Answering Complex Questions via Exploiting Social Q&A Collection

Youzheng Wu Chiori Hori Hisashi Kawai Hideki Kashioka
Spoken Language Communication Group, MASTAR Project
National Institute of Information and Communications Technology (NiCT)
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
{youzheng.wu, chiori.hori, hisashi.kawai, hideki.kashioka}@nict.go.jp

Abstract

This paper regards social Q&A collections, such as Yahoo! Answer as a knowledge repository and investigates techniques to mine knowledge from them for improving a sentence-based complex question answering (QA) system. In particular, we present a question-type-specific method (QTSM) that studies at extracting question-type-dependent cue expressions from the social Q&A pairs in which question types are the same as the submitted question. The QTSM is also compared with question-specific and monolingual translation-based methods presented in previous work. Thereinto, the question-specific method (QSM) aims at extracting question-dependent answer words from social Q&A pairs in which questions are similar to the submitted question. The monolingual translation-based method (MTM) learns word-to-word translation probabilities from all social Q&A pairs without consideration of question and question type. Experiments on extension of the NTCIR 2008 Chinese test data set verify the performance ranking of these methods as: QTSM $>$ {QSM, MTM}. The largest F_3 improvements of the proposed QTSM over the QSM and MTM reach 6.0% and 5.8%, respectively.

1 Introduction

Research on the topic of QA systems has mainly concentrated on answering factoid, definitional, reason and opinion questions. Among the approaches proposed for answering these questions, machine learning techniques have been found more effective in constructing QA components from scratch. Yet these supervised techniques require a certain scale of question and answer (Q&A) pairs as training data. For example, Echihabi et al. (2003) and Sasaki (2005) respectively constructed 90,000 English and 2,000 Japanese Q&A pairs for their factoid QA systems. Cui et al. (2004) collected 76 term-definition pairs for their definitional QA system. Higashinaka and Isozaki (2008) used 4,849 positive and 521,177 negative examples in their reason QA system. Stoyanov et al. (2005) required a known subjective vocabulary for their opinion QA system. This paper is concerned with answering complex questions which answers generally consists of a list of nuggets (Voorhees, 2003; Mitamura et al., 2008). Apart from definitional and opinion (TAC, 2008) complex questions, many other types of complex questions have not yet to be thoroughly studied[1]. To answer these complex questions via supervised techniques, we need to collect training Q&A pairs for each type of complex question, though this is an extremely expensive and labor-intensive task.

This paper is to explore the possibility of automatic learning of training Q&A pairs and mining needed knowledge from social Q&A collections such as Yahoo! Answer[2]. That is to say, we are interested in whether or not millions of, possible noisy, user-generated Q&A pairs can be exploited for automatic QA system. This is a very important question because a positive answer can indicate that a plethora of training Q&A data is readily available to QA researchers.

Many studies, such as (Riezler et al., 2007; Surdeanu et al., 2008; Duan et al., 2008; Wang, 2010a) have addressed retrieving of similar Q&A pairs from social QA websites as answers to test questions; thus answers cannot be generated for questions that have not been answered on such

[1]Most complex questions have generally been called what-questions in previous studies. This paper argues that it is helpful to treat them discriminatively.

[2]http://answers.yahoo.com/

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 956–964,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

sites. Our study, however, regards social Q&A websites as a knowledge repository and aims at exploiting knowledge from them for synthesizing answers to questions, which have not been answered on these sites. Even for questions that have been answered, it is necessary to perform answer summarization as (Liu et al., 2008) indicated. Our approach can also be used for this purpose. To the best of our knowledge, there appears to be very little literature on this aspect.

Various kinds of knowledge can be mined from social Q&A collections for supporting complex QA system. In this paper, we present a question-type-specific method (QTSM) to mine question-type-specific knowledge and compare it with question-specific and monolingual translation-based methods proposed in related work. Given a question Q, the three methods can be summarized as follows: (1) The proposed QTSM studies at recognizing question type Q_t from the Q; collecting Q&A pair in which question types are the same as Q_t; extracting salient cue expressions that are indicative of answers to the question type Q_t; and using the expressions and Q&A pairs to train a binary classifier for removing noise candidate answers. (2) The question-specific method (QSM) tries to collect Q&A pairs that are similar to Q from social Q&A collection, and extract question-dependent (Q-specific in this case) answer words to improve complex QA system. (3) The monolingual translation-based method (MTM) employs all social Q&A pairs and learns word-to-word translation probabilities from them without consideration of question Q and question type Q_t to solve the lexical gap problem in complex QA system. The three methods are evaluated in terms of the extension of the NTCIR 2008 test data set. The Pourpre v.0c evaluation tool (Lin and Demner-Fushman, 2006) is employed, which is also adopted to evaluate TREC QA systems. The experiments show that the proposed QTSM is most effective, for instance, the largest F_3/NR improvements of QTSM over the baseline, QSM, and MTM models reach 8.6%/12.6%, 6.0%/6.7%, and 5.8%/7.1%, respectively. The ranking of the methods was: QTSM $>$ {QSM, MTM}.

2 Social Q&A Collection

Social QA websites such as Yahoo! Answer and Baidu Zhidao[3] provide an interactive platform for users to post questions and answers. After questions are answered by users, the best answer can be chosen by the asker or nominated by the community. Table 1 demonstrates an example of these Q&A pairs, the number of which has risen dramatically on such sites. The pairs could collectively form a source of training data needed in supervised machine-learning-based QA systems.

Question	What do you think is the main cause of global warming?
Best Answer	The primary cause of global warming is the emission of green house gases like carbon dioxide, methane, nitrous oxide...
Other Answer	...What is NOT at all clear is whether human-activity is causing for the current warming trend...
Other Answer	First of all, it is damaging outcome of man-made faults...

Table 1: Example of social Q&A pairs

This paper aims at exploiting such user-generated Q&A collections for improving complex QA systems via automatic learning of Q&A training pairs and mining needed knowledge from them. Social collections, however, have two salient characteristics: textual mismatch between questions and answers (i.e., question words are not necessarily used in answers), and user-generated spam or flippant answers, which are unfavorable factors in our study. We only crawl questions and their best answers to form Q&A pairs, wherein the best answers are longer than the empirical threshold (20 words). Finally, about 40 million Q&A pairs were crawled from Chinese social QA websites and will be used as a source of training data.

3 Complex QA System

The typical complex QA system architecture is a cascade of three modules. The Question Analyzer analyzes test question and identifies type of question. The Document Retriever & Answer Candidate Extractor retrieves documents related to questions from the given collection (*Xinhua* and *Lianhe Zaobao* newspapers from 1998-2001 were used in this study) for consideration, and segments them into sentences as answer candidates. The Answer Ranker applies state-of-the-art IR formulas (e.g., KL-divergence language model) to estimate "similarities" between sentences (we used 1,024 sentences) and question and ranks sentences according to their similarities. Finally, the top N sentences are deemed the final answers.

Given question Q_1 = "What are the hazards of global warming?" and its three answer candidates, a_1 = "Solutions to global warming range from changing a light bulb to engineering giant reflec-

[3] http://zhidao.baidu.com/

tors in space ...," a_2 = "Global warming will bring bigger storms and hurricanes that hold more water ...," and a_3 = "nuclear power has relatively low emission of carbon dioxide (CO_2), one of the major causes of global warming", it is hard for the above architecture to correctly select a_2 as answer, because the three candidates contain the same keywords in question Q_1. To improve this architecture, external knowledge must be incorporated. As introduced in section 2, social Q&A collection is a good choice for mining needed knowledge. In this paper, we propose a question-type-specific technique of exploiting social Q&A collection (as introduced in section 4) to mine the knowledge, and compare it with question-specific (section 5.1) and monolingual translation-based (section 5.2) methods in experiments.

4 QTSM

Based on our observation, that is, answers to a type of complex question usually contain question-type-dependent cue expressions that are helpful in answering complex questions, we propose the QTSM that aims to learn these cue expressions for each type of question and utilize them to improve complex QA systems.

For each test question, the QTSM performs the following steps: (1) Recognizing the type of test question by identifying the *question focus* of question. (2) Collecting positive and negative training Q&A pairs of the type of question from the social Q&A collection. (3) Extracting question-type-specific salient cue expressions from the Q&A pairs. (4) Utilizing the cue expressions and Q&A pairs to build a binary classifier of the type of the test question. (5) Employing the classifier to remove noise from candidate answers before using the Answer Ranker to select final answers to the question.

4.1 Question Type

Earlier work on factoid QA systems tried to recognize question types via classification techniques (Li, et al., 2002), which require taxonomy of question types such as location, organization, person and training instances for each type. This algorithm may be inappropriate to complex QA systems due to there are hundreds of question types and we have little prior knowledge about defining complex QA-oriented taxonomy. This paper recognizes type of complex question by identifying its question focus. Question focus is defined as a short subsequence of tokens (typically 1-3 words) in a question that are adequate for indicating its question type. Take Q_1 = "What are the hazards of global warming?" and Q_2 = "What disasters are caused due to global warming?" as examples, *hazard* and *disaster* are their corresponding question focuses.

To recognize question type, we simply assume that type of complex question is only determined by its question focus; that is to say, question-type and question focus can be used interchangeably in this paper. Based on this assumption, question Q_1 and Q_2 belong to the hazard-type and disaster-type questions, respectively. Krishnan (2005) has showed that (a) the accuracy of recognizing question types reached 92.2% by using only question focuses and (b) the accuracy of recognizing question focuses was 84.6%. This indicates that most questions contain question focuses and it is practicable to represent question types by question focuses. Thereby, the task of recognizing question types shifts to recognizing question focuses from questions.

We regard question focus recognition as a sequence-tagging problem and employ conditional random fields (CRFs) because many studies have proven a consistent advantage of CRFs in sequence tagging. We manually annotate 4,770 questions with question focuses to train a CRF model, which classifies each question word into a set of tags $O = \{I_B, I_I, I_O\}$: I_B for a word that begins a focus, I_I for a word occuring in the middle of a focus and I_O for a word outside of a focus. In the following feature templates used in the CRF model, w_n and t_n refer to word and part-of-speech (PoS), respectively, and n refers to the relative position from the current word n=0. The feature templates contain four types: unigrams of w_n and t_n, where $n = -2, -1, 0, 1, 2$; bigrams of $w_n w_{n+1}$ and $t_n t_{n+1}$, where n=$-1, 0$; trigrams of $w_n w_{n+1} w_{n+2}$ and $t_n t_{n+1} t_{n+2}$, where $n = -2, -1, 0$; and bigrams of $O_n O_{n+1}$, where n=$-1, 0$.

Among 4,770 questions, 1,500 are held out as test set, the others are used for training. The experiment shows that precision of the CRF model on the test set is 89.5%. At offline, the CRF model is used to recognize question focuses from questions of social Q&A pairs. Finally, we recognize 103 question focuses for which frequencies are larger

than 10,000. Moreover, the numbers of question focuses for which frequencies are larger than 100, 1,000, and 5,000 are 4,714, 807, and 194, respectively. Among 4,714 recognized question focuses, 87% are not included in the question focus training questions. At online phrase, the CRF model is used to identify question focus of test question.

4.2 Q&A Pairs

It is necessary to manually annotate question focuses for identifying question types, however, training Q&A pairs for the question types can be automatically leant as follows once question types are determined.

4.2.1 Basic Positive Q&A Pairs

For question-type X, social Q&A pairs for which question focuses are the same as X are regarded as basic positive Q&A pairs QA_{basic} of X-type questions. Formally, $QA_{basic} = \{QA_i | AT_i = X\}$, where QA_i denotes a Q&A pair, and AT_i denotes question focus of QA_i. Table 2[4] reports the number of Q&A pairs for each type of question in the extension of the NTCIR 2008 test set (discussed in the experimental section). For example, 10,362 Q&A pairs are leant for answering hazard-type questions. Table 3 lists questions which, together with their best answers, are utilized as basic positive training pairs of the corresponding type of complex questions.

Qtype	#	Qtype	#
Hazard-type	10,362	Function-type	41,005
Impact-type	35,097	Significance-type	14,615
Attitude-type	1,801	Measure-type	3,643
Reason-type	50,241	Casualty-type	102
Event-type	5,871	Scale-type	642

Table 2: Numbers of basic positive Q&A pairs learned (#)

4.2.2 Bootstrapping Positive Q&A Pairs

For question types like casualty(伤亡)-type for which only a few basic positive Q&A pairs are learnt, Q&A pairs for similar question types like fatality(死伤)-type can be used. Hownet (Dong, 1999), a lexical knowledge base with rich semantic information and which serves as a powerful tool for meaning computation, is adopted for bootstrapping the basic positive Q&A pairs. In Hownet, a word may represent multiple concepts, and each concept consists of a group of sememes. For example, the Chinese word for "伤亡(casualty)" is described as: "phenomena|现象, wounded|受伤, die|死, undesired|莠". The similarity between two words can be estimated by,

$$sim(w_1, w_2) = \max_{1 \le i \le |w_1|; 1 \le j \le |w_2|} sim(c_i, c_j)$$

$$sim(c_i, c_j) = \frac{\sum_{1 \le k \le |c_i|} \max_{1 \le z \le |c_j|} sim(se_{i,k}, se_{j,z})}{|c_i|}$$

where c_i and c_j represent the i-th and j-th concept of word w_1 and w_2, respectively, $|w_1|$ is the number of concepts that w_1 represents, $se_{i,k}$ denotes the k-th sememe of concept c_i, $|c_i|$ is the number of sememes of concept c_i, and $sim(se_{i,k}, se_{j,z})$ is 1 if they are same, otherwise the value is set to 0.

Accordingly, the bootstrapping positive Q&A pairs QA_{boot} of X-type questions is composed of the Q&A pairs in which question focuses are similar to X. Formally, $QA_{boot} = \{QA_j | sim(AT_j, X) > \theta_1\}$, where, AT_j is question focus of QA_j, θ_1 is the similarity threshold.

Qtype	Questions of Q&A pairs
Hazard -type	What are the **hazards** of the trojan.psw.misc.kah virus? List the **hazards** of smoking. What are the **hazards** of contact lenses?
Casualty -type	What were the **casualties** of the Sino-French War? What were the **casualties** of the Sichuan earthquake? What were the **casualties** of the Indonesian Tsunami?
Reason -type	What are the main **reasons** for China's water shortage? What are the **reasons** for asthma? What are the **reasons** for air pollution?

Table 3: Questions (translated from Chinese) of Q&A pairs (words in bold are question focuses).

4.2.3 Negative Pairs & Preprocessing

For each type of question, we also randomly select some Q&A pairs that do not contain question focuses and their similar words in questions as negative training Q&A pairs.

Preprocessing of the training data, including word segmentation, PoS tagging, named entity (NE) recognization (Wu et al., 2005), and dependency parsing (Chen, 2009), is conducted. We also replace each NE with its tag type.

4.3 Extracting Cue Expressions and Building Classifiers

In this paper, we extract two kinds of cue expressions: n-grams at the sequential level and depen-

[4]Function-type: What are the **functions** of the United Nations? Impact-type: List the **impact** of the 911 attacks on the United States. Significance-type: List the **significance** of China's accession to the WTO. Attitude-type: List the **attitudes** of other countries toward the Israel-Palestine conflict. Measure-type: What **measures** have been taken for energy-saving in Japan? Event-type: List the **events** in the Northern Ireland peace process. Scale-type: Give information about the **scale** of the Kunming World Horticulture Exposition. Refer to Table 3 for other types of questions.

dency patterns at the syntactic level. The purpose of cue expression mining is to extract a set of frequent lexical and PoS-based subsequences that are indicative of answers to a type of question.

The n-gram cue expressions include (1) 3,000 lexical unigrams selected using the formula: $score_w = tf_w \times log(\frac{N}{df_w})$, where tf_w denotes the frequency of word w, df_w denotes the frequency of Q&A pairs in which w appears, and N is the total number of the Q&A pairs; (2) lexical bigrams and trigrams that contain the selected unigrams and their frequencies are larger than the empirical thresholds; (3) PoS-based unigrams; and (4) PoS-based bigrams with frequencies larger than the threshold. The dependency pattern is defined as relation between words of a dependency tree. Figure 1 shows an example. Both lexical and PoS patterns with frequencies larger than the threshold are selected.

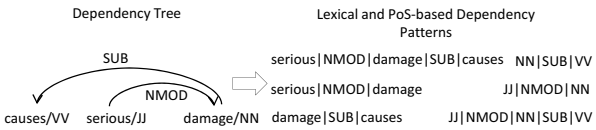

Figure 1: Example of dependency patterns

We also assign each extracted cue expression ce_i a weight calculated using the equation $weight_{ce_i} = c_1^{ce_i}/(c_1^{ce_i} + c_2^{ce_i})$, where, $c_1^{ce_i}$ and $c_2^{ce_i}$ denote its frequencies in positive and negative training Q&A pairs, respectively. The weights are used as values of features in SVM classifier.

The extracted cue expressions and collected Q&A pairs are used to build a question-type-specific classifier for each type of question, which is then used to remove noise sentences from answer candidates. For classifiers, we employ multivariate classification SVMs (Thorsten Joachims, 2005) that can directly optimize a large class of performance measures like F_1-Score, prec@k (precision of a classifier that predicts exactly k = 100 examples to be positive) and error-rate (percentage of errors in predictions).

5 Comparison Models

5.1 QSM

The QSM (question-specific method) first learns potential answer words to the question, and then re-ranks candidates by incorporating their "similarities" to the answer words. For each submitted question, the following four steps are performed. (1) An IR algorithm is used to retrieve the most similar Q&A pairs (top 50 in our experiments) to the question from the social Q&A collection. (2) All non-stop words from the retrieved Q&A pairs are weighted using a TFIDF score and the top M words are selected to form an answer profile Ap. (3) Answer candidates are re-ranked according to the similarity formula $sim(a_i) = \gamma sim(Q, a_i) + (1-\gamma)sim(a_i, Ap)$, where $sim(Q, a_i)$ denotes the similarity between question Q and candidates a_i, $sim(a_i, Ap)$ means the similarity between candidates and the answer profile Ap, γ is the weight. Both $sim(Q, a_i)$ and $sim(a_i, Ap)$ are estimated using cosine similarity in this paper. (4) Finally, the top N candidates are selected as answers to Q.

QSM is also widely used in answering definitional questions and TREC QA "other" questions (Kaisser et al., 2006; Chen, et al., 2006), which, however, learn answer words from the most relevant snippets returned by a Web search engine. Section 6 compares QSM based on 50 most relevant social Q&A pairs and that based on 50 most relevant snippets returned by Yahoo!.

5.2 MTM

The MTM learns word-to-word translation probability from all social Q&A pairs without consideration of the question and question type to improve complex QA system. The monolingual translation-based method treats Q&A pairs as the parallel corpus, with questions corresponding to the "source" language and answers to the "target" language. Monolingual translation models have recently been introduced to solve the lexical gap problem in IR and QA systems (Berger et al., 1999; Riezler et al., 2007; Xue, et al., 2008; Bernhard et al., 2009). A monolingual translation-based method for our complex QA system can be expressed by:

$$P(Q|a_i) = \prod_{w \in Q} ((1-\gamma)P_{mx}(w|a_i) + \gamma P_{ml}(w|C))$$
$$P_{mx}(w|a_i) = (1-\zeta)P_{ml}(w|a_i) + \zeta \sum_{t \in S} P(w|t)P_{ml}(t|a_i) \quad (1)$$

where Q is the question, a_i the candidate answer, γ the smoothing parameter for the whole Q&A collection, $P(w|t)$ the probability of translating an answer term t to a question term w, which is obtained by using the GIZA++ (Och and Ney, 2003),

the impact of the translation probabilities is controlled by ζ (=0.6 in this paper).

As in the common practice in translation-based retrieval, we utilize IBM model 1 for obtaining word-to-word probability $P(w|t)$ from 6.0 million social Q&A pairs. Preprocessing of the Q&A pairs only involves word segmentation (Wu et al., 2005) and stop word removal.

6 Experiments

As Section 4.1 shows, there exist hundreds of types of complex questions, it is hard to evaluate our approach on all of them. In this paper, question types contained in the NTCIR 2008 test set (Mitamura et al., 2008) are used. The NTCIR 2008 test data set contains 30 complex questions[5] that we discuss here. However, a small number of test questions are included for certain question types; e.g., it contains only one hazard-type, one scale-type, and three significance-type questions. To form a more complete test set, we create another 57 test questions to be released with this paper. The test data used in this paper therefore includes 87 questions and is called an extension of the NTCIR 2008 test data set. For each test question we also provide a list of weighted answer nuggets, which are used as the gold standard answers for evaluation. The evaluation is conducted by employing Pourpre v1.0c tool that uses the standard scoring methodology for TREC "other" questions (Voorhees, 2003). Each question is scored using nugget recall NR, nugget precision NP, and a combination score F_3 of NR and NP. Refer to (Lin and Demner-Fushman, 2006) for the detailed computation. The final score of a system run is the mean of the scores across all test questions.

6.1 Overall Results

Table 4 summarizes the evaluation results of the systems. The baseline refers to the conventional method in which the similarity is the same as $sim(Q, a_i)$ in section 5.1. QSM$_{web}$ and QSM$_{qa}$ indicate QSM that learns answer words from the Web and the social Q&A pairs, respectively. QTSM$_{prec}$ denotes QTSM based on the classifier optimizing performance prec@k.

This table indicates that the complex QA performance can be clearly improved by exploiting social Q&A collection. In particular, we observe that: 1) QTSM obtains the best performance; e.g., the F_3 improvements of QTSM$_{prec}$ over MTM and QSM$_{qa}$ in terms of N=10 are 5.8% and 6.0%, respectively. 2) QSM$_{qa}$ outperforms QSM$_{web}$ by 2.0% when N=10. Further analysis shows that the average number of the gold standard answer words learned in QSM$_{web}$ (42.9%) are fewer than that learned in QSM$_{qa}$ (58.1%). The reason may lie in: Q&A pairs are more complete and complementary than snippets that only contain length-limited contexts of question words. This proves that learning answer words from social Q&A pairs is superior to that from the snippets returned by a Web search engine. 3) The performance ranking of these models is: QTSM$_{prec}$ > {MTM, QSM$_{qa}$} > QSM$_{web}$. QSM$_{qa}$ depends on very specific knowledge, i.e. answer words to each question, which may fail when social Q&A collection does not contain similar Q&A pairs, or similar Q&A pairs do not contain answer words to the question. MTM learns very general knowledge from social Q&A collection, i.e., word-to-word translation probability, which is not apt to any question, any type of question, or any domain question. QTSM$_{prec}$, however, learns question-type-specific salient expressions, which granularity is between QSM$_{qa}$ and MTM. This may be the reason that QTSM$_{prec}$ achieves better performance.

Figure 2 displays how well QTSM$_{prec}$ performs for each type of question when N=10 for further comparison. This figure indicates that our method improves QSM$_{qa}$ on most types of test questions; e.g., the F_3 improvements on function-type and hazard-type questions are 20.0% and 14%, respectively. It is noted that QSM$_{qa}$ achieves better performance than QTSM$_{prec}$ on event-type questions. We interpret this to mean that the extracted salient cue expressions may not characterize answers to event-type questions. More complex features such as templates used in MUC-3 (MUC, 1991) may be needed. Figure 3 shows NR recall curves of the three models, which characterize the amount of relevant information contained within a fixed-length text segment (Lin, 2007). We observe that QTSM$_{prec}$ can greatly improve MTM and QSM$_{qa}$ at every answer length. For example, the improvement of QTSM$_{prec}$ over MTM is about 10.0% when the answer length is 400 words. Yet there is no distinct difference between MTM and QSM$_{qa}$.

[5]Because definitional, biography, and relationship questions in the NTCIR 2008 test set are not discussed here.

	F_3 (%)		NR (%)		NP (%)	
	$N = 5$	$N = 10$	$N = 5$	$N = 10$	$N = 5$	$N = 10$
Baseline	18.18	21.95	19.85	27.64	25.32	18.96
QSM$_{web}$	20.36	22.57	23.47	29.63	22.30	13.57
QSM$_{qa}$	21.28†	24.63†	24.60	33.49	22.99	15.47
MTM	20.47	24.76†	19.85	33.10	21.73	13.57
QTSM$_{prec}$	23.47$^{\flat}$	30.58$^{\flat}$	26.68	40.22	27.65	20.33

Table 4: Overall performance for the test data when outputting the top N sentences as answers. Significance tests are conducted on the F_3 scores. †: significantly better than Baseline at the $p = 0.1$ level using two-sided t-tests; $^{\flat}$: significantly better than QSM$_{qa}$ at the 0.005 level.

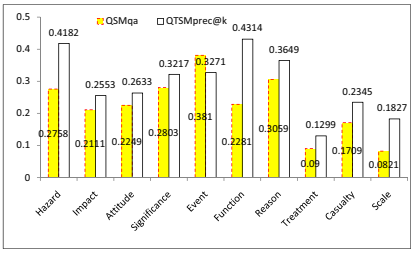

Figure 2: F_3 performance by type of question

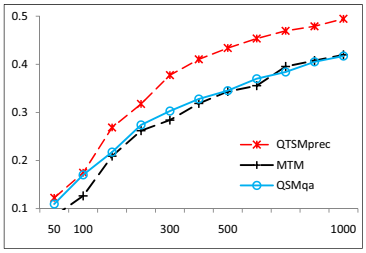

Figure 3: Recall over various answer lengths

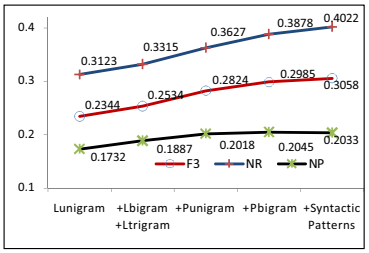

Figure 4: Impact of features on QTSM$_{prec}$

6.2 Impact of Features

To evaluate the contributions of individual features to the QTSM, this experiment gradually adds them. Figure 4 shows the performance of QTSM$_{prec}$ on different sets of features, L and P represent lexical and PoS-based n-gram cue expressions, respectively. This table demonstrates that all the lexical and PoS features can positively impact QTSM$_{prec}$. The contribution from dependency patterns is, however, not significant, which may be due to the limited number of dependency patterns learned.

6.3 Subjective evaluation

Pourpre v1.0c evaluation is based on n-gram overlap between the automatically produced answers and human-generated reference answers. Thus, it is not able to measure the conceptual equivalent. In subjective evaluation, the answer sentences returned by QA systems are labeled by two native Chinese assessors. Given a pair of answers for each question, the assessors are asked to determine which summary has better content for the question, or whether both are equally responsive. If their judgements are different, they will discuss a final judgement. This kind of evaluation is also used in (Biadsy et al., 2008; Liu et al., 2008).

Table 6 indicates that QTSM$_{prec}$ is much better than MTM and QSM$_{qa}$. For example, 56.3% of these judgements preferred the answers produced by QTSM$_{prec}$ over those produced by MTM. Table 5 compares the top 3 answers to question Q_1 answered by MTM and QTSM$_{prec}$.

QTSM$_{prec}$	Better	Equal	Worse
QTSM$_{prec}$ vs. MTM	56.3%	12.6%	31.1%
QTSM$_{prec}$ vs. QSM$_{qa}$	55.2%	13.8%	31.0%

Table 6: Results of subjective evaluation

7 Related Work

Some pioneering studies on social Q&A collection have recently been conducted. Much of the research aims at retrieving answers to queried questions from social Q&A collection. For example, Surdeanu et al. (2008) proposed an answer-ranking engine for non-factoid questions by incorporating textual features into a machine learning approach. Duan et al. (2008) retrieved questions semantically equivalent or close to the queried question for a question recommendation system. Agichtein et al. (2008) investigated techniques for finding high-quality content in social Q&A collection, and indicated that 94% of answers to questions have high quality. Xue, et al. (2008) proposed a retrieval model that combines a translation-based language model for the question part with a query likelihood approach for the

MTM	...非洲将是最易受全球气候变暖危害冲击的大陆.../...Africa will be most vulnerable to the impacts of global warming...
	...全球气候变暖将会对全球不同地区气候变化产生更为严重的影响。/...global warming will more seriously impact climate change in different regions of the world...
	...气候变暖对非洲造成的严重负面影响.../...global warming had a serious negative impact on Africa...
QTSM$_{prec}$	...全球气候变暖将给生态环境带来严重破坏, 致使自然灾害频繁发生, 而受其影响最为严重的当属非洲大陆。/...global warming will bring serious damage to the ecological environment, and result in frequent occurrences of natural disasters. There is no doubt that Africa is the most seriously impacted continent.
	全球气候变暖将会使非洲大陆的干旱地区, 特别是非洲中部和南部的干旱、半干旱地区更加缺水, 耕地退化和荒漠化现象越来越严重。/Global warming will cause more serious water shortages in the arid areas of the African continent, especially in central and southern arid and semi-arid areas. Land degradation and desertification will become increasingly serious.
	此外, 全球变暖还会导致极端气候现象频繁发生, 如寒潮、热浪、暴雨、龙卷风等, 对人类社会构成极大威胁。/Global warming will also lead to frequent extreme weather phenomena such as cold waves, heat waves, storms, and tornados, which poses a great threat to human beings.

Table 5: Top 3 answers to question, "What are the hazards of global warming?" returned by MTM and QTSM$_{prec}$

answer part. Wang (2010a) proposed an effective question retrieval in social Q&A collections.

Another category of study regards social Q&A collection as a kind of knowledge repository and aims to mine knowledge from it for generating answers to questions. To the best of our knowledge, there appears to be very limited work addressing this aspect. Mori et al. (2008) proposed a QSM method for improving complex Japanese QA systems, which collect Q&A pairs using 7-grams for which centers are interrogatives.

This paper is also related to query-based summarization of DUC (Dang, 2006; Harabagiu et al., 2006), which aims at synthesizing a fluent, well-organized 250-word summary for a given topic description and a collection of relevant documents generated manually. The topic descriptions usually consist of several complex questions such as "*Describe theories on the causes and effects of global warming and arguments against them.*" Thus, many approaches such as LexRank (Erkan and Radev, 2004) focus on compressing the relevant documents. We implement a LexRank method for our task, for which performance is even worse than the baseline. Our observation is that a query-based summarization task is given a set of manually generated relevant documents, but our QA systems need to retrieve relevant documents automatically, and there exist a great deal of noise.

8 Conclusion

This paper investigated techniques for mining knowledge from social Q&A websites for improving a sentence-based complex QA system. The proposed QTSM (question-type-specific method) explored social Q&A collection to automatically learn question-type-specific training Q&A pairs and cue expressions, and create a question-type-specific classifier for each type of question to filter out noise sentences before answer selection. Experiments on the extension of NTCIR 2008 test questions indicate that QTSM is more effective than QSM (question-specific method) and MTM (monolingual translation-based method) methods; e.g., the largest improvements in F_3 over QSM and MTM reaches 6.0% and 5.8%, respectively.

In the future, we will endeavor to: (1) reduce noise in the training Q&A pairs, and design more characteristic cue expressions to various types of questions such as event-templates for event-type question (MUC, 1991); (2) adapt QTSM to summarize answers in social QA sites (Liu et al., 2008); (3) learn paraphrases to recognize types of questions that do not contain question focuses such as "What causes global warming?"; (4) adapt the QA system to a topic-based summarization system, which will, for instance, summarize accidents according to "casualty" and"reason", and events according to "reason", "measure" and "impact".

References

Abdessamad Echihabi and Daniel Marcu. 2003. A Noisy-Channel Approach to Question Answering. In *Proc. of ACL 2003*, Japan.

Adam Berger and John Lafferty. 1999. Information Retrieval as Statistical Translation. In *Proc. of SIGIR 1999*, pp222-229.

Delphine Bernhard and Iryna Gurevych. 2009. Combining Lexical Semantic Resources with Question & Answer Archives for Translation-based Answer Finding. In *Proc. of ACL-IJCNLP 2009*, pp728-736.

Ellen M. Voorhees. 2003 Overview of the TREC 2003 Question Answering Track. In *Proc. of TREC 2003*, pp54-68, USA.

Eugene Agichtein, Carlos Castillo, Debora Donato. 2008 Finding High-Quality Content in Social Media. In *Proc. of WSDM 2008*, California, USA.

Fadi Biadsy, Julia Hirschberg, Elena Filatova. 2008. An Unsupervised Approach to Biography Production using Wikipedia. In *Proc. of ACL2008*.

Franz J. Och and Hermann Ney. 2003. A systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, 29(1):19-51.

Gunes Erkan and Dragomir Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text. In *Journal of Artificial Intelligence Research*,22:457-479.

Hang Cui, Min Yen Kan, and Tat Seng Chua. 2004. Unsupervised Learning of Soft Patterns for Definition Question Answering. In *Proc. of WWW 2004*.

Hoa Trang Dang. 2006. Overview of DUC 2006. In *Proc. of TREC 2006*.

Huizhong Duan, Yunbo Cao, Chin Yew Lin, and et al. 2008. Searching Questions by Identifying Question Topic and Question Focus. In *Proc. of ACL 2008*.

Jimmy Lin. 2007. Is Question Answering Better Than Information Retrieval? Toward a Task-based Evaluation Framework for Question Series. In *Proc. of HLT/NAACL2007*, pp 212-219.

Jimmy Lin and Dina Demner-Fushman. 2006. Will Pyramids Built of Nuggets Topple Over. In *Proc. of HLT/NAACL2006*, pp 383-390.

Kai Wang, Zhao-Yan Ming, Xia Hu, Tat-Seng Chua. 2010a. Segmentation of Multi-Sentence Questions: Towards Effective Question Retrieval in cQA Services. In *Proc. of SIGIR 2010*, pp 387-394.

Michael Kaisser, Silke Scheible, and Bonnie Webber. 2006. Experiments at the University of Edinburgh for the TREC 2006 QA track. In *Proc. of TREC2006*.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In *Proc. of ACL 2008*.

Ryuichiro Higashinaka and Hideki Isozaki. 2008. Corpus-based Question Answering for why-Questions. In *Proc. of IJCNLP 2008*, pp 418-425.

Sanda Harabagiu, Dan Moldovan, Christine Clark, Mitchell Bowden, Andrew Hickl, 2005. Employing Two Question Answering Systems in TREC-2005. In *Proc. of TREC2005*.

Sanda Harabagiu, Finley Lacatusu, Andrew Hickl. 2006. Answering Complex Questions with Random Walk Models. In *Proc. of SIGIR 2006*, pp 220-227.

Stefan Riezler, Er Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, Yi Liu. 2007. Statistical Machine Translation for Query Expansion in Answer Retrieval. In *Proc. of the ACL-2007*, pp 464-471.

Tatsunori Mori, Takuya Okubo, and Madoka Ishioroshi. 2008. A QA system that can answer any class of Japanese non-factoid questions and its application to CCLQA EN-JA task. In *Proc. of NTCIR2008*, Tokyo, pp 41-48.

Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, and et al. 2008. Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access. In *Proc. of NTCIR 2008*.

Thorsten Joachims. 2005. A Support Vector Method for Multivariate Performance Measures. In *Proc. of ICML2005*, pp 383-390.

Ves Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-Perspective Question Answering Using the OpQA Corpus. In *Proc. of HLT/EMNLP 2005*.

Vijay Krishnan, Sujatha Das and Soumen Chakrabarti. 2005. Enhanced Answer Type Inference from Questions using Sequential Models. In *Proc. of EMNLP 2005*, pp 315-322.

Vladimir Vapnik 1998. Statistical learning theory. John Wiley.

Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Improving Dependency Parsing with Subtrees from Auto-Parsed Data. In *Proc. of EMNLP 2009*, pp 570-579.

Xiaobing Xue, Jiwoon Jeon, and W.Bruce Croft. 2008. Retrieval Models for Question and Answer Archives. In *Proc. of SIGIR 2008*, pp 475-482.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proc. of COLING 2002*, pp 556-562.

Yi Chen, Ming Zhou, and Shilong Wang. 2006. Reranking Answers for Definitional QA Using Language Modeling. In Proc. of ACL/COLING2006.

Youzheng Wu, Jun Zhao, Bo Xu, and Hao Yu. 2005. Chinese Named Entity Recognition Model based on Multiple Features. In *Proc. of HLT/EMNLP 2005*.

Yuanjie Liu, Shasha Li, Yunbo Cao, and et al. 2008. Understanding and Summarizing Answers in Community-Based Question Answering Services. In *Proc. of COLING 2008*.

Yutaka Sasaki. 2005. Question Answering as Question-biased Term Extraction: A New Approach toward Multilingual QA. In *Proc. of ACL 2005*.

Dong Zhendong, Dong Qiang. 1999. Hownet. `http://www.keenage.com`.

Message Understanding Conference (MUC) `http://www-nlpir.nist.gov/related_projects/muc/index.html`.

TAC (Text Analysis Conference) 2008 Question Answering Track. `http://www.nist.gov/tac/2008/qa/`.