute semantic similarity of cross-part-of-speech words.

Malik et al. [13] have proposed a simplified variation of semantic similarity measure (sim_{sem}) by determining sentence similarity based on the sum of maximum word similarity scores of words in the same part-of-speech class normalized by the sum of sentence's lengths.

3.3.2 Word Order Similarity

Apart from lexical semantics, word composition also plays a role in sentence understanding. Basic syntactic information, such as word order, can provide useful information to distinguish the meaning of two sentences. This is particularly important in many similarity measures where a single word token was used as a basic lexical unit when computing similarity of sentences. Without syntactic information, it is impossible to discriminate sentences that share the similar bag-of-word representations. For example, "the sale manager hits the office worker" and "the office manager hits the sale worker" will be judged as identical sentences because they have the same surface text. However, their meanings are very different.

To utilize word order in similarity calculation, Li et al. [10] defines word order similarity measure as the normalized difference of word order between the two sentences. The formulation for word order similarity is described in equation 4 below:

$$sim_{wo}(s_1,s_2)=1-\frac{\|r_1-r_2\|}{\|r_1+r_2\|} \quad (4)$$

where r_1 and r_2 is a word order vector of sentence s_1 and s_2, respectively. Word order vector is a feature vector whose feature set comes from words that appear in a sentence pair. The index position of the words in the corresponding sentence are used as term weights for the given word features. That is, each entry in the word order vector r_i is derived from computing a word similarity score between a word feature *w* with all the words in the sentence s_i. An index position of the word in s_i that gives the maximum word similarity score to *w* is selected as *w*'s term weight.

3.3.3 The Combined Semantic and Syntactic Measures

Using the notion that both semantic and syntactic information contribute to the understanding of a sentence, Li et al. [10] defined a sentence similarity measure as a linear combination of semantic vector similarity and word order similarity (equation 5). The relative contribution of semantic and syntactic measures is controlled by a coefficient alpha. It has been empirically proved [10][1] that a sentence similarity measure performs the best when semantic measure is weighted more than syntactic measure (alpha = 0.8). This follows the conclusion from a psychological experiment conducted by [10] which emphasizes the role of semantic information over syntactic information in passage understanding. In this study, we also introduce a minor variation of the combined sentence similarity formulation by substituting the semantic vector similarity measure in equation 5 with Malik et al.'s measure (equation 6). The same semantic coefficient value (alpha = 0.8) is applied.

$$sim_{ssv+wo}(s_1, s_2) = \alpha sim_{ssv}(s_1, s_2) + (1-\alpha) sim_{wo}(s_1, s_2) \tag{5}$$

$$sim_{sem+wo}(s_1, s_2) = \alpha sim_{sem}(s_1, s_2) + (1-\alpha) sim_{wo}(s_1, s_2) \tag{6}$$

4 Experimental Evaluation

4.1 Evaluation Criteria

We define six evaluation metrics based on the general notion of positive and negative judgments in information retrieval and text classification as follows.

Recall is a proportion of correctly predicted similar sentences compared to all similar sentences. *Precision* is a proportion of correctly predicted similar sentences compared to all predicted similar sentences. *Rejection* is a proportion of correctly predicted dissimilar sentences compared to all dissimilar sentences. *Accuracy* is a proportion of all correctly predicted sentences compared to all sentences. F_1 is a uniform harmonic mean of precision and recall. Lastly, we define f_1 as a uniform harmonic mean of rejection and recall. A scoring threshold for similar pairs is defined at 0.5. In this work, we include rejection and f_1 metrics in addition to the standard precision-recall based metrics as it presents another aspect of the performance based on the tradeoff between true positive and true negative judgments.

4.2 Data Sets

Three publicly-available sentence pair data sets are used to evaluate the performance of the sentence similarity measures. The data sets are TREC9 question variants key (*TREC9*) [1], Microsoft Research paraphrase corpus (*MSRP*) [7], and the third recognising textual entailment challenge (*RTE3*) data set [6].

TREC9 comprises 193 paraphrased pairs used in TREC9 Question Answering experiment. The original questions were taken from a query log of user submitted questions while the paraphrased questions were manually constructed by human assessors. For this study, we randomly pair original questions with non-paraphrased questions to create additional 193 pairs of dissimilar questions. Despite its semi-artificial nature, the data set contains adequate complexity to reflect the variability of nature language expression judging from its various compositions of paraphrasing categories [21].

MSRP contains 1,725 test pairs automatically constructed from various web new sources. Each sentence pair is judged by two human assessors whether they are semantically equivalent or not. Overall, 67% of the total sentence pairs are judged to be the positive examples. Semantically equivalent sentences may contain either identical information or the same information with minor differences in detail according to the principal agents and the associated actions in the sentences. Sentence that describes the same event but is a superset of the other is considered to be a dissimilar pair. Note that this rule is similar to the one used in text entailment task.

RTE3 consists of 800 sentence pairs in the test set. Each pair comprises two small text segments, which are referred to as *text* and *hypothesis*. The text-hypothesis pairs are collected by human assessors and can be decomposed into four subsets corresponding to the application domains: information retrieval, multi-document summarization, question answering, and information extraction. Similarity judgment between sentence pairs is based on directional inference between text and hypothesis. If the hypothesis can be entailed by the text, then that pair is considered to be a positive example.

From the complexity standpoint, we consider TREC9 to be the lowest complexity data set for its smallest vocabulary space and relatively simple sentence construction. On the other hand, MSRP and RTE3 are considered to be higher complexity data sets due to larger vocabulary space and longer sentence lengths and differences.

Table 1. Summary of three sentence pair data sets used in the experiment

Summary	TREC9	MSRP	RTE3
Number of sentence pairs	386	1,725	800
Number of unique words	252	8,256	5,700
Percentage of unique words covered by WordNet	84.5%	64.5%	70.1%
Average sentence length (in characters)	39.35	115.30	227.87
Average difference in length between two comparing sentences (in characters)	4.32	9.68	132.81
Linguistic complexity	Low	High	High

4.3 Preprocessing

For each data set, we perform a part-of-speech tagging on a sentence using LingPipe libraries (http://alias-i.com/lingpipe/). Next, single word tokens in the sentences are extracted. Then, we remove functional words, such as articles, pronouns, prepositions, conjunctions, auxiliary verbs, modal verbs, and punctuations from the sentence since they do not carry semantic content, but keep the cardinal numbers. Stemming is not applied in the case of linguistic measures to preserve the original meaning of the words. Information about word relations is obtained from WordNet.

5 Results and Discussion

5.1 Question Paraphrase Identification

Table 2 displays the performance of sentence similarity measures on TREC 9 data set. Overall, linguistic measure is the best performer according to F_1, f_1, and accuracy metrics. Within this class of measures, sentence semantic similarity (sim_{sem}) and combined similarity measures (sim_{sem+wo}) perform significantly better than other measures at $p<0.05$. Phrasal overlap measure is the best performer in word overlap category and standard TF-IDF vector and identity measure perform equally well in TF-IDF measures. Most word order measures and TF-IDF measures exhibit a strong rejection rate. This is to be expected, as the dissimilar pairs in TREC9 contain a relatively small number of word overlaps.

Table 2. Comparison of the performance of sentence similarity measures on TREC9 data set. Results with * indicate that the differences are not statistically significant.

Sentence Similarity Measures	Prec.	Rec.	Rej.	F_1	f_1	Acc.
$sim_{jaccard}$	1	0.383	1	0.554	0.554	0.691
$sim_{overlap}$	0.99	0.362	0.995	0.53	0.532	0.679
$sim_{overlap,IDF}$	0.978	0.233	0.995	0.377	0.378	0.614
$sim_{overlap,phrase}$	1	0.637	1	**0.778**	**0.778**	**0.819**
$sim_{TF,vector}$	0.993	0.689	0.995	0.813	0.814	0.842
$sim_{TFIDF,vector}$	1	0.762	1	**0.865***	**0.865***	**0.881***
$sim_{TFIDF,nov}$	1	0.192	1	0.322	0.322	0.6
$sim_{identity}$	0.98	0.767	0.984	**0.86***	**0.862***	**0.876***
sim_{ssv}	0.67	0.969	0.523	0.79	0.68	0.746
sim_{sem}	0.983	0.912	0.984	**0.946***	**0.947***	**0.948***
$sim_{simsem,IDF}$	0.949	0.575	0.969	0.716	0.722	0.772
sim_{wo}	0.644	0.487	0.731	0.555	0.584	0.609
sim_{ssv+wo}	0.68	0.979	0.539	0.803	0.695	0.759
sim_{sem+wo}	0.963	0.933	0.964	**0.948***	**0.948***	**0.948***

5.2 Paraphrase Recognition

Similar to TREC9 result, linguistic measure is also the overall best performer according to F_1 metric on MSRP data set. Many best linguistic measures perform at an equal F_1 score of 80%. Word overlap and TF-IDF measures perform at a lower F_1 score but the performance gap is very minimal. The performance on f_1 metric, on the other hand, is different from that of TREC9. Due to the fact that most linguistic measures have a very low rejection rate compared to word overlap and TF-IDF measures, they perform poorly on f_1 metric. In this case, the best performer in f_1 category is Jaccard similarity coefficient ($sim_{jaccard}$). A further analysis has shown that several false positive cases are in a "difficult" subset which requires entailment judgment. For example, the following non paraphrase pair produces an average 85% similarity score from the linguistics measures which results in a false positive judgment:

Sentence 1: Russian stocks fell after the arrest last Saturday of Mikhail Khodorkovsky, chief executive of Yukos Oil, on charges of fraud and tax evasion.
Sentence 2: The weekend arrest of Russia's richest man, Mikhail Khodorkovsky, chief executive of oil major YUKOS, on charges of fraud and tax evasion unnerved financial markets.

According to the above example, sentence 1 and sentence 2 describe a parallel event with slightly different detail (generic vs. specific information). Moreover, it requires a semantic inference to relate the two phrases "Russian stocks fell" and "unnerved financial markets." In the cases of superset-subset relationship, all classes of similarity measures fail to make a correct prediction, for example:

Sentence 1: He said the attackers left behind leaflets urging staff at the Ishtar Sheraton to stop working at the hotel and demanding U.S. forces leave Iraq.
Sentence 2: He said the attackers left behind leaflets urging workers at the Ishtar Sheraton to stop working at the hotel.

Table 3. Comparison of the performance of sentence similarity measures on MSRP data set. Results with * indicate that the differences are not statistically significant.

Sentence Similarity Measures	Prec.	Rec.	Rej.	F_1	f_1	Acc.
$sim_{jaccard}$	0.835	0.603	0.763	0.7	**0.674**	0.657
$sim_{overlap}$	0.76	0.678	0.574	0.717	0.622	0.643
$sim_{overlap,IDF}$	0.829	0.325	0.867	0.467	0.473	0.507
$sim_{overlap,phrase}$	0.7	0.892	0.244	**0.785**	0.383	**0.675**
$sim_{TF,vector}$	0.713	0.881	0.298	0.789	0.445	**0.686***
$sim_{TFIDF,vector}$	0.734	0.836	0.398	0.782	**0.539**	**0.69***
$sim_{TFIDF,nov}$	0.858	0.283	0.907	0.426	0.431	0.492
$sim_{identity}$	0.665	1	0	**0.798**	0.01	0.664
sim_{ssv}	0.669	0.989	0.031	**0.798***	0.06	0.668
sim_{sem}	0.674	0.99	0.052	**0.802***	0.099	**0.675***
$sim_{simsem,IDF}$	0.714	0.835	0.337	0.77	0.48	0.668
sim_{wo}	0.681	0.619	0.424	0.648	**0.503**	0.554
sim_{ssv+wo}	0.673	0.983	0.052	**0.799***	0.099	**0.671***
sim_{sem+wo}	0.674	0.977	0.064	**0.8***	0.12	**0.671***

5.3 Textual Entailment Recognition

The performance comparison of sentence similarity measures on RTE3 data is shown in table 4. Overall, linguistic measures outperform other classes of measures in F_1, f_1, and accuracy metrics. Most linguistic measures perform equally well on F_1 metric while the combined sentence semantic and word order measure (sim_{sen+wo}) significantly outperforms other linguistic measures on f_1 and accuracy. Word overlap measures other than phrasal overlap are not viable for text entailment task at all due to low F_1 and f_1 scores. Since sentence length in RTE3 is relatively long compared to the other two data sets, and text length is much greater than hypothesis length, measures that rely on the proportion of word overlap or word distribution are penalized by the unequal sentence lengths. Like MSRP result, linguistic measures produce a significantly lower rejection rate than word overlap and TF-IDF measures. The example of false positive judgment, where no similarity measures are able to correctly reject the above sentence pair, is as follow:

> **Sentence 1 (text):** It's very difficult to get teams from China the right to stay here for a longer period of time.
>
> **Sentence 2 (hypothesis):** It is difficult to get the right to stay in China for a long period of time.

5.4 The Effect of Word Specificity

There are no clear advantages of word specificity measure such as IDF on the overall performance of sentence similarity measures. Apart from the result of TREC9 evaluation, where an IDF measure, $sim_{TFIDF,vector}$, performs significantly better across all evaluation metrics compared to its non-IDF counterpart, $sim_{TF,vector}$, other IDF-based measures perform poorer on recall, accuracy, F_1, and f_1 metrics. Note that IDF does help improve precision and rejection scores in most measures. This indicates its relative effectiveness in handling false positive cases. However, the loss in recall far

Table 4. Comparison of the performance of sentence similarity measures on RTE3 data set. Results with * indicate that the differences are not statistically significant.

Sentence Similarity Measures	Prec.	Rec.	Rej.	F_1	f_1	Acc.
$sim_{jaccard}$	0.579	0.027	0.979	0.051	0.052	0.491
$sim_{overlap}$	0.565	0.032	0.974	0.06	0.061	0.491
$sim_{overlap,IDF}$	0.6	0.007	0.995	0.014	0.015	0.489
$sim_{overlap,phrase}$	0.638	0.417	0.751	**0.504**	**0.536**	**0.58**
$sim_{TF,vector}$	0.652	0.324	0.812	0.433	0.465	**0.565**
$sim_{TFIDF,vector}$	0.644	0.283	0.836	0.393	0.423	0.553
$sim_{TFIDF,nov}$	0.69	0.141	0.933	0.235	0.246	0.528
$sim_{identity}$	0.539	0.471	0.577	**0.503**	**0.518**	0.523
sim_{ssv}	0.52	0.893	0.133	**0.657***	0.232	0.523
sim_{sem}	0.592	0.727	0.474	**0.653***	0.574	0.604
$sim_{simsem,IDF}$	0.602	0.585	0.592	0.593	0.589	0.589
sim_{wo}	0.569	0.424	0.661	0.486	0.517	0.54
sim_{ssv+wo}	0.532	0.863	0.203	**0.659***	0.328	0.541
sim_{sem+wo}	0.614	0.695	0.541	**0.652***	**0.608**	**0.62**

outweighs the gain in precision and rejection. The results offer a contradicting implication to the previous work [18] where IDF has been empirically proven to be an optimal weight for document retrieval and reinforce the challenge of sentence similarity task. The inclusion of word specificity into the similarity calculation might provide a significant improvement to the task of identifying topically related documents. However, it does have the same effect in the case of paraphrase recognition and entailment identification.

5.5 WordNet Coverage and Linguistic Measures

The effectiveness of linguistic measures depends on a heuristic to compute semantic similarity between words as well as the comprehensiveness of the lexical resource. As WordNet is used as a primary lexical resource in this study, its comprehensiveness is determined by the proportion of words in the text collections that are covered by its knowledge base. In general, a major criticism of WordNet-based similarity measures is in its limited word coverage to handle a large text collection, particularly on the named entities coverage. As indicated in table 1, the percentage of word coverage in WordNet decreases as the size of test collection and vocabulary space increases. Thus, the effectiveness of linguistic measures is likely to be effected because word-to-word similarity calculation will inevitably produce many "misses". One solution is to resort to approaches that utilize other knowledge resources, such as Wikipedia [19] or web search results [20], to derive semantic similarity between words.

6 Conclusions

We have investigated the performance of several classes of sentence similarity measures on multiple sentence pair data sets. In a low-complexity data set, linguistic measures are superior in identifying paraphrases than word overlap and TF-IDF measures.

They are also the best performer in the higher-complexity data sets but the performance gap between measures diminishes depending on the characteristics of the test data. Several factors influence the result. First, MSRP data set contains a high degree of word overlap. Therefore, overlap-based measures are able to produce a reasonable result. Second, linguistics measures perform relatively poor in judging dissimilar pairs in high-complexity data sets. Thus, it adversely affects the overall accuracy. Keep in mind that word overlap and TF-IDF measures tend to reject many dissimilar sentence pairs since their proportion of overlap or the word occurrence is likely to be smaller in high-complexity data sets due to the difference in sentence pair lengths. For "harder" test pairs, such as those in RTE3 or part of MSRP, which require even more specific judgment such as textual entailment, most sentence similarity measures do not produce a satisfactory result.

We are aware of other factors apart from the similarity measure itself which contribute to the application performance. Many of which are considered in our future work. For example, instead of representing a sentence as a bag of words, a graph-based representation can be used. Next, different lexical unit that is more meaningful, such as multi-word phrase, can be used as opposed to a single word. Different heuristics to compute semantic similarity between words and different lexical resources can be used, etc. Nevertheless, we strongly believe that the comparative evaluation of sentence similarity in this study offers an interesting and useful insight into the performance of these similarity measures which are crucial to any sentence-level text applications.

Acknowledgments. This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196) and PA Dept of Health Grant (No. 239667).

References

1. Achananuparp, P., Hu, X., Zhou, X., Zhang, X.: Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community. In: Proceedings of QAWeb 2008 Workshop, Beijing, China (to appear, 2008)
2. Allan, J., Bolivar, A., Wade, C.: Retrieval and novelty detection at the sentence level. In: Proceedings of SIGIR 2003, pp. 314–321 (2003)
3. Balasubramanian, N., Allan, J., Croft, W.B.: A comparison of sentence retrieval techniques. In: Proceedings of SIGIR 2007, Amsterdam, The Netherlands, pp. 813–814 (2007)
4. Banerjee, S., Pedersen, T.: Extended gloss overlap as a measure of semantic relatedness. In: Proceedings of IJCAI 2003, Acapulco, Mexico, pp. 805–810 (2003)
5. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of semantic distance. Computational Linguistics 32(1), 13–47 (2006)
6. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: Proceedings of the PASCAL Workshop (2005)
7. Dolan, W., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel new sources. In: Proceedings of the 20th International Conference on Computational Linguistics (2004)
8. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)

9. Hoad, T., Zobel, J.: Methods for identifying versioned and plagiarized documents. Journal of the American Society of Information Science and Technology 54(3), 203–215 (2003)
10. Landauer, T.K., Laham, D., Rehder, B., Schreiner, M.E.: How Well Can Passage Meaning Be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. In: Proc. 19th Ann. Meeting of the Cognitive Science Soc., pp. 412–417 (1997)
11. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering 18(8), 1138–1150 (2006)
12. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proceedings of the Fifteenth international Conference on Machine Learning, San Francisco, CA, pp. 296–304 (1998)
13. Malik, R., Subramaniam, V., Kaushik, S.: Automatically Selecting Answer Templates to Respond to Customer Emails. In: Proceedings of IJCAI 2007, Hyderabad, India, pp. 1659–1664 (2007)
14. Metzler, D., Bernstein, Y., Croft, W., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: Proceedings of CIKM, pp. 517–524 (2005)
15. Metzler, D., Dumais, S.T., Meek, C.: Similarity Measures for Short Segments of Text. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 16–27. Springer, Heidelberg (2007)
16. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In: Proceedings of AAAI 2006, Boston (July 2006)
17. Murdock, V.: Aspects of sentence retrieval. Ph.D. Thesis, University of Massachusetts (2006)
18. Papineni, K.: Why inverse document frequency? In: Proceeding of the North American Chapter of the Association for Computational Linguistics, pp. 25–32 (2001)
19. Ponzetto, S.P., Strube, M.: Knowledge Derived From Wikipedia for Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30, 181–212 (2007)
20. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of WWW 2006, Edinburgh, Scotland, pp. 377–386 (2006)
21. Tomuro, N.: Interrogative Reformulation Patterns and Acquisition of Question Paraphrases. In: Proceedings of the 2nd international Workshop on Paraphrasing, pp. 33–40 (2003)