Word Sense Disambiguation-based Sentence Similarity

Chukfong Ho[1], Masrah Azrifah Azmi Murad[2]
Department of Information System
University Putra Malaysia
hochukfong@yahoo.com[1], masrah@fsktm.upm.edu.my[2]

Rabiah Abdul Kadir, Shyamala C. Doraisamy
Department of Multimedia
University Putra Malaysia
{rabiah, shyamala}@fsktm.upm.edu.my

Abstract

Previous works tend to compute the similarity between two sentences based on the comparison of their nearest meanings. However, the nearest meanings do not always represent their actual meanings. This paper presents a method which computes the similarity between two sentences based on a comparison of their actual meanings. This is achieved by transforming an existing most-outstanding corpus-based measure into a knowledge-based measure, which is then integrated with word sense disambiguation. The experimental results on a standard data set show that the proposed method outperforms the baseline and the improvement achieved is statistically significant at 0.025 levels.

1 Introduction

Although measuring sentence similarity is a complicated task, it plays an important role in natural language processing applications. In text categorization (Yang and Wen, 2007), documents are retrieved based on similar or related features. In text summarization (Zhou et al., 2006) and machine translation (Kauchak and Barzilay, 2006), summaries comparison based on sentence similarity has been applied for automatic evaluation. In text coherence (Lapata and Barzilay, 2005), different sentences are linked together based on the sequence of similar or related words.

Two main issues are investigated in this paper: 1) the performance between corpus-based measure and knowledge-based measure, and 2) the influence of word sense disambiguation (WSD) on measuring sentence similarity. WSD is the task of determining the sense of a polysemous word within a specific context (Wang et al., 2006). Corpus-based methods typically compute sentence similarity based on the frequency of a word's occurrence or the co-occurrence between collocated words. Although these methods benefit from the statistical information derived from the corpus, this statistical information is closer to syntactic representation than to semantic representation. In comparison, knowledge-based methods compute the similarity between two sentences based on the semantic information collected from knowledge bases. However, this semantic information is applied in a way that, for any two sentences, the comparison of their nearest meanings is taken into consideration instead of the comparison of their actual meanings. More importantly, the nearest meaning does not always represent the actual meaning. In this paper, a solution is proposed that seeks to address these two issues. Firstly, the most outstanding existing corpus-based sentence similarity measure is transformed into a knowledge-based measure. Then, its underlying concept, which is the comparison of the nearest meanings, is replaced by another underlying concept, the comparison of the actual meanings.

The rest of this paper is organized into five sections. Section 2 presents an overview of the related works. Section 3 details the problem of the existing method and the improvement of the proposed method. Section 4 describes the experimental design. In Section 5, the experimental results are discussed. Finally, the implications and contributions are addressed in Section 6.

Coling 2008: Poster Volume, pages 418–426,
Beijing, August 2010

2 Related Work

In general, related works can be categorized into corpus-based, knowledge-based and hybrid-based methods. Islam and Inkpen (2008) proposed a corpus-based sentence similarity measure as a function of string similarity, word similarity and common word order similarity (CWO). They claimed that a corpus-based measure has the advantage of large coverage when compared to a knowledge-based measure. However, the judgment of similarity is situational and time dependent (Feng et al., 2008). This suggests that the statistical information collected from the past corpus may not be relevant to sentences present in the current corpus. Apart from that, the role of string similarity is to identify any misspelled word. A malfunction may occur whenever string similarity deals with any error-free sentences because the purpose for its existence is no longer valid.

For knowledge-based methods, Li et al. (2009) adopted an existing word similarity measure to deal with the similarities of verbs and nouns while the similarities of adjectives and adverbs were measured only based on simple word overlaps. However, Achananuparp et al. (2008) previously showed that the word overlap-based method performed badly in measuring text similarity. Liu et al. (2007) integrated the Dynamic Time Warping (DTW) technique into the similarity measure to identify the distance between words. The main drawback of DTW is that the computational cost and time will increase proportionately with the sentence's length. Wee and Hassan (2008) proposed a method that takes into account the directionality of similarity in which the similarity of any two words is treated as asymmetric. The asymmetric issue between a pair of words was resolved by considering both the similarity of the first word to the second word, and vice versa.

Corley and Mihalcea (2005) proposed a hybrid method by combining six existing knowledge-based methods. Mihalcea et al. (2006) further combined those six knowledge-based methods with two corpus-based methods and claimed that they usually achieved better performance in terms of precision and recall respectively. However, those methods were only combined by using simple average calculation.

Perhaps the most closely related work is a recently proposed query extension technique. Perez-Agüera and Zaragoza (2008) made use of WSD information to map the original query words and the expansion words to WordNet senses. However, without the presence of or considering the surrounding words, the meaning of the expansion words alone tend to be represented by their most general meanings instead of the disambiguated meanings, which results in the possibility of WSD information not being useful for word expansions. In contrast to their work, which is more suitable to be applied on word-to-word similarity task, the method proposed in this paper is more suitable for application on sentence-to-sentence similarity tasks.

Overall, the above-mentioned related works compute similarity based either on statistical information or on a comparison of the nearest meanings in terms of words. None of them compute sentence similarity based on the comparison of actual meanings. Our proposed method, which is a solution to this issue, will be explained in detail in the next section.

3 Sentence Similarity

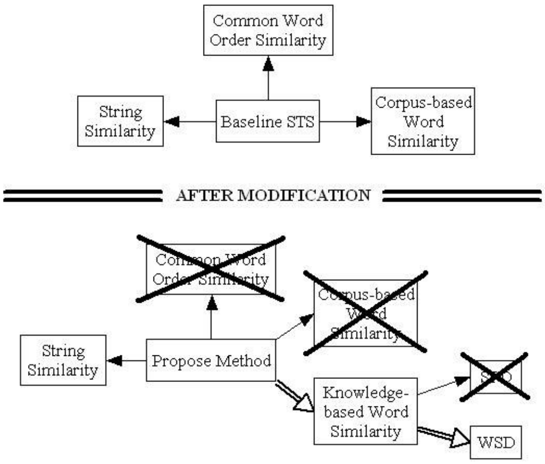

Common Word Order Similarity
String Similarity
Baseline STS
Corpus-based Word Similarity
AFTER MODIFICATION
~~Common Word Order Similarity~~
~~Corpus-based Word Similarity~~
String Similarity
Propose Method
Knowledge-based Word Similarity
WSD

Figure 1. The proposed method

Our proposed method shown in Figure 1, is the outcome of some modifications on an existing method, which is also the most outstanding method, the Semantic Text Similarity (STS) model (Islam and Inkpen, 2008). First of all, CWO is removed from STS as the previous works (Islam and Inkpen, 2007; Islam and Inkpen, 2008) have shown that the presence of CWO has no influence on the outcome. Then, the corpus-based word similarity function of STS is replaced by an existing knowledge-based word similarity measure called YP (Yang and Powers, 2005). Finally, the underlying concept of YP is modified by the integration of WSD and is based on the assumption that any disambiguated sense of a word represents its actual meaning. Thus, the proposed method is also called WSD-STS.

3.1 String similarity measure

The string similarity between two words is measured by using the following equations:

$$v_1 = NLCS(w_i^a, w_j^b) = \frac{l(LCS(w_i^a, w_j^b))^2}{l(w_i) \times l(w_j)} \quad (1)$$

$$v_2 = NMCLCS_1(w_i^a, w_j^b) = \frac{l(MCLCS_1(w_i^a, w_j^b))^2}{l(w_i) \times l(w_j)} \quad (2)$$

$$v_3 = NMCLCS_n(w_i^a, w_j^b) = \frac{l(MCLCS_n(w_i^a, w_j^b))^2}{l(w_i) \times l(w_j)} \quad (3)$$

$$Sim_{string}(X,Y) = 0.33 v_1 + 0.33 v_2 + 0.33 v_3 \quad (4)$$

where $l(x)$ represents the length of x; a and b represent the lengths of sentences X and Y respectively after removing stop words; w_i represents the i-th word in sequence a; w_j represents the j-th word in sequence b; and $Sim_{string}(X,Y)$ represents the overall string similarity. The underlying concept of string similarity is based on character matching. *NLCS* represents the normalized version of the traditional longest common subsequence (LCS) technique in which the lengths of the two words are taken into consideration. $MCLCS_1$ represents the modified version of the traditional LCS in which the string matching must start from the first character while $MCLCS_n$ represents the modified version of the traditional LCS in which the string matching may start from any character. $NMCLCS_1$ and $NMCLCS_n$ represent the normalized versions of $MCLCS_1$ and $MCLCS_n$ respectively. More detailed information regarding string similarity measure can be found in the original paper (Islam and Inkpen, 2008).

3.2 Adopted word similarity measure

Yang and Powers (2005) proposed YP based on the assumptions that every single path in the hierarchical structure of WordNet 1) is identical; and 2) represents the shortest distance between any two connected words. The similarity between two words in sequence a and sequence b can be represented by the following equation:

$$Sim_{word}(w_i^a, w_j^b) = \begin{cases} \alpha_t \prod_{i=1}^{l-1} \beta_{t_i}, & l < \gamma \\ 0, & l \geq \gamma \end{cases} \quad (5)$$

where $0 \leq Sim_{word}(w_i^a, w_j^b) \leq 1$; d is the depth of LCS; l is the length of path between disambiguated w_i^a and w_j^b; t represents the type of path (hypernyms/hyponym, synonym or holonym/meronym) which connects them; α_t represents their path type factor; β_t represents their path distance factor; and γ represents an arbitrary threshold on the distance introduced for efficiency, representing human cognitive limitations. The values of α_t, β_t and γ have already been empirically tuned as 0.9, 0.85 and 12 respectively. More detailed information regarding YP can be found in the original paper (Yang and Powers, 2005).

In order to adapt a different underlying concept, which is the comparison of actual meanings, l has to be redefined as the path distance between disambiguated words, w_i^a and w_j^b. Since YP only differs from the modified version of YP (MYP) in terms of the definition of l, MYP can also be represented by equation (5).

3.3 The proposed measure

The gap

Generally, all the related works in Section 2 can be abstracted as a function of word similarity. This reflects the importance of a word similarity measure in measuring sentence similarity. However, measuring sentence similarity is always a more complicated task than measuring word similarity. The reason is that while a word similarity measure only involves a single pair of words, a sentence similarity measure has to deal with multiple pairs of words. In addition, due to the presence of the surrounding words in a sentence, the possible meaning of a word is always being restricted (Kolte and Bhirud, 2008). Thus, without some modifications, the traditional word similarity measures, which are based on the concept of a comparison of the nearest meanings, are inapplicable in the context of sentence similarity measures.

The importance of WSD in reducing the gap

Before performing the comparison of actual meanings, WSD has to be integrated so that the most suitable sense can be assigned to any polysemous word. The importance of WSD can be investigated by using a simple example. Consider a pair of sentences, collected from WordNet 2.1, which use two words, "*dog*" and "*cat*":

X: The dog barked all night.

Y: What a cat she is!

Based on the definition in WordNet 2.1, the word "*dog*" in *X* is annotated as the first sense which means "*a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times*". Meanwhile, the word "*cat*" in *Y* is annotated as the third sense with the definition of "*a spiteful woman's gossip*". The path distance between "*cat*" and "*dog*" based on their actual senses is equal to 7. However, their shortest path distance (SPD), which is based on their nearest senses, is equal to 4. SPD is the least number of edges connecting two words in the hierarchical structure of WordNet. In other words, "*cat*" and "*dog*" in *X* and *Y* respectively, are not as similar as the one measured by using SPD. The presence of the additional path distances is significant as it is almost double the actual path distance between "*cat*" and "*dog*".

WSD-STS

The adopted sentence similarity measure, STS, can be represented by the following equations:

$$Sim_{semantic}(X,Y) = \frac{(\delta + \sum_{i=1}^{c} \tau_i) \times (a+b)}{2ab} \quad (6)$$

$$SIM(X,Y) = \frac{Sim_{smeantic}(X,Y) + Sim_{string}(X,Y)}{2} \quad (7)$$

where for equation (6): δ represents the number of overlapped words between the words in sequence a and sequence b; c represents the number of semantically matched words between the words in sequence a and sequence b, in which $c = a$ if $a < b$ or $c = b$ if $b < a$, τ_i represents the highest matching similarity score of i-th word in the shorter sequence with respect to one of the words in the longer sequence; and $\Sigma\tau$ represents the sum of the highest matching similarity score between the words in sequence a and sequence b.

For STS, the similarity between two words is measured by using a corpus-based measure. For WSD-STS, this corpus-based measure is replaced by MYP. Finally, the overall sentence similarity is represented by equation (7).

4 Experimental Design

4.1 Data set

Li et al., (2006) constructed a data set which consists of 65 pairs of human-rated sentences by applying the similar experimental design for creating the standard data set for the word similarity task (Rubenstein and Goodenough, 1965). These 65 sentence pairs were the definitions collected from the Collin Cobuild Dictionary. Out of these, 30 sentence pairs with rated similarity scores that ranged from 0.01 to 0.96 were selected as test data set. The corresponding 30 word pairs for these 30 sentence pairs are shown in the second column of Table 1. A further set of 66 sentence pairs is still under development and it will be combined with the existing data set in the future (O'Shea et al., 2008b).

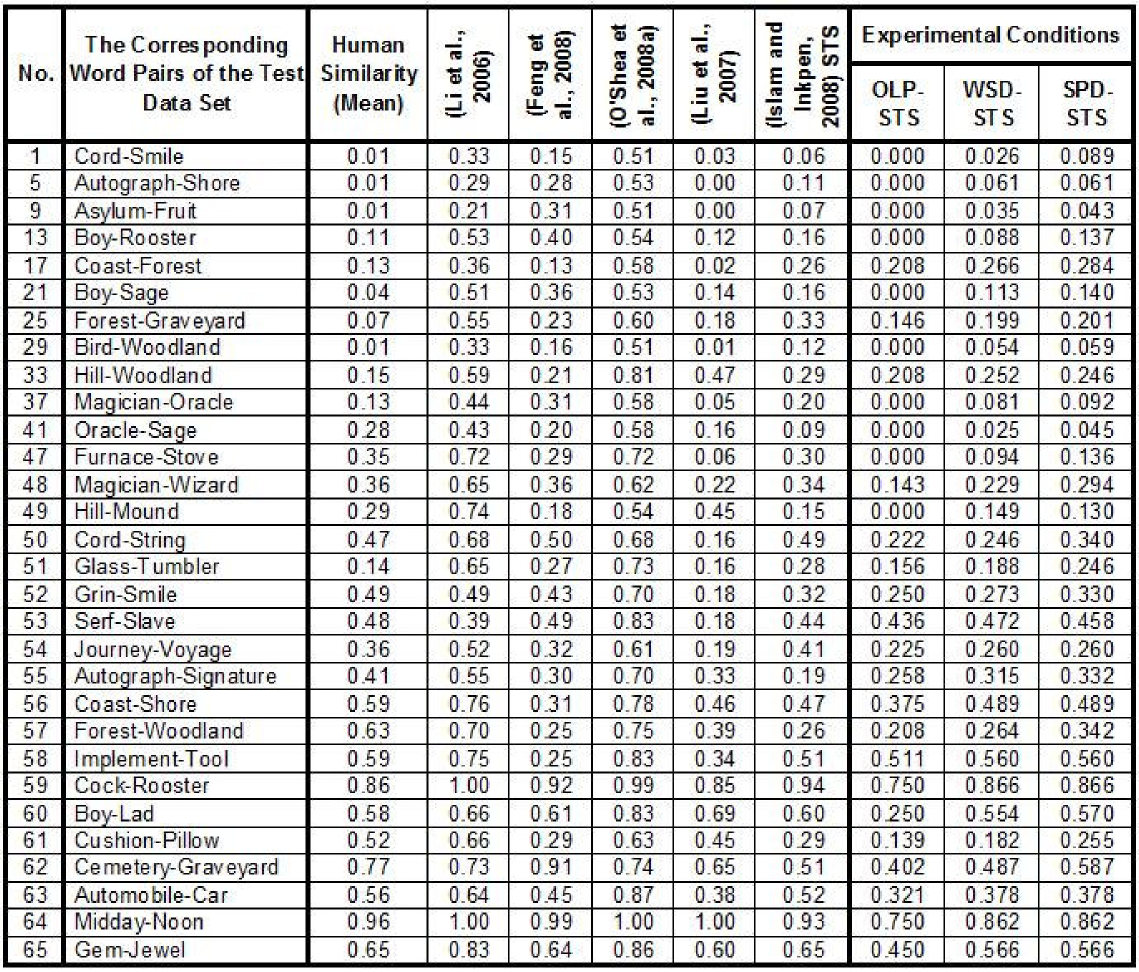
4.2 Procedure

Firstly, Stanford parser[1] is used to parse each sentence and to tag each word with a part of speech (POS). Secondly, Structural Semantic Interconnections[2] (SSI), which is an online WSD system, is used to disambiguate and to assign a sense for each word in the 30 sentences based on the assigned POS. SSI is applied based on the assumption that it is able to perform WSD correctly. The main reason for choosing SSI to perform WSD is its promising results reported in a study by Navigli and Verladi (2006). Thirdly, all the stop words which exist in these 30 pairs of sentences are removed. It is important to note that the 100 most frequent words collected from British National Corpus (BNC) were applied as the stop words list on the baseline, STS. However, due to the limited accessibility to BNC, a different stop words list[3], which is available online, is applied in this paper.

[1] http://nlp.stanford.edu/software/lex-parser.shtml

[2] http://lcl.uniroma1.it/ssi

[3] http://www.translatum.gr/forum/index.php?topic=2476.0



No.	The Corresponding Word Pairs of the Test Data Set	Human Similarity (Mean)	(Li et al., 2006)	(Feng et al., 2008)	(O'Shea et al., 2008a)	(Liu et al., 2007)	(Islam and Inkpen, 2008) STS	Experimental Conditions		
								OLP-STS	WSD-STS	SPD-STS
1	Cord-Smile	0.01	0.33	0.15	0.51	0.03	0.06	0.000	0.026	0.089
5	Autograph-Shore	0.01	0.29	0.28	0.53	0.00	0.11	0.000	0.061	0.061
9	Asylum-Fruit	0.01	0.21	0.31	0.51	0.00	0.07	0.000	0.035	0.043
13	Boy-Rooster	0.11	0.53	0.40	0.54	0.12	0.16	0.000	0.088	0.137
17	Coast-Forest	0.13	0.36	0.13	0.58	0.02	0.26	0.208	0.266	0.284
21	Boy-Sage	0.04	0.51	0.36	0.53	0.14	0.16	0.000	0.113	0.140
25	Forest-Graveyard	0.07	0.55	0.23	0.60	0.18	0.33	0.146	0.199	0.201
29	Bird-Woodland	0.01	0.33	0.16	0.51	0.01	0.12	0.000	0.054	0.059
33	Hill-Woodland	0.15	0.59	0.21	0.81	0.47	0.29	0.208	0.252	0.246
37	Magician-Oracle	0.13	0.44	0.31	0.58	0.05	0.20	0.000	0.081	0.092
41	Oracle-Sage	0.28	0.43	0.20	0.58	0.16	0.09	0.000	0.025	0.045
47	Furnace-Stove	0.35	0.72	0.29	0.72	0.06	0.30	0.000	0.094	0.136
48	Magician-Wizard	0.36	0.65	0.36	0.62	0.22	0.34	0.143	0.229	0.294
49	Hill-Mound	0.29	0.74	0.18	0.54	0.45	0.15	0.000	0.149	0.130
50	Cord-String	0.47	0.68	0.50	0.68	0.16	0.49	0.222	0.246	0.340
51	Glass-Tumbler	0.14	0.65	0.27	0.73	0.16	0.28	0.156	0.188	0.246
52	Grin-Smile	0.49	0.49	0.43	0.70	0.18	0.32	0.250	0.273	0.330
53	Serf-Slave	0.48	0.39	0.49	0.83	0.18	0.44	0.436	0.472	0.458
54	Journey-Voyage	0.36	0.52	0.32	0.61	0.19	0.41	0.225	0.260	0.260
55	Autograph-Signature	0.41	0.55	0.30	0.70	0.33	0.19	0.258	0.315	0.332
56	Coast-Shore	0.59	0.76	0.31	0.78	0.46	0.47	0.375	0.489	0.489
57	Forest-Woodland	0.63	0.70	0.25	0.75	0.39	0.26	0.208	0.264	0.342
58	Implement-Tool	0.59	0.75	0.25	0.83	0.34	0.51	0.511	0.560	0.560
59	Cock-Rooster	0.86	1.00	0.92	0.99	0.85	0.94	0.750	0.866	0.866
60	Boy-Lad	0.58	0.66	0.61	0.83	0.69	0.60	0.250	0.554	0.570
61	Cushion-Pillow	0.52	0.66	0.29	0.63	0.45	0.29	0.139	0.182	0.255
62	Cemetery-Graveyard	0.77	0.73	0.91	0.74	0.65	0.51	0.402	0.487	0.587
63	Automobile-Car	0.56	0.64	0.45	0.87	0.38	0.52	0.321	0.378	0.378
64	Midday-Noon	0.96	1.00	0.99	1.00	1.00	0.93	0.750	0.862	0.862
65	Gem-Jewel	0.65	0.83	0.64	0.86	0.60	0.65	0.450	0.566	0.566

Table 1. Data Set Results

Finally, the remaining content words are lemmatized by using Natural Language Toolkit[4] (NLTK). Nevertheless, those words which can be found in WordNet and which have different definitions from their lemmatized form will be excluded from lemmatization. For instance,

Cooking[NN] can be a great art.

The word in the bracket represents the tagged POS for its corresponding word. Since based on the definitions provided by WordNet, "*cooking*", which is tagged as a noun, has a different meaning from its lemmatized form "*cook*", which is also tagged as a noun. Therefore, "*cooking*" is excluded from lemmatization.

4.3 Experimental conditions

Sentence similarity is measured under the following three conditions:

- **OLP-STS**: A modified version of the baseline, STS (Islam and Inkpen, 2008), in which it only relies on the presence of overlapped words. This means that the component $\sum_{i=1}^{c} \tau_i$, which represents the word similarity, is removed from equation (6).
- **SPD-STS**: The corpus-based word similarity measure of the baseline, STS, which is represented by $\sum_{i=1}^{c} \tau_i$ in equation (6), is replaced by a knowledge-based word similarity measure, YP.
- **WSD-STS**: A modified version of SPD-STS in which the knowledge-based measure, YP, is replaced by MYP.

As mentioned in Section 4.2, different stop words lists were applied between the baseline and the proposed methods under different ex-

[4] http://www.nltk.org/

perimental conditions in this paper. Since this issue may be questioned due to the unfair comparison, the performance of WSD-STS is evaluated on top of a number of different stop words lists which are available online in order to investigate any influence which may be caused by stop words list.

5 Results and Discussion

Table 1 presents the similarity scores obtained from the mean of human ratings, the benchmarks, and different experimental conditions of the proposed methods. Figure 2 presents the corresponding Pearson correlation coefficients of various measures as listed in Table 1.

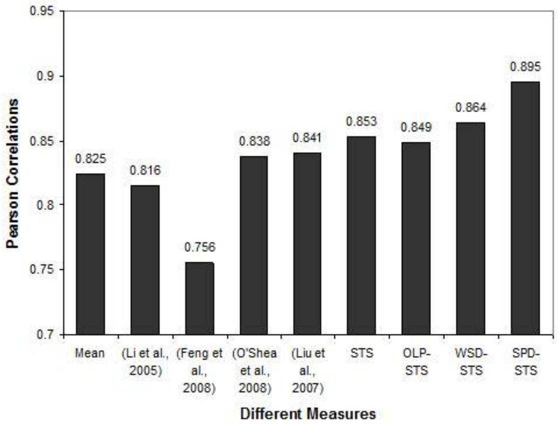


Figure 2. Pearson Correlation Coefficient

Figure 2 shows that STS appears to be the most outstanding measure among the existing works with a correlation coefficient of 0.853. However, Figure 2 also shows that both the proposed methods in this paper, WSD-STS and SPD-STS, outperform STS. This result indicates that knowledge-based method tends to perform better than a corpus-based method. The reason is that a knowledge base is much closer to human representation of knowledge (WordNet is the knowledge base applied in this paper) than a corpus. A corpus only reflects the usage of languages and words while WordNet is a model of human knowledge constructed by many expert lexicographers (Li et al., 2006). In other words, a corpus is more likely to provide unprocessed raw data while a knowledge base tends to provide ready-to-use information.

The results of the performance of the two proposed methods are as expected. SPD-STS achieved a bigger but statistically insignificant improvement while WSD-STS achieved a smaller but statistically significant improvement at 0.01 levels. The significance of a correlation is calculated by using an online calculator, *VassarStats*[5]. The reason for the variance in the outcomes between SPD-STS and WSD-STS is obvious; it is the difference in terms of their underlying concepts. In other words, sentence similarity computation, which is based on a comparison of the nearest meanings, results in insignificant improvement while sentence similarity computation, which is based on a comparison of actual meanings, achieves statistically significant improvement. These explanations indicate that WSD is essential in confirming the validity of the task of measuring sentence similarity.

Figure 2 also reveals that a relatively low correlation is achieved by OLP-STS. This is not at all surprising since Achananuparp et al. (2008) has already demonstrated that the overlapped word-based method tends to perform badly in measuring sentence similarity. However, it is interesting to find that the difference in performance between STS and OLP-STS is very small. This indirectly suggests that the presence of the string similarity measure and the corpus-based word similarity measure has only a slight improvement on the performance of OLP-STS.

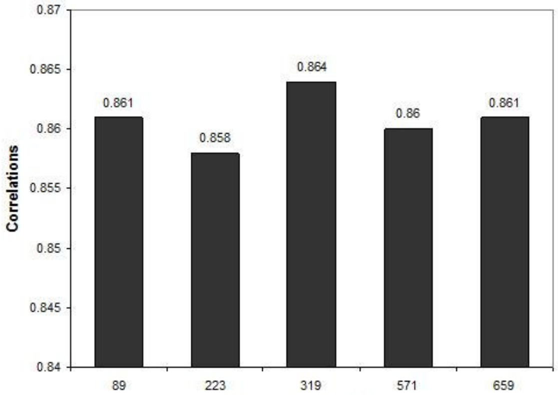


Figure 3. The performance of the WSD-SPD versus different stop words lists

Next, in order to address the issue of unfair comparison due to the usage of different stop words lists, the performance of WSD-SPD has been evaluated on top of a number of different

[5] http://faculty.vassar.edu/lowry/rdiff.html?

stop words lists. A total of five stop words lists with different lengths (89[6], 223[7], 319, 571[8] and 659[9]) of stop words were applied. The performances of WSD-SPD with respect to these stop words lists are portrayed in Figure 3. They are found to be in a comparable condition. This result connotes that the influence caused by the usage of different stop words lists is small and can be ignored. Hence, the unfair comparison between our proposed method and the baseline should not be treated as an issue for the benchmarking purpose of this paper.

On the other hand, although an assumption is made that SSI performs WSD correctly, we noticed that not all the words were disambiguated confidently. The confident scores which were assigned to the disambiguated words by SSI range between 30% and 100%. These confident scores reflect the confidence of SSI in performing WSD. Thus, it is possible that some of those words which were assigned with low confident scores were disambiguated incorrectly. Consequently, the final sentence similarity score is likely to be affected negatively. In order to reduce the negative effect which may be caused by incorrect WSD, any words pair which is not confidently disambiguated is assigned the similarity score based on the concept of comparing the nearest meanings instead of comparing the actual meanings. In other words, WSD-STS and SPD-STS are combined and results in WSD-SPD. The performance of WSD-SPD across a range of confident scores is essential in revealing the impact of WSD and SPD on the task of measuring sentence similarity.

Figure 4 outlines the performance achieved by WSD-SPD across different confident scores assigned by SSI. The confident score of at least 0.7 is identified as the threshold in which SSI optimizes its performance. The performance of WSD-SPD is found to be statistically insignificant for those confident scores above the threshold. The explanation for this phenomenon can be found in Figure 5. Figure 5 illustrates the percentage of the composition between WSD and SPD in WSD-SPD. It is obvious that once the portion of WSD exceeds the portion of SPD, the performance of WSD-SPD is found to be statistically insignificant. This finding suggests that SPD, which reflects the application of the concept of nearest meaning comparison, is likely to decrease the validity of sentence similarity measurement while WSD, which reflects the application of the concept of actual meaning comparison, is essential in confirming the validity of sentence similarity measurement.

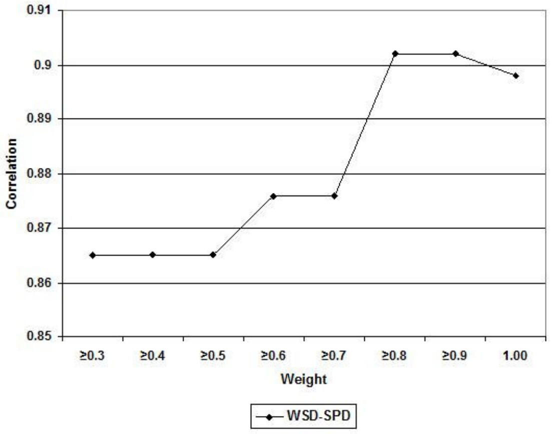


Figure 4. The performance of WSD-SPD versus confident scores

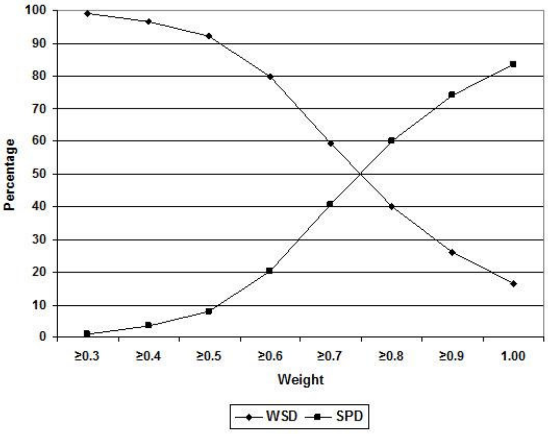


Figure 5. The percentage of WSD/SPD versus confident score

The trend of the performance of string similarity measure and word similarity measure with respect to different weight assignments is delineated in Figure 6. The lowest correlation of 0.856 is obtained when only the string similarity function is considered while the word similarity

[6] http://msdn.microsoft.com/en-us/library/bb164590.aspx

[7] http://snowball.tartarus.org/algorithms/english/stop.txt

[8] http://truereader.com/manuals/onix/stopwords2.html

[9] http://www.link-assistant.com/seo-stop-words.html

function is excluded. A better performance is achieved by taking the two measures into consideration where more weight is given to the measure of word similarity. This trend intimates that the string similarity measure offers a smaller contribution in measuring sentence similarity than word similarity measure. In contrast to a word similarity measure, a string similarity measure is purposely proposed to address the issue of misspelled words. Since the data set applied in this experiment does not contain any misspelled words, it is obvious that a string similarity measure performs badly. In addition, the underlying concept of string similarity is questionable. Does it make sense to determine the similarity of two words based on the matching between their characters or the matching of the sequence of characters? Consider four pairs of words: "*play*" versus "*pray*", "*plant*" versus "*plane*", "*plane*" versus "*plan*" and "*stationary*" versus "*stationery*". These word pairs are highly similar in terms of characters but they are semantically dissimilar or unrelated.

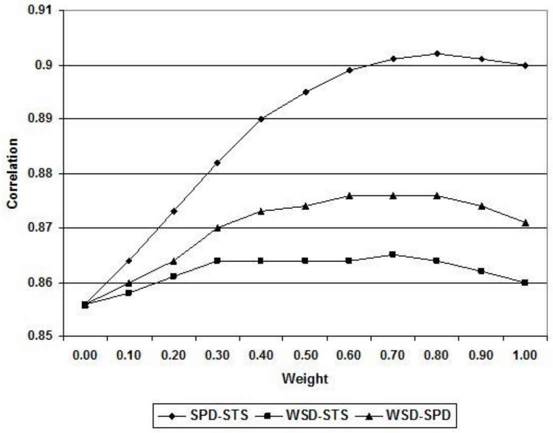


Figure 6. The performance of the different measures versus the weight between string similarity and word similarity

Figure 6 also depicts that the combination of word similarity measure (70%) and string similarity measure (30%) performs better than the measure which is solely based on word similarity function. It is obvious that the difference is caused by the presence of string similarity measure. The combination assigns similarity scores to all word pairs while the word similarity measure only assigns similarity scores to those word pairs which fulfill two requirements: 1) any two words which share an identical POS, and 2) any two words which must either be a pair of nouns or a pair of verbs. In fact, adjectives and adverbs do contribute to representing the meaning of a sentence although their contribution is relatively smaller than the contribution of nouns and verbs (Liu et al., 2007; Li et al., 2009). Therefore, by ignoring the presence of adjectives and adverbs, the performance will definitely be affected negatively.

6 Conclusion

This paper has presented a knowledge-based method which measures the similarity between two sentences based on their actual meaning comparison. The result shows that the proposed method, which is a knowledge-based measure, performs better than the baseline, which is a corpus-based measure. The improvement obtained is statistically significant at 0.025 levels. This result also shows that the validity of the output of measuring the similarity of two sentences can be improved by comparing their actual meanings instead of their nearest meanings. These are achieved by transforming the baseline into a knowledge-based method and then by integrating WSD into the adopted knowledge-based measure.

Although the proposed method significantly improves the quality of measuring sentence similarity, it has a limitation. The proposed method only measures the similarity between two words with an identical part of speech (POS) and these two words must either be a pair of nouns or a pair of verbs. By ignoring the importance of adjectives and adverbs, and the relationship between any two words with different POS, a slight decline is observed in the obtained result. In future research, these two issues will be addressed by taking into account the relatedness between two words instead of only considering their similarity.

References

Achananuparp, Palakorn, Xiao-Hua Hu, and Xiao-Jiong Shen. 2008. The Evaluation of Sentence Similarity Measures. *In Proceedings of the 10th International Conference on Data Warehousing* *and Knowledge Discovery (DaWak)*, pages 305-316, Turin, Italy.

Corley, Courtney, and Rada Mihalcea. 2005. Measuring the Semantic Similarity of Texts. *In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 48-55, Ann Arbor.

Feng, Jin, Yi-Ming Zhou, and Trevor Martin. 2008. Sentence Similarity based on Relevance. *In Proceedings of IPMU*, pages 832-839.

Islam, Aminul, and Diana Inkpen. 2007. Semantic Similarity of Short Texts. *In Proceedings of RANLP*, pages 291-297.

Islam, Aminul, and Diana Inkpen. 2008. Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):10.

Kauchak, David, and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. *In Proceedings of HLT-NAACL*, pages 455-462, New York.

Kolte, Sopan Govind, and Sunil G. Bhirud. 2008. Word Sense Disambiguation using WordNet Domains. *In The First International Conference on Emerging Trends in Engineering and Technology*, pages 1187-1191.

Lapata, Mirella, and Regina Barzilay. 2005. Automatic Evaluation of Text Coherence: Models and Representations. *In Proceedings of the 19th International Joint Conference on Artificial Intelligence*.

Li, Lin, Xia Hu, Bi-Yun Hu, Jun Wang, and Yi-Ming Zhou. 2009. Measuring Sentence Similarity from Different Aspects. *In Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, pages 2244-2249.

Li, Yu-Hua, David McLean, Zuhair A. Bandar, James D.O'Shea, and Keeley Crockett. 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138-50.

Liu, Xiao-Ying, Yi-Ming Zhou, and Ruo-Shi Zheng. 2007. Sentence Similarity based on Dynamic Time Warping. *In The International Conference on Semantic Computing*, pages 250-256.

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *In Proceedings of the American Association for Artificial Intelligence.*

Navigli, Roberto, and Paola Velardi. 2005. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(7):1075-86.

O'Shea, James, Zuhair Bandar, Keeley Crockett, and David McLean. 2008a. A Comparative Study of Two Short Text Semantic Similarity Measures. *In KES-AMSTA, LNAI: Springer Berlin / Heidelberg.*

O'Shea, James, Zuhair Bandar, Keeley Crockett, and David McLean. 2008b. Pilot Short Text Semantic Similarity Benchmark Data Set: Full Listing and Description.

Perez-Aguera, Jose R., and Hugo Zaragoza. 2008. UCM-Y!R at Clef 2008 Robust and WSD Tasks. *In Working Notes for CLEF Workshop.*

Rubenstein, Herbert, and John B. Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, pages 627-633.

Wang, Yao-Feng, Yue-Jie Zhang, Zhi-Ting Xu, and Tao Zhang. 2006. Research on Dual Pattern of Unsupervised and Supervised Word Sense Disambiguation. *In Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, pages 2665-2669.

Wee, Leong Chee, and Samer Hassan. 2008. Exploiting Wikipedia for Directional Inferential Text Similarity. *In Proceedings of Fifth International Conference on Information Technology: New Generations*, pages 686-691.

Yang, Cha, and Jun Wen. 2007. Text Categorization Based on Similarity Approach. *In Proceedings of International Conference on Intelligence Systems and Knowledge Engineering (ISKE).*

Yang, Dong-Qiang, and David M.W. Powers. 2005. Measuring Semantic Similarity in the Taxonomy of WordNet. *In Proceedings of the 28th Australasian Computer Science Conference*, pages 315-332, Australia.

Zhou, Liang, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. ParaEval: Using Paraphrases to Evaluate Summaries Automatically. *In Proceedings of Human Language Technology Conference of the North American Chapter of the ACL*, pages 447-454, New York.