Addressing the Variability of Natural Language Expression in Sentence Similarity with Semantic Structure of the Sentences

Palakorn Achananuparp, Xiaohua Hu, and Christopher C. Yang

College of Information Science and Technology
Drexel Univeristy, Philadelphia PA 19104
pkorn@drexel.edu, thu@cis.drexel.edu,
chris.yang@ischool.drexel.edu

Abstract. In this paper, we present a new approach that incorporates semantic structure of sentences, in a form of verb-argument structure, to measure semantic similarity between sentences. The variability of natural language expression makes it difficult for existing text similarity measures to accurately identify semantically similar sentences since sentences conveying the same fact or concept may be composed lexically and syntactically different. Inversely, sentences which are lexically common may not necessarily convey the same meaning. This poses a significant impact on many text mining applications' performance where sentence-level judgment is involved. The evaluation has shown that, by processing sentence at its semantic level, the performance of similarity measures is significantly improved.

Keywords: Sentence similarity, structural similarity, sentence semantics, verb-argument structure, semantic equivalence, textual entailment.

1 Introduction

A major issue that many text mining applications have to deal with is the variability of natural language expression. Due to flexibility of human language, the same information can be formulated in numerous ways. This has great implication in text summarization and question answering applications where the identification of redundant information is crucial to system performance. Another important issue pertains to the notion of similarity between sentences. Recent development has been made on this issue. In information retrieval field, different levels of topical similarity between sentences are proposed [12][15]. Relevant sentences either address the same specific topics or they might talk about the similar general topics. In natural language processing (NLP), two notions of sentence similarity, semantic equivalence and entailment [9], are understudied. Because they are strongly related, thus, a clear distinction is sometimes difficult to define. In general, two sentences are said to be semantically equivalent if they share the exact same meaning. Following this definition, paraphrase sentences are considered the most common form of semantic equivalency. On the other hand, entailment focuses on unidirectional inference. If the

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 548–555, 2009.

meaning of one sentence can be inferred from the other sentence, the two sentences are said to be an entailment pair.

In this paper, we propose the method that takes into account semantic structure of sentences to measure sentences similarity. Traditionally, sentences are transformed into a bag of words for sentence similarity computation. This results in “semantic loss” because syntactic construction of the sentences is discarded. In particular, this has a crucial consequence to the identification of semantic equivalence or entailment sentences in which a specific inference has to be made. Our proposed method aims to deal with this issue by utilizing semantic roles of constituents in the sentences and computing sentence similarity at verb-argument structure level.

2 Related Work

The issue of measuring similarity of sentences is gaining more attention from various research communities. Although the perceived notions of text similarity may differ depending on the application domains, many of them share a common goal of matching up semantically similar sentences. To that end, various techniques have been proposed. First, probabilistic approaches have been adopted to identify topically related sentences [12] [15] [13] in sentence retrieval application. Next, several unsupervised approaches have been proposed for paraphrase recognition tasks [4][9][7] [14]. Recently, natural language processing community has increasingly focused on developing NLP systems to recognize entailment between sentences [8]. For this task, systems that employ extensive linguistic tools, such as logical inference engine and anaphora resolution [11][20], have started to show significant improvement in result over relatively “shallower” approaches. Nevertheless, this comes with a trade off in computational cost which makes comprehensive NLP systems currently impractical for a large text collection. Motivated by works in related areas [5][19], our work aims to address the shortcoming of existing text similarity measures when applying to sentence similarity task. That is, most techniques either simply treat sentences as single text unit or transform them into a bag of words representation; thus ignoring syntactic and semantic relations between constituents. Specifically, we are not aware of any methods that compute the similarity of sentence pairs at their semantic structure level.

3 Method

The semantic structure of sentences, referred to as *verb-argument structure*, encodes the relations between individual components and their semantic roles with respect to a given verb in a sentence. The labels of semantic roles are varied depending on the annotation scheme [3] [16]. Generally, *rel* defines a verb or relation between two or more arguments. Arg0 denotes a prototypical agent, Arg1 indicates a prototypical patient or theme of a given verb, and ArgM represents adjunctive argument (e.g. ArgM-LOC specifies location-related argument). For example, a simple sentence “John broke a glass” can be decomposed into one verb-argument structure [$_{\text{Arg0}}$ John]

[$_{\text{rel}}$ broke] [$_{\text{Arg1}}$ a glass]. It consists of two arguments that correspond to a verb *broke*. *John* is labeled with Arg0 indicating that it is a prototypical agent and has a semantic role of *breaker* while *glass* is annotated with Arg1 denoting a prototypical patient, with *thing broken* as its semantic role. Notice that this is parallel to a grammatical relation where John is a subject and glass is a direct object of this sentence. By measuring semantic similarity of verb-argument structures, we can improve the effectiveness of sentence similarity measures despite the syntactic variability of language expression.

Based on the motivation that sentences that express the same meaning, in terms of event or idea, should share similar verb-argument structures, we decompose sentences into a set of verb-argument structures instead of comparing two unstructured text segments. Given two sentences s_i and s_j, the similarity score between verb-argument structures k of sentence s_i (v_{ik}) and verb-argument structure l of sentence s_j (v_{jl}) is determined by the similarity between their verbs $(sim_{rel}(v_{ik}, v_{jl}))$ and the sum of similarities between the corresponding arguments $(sim_{arg_n}(v_{ik}, v_{jl}))$. A scaling factor[1] $e^{-k \cdot score(v_{ik}, v_{jl})}$is applied to normalize similarity score to [0,1] in equation 2.

$$score(v_{ik}, v_{jl}) = \alpha \cdot sim_{rel}(v_{ik}, v_{jl}) + \beta \cdot \sum_n sim_{arg_n}(v_{ik}, v_{jl}) \tag{1}$$

$$sim_{va}(v_{ik}, v_{jl}) = \begin{cases} score(v_{ik}, v_{jl}) & \text{if } score(v_{ik}, v_{jl}) \leq 1 \\ e^{-k \cdot score(v_{ik}, v_{jl})} & \text{otherwise} \end{cases} \tag{2}$$

where α and β are coefficients that control the influence of semantic similarity between verbs and the sum of semantic similarity between arguments in verb-argument structures. The question is what is a reasonable weight for verb similarity and argument similarity components? Following a well-known psychological experiment of sentence sorting by [10] which suggests that verb is one of the main determinants of sentence meaning, we allocate higher importance factor to verb component than argument components[2]. For $sim_{rel}(v_{ik}, v_{jl})$, we employ WordNet-based gloss overlap measure [1] to compute word similarity score between verbs or verb phrases. To compute $sim_{arg_n}(v_{ik}, v_{jl})$, we treat argument text as phrase and utilize sentence-level similarity measures [2], e.g. Jaccard coefficient, phrasal overlap measure, etc., to determine their similarity.

$$sim(s_i, s_j) = \max_{k,l}[sim_{va}(v_{ik}, v_{jl})] \tag{3}$$

Finally, the similarity between sentence s_i and s_j is derived from verb-argument pair which produces the maximum similarity score as specified in equation 3. According to internal validation, deriving sentence similarity score by maximizing sim_{va} consistently produces better performance than linearly combining or averaging sim_{va}. This explains that sentence meaning tends to be dominated by one major verb-argument structure.

The whole computation process can be described as follows. First, we use semantic role labeler *SENNA* [6] to annotate each constituent in the sentences with their

[1] In this study, we consider k=0.05 as the optimal value.

[2] Based on our experiment, we find that weighting α and β at 0.5 gives the best result.

semantic roles. After the verb-argument structures are extracted by semantic role labeler, verb-argument text is parsed to identify the verb and individual arguments. During the parsing step, we apply syntactic rules to replace single verb with verb phrase if one exists in verb-argument text. In addition, we remove any words that are not part of the longest noun phrases in argument components. For example, given a verb-argument structure "[$_{\text{Arg1}}$ BBC] [$_{\text{rel}}$ stands] [$_{\text{Arg2}}$ for British Broadcasting Corporation]", a single verb "stands" will be expanded into a verb phrase "stands for". Arg1 text contains "BBC" and Arg2 text contains "British Broadcasting Corporation". Moreover, we perform noun denominalization on verb-argument structures of generic sentences, those which contain auxiliary verb, by expanding them into a new verb-argument structure. The expanded structure shares the same set of arguments as the original one while the auxiliary verb is replaced with a verb form of its forward adjacent noun. After that, the denominalized noun is removed from the corresponding argument. For instance, a verb-argument "[$_{\text{Arg1}}$ BBC] [$_{\text{rel}}$ is] [$_{\text{Arg2}}$ the abbreviation of British Broadcasting Corporation] will be expanded into "[$_{\text{Arg1}}$ BBC] [$_{\text{rel}}$ abbreviates] [$_{\text{Arg2}}$ of British Broadcasting Corporation].

After verb and argument components are identified, the next step is to determine the value of verb similarity component ($sim_{rel}(v_{ik}, v_{jl})$) and argument similarity component ($sim_{arg_n}(v_{ik}, v_{jl})$).WordNet-based gloss overlap measure is employed to compute word similarity score between verbs or verb phrases in $sim_{rel}(v_{ik}, v_{jl})$. Next, different sentence-level similarity measures are utilized to calculate $sim_{arg_n}(v_{ik}, v_{jl})$. For ArgM, we collapse all of its subtypes (e.g. ArgM-LOC, ArgM-TMP, ArgM-DIR, etc.) into single category and find its maximum score from all possible cross-argument pair comparison (ArgM vs. Arg0, ArgM vs. Arg1, etc.) This is done to deal with the cases where ArgM is not accurately tagged by semantic role labeler.

4 Experimental Evaluation

4.1 Data Sets

We use two publicly-available sentence pair data sets, Microsoft Research paraphrase corpus (MSRP) [9] and the third PASCAL recognising textual entailment challenge (RTE3) data set [8], to evaluate the performance of the similarity measures.

MSRP contains 5,801 sentence pairs (4,076 training pairs and 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. 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.

RTE3 consists of 800 sentence pairs from the development set and 800 sentence pairs from the test set. Each pair comprises two small text segments, which are referred to as *text* and *hypothesis*. 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.

4.2 Evaluation Setting

We define five 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. *F_1* is a uniform harmonic mean of precision and recall. *Accuracy* is a proportion of all correctly predicted sentences compared to all sentences. A scoring threshold for positive pairs is defined at 0.5 as it is used in the literature [14].

Three baseline similarity measures are employed to compare the performance with the proposed method. These include Jaccard coefficient (sim_{jac}), Sumo-metric (sim_{sumo}) [7], and n-gram phrasal overlap measure ($sim_{overlap}$). For a detail description of these measures, please see [2]. We first compute baseline similarity scores between unstructured sentence pairs. After that, we generate structural similarity scores where each baseline sentence similarity measure is utilized in argument similarity (sim_{arg_n}) computation. For example, sim_{va}^{jac} denotes the structural similarity measure having Jaccard coefficient as the underlying method for computing argument similarity scores.

5 Results and Discussion

5.1 Experimental Results

Table 1. Comparison of the performance of structural similarity measures on paraphrase recognition task. Results with * indicate that the differences are not statistically significant compared to baseline.

Measure	Recall	Precision	F_1	Accuracy
sim_{va}^{jac}	0.9371	0.6806	0.7887	0.6701
sim_{va}^{sumo}	0.9305	0.6798	0.7856	0.6667
$sim_{va}^{overlap}$	0.9758	0.6753	**0.7982**	**0.6756***
	Baseline			
sim_{jac}	0.6033	0.8347	0.7000	0.6568
sim_{sumo}	0.3967	0.5398	0.4872	0.4446
$sim_{overlap}$	0.8919	0.7001	0.7848	0.6748

Table 1 presents a comparison of structural similarity approaches to the traditional sentence similarity approaches. In this experiment, the best measures ($sim_{va}^{overlap}$)

metric performs slightly better than its baseline counterpart ($sim_{overlap}$). Comparisons between related measures (e.g. sim_{va}^{jac} and sim_{jac}, etc.) reveal that structural similarity approaches improve the performance of the corresponding sentence similarity measures. For example, sim_{jac} and sim_{sumo} perform significantly better (F$_1$ scores increase by 12.67% and 61.25%, respectively) when applying to verb-argument structures as opposed to unstructured sentences.

Table 2. Comparison of the performance of structural similarity measures on textual entailment recognition task. Results with * indicate that the differences are not statistically significant compared to baseline.

Measure	Recall	Precision	F$_1$	Accuracy
sim_{va}^{jac}	0.6724	0.5945	0.6310	**0.5911***
sim_{va}^{sumo}	0.7189	0.5424	0.6183	0.5413
$sim_{va}^{overlap}$	0.7734	0.5688	**0.6555**	0.5767
		Baseline		
sim_{jac}	0.0512	0.6363	0.0948	0.4988
sim_{sumo}	0.2146	0.6069	0.3171	0.5263
$sim_{overlap}$	0.4561	0.6493	0.5358	0.5950

According to Table 2, the best performance is achieved by $sim_{va}^{overlap}$ at F$_1$ score of 0.6555 and sim_{va}^{jac} at accuracy level of 0.5911. A closer look at each related measures also shows a trend similar to paraphrase recognition task. The use of verb-argument structure has improved the performance of many naive sentence similarity measures. For instance, on F$_1$ metric, the performance of sim_{jac}, sim_{sumo}, and $sim_{overlap}$ has substantially increased by 560% (from 9.48% to 63.10%), 94.99% (from 31.71% to 61.83%), and 22.34% (from 53.58% to 65.55%), respectively. According to the result, we conclude that structural approach offers greater benefit to similarity computation of highly asymmetric sentences (those in RTE3 data) than those which are more symmetric in length.

5.2 Shallow vs. Deep Semantic Processing

The overall result in section 5.1 differs from that of text categorization task [19] where concept-based weighting has significantly improved classification performance over the traditional *tf-idf* scheme. One reason is that conceptual term frequency aims to capture the importance of a given concept in a document by leveraging the frequency of a concept in verb-argument structures. This approach is better suited for text categorization mechanism in which documents are classified according to their topics represented by terms or concepts in documents. On the other hand, the task of identifying semantic equivalence or entailment pairs requires a deeper understanding of sentence meaning. As shown in section 5.1, deeper semantic measures are able to recognize at least the same or greater number of positive pairs according to F$_1$ scores than those of vector space approach. The magnitude of improvement is even more apparent in entailment task in which specific relations between constituents have to be identified.

5.3 The Advantage of Structural Approach over Linguistic Measures

Linguistic measures, those that employ natural language resources such as WordNet, has been proven to be highly effective in sentence similarity task [14]. However, a major criticism of such approaches is the lack of computational efficiency due to the exhaustive calculation of semantic similarity between word pairs. Therefore, they might not be as robust to employ in the real-world text mining applications as most naïve measures. In this regard, our approach offers a greater benefit over linguistic measures as it greatly improves the effectiveness of naïve measures while maintaining their computational efficiency, particularly at sentence processing time.

6 Conclusions

In this paper, we present the method that integrates semantic structure of the sentences to handle variability of natural language expression in sentence similarity task. Traditional similarity measures, which represent sentences as a bag of words, simply judge similarity between sentences according to their common word occurrences. However, due to the complexity of many text mining applications where the similarity judgment at semantic level is required, the performance of naïve measures are likely to degrade because of their disregard of syntactic construction. Our proposed method aims to address the issue by computing sentence similarity at verb-argument structure level. By annotating sentences with semantic roles, we can better perform similarity calculation between semantically related components. The evaluation results confirm that the inclusion of sentence semantics significantly improves the effectiveness sentence similarity tasks. Given the encouraging result, we plan to evaluate the effectiveness of the proposed measure in several application-specific contexts, such as text summarization and question answering.

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., Han, H., Nasraoui, O., Johnson, R.: Semantically enhanced user modeling. In: Proceedings of the 2007 ACM Symposium on Applied Computing, pp. 1335–1339. ACM Press, New York (2007)
2. Achananuparp, P., Hu, X., Xiajiong, X.: The evaluation of sentence similarity measures. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 305–316. Springer, Heidelberg (2008)
3. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proceedings of the COLING-ACL, Montreal, Canada (1998)

4. Barzilay, R., Elhadad, N.: Sentence Alignment for Monolingual Comparable Corpora. In: Proceedings of EMNLP, Sapporo, Japan, pp. 25–33 (2003)
5. Bilotti, M.W., Ogilvie, P., Callan, J., Nyberg, E.: Structured retrieval for question answering. In: Proceedings SIGIR 2007, pp. 351–358. ACM, New York (2007)
6. Collobert, R., Weston, J.: Fast Semantic Extraction Using a Novel Neural Network Architecture. In: Proceedings of ACL 2007, Prague, Czech Republic, June 23–30 (2007)
7. Cordeiro, J., Dias, G., Brazdil, P.: A Metric for Paraphrase Detection. In: Proceedings ICCGI 2007, p. 7. IEEE Computer Society, Washington (2007)
8. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: Proceedings of the PASCAL Workshop (2005)
9. 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)
10. Healy, A., Miller, G.: The verb as the main determinant of sentence meaning. Pschonomic Science 20, 372 (1970)
11. Hickl, A., Bensley, J.: A Discourse Commitment-Based Framework for Reconizing Textual Entailment. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, pp. 171–176 (2007)
12. Metzler, D., Bernstein, Y., Croft, W., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: Proceedings of CIKM 2005, pp. 517–524 (2005)
13. 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)
14. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In: Proceedings of AAAI 2006, Boston (July 2006)
15. Murdock, V.: Aspects of sentence retrieval. Ph.D. Thesis, University of Massachusetts (2006)
16. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles. Comput. Linguist. 31(1), 71–106 (2005)
17. Park, Y., Byrd, R.J., Boguraev, B.K.: Automatic glossary extraction: beyond terminology identification. In: Proceedings of the 19th international Conference on Computational Linguistics, Taipei, Taiwan, August 24 - September 01, pp. 1–7 (2002)
18. Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., Jurafsky, D.: Shallow Semantic Parsing using Support Vector Machines. In: Proceedings of HLT/NAACL 2004, Boston, MA, May 2-7 (2004)
19. Shehata, S., Karray, F., Kamel, M.: A concept-based model for enhancing text categorization. In: Proceedings of KDD 2007, pp. 629–637. ACM, New York (2007)
20. Tatu, M., Moldovan, D.: COGEX at RTE3. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, pp. 22–27 (2007)