A Utility-Driven Approach to Question Ranking in Social QA

Razvan Bunescu
School of EECS
Ohio University
bunescu@ohio.edu

Yunfeng Huang
School of EECS
Ohio University
yh324906@ohio.edu

Abstract

We generalize the task of finding question paraphrases in a question repository to a novel formulation in which known questions are ranked based on their utility to a new, reference question. We manually annotate a dataset of 60 groups of questions with a partial order relation reflecting the relative utility of questions inside each group, and use it to evaluate meaning and structure aware utility functions. Experimental evaluation demonstrates the importance of using structural information in estimating the relative usefulness of questions, holding the promise of increased usability for social QA sites.

1 Introduction

Open domain Question Answering (QA) is one of the most complex and challenging tasks in natural language processing. While building on ideas from Information Retrieval (IR), question answering is generally seen as a more difficult task due to constraints on both the input representation (natural language questions vs. keyword-based queries) and the form of the output (focused answers vs. entire documents). Recently, community-driven QA sites such as Yahoo! Answers and WikiAnswers have established a new approach to question answering in which the burden of dealing with the inherent complexity of open domain QA is shifted from the computer system to volunteer contributors. The computer is no longer required to perform a deep linguistic analysis of questions and generate corresponding answers, and instead acts as a mediator between users submitting questions and volunteers providing the answers. In most implementations of community-driven QA, the mediator system has a well defined strategy for enticing volunteers to post high quality answers on the website. In general, the overall objective is to minimize the response time and maximize the accuracy of the answers, measures that are highly correlated with user satisfaction. For any submitted question, one useful strategy is to search the QA repository for similar questions that have already been answered, and provide the corresponding ranked list of answers, if such a question is found. The success of this approach depends on the definition and implementation of the question-to-question similarity function. In the simplest solution, the system searches for previously answered questions based on exact string matching with the reference question. Alternatively, sites such as WikiAnswers allow the users to mark questions they think are rephrasings ("alternate wordings", or paraphrases) of existing questions. These question clusters are then taken into account when performing exact string matching, therefore increasing the likelihood of finding previously answered questions that are semantically equivalent to the reference question. Like the original question answering task, the solution to question rephrasing is also based on volunteer contributions. In order to lessen the amount of work required from the contributors, an alternative solution is to build a system that automatically finds rephrasings of questions, especially since question rephrasing seems to be computationally less demanding than question answering. The question rephrasing subtask has spawned a diverse set of approaches. (Herm-

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 125–133,
Beijing, August 2010

jakob et al., 2002) derive a set of phrasal patterns for question reformulation by generalizing surface patterns acquired automatically from a large corpus of web documents. The focus of the work in (Tomuro, 2003) is on deriving reformulation patterns for the interrogative part of a question. In (Jeon et al., 2005), word translation probabilities are trained on pairs of semantically similar questions that are automatically extracted from an FAQ archive, and then used in a language model that retrieves question reformulations. (Jijkoun and de Rijke, 2005) describe an FAQ question retrieval system in which weighted combinations of similarity functions corresponding to questions, existing answers, FAQ titles and pages are computed using a vector space model. (Zhao et al., 2007) exploit the Encarta logs to automatically extract clusters containing question paraphrases and further train a perceptron to recognize question paraphrases inside each cluster based on a combination of lexical, syntactic and semantic similarity features. More recently, (Bernhard and Gurevych, 2008) evaluated various string similarity measures and vector space based similarity measures on the task of retrieving question paraphrases from the WikiAnswers repository.

According to previous work in this domain, a question is considered a rephrasing of a reference question Q_0 if it uses an alternate wording to express an identical information need. For example, Q_0 and Q_1 below may be considered rephrasings of each other, and consequently they are expected to have the same answer.

Q_0 What should I feed my turtle?

Q_1 What do I feed my pet turtle?

Community-driven QA sites are bound to face situations in which paraphrasings of a new question cannot be found in the QA repository. We believe that computing a ranked list of existing questions that partially address the original information need could be useful to the user, at least until other users volunteer to give an exact answer to the original, unanswered reference question. For example, in the absence of any additional information about the reference question Q_0, the expected answers to questions Q_2 and Q_3 above may be seen as partially overlapping in information content with the expected answer for the reference question. An answer to question Q_4, on the other hand, is less likely to benefit the user, even though it has a significant lexical overlap with the reference question.

Q_2 What kind of fish should I feed my turtle?

Q_3 What do you feed a turtle that is the size of a quarter?

Q_4 What kind of food should I feed a turtle dove?

In this paper, we propose a generalization of the question paraphrasing problem to a question ranking problem, in which questions are ranked in a partial order based on the relative information overlap between their expected answers and the expected answer of the reference question. The expectation in this approach is that the user who submits a reference question will find the answers of the highly ranked question to be more useful than the answers associated with the lower ranked questions. For the reference question Q_0 above, the system is expected to produce a partial order in which Q_1 is ranked higher than Q_2, Q_3 and Q_4, whereas Q_2 and Q_3 are ranked higher than Q_4. In Section 2 we give further details on the question ranking task and describe a dataset of questions that have been manually annotated with partial order information. Section 3 presents a set of initial approaches to question ranking, followed by their experimental evaluation in Section 4. The paper ends with a discussion of future work, and conclusion.

2 A Partially Ordered Dataset for Question Ranking

In order to enable the evaluation of question ranking approaches, we created a dataset of 60 groups of questions. Each group consists of a reference question (e.g. Q_0 above) that is associated with a partially ordered set of questions (e.g. Q_1 to Q_4 above). The 60 reference questions have been selected to represent a diverse set of question categories from Yahoo! Answers. For each reference questions, its corresponding partially ordered set is created from questions in Yahoo! Answers

REFERENCE QUESTION (Q_r)
Q_5 What's a good summer camp to go to in FL?
PARAPHRASING QUESTIONS ($\mathcal{P}$)
Q_6 What camps are good for a vacation during the summer in FL?
Q_7 What summer camps in FL do you recommend?
USEFUL QUESTIONS ($\mathcal{U}$)
Q_8 Does anyone know a good art summer camp to go to in FL?
Q_9 Are there any good artsy camps for girls in FL?
Q_{10} What are some summer camps for like singing in Florida?
Q_{11} What is a good cooking summer camp in FL?
Q_{12} Do you know of any summer camps in Tampa, FL?
Q_{13} What is a good summer camp in Sarasota FL for a 12 year old?
Q_{14} Can you please help me find a surfing summer camp for beginners in Treasure Coast, FL?
Q_{15} Are there any acting summer camps and/or workshops in the Orlando, FL area?
Q_{16} Does anyone know any volleyball camps in Miramar, FL?
Q_{17} Does anyone know about any cool science camps in Miami?
Q_{18} What's a good summer camp you've ever been to?
NEUTRAL QUESTIONS ($\mathcal{N}$)
Q_{19} What's a good summer camp in Canada?
Q_{20} What's the summer like in Florida?

Table 1: A question group.

and other online repositories that have a high cosine similarity with the reference question. Due to the significant lexical overlap between the questions, this is a rather difficult dataset, especially for ranking methods that rely exclusively on bag-of-words measures. Inside each group, the questions are manually annotated with a partial order relation, according to their utility with respect to the reference question. We shall use the notation $\langle Q_i \succ Q_j | Q_r \rangle$ to encode the fact that question Q_i is *more useful than* question Q_j with respect to the reference question Q_r. Similarly, $\langle Q_i = Q_j \rangle$ will be used to express the fact that questions Q_i and Q_j are reformulations of each other (the reformulation relation is independent of the reference question). The partial ordering among the questions Q_0 to Q_4 above can therefore be expressed concisely as follows: $\langle Q_0 = Q_1 \rangle$, $\langle Q_1 \succ Q_2 | Q_0 \rangle$, $\langle Q_1 \succ Q_3 | Q_0 \rangle$, $\langle Q_2 \succ Q_4 | Q_0 \rangle$, $\langle Q_3 \succ Q_4 | Q_0 \rangle$. Note that we do not explicitly annotate the relation $\langle Q_1 \succ Q_4 | Q_0 \rangle$, since it can be inferred based on the transitivity of the *more useful than* relation: $\langle Q_1 \succ Q_2 | Q_0 \rangle \wedge \langle Q_2 \succ Q_4 | Q_0 \rangle \Rightarrow \langle Q_1 \succ Q_4 | Q_0 \rangle$. Also note that no relation is specified between Q_2 and Q_3, and similarly no relation can be inferred between these two questions. This reflects our belief that, in the absence of any additional information regarding the user or the "turtle" referenced in Q_0, we cannot compare questions Q_2 and Q_3 in terms of their usefulness with respect to Q_0.

Table 1 shows another reference question Q_5 from our dataset, together with its annotated group of questions Q_6 to Q_{20}. In order to make the annotation process easier and reproducible, we divide it into two levels of annotation. During the first annotation stage (L_1), each question group is partitioned manually into 3 subgroups of questions:

- $\mathcal{P}$ is the set of *paraphrasing* questions.
- $\mathcal{U}$ is the set of *useful* questions.
- $\mathcal{N}$ is the set of *neutral* questions.

A question is deemed useful if its expected answer may overlap in information content with the expected answer of the reference question. The expected answer of a neutral question, on the other

hand, should be irrelevant with respect to the reference question. Let Q_r be the reference question, $Q_p \in \mathcal{P}$ a paraphrasing question, $Q_u \in \mathcal{U}$ a useful question, and $Q_n \in \mathcal{N}$ a neutral question. Then the following relations are assumed to hold among these questions:

1. $\langle Q_p \succ Q_u | Q_r \rangle$: a *paraphrasing* question is more useful than a *useful* question.

2. $\langle Q_u \succ Q_n | Q_r \rangle$: a *useful* question is more useful than a *neutral* question.

We also assume that, by transitivity, the following ternary relations also hold: $\langle Q_p \succ Q_n | Q_r \rangle$, i.e. a *paraphrasing* question is more useful than a *neutral* question. Furthermore, if $Q_{p_1}, Q_{p_2} \in \mathcal{P}$ are two paraphrasing questions, this implies $\langle Q_{p_1} = Q_{p_2} | Q_r \rangle$.

For the vast majority of questions, the first annotation stage is straightforward and non-controversial. In the second annotation stage (L_2), we perform a finer annotation of relations between questions in the middle group $\mathcal{U}$. Table 1 shows two such relations (using indentation): $\langle Q_8 \succ Q_9 | Q_5 \rangle$ and $\langle Q_8 \succ Q_{10} | Q_5 \rangle$. Question Q_8 would have been a rephrasing of the reference question, were it not for the noun "art" modifying the focus noun phrase "summer camp". Therefore, the information content of the answer to Q_8 is strictly subsumed in the information content associated with the answer to Q_5. Similarly, in Q_9 the focus noun phrase is further specialized through the prepositional phrase "for girls". Therefore, (an answer to) Q_9 is less *useful* to Q_5 than (an answer to) Q_8, i.e. $\langle Q_8 \succ Q_9 | Q_5 \rangle$. Furthermore, the focus "art summer camp" in Q_8 conceptually subsumes the focus "summer camps for singing" in Q_{10}, therefore $\langle Q_8 \succ Q_{10} | Q_5 \rangle$.

Table 2 below presents the following statistics on the annotated dataset: the number of reference questions (Q_r), the total number of paraphrasings ($\mathcal{P}$), the total number of useful questions ($\mathcal{U}$), the total number of neutral questions ($\mathcal{N}$), and the total number of *more useful than* ordered pairs encoded in the dataset, either explicitly or through transitivity, in the two annotation levels L_1 and L_2.

Q_r	$\mathcal{P}$	$\mathcal{U}$	$\mathcal{N}$	L_1	L_2
60	177	847	427	7,378	7,639

Table 2: Dataset statistics.

3 Question Ranking Methods

An ideal question ranking method would take an arbitrary triplet of questions Q_r, Q_i and Q_j as input, and output an ordering between Q_i and Q_j with respect to the reference question Q_r, i.e. one of $\langle Q_i \succ Q_j | Q_r \rangle$, $\langle Q_i = Q_j | Q_r \rangle$, or $\langle Q_j \succ Q_i | Q_r \rangle$. One approach is to design a *usefulness* function $u(Q_i, Q_r)$ that measures how useful question Q_i is for the reference question Q_r, and define the *more useful than* ($\succ$) relation as follows:

$$\langle Q_i \succ Q_j | Q_r \rangle \Leftrightarrow u(Q_i, Q_r) > u(Q_j, Q_r)$$

If we define $I(Q)$ to be the information need associated with question Q, then $u(Q_i, Q_r)$ could be defined as a measure of the relative overlap between $I(Q_i)$ and $I(Q_r)$. Unfortunately, the information need is a concept that, in general, is defined only intensionally and therefore it is difficult to measure. For lack of an operational definition of the information need, we will approximate $u(Q_i, Q_r)$ directly as a measure of the similarity between Q_i and Q_r. The similarity between two questions can be seen as a special case of text-to-text similarity, consequently one possibility is to use a general text-to-text similarity function such as *cosine similarity* in the vector space model (Baeza-Yates and Ribeiro-Neto, 1999):

$$cos(Q_i, Q_r) = \frac{Q_i^T Q_r}{\|Q_i\| \|Q_r\|}$$

Here, Q_i and Q_r denote the corresponding *tf*×*idf* vectors. As a measure of question-to-question similarity, cosine has two major drawbacks:

1. As an exclusively lexical measure, it is oblivious to the meanings of words in each question.

2. Questions are treated as bags-of-words, and thus important structural information is missed.

3.1 Meaning Aware Measures

The three questions below illustrate the first problem associated with cosine similarity. Q_{22} and Q_{23} have the same cosine similarity with Q_{21}, they are therefore indistinguishable in terms of their usefulness to the reference question Q_{21}, even though we expect Q_{22} to be more useful than Q_{23} (a place that sells hydrangea often sells other types of plants too, possibly including cacti).

Q_{21} Where can I buy a hydrangea?

Q_{22} Where can I buy a cactus?

Q_{23} Where can I buy an iPad?

To alleviate the lexical chasm, we can redefine $u(Q_i, Q_r)$ to be the similarity measure proposed by (Mihalcea et al., 2006) as follows:

$$mcs(Q_i, Q_r) = \frac{\sum_{w \in \{Q_i\}} (maxSim(w, Q_r) * idf(w))}{\sum_{w \in \{Q_i\}} idf(w)} + \frac{\sum_{w \in \{Q_r\}} (maxSim(w, Q_i) * idf(w))}{\sum_{w \in \{Q_r\}} idf(w)}$$

Since scaling factors are immaterial for ranking, we have ignored the normalization constant contained in the original measure. For each word $w \in Q_i$, $maxSim(w, Q_r)$ computes the maximum semantic similarity between w and any word $w_r \in Q_r$. The similarity scores are then weighted by the corresponding *idf*'s, and normalized. A similar score is computed for each word $w \in Q_r$. The score computed by $maxSim$ depends on the actual function used to compute the word-to-word semantic similarity. In this paper, we evaluated four of the knowledge-based measures explored in (Mihalcea et al., 2006): *wup* (Wu and Palmer, 1994), *res* (Resnik, 1995), *lin* (Lin, 1998), and *jcn* (Jiang and Conrath, 1997). Since all these measures are defined on pairs of WordNet concepts, their analogues on word pairs (w_i, w_r) are computed by selecting pairs of WordNet synsets (c_i, c_r) such that w_i belongs to concept c_i, w_r belongs to concept c_r, and (c_i, c_r) maximizes the similarity function. The measure introduced in (Wu and Palmer, 1994) finds the *least common subsumer (LCS)* of the two input concepts in the WordNet hierarchy, and computes the ratio between its depth and the sum of the depths of the two concepts:

$$wup(c_i, c_r) = \frac{2 * depth(lcs(c_i, c_r))}{depth(c_i) + depth(c_r)}$$

Resnik's measure is based on the Information Content (IC) of a concept c defined as the negative log probability $-\log P(c)$ of finding that concept in a large corpus:

$$res(c_i, c_r) = IC(lcs(c_i, c_r))$$

Lin's similarity measure can be seen as a normalized version of Resnik's information content:

$$lin(c_i, c_r) = \frac{2 * IC(lcs(c_i, c_r))}{IC(c_i) + IC(c_r)}$$

Jiang & Conrath's measure is closely related to *lin* and is computed as follows:

$$jcn(c_i, c_r) = [IC(c_i) + IC(c_r) - 2 * IC(lcs(c_i, c_r))]^{-1}$$

3.2 Structure Aware Measures

Cosine similarity, henceforth referred as *cos*, treats questions as bags-of-words. The meta-measure proposed in (Mihalcea et al., 2006), henceforth called *mcs*, treats questions as bags-of-concepts. Consequently, both *cos* and *mcs* may miss important structural information. If we consider the question Q_{24} below as reference, question Q_{26} will be deemed more useful than Q_{25} when using *cos* or *mcs* because of the higher relative lexical and conceptual overlap with Q_{24}. However, this is contrary to the actual ordering $\langle Q_{25} \succ Q_{26} | Q_{24} \rangle$, which reflects that fact that Q_{25}, which expects the same answer type as Q_{24}, should be deemed more useful than Q_{26}, which has a different answer type.

Q_{24} What are some good thriller *movies*?

Q_{25} What are some thriller *movies* with happy ending?

Q_{26} What are some good *songs* from a thriller movie?

The analysis above shows the importance of using the answer type when computing the similarity between two questions. However, instead of relying exclusively on a predefined hierarchy of answer types, we have decided to identify the *question focus* of a question, defined as the set of maximal noun phrases in the question that corefer with the expected answer. Focus nouns such as *movies* and *songs* provide more discriminative information than general answer types such as *products*. We use answer types only for questions such as Q_{27} or Q_{28} below that lack an explicit question focus. In such cases, an artificial question focus is created from the answer type (e.g. *location* for Q_{27}, or *method* for Q_{28}) and added to the set of question words.

Q_{27} *Where* can I buy a good coffee maker?

Q_{28} *How* do I make a pizza?

Let *qsim* be a general bag-of-words question similarity measure (e.g. *cos* or *mcs*). Furthermore, let *wsim* by a generic word meaning similarity measure (e.g. *wup*, *res*, *lin* or *jcn*). The equation below describes a modification of *qsim* that makes it aware of the questions focus:

$$qsim_f(Q_i, Q_r) = wsim(f_i, f_r) * qsim(Q_i - \{f_i\}, Q_r - \{f_r\})$$

Here, Q_i and Q_r refer both to the questions and their sets of words, while f_i and f_r stand for the corresponding focus words. We define $qsim$ to return 1 if one of its arguments is an empty set, i.e. $qsim(\emptyset, \_) = qsim(\_, \emptyset) = 1$. The new similarity measure $qsim_f$ multiplies the semantic similarity between the two focus words with the bag-of-words similarity between the remaining words in the two questions. Consequently, the word "movie" in Q_{26} will not be compared with the word "movies" in Q_{24}, and therefore Q_{26} will receive a lower utility score than Q_{25}.

In addition to the question focus, the *main verb* of a question can also provide key information in estimating question-to-question similarity. We define the main verb to be the content verb that is highest in the dependency tree of the question, e.g. *buy* for Q_{27}, or *make* for Q_{28}. If the question does not contain a content verb, the main verb is defined to be the highest verb in the dependency tree, as for example *are* in Q_{24} to Q_{26}. The utility of a question's main verb in judging its similarity to other questions can be seen more clearly in the questions below, where Q_{29} is the reference:

Q_{29} How can I *transfer* music from iTunes to my iPod?

Q_{30} How can I *upload* music to my iPod?

Q_{31} How can I *play* music in iTunes?

The fact that *upload*, as the main verb of Q_{30}, is more semantically related to *transfer* (*upload* is a hyponym of *transfer* in WordNet) is essential in deciding that $\langle Q_{30} \succ Q_{31} | Q_{29} \rangle$, i.e. Q_{30} is more useful than Q_{31} to Q_{29}.

Like the focus word, the main verb can be incorporated in the question similarity function as follows:

$$qsim_{fv}(Q_i, Q_r) = wsim(f_i, f_r) * wsim(v_i, v_r) * qsim(Q_i - \{f_i, v_i\}, Q_r - \{f_r, v_r\})$$

The new measure $qsim_{fv}$ takes into account both the focus words and the main verbs when estimating the semantic similarity between questions. When decomposing the questions into focus words, main verbs and the remaining words, we have chosen to multiply the corresponding similarities instead of, for example, summing them. Consequently, a close to zero score in each of them would drive the entire similarity to zero. This reflects the belief that question similarity is sensitive to each component of a question.

4 Experimental Evaluation

We use the question ranking dataset described in Section 2 to evaluate the two similarity measures cos and mcs, as well as their structured versions cos_f, cos_{fv}, mcs_f, and mcs_{fv}. We report one set of results for each of the four word similarity measures *wup*, *res*, *lin* or *jcn*. Each question similarity measure is evaluated in terms of its accuracy on the set of ordered pairs for each of the two annotation levels described in Section 2. Thus, for the first annotation level (L_1) , we evaluate only over the set of relations defined across the three

Question similarity ($qsim$)	Word similarity ($wsim$)							
	wup		*res*		*lin*		*jcn*	
	L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2
cos	69.1	69.3	69.1	69.3	69.1	69.3	69.1	69.3
cos_f	69.9	70.1	**72.5**	**72.7**	71.0	71.2	69.6	69.8
cos_{fv}	69.9	70.1	**72.5**	**72.6**	71.0	71.2	69.6	69.8
mcs	62.6	62.5	65.0	65.0	65.6	65.7	66.8	66.9
mcs_f	64.2	64.4	68.5	68.5	**68.8**	**68.9**	67.2	67.4
mcs_{fv}	65.8	66.0	68.8	68.8	**69.7**	**69.8**	67.7	67.8

Table 3: Accuracy results, with and without meaning and structure information.

sets $\mathcal{R}$, $\mathcal{U}$, and $\mathcal{N}$. If $\langle Q_i \succ Q_j | Q_r \rangle$ is a relation specified in the annotation, we consider the tuple $\langle Q_i, Q_j, Q_r \rangle$ correctly classified if and only if $u(Q_i, Q_r) > u(Q_j, Q_r)$, where u is the question similarity measure (Section 3). For the second annotation level (L_2), we also consider the relations annotated between *useful* questions inside the group $\mathcal{U}$.

We used the NLTK [1] implementation of the four similarity measures *wup*, *res*, *lin* or *jcn*. The *idf* values for each word were computed from frequency counts over the entire Wikipedia. For each question, the *focus* is identified automatically by an SVM tagger trained on a separate corpus of 2,000 questions manually annotated with focus information. The SVM tagger uses a combination of lexico-syntactic features and a quadratic kernel to achieve a 93.5% accuracy in a 10-fold cross validation evaluation on the 2,000 questions. The *main verb* of a question is identified deterministically using a breadth first traversal of the dependency tree.

The overall accuracy results presented in Table 3 show that using the focus word improves the performance across all 8 combinations of question and word similarity measures. For cosine similarity, the best performing system uses the focus words and Resnik's similarity function to obtain a 3.4% increase in accuracy. For the meaning aware similarity *mcs*, the best performing system uses the focus words, the main verb and Lin's word similarity to achieve a 4.1% increase in accuracy. The improvement due to accounting for focus words is consistent, whereas adding the main verb seems to improve the performance only for *mcs*, although not by a large margin. The second level of annotation brings 261 more relations in the dataset, some of them more difficult to annotate when compared with the three groups in the first level. Nevertheless, the performance either remains the same (somewhat expected due to the relatively small number of additional relations), or is marginally better. The random baseline – assigning a random similarity value to each pair of questions – results in 50% accuracy. A somewhat unexpected result is that *mcs* does not perform better than *cos* on this dataset. After analysing the result in more detail, we have noticed that *mcs* seems to be less resilient than *cos* to variations in the length of the questions. The Microsoft paraphrase corpus was specifically designed such that "the length of the shorter of the two sentences, in words, is at least 66% that of the longer" (Dolan and Brockett, 2005), whereas in our dataset the two questions in a pair can have significantly different lengths [2].

The questions in each of the 60 groups have a high degree of lexical overlap, making the dataset especially difficult. In this context, we believe the results are encouraging. We expect to obtain further improvements in accuracy by allowing relations between all the words in a question to influence the overall similarity measure. For example, question Q_{19} has the same focus word as the reference question Q_5 (repeated below), yet the difference between the focus word prepositional modifiers makes it a neutral question.

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

[2] Our implementation of *mcs* did performed better than *cos* on the Microsoft dataset.

Q_5 What's a good summer camp to go to in FL?

Q_{19} What's a good summer camp in Canada?

Some of the questions in our dataset illustrate the need to design a word similarity function specifically tailored to reflect how words change the relative usefulness of a question. In the set of questions below, in deciding that Q_{33} and Q_{34} are more useful than Q_{36} for the reference question Q_{32}, an ideal question ranker needs to know that the "Mayflower Hotel" and the "Queensboro Bridge" are in the proximity of "Midtown Manhattan", and that proximity relations are relevant when asking for directions. A coarse measure of proximity can be obtained for the pair ("Manhattan", "Queensboro Bridge") by following the *meronymy* links connecting the two entities in WordNet. However, a different strategy needs to be devised for entities such as "Mayflower Hotel", "JFK", or "La Guardia" which are not covered in WordNet.

Q_{32} What is the best way to get to Midtown Manhattan from JFK?

Q_{33} What's the best way from JFK to Mayflower Hotel?

Q_{34} What's the best way from JFK to Queensboro Bridge?

Q_{35} How do I get from Manhattan to JFK airport by train?

Q_{36} What is the best way to get to LaGuardia from JFK?

Finally, to realize why question Q_{35} is useful one needs to know that, once directions on how to get by train from location X to location Y are known, then normally it suffices to reverse the list of stops in order to obtain directions on how to get from Y back to X.

5 Future Work

We plan to integrate the entire dependency structure of the question in the overall similarity measure, possibly by defining kernels between questions in a maximum margin model for ranking. We also plan to extend the word similarity functions to better reflect the types of relations that are relevant when measuring question utility, such as proximity relations between locations. Furthermore, we intend to take advantage of databases of interrogative paraphrases and paraphrase patterns that were created in previous research on question reformulation.

6 Conclusion

We presented a novel question ranking task in which previously known questions are ordered based on their relative utility with respect to a new, reference question. We created a dataset of 60 groups of questions [3] annotated with a partial order relation reflecting the relative utility of questions inside each group, and used it to evaluate the ranking performance of several meaning and structure aware utility functions. Experimental results demonstrate the importance of using structural information in judging the relative usefulness of questions. We believe that the new perspective on ranking questions has the potential to significantly improve the usability of social QA sites.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions.

References

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press, New York.

Bernhard, Delphine and Iryna Gurevych. 2008. Answering learners' questions by retrieving question paraphrases from social Q&A sites. In *EANL '08: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 44–52, Morristown, NJ, USA. Association for Computational Linguistics.

Dolan, William B. and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pages 9–16.

[3] The dataset will be made publicly available.

Hermjakob, Ulf, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural language based reformulation resource and web exploitation for question answering. In *Proceedings of TREC-2002*.

Jeon, Jiwoon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM'05)*, pages 84–90, New York, NY, USA. ACM.

Jiang, J.J. and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, pages 19–33.

Jijkoun, Valentin and Maarten de Rijke. 2005. Retrieving answers from frequently asked questions pages on the Web. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM'05)*, pages 76–83, New York, NY, USA. ACM.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on Artificial intelligence (AAAI'06)*, pages 775–780. AAAI Press.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tomuro, Noriko. 2003. Interrogative reformulation patterns and acquisition of question paraphrases. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 33–40, Morristown, NJ, USA. Association for Computational Linguistics.

Wu, Zhibiao and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA. Association for Computational Linguistics.

Zhao, Shiqi, Ming Zhou, and Ting Liu. 2007. Learning question paraphrases for QA from Encarta logs. In *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI'07)*, pages 1795–1800, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.