Utilizing Semantic, Syntactic, and Question Category Information for Automated Digital Reference Services

Palakorn Achananuparp, Xiaohua Hu, Xiaohua Zhou, and Xiaodan Zhang

College of Information Science and Technology
Drexel University, Philadelphia, PA 19104
pkorn@drexel.edu, thu@cis.drexel.edu, xiaohua.zhou@drexel.edu, xzhang@cis.drexel.edu

Abstract. Digital reference services normally rely on human experts to provide quality answers to the user requests via online communication tools. As the services gain more popularity, more experts are needed to keep up with a growing demand. Alternatively, automated question answering module can help shorten the question-answering cycle. When the system receives a new user submitted question, the similarity of the user's request and the existing questions in the archive can be compared. If the appropriate match is found, the system then uses the associated answer to response to such request. Since a question is relatively short and two questions might contain very few words in common, the challenge is how to effectively identify the similarity of questions. In this paper, we focus on the problem of identifying questions that convey the similar information need. That is, our goal is to find paraphrases of the original questions. To achieve this, we propose a hybrid approach that combines semantic, syntactic, and question category to judge question similarity. Semantic and syntactic information is measured by taking into account word similarity, word order, and part of speech information. Information about the types of question is derived from a Support Vector Machine classifier. The experimental results demonstrate that our combined measures are highly effective in distinguishing original questions and their paraphrases, thus improving the potency of question matching task.

Keywords: Question similarity, sentence similarity, question categories, answer reuse, question answering, factoid questions, semantic and syntactic techniques.

1 Introduction

Digital reference services have gradually becoming a major part of the digital library services due to the popularity of well-known online services, such as Internet Public Library (IPL) and Ask Dr. Math. In a typical digital reference service, librarians are responsible for answering the users' requests via online communication tools, e.g. email, web form, etc. We believe the process of answering questions in digital reference services can be significantly expedited by automated question answering module. If the system can determine whether a submitted question has been asked before, it can match the users' request with the existing question & answer pairs in the archive. This approach provides a tremendous value to the service as it offers a real-time response to

G. Buchanan, M. Masoodian, S.J. Cunningham (Eds.): ICADL 2008, LNCS 5362, pp. 203–214, 2008.

the redundant questions and increase the librarians' availability to assist the users with a truly unique need. Since a question is normally represented in a short sentence text, the challenge is how to effectively identify the similarity of questions, thus improving the potency of question matching task. Due to the variability of natural language expression, the same information can be formulated in numerous ways. Therefore, most document similarity approaches are likely to assign a low similarity score to those questions.

In this paper, we propose a method to measure the similarity between questions by utilizing semantic, syntactic, and question category information. Semantic information was derived from a lexical resource while syntactic information was derived from word order and part of speech information. Categorical information of questions was provided by a trained question classifier. Additionally, we are interested in the investigating an agreement between different evaluation metrics. Specifically, we investigate whether a similarity measure that correlates higher with human judgment also leads to a better performance on precision/recall based metric. Next, we examine how different word similarity measures and their similarity threshold affect the overall performance of sentence-level similarity measure.

The paper is organized as follows. First, we describe our approach to determine question similarity in section 2. In section 3, we describe the experimental set up and discuss about the results in section 4. Then, we review related work in section 5. Finally, we conclude the paper in section 6.

2 The Hybrid Approach

Our proposed method is a hybrid one based on the combinations of three different components: *semantic similarity*, *syntactic similarity* and *question category similarity*. The combination of the first two components (*semantic* + *syntactic*) represent the *sentence similarity* component [11] while the addition of the third component, *question category*, transform the similarity measure into question similarity measure. To quantify the similarity between words in the sentence, semantic information was obtained from WordNet [5]. A part of speech tagger was used to acquire the structural information of the question phrases, i.e. word order and part of speech labels. While a deep NLP technique, e.g. complete parse trees, might provide greater syntactic information of the sentences, our reason to use shallow NLP technique, i.e. part of speech tagging, was to balance the trade offs between the effectiveness and efficiency of the similarity measure. Equation 1 below describes the question similarity function (S) between two questions q_1 and q_2 as follow:

$$S(q_1, q_2) = \alpha \cdot (\gamma \cdot S_s(q_1, q_2) + \delta \cdot S_t(q_1, q_2)) + \beta \cdot S_c(q_1, q_2) \quad (1)$$

Four component coefficients were used to fine tune three similarity components. First, we optimized two sub-components within the sentence similarity component: *semantic similarity* (S_s) and *syntactic similarity* (S_t), through γ and δ, respectively Then, we controlled the influence of sentence similarity and question category similarity (S_c) components via α and β, respectively. All component coefficients have a real-number value ranging from 0 to 1.

To produce the actual question similarity score, each component will be replaced by the appropriate sentence similarity measures, which are described in section 2.2, and question category similarity measure, described in section 2.3. For example, either *sentence vector similarity* or *part-of-speech semantic similarity* measures can be plugged into the semantic similarity component. This results in a number of similarity measure combinations, in which we described them in section 2.4. Finally, most sentence similarity measures in section 2.4 rely on the comparison of individual words between two sentences. Such comparison requires word similarity measures which is described in the next section.

2.1 Word Similarity Measures

We adapted two existing measures to compute word similarity scores: Lin's universal similarity [12] and gloss overlap measures [1]. The two measures were chosen because of their superior performance to the conventional path-based similarity measures and their distinct approach to compute word similarity. Mainly, Lin's measure combines local similarity judgment with global term information from information content value while gloss overlap measure only computes word similarity on a local basis. The similarity value produced by both measures has a real-number value ranging from 0 (not similar) to 1 (identical).

2.1.1 Universal Similarity Measure

In this measure, the similarity between two words, w_1 and w_2 is determined by their information content and the path distance in WordNet hierarchies. Here, we used Resnik's formulation [19] of information content which defines the information content of concept c as the negative log likelihood function $-log(p(c))$, where $p(c)$ is the probability of encountering such concept c.

2.1.2 Gloss Overlap Measure

The Gloss overlap approach for measuring word similarity was first introduced by [9]. Our variation of gloss overlap similarity between two words is defined as the overlap between their glosses (dictionary definition) and their direct hypernym and hyponym in WordNet hierarchies [1]. The overall similarity measure is formulated as follow:

We empirically tested the correlation with human judgment for both measures on the selected noun pairs from the standard Rubenstein and Goodenough (R&G) data set used in [11] and found that both correlated highly with human judgment. sim_{ic} performed slightly better than sim_{gloss} (r_{lin}=0.924 and r_{gloss}=0.901).

2.2 Sentence Similarity Measures

We adopted the similarity measures used in [11] and [14] due to their efficiency in representing sentence-level text. All similarity measures used in this work rely on a pair-wise comparison between words in the two sentences. To select the best score for each word pairs, we performed a simple word sense disambiguation by choosing the maximum similarity score. The similarity score generated by all three measures has a real-number value ranging from 0 (not similar) to 1 (identical).

2.2.1 Sentence Vector Similarity

The motivation behind semantic measure was to distinguish sentences beyond their surface-text form by utilizing semantic similarity between words. From the vectorial model perspective, this means regular term weights, e.g. word frequency or TF-IDF, are replaced by semantic similarity scores. The process to compute the sentence vector similarity is described as follows. First, each sentence is converted into a sentence vector. Then the similarity between two sentences is derived from the cosine coefficient between the two sentence vectors. Each entry in the sentence vector is derived from computing word similarity score between word feature w_i and each word in the sentence. After that, the maximum score from the matching word that exceeds certain similarity threshold will be chosen. The idea behind sentence vector representation is very simple yet effective solution for a pair-wise comparison of sentences. The following example demonstrates the process to construct a sentence vector. Suppose sentence s_1 and s_2 are the two sentences to be compared; $s_1 = \{w_1,w_2,w_3\}$ and $s_2 = \{w_1,w_3,w_4\}$, the sentence vector sv_1 and sv_2 are shown below:

	w_1	w_2	w_3	w_4
sv_1	1	1	1	$sim_m(w_4,s_1)$
sv_2	1	$sim_m(w_2,s_2)$	1	$sim_m(w_4,s_2)$

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

2.2.2 Word Order Similarity

The ability to deal with different word compositions or morpho-syntactic variations in sentences is crucial for determining sentence similarity. Basic information, such as word order, can provide useful information to distinguish the meaning of two sentences. This is particular important in our case where single word token was used as a basic lexical unit. Without syntactic information, it is impossible to discriminate the sentence containing words "sale manager" and "office worker" from another sentence containing "office worker" and "sale manager" since both of them essentially share the same bag-of-word representation. Word order similarity is defined as the normalized difference of word order between the two sentences. It has been proved in [11] to be an efficient method to compute the similarity of word order. The formulation for word order similarity is defined as follow:

$$sim_{wo}(s_1, s_2) = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \tag{2}$$

r_1 and r_2 is a word order vector of sentence s_1 and s_2, respectively. The steps to build a word order vector are similar to sentence vector's process. That is, a feature set of word order vector is taken from the individual words of the two sentences. Each entry in the word order vector is derived by comparing word feature w_i with each word in the sentence. If the two are identical, then we fill the entry of w_i with an index number (word position) of the corresponding word. Otherwise, we calculate word similarity score between w_i and the remaining words in the sentence and fill w_i entry with an index number of a matching word that gives a maximum similarity score.

2.2.3 Part-of-Speech Semantic Similarity

Unlike sentence vector similarity where word similarity is exhaustively computed over the possible word pairs, another way to measure semantic information between sentences is to compute word similarity between words of the same part of speech [5]. That is, a simple syntactic analysis (part of speech) of words is included in this measure. As such, this approach intuitively fits to handle a common form of paraphrase – lexical substitution. The Part-of-speech semantic similarity between two sentences is defined as a combined maximum similarity score from all word pairs in each part of speech class. In this formulation, we consider noun, verb, adjective, and adverb as the major part-of-speech classes. The overall sentence similarity is defined as follow:

$$sim_{ps}(s_1, s_2) = \frac{\sum_{w \in \{s_1\}} \max Sim(w, s_2) + \sum_{w \in \{s_2\}} \max Sim(w, s_1)}{|s_1| + |s_2|} \quad (3)$$

where *max Sim* (w,s_2) is derived from selecting the maximum similarity score of w and the matching word in s_2, while *max* $Sim_m(w,s_1)$ is derived from selecting the maximum similarity score of w and the matching word in s_1. Then $sim_{ps}(s_1,s_2)$ is computed by combining the sum of $sim_m(w_i,s_2)$ and $sim_m(w_j,s_1)$ from all four part of speech classes normalized by the length of s_1 and s_2.

2.3 Question Category Similarity Measure

Questions differ from regular sentences in that they contain interrogative words. These words can be used to determine the *aboutness* of the questions. Given two questions with almost exact same words, the interrogative part acts as a surrogate of the categorical information that helps distinguish them. For instance, we can say that "*where was JFK assassinated?*" and "*when was JFK assassinated?*" are two different questions judging by different *wh*-pronouns: *where* (location) and *when* (time). Thus, we built our question category similarity measure around the idea that similar questions share the same interrogative words or question categories.

We define question category similarity as a cosine similarity between the question category vectors. Thus, the major step in our approach is the construction of question category vector. For the task of classifying questions into different types, we chose Support Vector Machine (SVM) as the underlying classifier as it has been shown in many literatures to be the best performer for question classification task [23]. In this work, we used SVMLight [8] as the implementation of the classifier.

We developed SVM classifier using linear kernel to predict the question categories. The features for the classifier include unigram, multiword collocations, and the hypernyms of the head nouns (the head of the noun phrases). Specifically, we restricted the head nouns to those following the interrogative words. For instance, a head noun of the question "*What tourist attractions are there in Reims*?" is *tourist*. A list of multiword collocations, including interrogative words, was compiled from the training example. For example, "how many", "how much", "what is a", "what is the" were automatically identified and extracted by the aforementioned tool. In the testing stage, we simply used exact string match to identify multiword collocations. The hypernyms of the head noun serve as semantic features which increase the chance of semantically-similar concepts sharing common features. The method of extracting head nouns and their hypernyms are the same as the one in [15]. The classifier was built on the

UIUC dataset which is a superset of the TREC QA track [10] dataset. The UIUC dataset contains 5,500 training questions and 500 TREC-10 questions for testing. Their question class taxonomy contains two levels. The coarse level has six categories whereas the fine level has fifty categories. The classification precisions for coarse-grained and fine-grained taxonomies are 81.8% and 89.2%, respectively. In this study, we classified questions based on the fine-grained categories due to their superior performance. Moreover, we took a multi-label classification approach to categorize questions. As such, a question was classified into multiple categories.

We experimented with two approaches to build the question category vector. First, we constructed the vector based on *the ranked category*. In this approach, we use the ranks of predicted category as a feature set, starting from rank 1, 2, 3, etc. The other approach utilizes *the classification probability*. Here, we used the question categories as a feature set in the vector, starting from category 1, 2, 3, etc. Each entry in the question category vector contains the probability that a question will be classified into a given category.

2.4 The Combinations of Measures

By replacing the similarity components with appropriate similarity measures described in section 2.2 and 2.3, we derived a set of similarity measure combinations. First, three possible combinations of sentence similarity measures are shown in table 1. To simplify the task of tuning the semantic and syntactic components, we employed *ps* in a syntactic similarity component in one combination and a semantic similarity component in another since it considers both semantic and syntactic information in comparing sentence similarity. As a result, each combination represents a different flavor of similarity notions. For example, *sv+wo* judges similar questions on their word semantics and word order, *sv+ps* uses word semantics and part of speech information, and finally, and *ps+wo* employs word semantics from specific part of speech classes and word order. Next, we combined the each sentence similarity combination in table 1 with two variations of question category similarity measure described in the previous section. Each variation is based on the different approaches to construct question category vector. For example, *sv+wo+rank* represents sentence vector similarity + word order similarity + ranked category vector similarity. Ultimately, a total number of six combinations of question similarity measures were derived.

Table 1. The combinations of sentence similarity measures that represent sentence similarity component

Combination of Measures	Semantic Component	Syntactic Component
sv+wo	Sentence vector	Word order
sv+ps	Sentence vector	Part of speech
ps+wo	Part of speech	Word order

3 Experimental Evaluation

The experiment was structured into two parts. First, we investigated the effectiveness of the sentence similarity measures in terms of correlation with human judgment. In

the second part, we compared the performance of sentence similarity-only measures and the combined sentence and question category similarity measures on a set of paraphrased questions.

Baseline: Three measures were selected as the baseline comparisons: *Jaccard* coefficient, which is traditionally used as a distance measure when comparing the two strings, a standard *TF-IDF* term vector similarity, and a text semantic similarity measure based on the combined semantic similarity of words in the same part of speech and their IDF scores (henceforth *ps-IDF*) [5].

3.1 Sentence Similarity Experiment

The goal of sentence similarity experiment was to compare the performance of sentence similarity measures on the correlation with human judgment. To achieve that, we computed sentence similarity scores between sentence pairs using the three combinations of sentence similarity measures described in the previous section. Moreover, we investigated the three factors and their impact on the performance of each sentence similarity measure. These are *the underlying word similarity measure* (*lin* vs. *gloss*), *its threshold level* (from 0 to 1), and *the relative contribution of the semantic and syntactic components* (γ and δ, respectively) indicated by the combinations of their coefficient values from 0 to 1.

3.2 Question Similarity Experiment

In this part of the experiment, we compared the performance between the sentence similarity component (semantic + syntactic) and the combined sentence similarity and question category similarity components (semantic + syntactic + question category). Additionally, we explored the effect of two variations of question category similarity measure (*rank* and *conf*) described previously on performance of overall question similarity measure. Furthermore, we investigated how the performance of question similarity measure is affected by the relative contribution of sentence similarity and question similarity components (α and β, respectively). For cross-validation purpose, we also compared γ and δ obtained from sentence pairs data set used in 3.1 with the optimal values derived from question pair data set in this experiment.

3.3 Data Sets

We conducted a sentence similarity evaluation on thirty sentence-pair data set published in [11]. Each sentence pair was derived from a definitive sentence of a subset of noun pairs from Rubenstein and Goodenough (R&G) data set. To evaluate the performance of the question similarity measures, we selected a set of 193 question pairs from TREC-9 question variants key. The variants key consists of fifty four original questions and their variants. The original questions are a subset of test questions used in TREC-9 QA experiment and were taken from the actual users' submissions. The question variants are the paraphrased questions that were constructed by human assessors to be semantically identical but syntactically different from the original questions. The total number of question pairs used in the experiment is 386 -- 193 pairs for testing paraphrased questions and another 193 pairs for testing

non-paraphrased questions. Although the data set is semi-artificial, it contains sufficient linguistic complexity to reflect the variability of nature language expressions. That is, there are various types of paraphrasing strategies [22] exhibited in the question variants, e.g., lexical substitution (*what kind of animal was Winnie the Pooh?* vs. *what species was Winnie the Pooh?*, morpho-syntactic variations (*what kind of animal was Winnie the Pooh?* vs. *Winnie the Pooh is what kind of animal?*, *who owns CNN?* vs. *CNN is owned by whom?*), interrogative reformulation (*how did Bob Marley die?* vs. *what killed Bob Marley?*), semantic inference (*What tourist attractions are there in Reims?* vs. *What do most tourists visit in Reims?*), with more than 50% of the paraphrases categorized into multiple categories.

Table 2. The composition of paraphrase categories in TREC-9 question variants

Paraphrase Category	Lexical Substitution	Morpho-Syntactic Variation	Interrogative Reformulation	Semantic Inference
# of questions	63	97	112	31

3.4 Preprocessing

The preprocessing procedure is described as follows. First, individual words in sentence/question text were extracted, part-of-speech tagged, but not stemmed to preserve their meaning. A set of functional words -- words that do not contain semantic content such as articles, pronouns, prepositions, conjunctions, auxiliary verbs, modal verbs, and punctuations, was removed. Cardinal numbers were not discarded. Then, word similarity scores for all possible word pairs were computed and the results were cached for later use.

3.5 Evaluation Criteria

Pearson's correlation coefficient was used to measure the correlation between human-judgment scores and algorithmic scores in the sentence similarity experiment. The correlation coefficients were tested at the significant level of $p<0.01$. To evaluate the performance of our question similarity measures, we adapted the notion of rejection/recall used in [12] as it is a better representation of the task's performance. *Recall* is defined as the proportion of question pairs correctly judged to be similar compared to the total number of similar question pairs. *Rejection* is defined as the proportion of question pairs correctly judged to be dissimilar compared to the total number of dissimilar question pairs. Finally, to evaluate the combined performance of recall-rejection, we defined the harmonic mean of unigram recall and rejection (F_1) similar to the one used in standard information retrieval evaluation.

4 Results and Discussion

4.1 Sentence Similarity

According to table 3, sentence similarity measures significantly outperformed the baseline measure (r = 0.85-0.88) on the measures using *lin* as the word similarity

measure while sentence similarity measures that employed *gloss* as the word similarity performed poorer (r = 0.72-0.81). Since R&G experiment is based on synonymy evaluation, *lin*'s notion of similarity fits the human judgment better. The baseline Jaccard coefficient and TF-IDF measures correlated reasonably well with the sentence data set ($r_{Jaccard}$ = 0.81 and $r_{TF\text{-}IDF}$ = 0.87) while text semantic similarity measure correlated the lowest ($r_{ps\text{-}IDF}$ = 0.75). This comes as no surprise since most similar sentences in the sentence pair data set contain the reasonable numbers of word overlaps, while the dissimilar sentences contain fewer common words, the naïve methods that operate at a surface text level were expected to generate a good result.

Table 3. The Pearson's correlation coefficient of each similarity measure with subject to human judgment on R&G-based sentence pair data set

Word Similarity Measure	*Lin*			*Gloss*		
Similarity Measure	*sv+wo*	*sv+ps*	*ps+ws*	*sv+wo*	*sv+ps*	*ps+wo*
Correlation Coefficient	0.85	0.87	0.88	0.72	0.79	0.81

Next, the effect of semantic/syntactic contribution differs on each similarity measure combination. *sv+wo* and *ps+wo* correlated the highest when the semantic component was weighted higher than the syntactic component (γ=0.8 and δ = 0.2). In the case of *sv+ps*, the optimal result was met when the syntactic component was weighted higher (γ= 0.3 and δ = 0.7). The result in *sv+wo* and *ps+wo* combinations are similar to the one reported in [11]. That is, in general, the semantic component plays a greater role than the syntactic component in determining the similarity between sentences. Furthermore, *sv+wo* correlated the lowest with human judgment compared to the other two combinations. Both *lin* and *gloss* correlated relatively well with human judgment at high word similarity threshold levels (greater than 0.5). Again, *lin* consistently outperformed *gloss* in all combinations, having *sv+ps* and *ps+wo* as the best overall measures.

4.2 Question Similarity

First, we tried to find the word similarity measure and threshold level that produces the best performance for the combined sentence similarity measures (*sv+wo*, *sv+ps*, and *ps+wo*). Any question pairs with similarity score exceed a threshold of 0.7 were considered to be a paraphrased pair. The result indicated that all sentence similarity combinations significantly outperformed all three baselines. Next, similarity measures using *lin* as the word similarity measure did not significantly outperform those using *gloss* in both recall and rejection at $p<0.05$. Within the same word similarity measure, lower word semantic similarity threshold (0) performed better than higher word similarity threshold (0.6), $p<0.05$. The result offers an interesting contrast to that of the sentence similarity experiment. While the higher word similarity thresholds correlated higher with human judgment than the lower word similarity thresholds, it was the latter that performed significantly better on recall and rejection metrics.

Table 4. The performance of the best overall measures and baselines on identifying TREC-9 question variant

Combination of Measures	*sv+ps*	*sv+ps +rank*	*ps+wo +conf*	*Jaccard*	*TF-IDF*	*ps-IDF*
Recall	0.79	0.88	0.98	0.24	0.50	0.30
Rejection	1.00	1.00	0.93	1.00	1.00	0.99
F_1	0.88	0.94	0.95	0.39	0.67	0.46

Next, we compared the effectiveness of the sentence similarity and question similarity measures at the optimal word similarity setting obtained from the above experiment. That is, we used *lin* as word similarity measure computed at the word similarity threshold of 0. The result is shown in table 4. Among the three sentence similarity combinations, sv+*ps* performed the best (F_1^{sv+ps}= 0.88). The optimal semantic/syntactic coefficient values were similar to those in 4.1. This result reaffirmed that the optimal coefficient settings were applicable across data sets. Then, we reapplied the coefficient settings to the corresponding components in the question similarity measure. The inclusion of question category similarity measures (*rank* and *conf*) has significantly improved the overall performance to identify paraphrased questions. Among six question similarity combinations, measures that employ *conf* as the question category vector have significantly produced greater recalls than *rank* measures, however, with greater expense on rejection. *sv+ps+rank* is the best overall measure among *rank* combinations ($F_1^{sv+ps+rank}$= 0.94) while *ps+wo+conf* is the best overall among *conf* combinations ($F_1^{ps+wo+conf}$= 0.95) at the similarity threshold of 0.7.

Overall, the analysis of the experimental result on various parameter settings has shown that the best similarity measure consistently performed well across different evaluation metrics Different types of word similarity measures did not produce a significantly different result in sentence and question similarity evaluation. Due to the fact that both measures utilize the same word coverage in WordNet, they ultimately produced a similar result regardless of their approaches. Specifically, WordNet contains 85% of the vocabulary space of the test data set, making it reasonably effective. Different word similarity thresholds yielded significantly different results on Pearson's correlation and F_1 metrics. Measures with higher word similarity threshold performed better in correlation metric while measures with lower word similarity threshold performed better in F_1 metric. This shows that rejection/recall tended to over-penalize the similarity scores at higher word similarity thresholds. Different types of question category vector generated significantly different results. Overall, *conf* combinations produced the highest recall but suffered from a minimal rejection rate. The significant loss in rejection eventually outweighed the gain in recall. Finally, the optimal results were achieved by approximately equal contribution of the sentence similarity and question category similarity components.

5 Related Work

Several approaches to measure sentence-level similarity have been proposed recently [2][6][10][12][15][[17][18]. Vector space model and lexical resources have been

applied to measure question similarity [4]. Although vector space model approaches work very well in document retrieval task, they are not suitable for short text matching because of small word overlaps, data sparseness, and lexical chasm problem [2]. The work by [13] is perhaps the most relevant to ours since they used semantic metrics and question type metric to judge the question similarity. Our approach is different from theirs in many aspects. First, we cover a broader range of similarity metrics (semantic, syntactic, and question category). Second, there are a number of differences between their question type similarity and ours. They treated question classification as a binary classification task while we consider the task as a multi-label classification. We believe this approach follows a more intuitive notion. Next, we automatically extracted multiword collocations and the hypernyms of the head nouns instead of manually constructing the feature set. Lastly, we used fine-grained question categories due to their superior accuracy in question classification task.

6 Conclusions

We have demonstrated that semantic, syntactic, and question category information is very effective in identifying paraphrased questions. Semantic and syntactic measures were helpful in handling synonyms, related words, and different word compositions. The addition of question category information has significantly improved the performance of the similarity measure by providing discriminative power from the interrogative words in the question sentences. We recognized certain shortcomings in the use of TREC-9 data set since it is partially artificial. Hence, it might be less noisy and contain fewer cases of lexical-syntactic variations. Moreover, most questions in TREC-9 data set are factoid questions which only cover a subset of those being queried a real-world reference service. The future works include improving the method to incorporate more contextual information into the similarity measure. Currently, we represented sentence and question phrases at the individual words level. We believe the performance can be improved by considering a more meaningful lexical unit such as multiword phrases. In addition, we plan to extend the word similarity measures to handle words that do not exist in WordNet taxonomy via other knowledge resources, e.g. web search, Wikipedia, etc. Furthermore, we plan to test our approach on other question-answering dataset.

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 SAC 2007, pp. 1335–1339. ACM Press, New York (2007)
2. 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 (2008)

3. Berger, A., Caruana, D., Cohn, D., Freitag, D., Mittal, V.: Bridging the lexical chasm: Statistical approaches to answer-finding. In: Proceedings of SIGIR, pp. 222–229 (2000)
4. Burke, R.D., Hammond, K.J., Kulyukin, V.A., Lytinen, S.L., Tomuro, N., Schoenberg, S.: Question answering from frequently asked question files: Experiences with the FAQ finder system. Technical report (1997)
5. Corley, C., Mihalcea, R.: Measuring the semantic similarity of texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, Michigan, pp. 13–18 (June 2005)
6. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
7. Jeon, J., Croft, W.B., Lee, J.H.: Finding similar questions in large question and answer archives. In: Proceedings of ACM CIKM, pp. 84–90 (2005)
8. Joachims, T.: Text categorization with support vector machines: Leaning with many relevant features. In: Proceedings of European Conference on Machine Learning, pp. 137–142 (1998)
9. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on Systems documentation, pp. 24–26 (1986)
10. Li, X., Roth, D.: Learning Question Classifiers. In: COLING 2002 (August 2002)
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. Lytinen, S., Tomuro, N.: The Use of Question Types to Match Questions in FAQFinder. In: 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, pp. 46–53. AAAI Press, Menlo Park (2002)
14. Malik, R., Subramaniam, V., Kaushik, S.: Automatically Selecting Answer Templates to Respond to Customer Emails. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 1659–1664 (2007)
15. Metzler, D., Bernstein, Y., Croft, W., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: Proceedings of CIKM, pp. 517–524 (2005)
16. Metzler, D., Croft, W.B.: Analysis of Statistical Question Classification for Fact-based Questions. Information Retrieval 8(3), 481–504 (2005)
17. 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)
18. Murdock, V.: Aspects of sentence retrieval. Ph.D. Thesis, University of Massachusetts (2006)
19. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: International Joint Conference for Artificial Intelligence (IJCAI 1995), pp. 448–453 (1995)
20. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international Conference on World Wide Web, Edinburgh, Scotland, pp. 377–386 (2006)
21. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics 19(1), 143–177 (1993)
22. Tomuro, N.: Interrogative Reformulation Patterns and Acquisition of Question Paraphrases. In: Proceedings of the Second international Workshop on Paraphrasing, pp. 33–40 (2003)
23. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of SIGIR 2003, pp. 26–32. ACM Press, New York (2003)