Constructing a Question Corpus for Textual Semantic Relations

Rui Wang[†] and Shuguang Li[†‡]

†LT Lab, DFKI GmbH, Saarbruecken, Germany
[†‡]Greedy Intelligence Ltd., China
ruiwang@dfki.de, lishuguang202@gmail.com

Abstract

Finding *useful* questions is a challenging task in Community Question Answering (CQA). There are two key issues need to be resolved: 1) what is a *useful* question to the given reference question; and furthermore 2) what kind of relations exist between a given pair of questions. In order to answer these two questions, in this paper, we propose a fine-grained inventory of textual semantic relations between questions and annotate a corpus constructed from the WikiAnswers website. We also extract large archives of question pairs with user-generated links and use them as labeled data for separating useful questions from neutral ones, achieving 72.2% of accuracy. We find such online CQA repositories valuable resources for related research.

Keywords: Community Question Answering, Question Usefulness, Textual Semantic Relations

1. Introduction

Community Question Answering (CQA) is a recently established type of question answering (QA), which shifts the inherent complexity of constructing sophisticated open-domain QA systems to volunteer contributors. A user can post a query (reference) question and wait for the right contributors to answer it. The user can also search the large archives of data for similar questions, in order to minimize the time between the submission of a question by a user and the subsequent posting of answers by volunteer contributors. CQA thus becomes similar to the traditional information retrieval (IR) task. The aim is to recommend a list of questions that can satisfy the user's information need.

However, retrieving questions from such large CQA repositories to satisfy the user's information need is by no means an easy task. One reason is that the user's information need is usually quite complex. The query questions are sometimes vague and ambiguous. Another reason is the lexical gap or word mismatch problem between questions; the same information need can be expressed by two questions with few common words. Conventional word-based retrieval models for question search would fail to capture the similarity between two questions. Consider the following example of questions that are semantically similar to each other:

1. Where can I get cheap airplane tickets?
2. Any travel website for low airfares?

This problem in CQA was firstly studied as a question usefulness ranking task in (Bunescu and Huang, 2010). The retrieved questions fell into three predefined classes representing different levels of "usefulness" for fulfilling the user's information need. However, their definitions of *usefulness* and *neutral* were too vague to be interpreted and the size of the manually labeled dataset was too small. Research on using CQA data to study the information needs of users is still in the early stages. A further study on the question usefulness ranking problem and how the question usefulness can benefit the question search is needed.

In another line of research, semantic relations between statements (compared to questions) have been popular topics in the recent years. For instance, Recognizing Textual Entailment (RTE) has attracted a lot of attention, from RTE-1 (Dagan et al., 2006) to RTE-5 (Bentivogli et al., 2009), as well as other textual semantic relations, paraphrase (Das and Smith, 2009), contradiction (de Marneffe et al., 2008), and some (Heilman and Smith, 2010) or all of them (Wang and Zhang, 2011). However, there is no detailed study on textual semantic relations between *questions* to our best knowledge.

Bernhard and Gurevych (2008) evaluate various string similarity measures and vector space based similarity measures on the task of retrieving question paraphrases from the WikiAnswers repository, which relies on the wiki technology used in Wikipedia[1]. We find that most of the "Reformulation" questions are related but not paraphrasing. However, the wealth of such questions available on the WikiAnswers website can still be used in research for finding *useful* questions. Furthermore, most "Useful" questions often have a textual entailment relation with the reference question. Therefore, CQA could benefit from the techniques used in the RTE community.

In this study, we make the first attempt to exploit the possible textual semantic relations between questions from an online CQA repository. The motivation behind is that once we define such relations, many applications can benefit from it. For instance, we can raise a question in alternative ways, find similar questions (i.e., paraphrase), decompose complex questions (i.e., entailment/subset relation), find out contrast/contradictory questions (i.e., contradiction), or link the related questions (i.e., usefulness). In particular, we focus on the following issues: 1) a simple way to collect question pairs from the WikiAnswers repository (Sec-

[1]Users of WikiAnswers can mark question reformulations, in order to prevent the duplication of questions asking the same thing in a different way. It should be noted that contributors get no reward, in terms of trust points, for providing or editing alternate wordings for questions.

tion 3.); 2) a finer-grained inventory of possible semantic relations between questions (Section 4.); and 3) an automatic approach to recognize such relations (Section 5.).

2. Related Work

2.1. Question Usefulness Ranking

Bunescu and Huang (2010) present a machine learning approach for the task of ranking previously answered questions in a CQA resource, with respect to their relevance to a new, unanswered reference question. They manually label 60 groups of questions. The relations between the reference question and the ranked questions are divided into three categories:

1. Reformulation: paraphrasing questions which use alternative words or expressions;
2. Useful: useful questions;
3. Neutral: other not useful questions.

Table 1 shows an example from this dataset. However, we find this categorization to be too coarse-grained. The definitions for "Usefulness" and "Neutral" are too vague to be interpreted. We also find that most of the "Useful" questions had a textual entailment relation with the reference question in this dataset.

Reference1:
What bike should I buy for city riding?
Reformulation1:
What is the best bike to buy to get myself around campus in the city ?
What is the best bike for traveling in a city ?
Useful1:
What bike should I buy for a starter bike ?
What bike should I get as a beginner ?
What bike should I get if I 'm a bigger person ?
What bike should I buy for free riding ?
What bike can I ride as a 16 year old ?
What bike should I buy to participate in my first triathlon ?
Neutral1:
What is a good bike to start mountain biking ?
What bike should I buy for hills ?
What bike should I buy my toddler ?
What bike should I buy for working out ?
What exercise bike should I buy, upright or recumbent ?
Which stationary bike Should I buy ?
What bike should I buy for working out ?
What exercise bike should I buy, upright or recumbent ?
What racing bike should I buy ?
What motorcycle jacket is better for a sport bike ?
What is the best tire to buy for a city bike ?

Table 1: Question usefulness ranking example

2.2. Question Paraphrasing

Question paraphrasing is critical in many natural language processing (NLP) applications, especially for question reformulation in QA. Zhao et al. (2007) present a novel method for automatically extracting question paraphrases from a search engine's logs and generating templates for question reformulation. A SVM model is trained, by making use of different features extracted from the questions and the most effective combination of features was identified. However, this method requires access to Microsoft's search engine logs. Sacaleanu et al. (2008) summarize a set of question templates. When there is an input question, the system compares it with each of the template in the set and finds the best match. Therefore, the corresponding expected answer type can be found as well. Bernhard and Gurevych (2008) evaluate various string similarity measures and vector space based similarity measures on the task of retrieving question paraphrases from the WikiAnswers repository. As discussed in the previous section, most of the gold standard question pairs are not paraphrases. However, we will show that the wealth of this CQA data repository is still valuable for information need satisfaction research, after further annotation and processing.

2.3. Textual Entailment

Textual entailment is another important research field in Natural Language Processing. Recognizing Textual Entailment (RTE) (Dagan et al., 2006) detects whether a Hypothesis (H) can be inferred (or entailed) by a Text (T). It is shown to be helpful for QA by Harabagiu and Hickl (2006). Conventional methods for RTE tasks rely on similarity measures between texts. Androutsopoulos and Malakasiotis (2010) nicely survey the field together with paraphrase acquisition. Here we just list a few of the previous works. Many approaches make use of the bag-of-words similarity, assisted by lexical resources like WordNet, for instance, (Corley et al., 2005). Kouylekov and Magnini (2005) exploit the use of syntactic features and proposed a syntactic tree editing distance measure to detect entailment relations. Wang and Neumann (2007) propose a subsequence kernel method approach, to incorporate structural features extracted from syntactic dependency trees for this task. Wang and Zhang (2009) combine syntactic and semantic features to capture the key information shared between texts.

As far as can be ascertained, no previous work has been done in Recognizing Textual Entailment from CQA question pairs. We will make the first attempt to exploit this direction using our annotated WikiAnswers dataset and the first fine-grained framework.

3. WikiAnswers and Data Collection

WikiAnswers is a social QA website similar to Yahoo! Answers. As of February 2008, it contained 1,807,600 questions, sorted into 2,404 categories. Compared with its competitors, the main originality of WikiAnswers is that it relies on the wiki technology used in Wikipedia, which means that answers can be edited and improved over time by the contributors (Bernhard and Gurevych, 2008). WikiAnswers allows users to mark question pairs that they think are

rephrasings ("alternate wordings", or paraphrases) of existing questions. For example, the following questions are marked as paraphrasing for the reference question "How do vaccines work?":

1. How does the flu shot work?
2. Is there a vaccine to protect against swine flu?
3. How does the body get rid of viruses like the cold or flu?
4. What steps involving the immune system and white blood cells help people with the swine flu recover?
5. What is an example of how a vaccine works?

After a thorough investigation of this repository, we found that if a question Q is marked for many other questions (Q appears on many other questions' pages as a "rephrasing" question), Q is usually a more general question. In contrast, if a question is rarely marked for any other questions, it is usually more specific. This is similar to PageRank[2], in which a webpage with more incoming hyperlinks tends to be more important. Based on this property, we collected more than 1,500 groups of questions from WikiAnswers. In each group, one reference question whose incoming hyperlinks are more than 20 has been marked as "rephrasing" for the other questions. Some examples are shown in Table 2. For instance, "How do you write a good concluding sentence or paragraph?" is marked as a "rephrasing" for 74 other questions in the WikiAnswers website; five out of the 74 are shown following it, while the Question "How do you write good beginnings and endings for paragraphs and essays?" has not been marked as a "rephrasing" on any pages.

How do you write a good concluding sentence or paragraph?	74
How do you write good beginnings and endings for paragraphs and essays?	0
What is a conclusion of managing conclusion?	0
What is a good way to close an essay about Robert Hooke?	0
What is a good closing paragraph on your educational goals?	0
How do you write conclusion for bottle bio mes?	0
Who should not get a swine flu vaccination?	54
Can swine flu vaccination be taken during pregnancy?	2
Can you get the H1N1 vaccine if you are currently sick with the swine flu?	2
Is it safe for a pregnant woman to get the H1N1 flu shot?	4
Is a flu shot safe when you are pregnant?	2
Is the swine flu shot active?	0
What should you feed a rabbit?	51
How often to feed a 45 days old rabbit?	0
Can bunnies eat cantaloupe?	0
What do Rex bunnies eat?	0
Can bunnies eat potatoes?	0
What do backyard bunnies eat?	0

Table 2: WikiAnswers dataset examples

4. Textual Semantic Relations between Questions

We perform some preliminary manual analysis on the CQA repositories and the data show that most of the useful questions often have a textual entailment relation with the reference question (or other types of semantic relations). Therefore, CQA could potentially benefit from the techniques used in the RTE community.

Given a user's question, we want to compare the retrieved questions with it, in terms of information need satisfaction (i.e., how *useful* it is). Following the work of (Bunescu and Huang, 2010) and our previous work (Wang and Sporleder, 2010), we propose a finer-grained question usefulness classification framework for the possible relations between the reference question and the input question(s). In the whole spectrum, the inventory of possible textual semantic relations between a question pair $< Q_r, Q_i >$ are listed as follows:

1. $Q_r = Q_i$ (E): the two questions are (almost) the same, asking about the same thing;
2. $Q_r < Q_i$ (G): the input question is more general than the reference question. For example, the reference question is asking about the apple, while the input question is asking about fruit. In other words, the answer to the reference question is a subset of the answer to the input question;
3. $Q_r > Q_i$ (S): similar to the previous label, but the other way around. The input question is asking about more specific things than the reference question. In particular, yes-no questions can be viewed as verifications of concrete facts, which are very specific;
4. $Q_r < - Q_i$ (P): the input question is asking about a presupposition of the reference question, for example, a definition of a concept mentioned in the reference question;
5. $Q_r - Q_i$ (R): the input question is related to the reference question, but the relation is not one of the above mentioned ones;
6. $Q_r \; ! = Q_i$ (N): the input question is unrelated to the reference question, although the topics of both questions may be the same. The answer to the input question is also not helpful for answering the reference question, e.g., when is the summer camp in Florida vs. when is the summer camp in Canada;

Within this fine-grained inventory of textual semantic relations between questions, we are able to formulate our task

[2] http://en.wikipedia.org/wiki/PageRank

as a question usefulness classification task. Although this inventory is much finer-grained than a binary classification, to annotate a large corpus is quite time consuming and laborious. Also, the classification performance will drop greatly when the number of classes increases.

Therefore, we annotate a rather small corpus with two annotators and use it as the main test set for our experiments instead of the training set. Some annotated examples are shown in Table 3.

Did Prince become a Jehovah's Witness?	Reference
Is Prince a Jehovah's Witness?	E
Is Prince still a practicing Jehovah's Witness?	E
Are there celebrity Jehovah's Witnesses?	G
What kingdom hall does prince go to?	N
What did dolphins evolve from?	Reference
When did dolphins evolve?	N
Where are dolphins located?	N
Did dolphins evolve into anything?	N
How did dolphins evolve from a dinosaur?	S
How does the flu shot work?	Reference
What is a flu shot?	P
Is there a vaccine to protect against swine flu?	N
What are the ingredients in the swine flu vaccine?	Reference
Is the swine flu shot active?	N
Does the flu shot have swine ingredients?	R

Table 3: WikiAnswers questions with annotations

This multi-class classification problem can be decomposed into binary classification problems depending on the specific applications. In this paper, we focus on the following two binary classification subtasks (reduced from the classes defined above): 1) recognizing textual entailment between questions: the questions in E, G, and S are treated as positive instances (G and S with different directions), while the others are negative; and 2) question usefulness recognition: this is a task to distinguish the *useless* questions (N) from the rest.

5. Experiments

In our question usefulness classification task, we use both syntactic and semantic similarity measures to "calculate" the relation between the reference question and other questions in the CQA data. The scores of these similarity measures are used as features for training the classifier.

Bag-of-Words (BOW) Similarity is a widely used similarity measure for textual entailment, which calculates the similarity based on the ratio of overlapping words and it is usually combined with lexical resources like WordNet.

Syntactic Dependency Similarity can be defined based on the matching of two dependency triple sets output by a parser, e.g., (Wang and Neumann, 2007).

Predicate-Argument Similarity proposed by Wang and Zhang (2009) is a text relatedness scoring method for textual entailment. Predicate-argument structures (PASes) were used to calculate the semantic similarity between texts.

TF-IDF is widely used in most IR systems, as it is both efficient and effective.

We randomly select 50 groups of questions from our WikiAnswers collection. From each group, we also randomly select 4-5 questions, as well as the reference question. The resulting dataset is similar to the examples shown in Table 2. Within each selected group, two annotators are employed to mark the relations between the reference question and the other questions. A total of 213 pairs of questions were annotated, with an initial inter-annotator agreement of 76.5% (163 out of 213). However, if we collapse the annotations into the binary labels (either *entailment* vs. others or *usefulness* vs. others), all the annotations are agreed. The 213 question pairs are named as the dataset $Wiki_s$.

Feature Sets	ALL - TFIDF	ALL
Entailment Recognition	56.1%	57.7%
Usefulness Recognition	64.0%	66.2%

Table 4: Results of RTE on the annotated dataset

For our experiment of RTE on questions, we run 5-fold cross validations using libsvm[3] on the $Wiki_s$ dataset. The results for question textual entailment is shown in Table 4. The average accuracy is 56.1% when the bag-of-words, syntactic, and semantic features are used. The average accuracy can be further improved to 57.7% if the TFIDF feature is added. The results are quite consistent with normal RTE challenges (Dagan et al., 2006) and our previous research (Wang and Zhang, 2009), as we use the similar feature model and classifier.

We run the other experiment on predicting whether a question is useful or not to the reference question by 'changing' the data annotation to *useful* vs. neutral. It appears that for the classifier, *Usefulness* recognition is an 'easier' task. In order to further investigate this issue, we extend our experiments on other datasets in the following.

We use two datasets created by Bunescu and Huang (2010), $QSimiple$ and $QComplex$, which contain 60 groups of questions spanning a wide range of topics. Each group consists of a reference question followed by a partially ordered set of questions. The latter questions are mainly divided into three categories: "Reformulation" questions are thought to be more useful than the other questions, and the "Useful" questions are deemed to better satisfy the user's information need than "Neutral" questions. In $QComplex$, unanswered questions (reference questions) tend to be longer, whereas other questions in the group are shorter. The $QSimple$ is the opposite. There are 1,329 question pairs in $QSimple$ and 1,970 pairs in $QComplex$.

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Training → Testing	ALL - TFIDF	ALL
$Wiki_s \rightarrow QSimple$	48.2%	61.5%
$Wiki_s \rightarrow QComplex$	43.1%	52.2%
$Wiki_W \rightarrow QSimple$	65.5% (17.3%↑)	65.5% (4.0%↑)
$Wiki_W \rightarrow QComplex$	70.3% (27.2%↑)	72.2% (20.0%↑)

Table 5: Question usefulness classification results

We manually relabeled the datasets using our inventory of relations (Section 4.). We found that most of the "Useful" questions have a textual entailment relation with the reference question. In order to run the binary classification experiment, the questions in "Reformulation" and "Useful" are treated as positive instances, while the questions in "Neutral" are treated as negative ones. For the questions annotated in $Wiki_s$, the questions labeled N are treated as negative examples, while the other pairs are positive. The results of the classifiers with different feature sets are shown in the first two rows of Table 5.

One significant observation is that on these two datasets, $QSimple$ and $QComplex$, TFIDF shows great importance for the classification. However, the overall performance is not optimal. As in $Wiki_s$, about half of the question pairs are paraphrases or have a textual entailment relation, we think that WikiAnswers might be a good resource for learning to retrieve useful questions. Therefore, we randomly select 6,574 question pairs from the WikiAnswers dataset, detailed in Section 3. as positive examples for training a two-class SVM classifier. We also randomly select 6,293 question pairs ($Wiki_W$) which are not marked as "rephrasing" questions on the WikiAnswers website as negative examples. The results are shown in the last two rows of Table 5.

Along with the increase of the data size, the difference between the feature models seems to have less impact on the final result. Although the data do not contain fine-grained manual annotations, the result we achieve is competitive with previous research (Bunescu and Huang, 2010)[4]. Comparing the latter results with the previous ones ($Wiki_s$), we conclude that $Wiki_W$, which has been 'user-annotated', is a valuable resource for learning to retrieve useful questions. All the data with annotations are publicly available[5] by the time of this publication.

6. Summary

In this paper, we presented our investigation on the textual semantic relations between questions. We collected large quantities of question pairs from WikiAnswers, which were shown to be a valuable resource for learning to retrieve useful questions. We formulated this problem as a fine-grained question pair classification task. The classification task was further decomposed into two binary classification subtasks, recognizing textual entailment between questions and recognizing useful questions. Furthermore, we compared the effectiveness of two feature models (in- or excluding TFIDF) using a manually-annotated gold standard. Both the annotated and unannotated corpora are publicly available. For further improvements, we would like to consider more features of the questions, e.g., the categories, in- and out-coming link information, etc. In addition, we would also like to make a full cycle of bootstrapping by adding more links to the related questions in the existing CQA repositories.

[4]Although we use the same data, the results are not directly comparable. Bunescu and Huang (2010) evaluated the pairwise accuracy of ranking a list of questions, while we cast it into a classification task and evaluate the classification accuracy. In addition, we do not want to claim a better feature model or classification approach in this work.

[5]http://www.coli.uni-saarland.de/~rwang/resources/LREC2012Data.zip

Acknowledgements

The first author was partially supported by the EXCITEMENT project (ICT-287923) funded by the European Community under the Seventh Framework Programme (FP7) for Research and Technological Development.

7. References

I. Androutsopoulos and P. Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

L. Bentivogli, B. Magnini, I. Dagan, H.T. Dang, and D. Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *Proceedings of TAC Workshop*. NIST.

Delphine Bernhard and Iryna Gurevych. 2008. Answering learners' questions by retrieving question paraphrases from social q&a sites. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Razvan Bunescu and Yunfeng Huang. 2010. Learning the relative usefulness of questions in community qa. In *Proceedings of EMNLP*.

C. Corley, A. Csomai, and R. Mihalcea. 2005. Text semantic similarity, with applications. In *Proceedings of Recent Advances in Natural Language Processing*.

I. Dagan, O. Glickman, and B. Magnini. 2006. The pascal recognising textual entailment challenge. *Lecture Notes in Computer Science: Machine Learning Challenges*.

D. Das and N.A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of ACL-IJCNLP*.

M.C. de Marneffe, A.N. Rafferty, and C.D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual*

Meeting of the Association for Computational Linguistics, pages 905–912, Sydney, Australia, July. Association for Computational Linguistics.

M. Heilman and N.A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Proceedings of NAACL-HLT*.

Milen Kouylekov and Bernardo Magnini. 2005. Recognizing textual entailment with tree edit distance algorithms. In *Proceedings of the PASCAL Challenges on RTE*, pages 17–20.

Bogdan Sacaleanu, Constantin Orasan, Christian Spurk, Shiyan Ou, Óscar Ferrándezandez, Milen Kouylekov, and Matteo Negri. 2008. Entailment-based question answering for structured data. In *In Proceedings of COLING 2008 (Posters and Demonstrations)*.

Rui Wang and Günter Neumann. 2007. Recognizing textual entailment using a subsequence kernel method. In *Proceedings of AAAI*.

Rui Wang and Caroline Sporleder. 2010. Constructing a textual semantic relation corpus using a discourse treebank. In *Proceedings of LREC*.

Rui Wang and Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structures. In *Proceedings of EMNLP*.

Rui Wang and Yi Zhang. 2011. A multi-dimensional classification approach towards textual semantic relation recognition. In *Proceedings of CICLing*.

Shiqi Zhao, Ming Zhou, and Ting Liu. 2007. Learning question paraphrases for qa from encarta logs. In *Proceedings of IJCAI*.