Exploiting Social Q&A Collection in Answering Complex Questions

Youzheng Wu **Hisashi Kawai**
Spoken Language Communication Group, MASTAR Project
National Institute of Information and Communications Technology
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
{youzheng.wu, hisashi.kawai}@nict.go.jp

Abstract

This paper investigates techniques to automatically construct training data from social Q&A collections such as Yahoo! Answer to support a machine learning-based complex QA system[1]. We extract cue expressions for each type of question from collected training data and build question-type-specific classifiers to improve complex QA system. Experiments on 10 types of complex Chinese questions verify that it is effective to mine knowledge from social Q&A collections for answering complex questions, for instance, the F_3 improvement of our system over the baseline and translation-based model reaches 7.9% and 5.1%, 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 to answer 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, answer), short for Q&A, pairs as training data. For example, (Echihabi et al., 2003) and (Sasaki, 2005) constructed 90,000 English Q&A pairs and 2,000 Japanese Q&A pairs, respectively for their factoid QA systems. (Cui et al., 2004) constructed 76 term-definition pairs for their definitional QA systems. (Stoyanov et al., 2005) required a known subjective vocabulary for their opinion QA. (Higashinaka and Isozaki, 2008) used 4,849 positive and 521,177 negative examples in their reason QA system. Among complex QA systems, many other types of questions have not been well studied, apart from reason and definitional questions. Appendix A lists 10 types of complex Chinese questions and their examples we discussed in this paper.

According to the related studies on QA, supervised machine-learning technique may be effective for answering these questions. To employ the supervised approach, we need to reconstruct training Q&A pairs for each type of question, though this is an extremely expensive and labor-intensive task. To deal with the acquisition problem of training Q&A pairs, we investigate techniques to automatically construct training data by utilizing social Q&A collections crawled from the Web, which contains millions of user-generated Q&A pairs. Many studies (Surdeanu et al., 2008) (Duan et al., 2008) have been done on retrieving similar Q&A pairs from social QA websites as answers to test questions. Our study, however, regards social Q&A websites as a knowledge repository and aims to mine knowledge from them for synthesizing answers to questions from multiple documents. There is very little literature on this aspect. Our work can be seen as a kind of query-based summarization (Dang, 2006) (Harabagiu et al., 2006) (Erkan and Radev, 2004), and can also be employed to answer questions that have not been answered in social Q&A websites.

[1]Complex questions cannot be answered by simply extracting named entities. In this paper complex questions do not include definitional questions.

This paper mainly focuses on the following three steps: (1) automatically constructing question - type-specific training Q&A pairs from the social Q&A collection; (2) extracting cue expressions for each type of question from the collected training data, and (3) building question-type-specific classifiers to filer out noise sentences before using a state-of-the-art IR formula to select answers.

We evaluate our system on 10 types of Chinese questions by using the Pourpre evaluation tool (Lin and Demner-Fushman, 2006). The experimental results show the effectiveness of our system, for instance, the F_3/NR improvement of our system over the baseline and translation-based model reaches 7.9%/11.1%, and 5.1%/5.6%, respectively.

2 Social Q&A Collection

Recently launched social QA websites such as Yahoo! Answer[2] 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. The number of Q&A pairs on such sites has risen dramatically. These pairs could collectively form a source of training data that is required in supervised machine-learning-based QA systems.

In this paper we aim to explore such user-generated Q&A collections to automatically collect Q&A training data. However, social collections 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. Thus, we only crawl questions and their best answers to form Q&A pairs, wherein the best answers are longer than the empirical threshold. Finally, 60.0 million Q&A pairs were crawled from Chinese social QA websites. These pairs will be used as the source of training data required in our study.

[2] http://answers.yahoo.com/
[3] http://zhidao.baidu.com/

3 Our Complex QA System

The typical complex QA system architecture is a cascade of three modules. The Question Analyzer analyzes test questions and identifies answer types of questions. 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 the documents into sentences as answer candidates. The Answer Extraction module applies state-of-the-art IR formulas (e.g., KL-divergence language model) to directly estimate similarities between sentences (1,024 sentences were used in our case) and questions, and selects the most similar sentences as the final answers. Given three answer candidates, s_1 = "*Solutions to global warming range from changing a light bulb to engineering giant reflectors in space ...*", s_2 = "*Global warming will bring bigger storms and hurricanes that will hold more water ...*", and s_3 = "*nuclear power is the relatively low emission of carbon dioxide (CO_2), one of the major causes of global warming,*" to the question of "What are the hazards of global warming?", however, it is hard for this architecture to select the correct answer, s_2, because the three candidates contain the same question words "global warming".

According to our observation, answers to a type of question usually contain some type-of-question dependent cue expressions ("will bring" in this case). This paper argues that the above QA system can be improved by using such question-type-specific cue expressions. For each test question, we perform the following three steps. (1) Collecting question-type-specific Q&A pairs from the social Q&A collection which question types are same as the test question to form positive training data. Similarly, negative Q&A pairs are also collected which question types are different from the test question. (2) Extracting and weighting question-type-specific cue expressions from the collected Q&A pairs. (3) Building a question-type-specific classifier by employing the cue expressions and the collected Q&A pairs, which re-

moves noise sentences from answer candidates before using the Answer Extraction module.

3.1 Collecting Q&A Pairs

We first introduce the notion of the *answer type informer* of the question as follows. In a question, a short subsequence of tokens (typically 1-3 words) that are adequate for question classification is considered an answer-type informer, e.g., "hazard" in the question of "What are the hazards of global warming?" This paper makes the following assumption: type of complex question is determined by its answer type informer. For example, the question of "What are the hazards of global warming?" belongs to hazard-type question, because its answer type informer is "hazard". Therefore, the task of recognizing question-types is shifted to identifying *answer type informer* of question.

In this paper, we regard answer-type informer recognition as a sequence tagging problem and adopt conditional random fields (CRFs) because many work has shown that CRFs have a consistent advantage in sequence tagging. We manually label 3,262 questions with answer-type informers to train a CRF, which classifies each question word into a set of tags $O = \{I_B, I_I, I_O\}$: I_B for a word that begins an informer, I_I for a word that occurs in the middle of an informer, and I_O for a word that is outside of an informer. In the following feature templates used in the CRF model, w_n and t_n, refer to word and PoS, respectively; n refers to the relative position from the current word n=0. The feature templates include the following 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=$-1, 0$.

The trained CRF model is then employed to recognize answer-type informers from questions of social Q&A pairs. Finally, we recognized 103 answer-type informers in which frequencies are larger than 10,000. Moreover, the numbers of answer type informers for which frequencies are larger than 100, 1,000, and 5,000 are 2,714, 807, and 194, respectively.

Based on answer-type informers of questions recognized, we can collect training data for each type of question as follows: (1) Q&A pairs are grouped together in cases in which the answer-type informers X of their questions are the same, and (2) Q&A pairs clustered by informers X are regarded as the positive training data of X-type questions. For instance, 10,362 Q&A pairs grouped via informer X (="hazard") are regarded as positive training data of answering hazard-type questions. Table 1 lists some questions, which, together with their best answers, are employed as the training data of the corresponding type of questions. For each type of question, we also randomly select some Q&A pairs that do not contain informers in questions as negative training data. Preprocessing of the training data, including word segmentation, PoS tagging, and named entity (NE) tagging (Wu et al., 2005), is conducted. We also replace each NE with its tag type.

Qtype	Questions of Q&A pairs
Hazard-type	What are the **hazards** of the trojan.psw.misc.kah virus? What are the **hazards** of RMB appreciation on China's economy? **Hazards** of smoke What are the **hazards** of contact lenses? What are the **hazards** of waste accumulation?
Casualty-type	What were the **casualties** on either side from the U.S.-Iraq war? What were the **casualties** of the Sino-French War? What were the **casualties** of the Sichuan earthquake in 2008? What were the **casualties** of highway accidents over the years? What were the **casualties** of the Ryukyu Islands tsunami?
Reason-type	What are the main **reasons** of China's water shortage? What are the **reasons** of asthma? What are the **reasons** of blurred photos? What are the **reasons** of air pollution? The **reasons** for the soaring prices!

Table 1: Questions (translated from Chinese) of social Q&A pairs (words in bold denote answer-type informers of questions). These questions and their best answers are regarded as positive training data for hazard-type question.

3.2 Cue Expressions

We extract lexical and PoS-based n-grams as cue expressions from the collected training data. To reduce the dimensionality of the cue expression space, we first select the top 3,000 lexical unigrams using the formula: $score_w = tf_w \times log(idf_w)$, where $tf(w)$ denotes the frequency of word w, and $idf(w)$ represents the inverted document frequency of w that indicates its global importance. Table 2 shows some of the learned unigrams. The top 300 unigrams are then used as seeds to learn lexical bigrams and trigrams iteratively. Only lexical bigrams and trigrams that contain seed unigrams with frequencies larger than the thresholds are retained as lexical features. Moreover, we extract PoS-based unigrams and bigrams as cue expressions.

Further, we assign each extracted feature s_i a weight calculated using the equation $weight_{s_i} = c_1^{s_i}/(c_1^{s_i} + c_2^{s_i})$, where, $c_1^{s_i}$ and $c_2^{s_i}$ denote its frequencies in positive and negative training Q&A pairs, respectively.

Qtype	Top Unigrams
Hazard-type	危害/hazard 导致/lead to 造成/cause 引起/give rise to 产生/bring about 影响/influence 损害/damage
Casualty-type	伤亡/casualty 死亡/death 受伤/hurt 失踪/missing 遇难/wrecked 阵亡/die in battle 负伤/wounded

Table 2: Top unigrams learned from hazard-type and casualty-type Q&A pairs

3.3 Classifiers

As mentioned above, we use the extracted cue expressions and the collected Q&A pairs to build question-type-specific classifiers, which is 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). Instead of learning a univariate rule that predicts the label of a single example in conventional SVMs (Vapnik, 1998), multivariate SVMs formulate the learning problem as a multivariate prediction of all examples in the data set. Considering hypotheses $\overline{h}$ that map a tuple $\overline{\mathbf{x}}$ of n feature vectors $\overline{\mathbf{x}} = (\mathbf{x_1}, ..., \mathbf{x_n})$ to a tuple $\overline{y}$ of n labels $\overline{y} = (y_1, ..., y_n)$, multivariate SVMs learn a classifier

$$\overline{h}_{\mathbf{w}}(\overline{\mathbf{x}}) = argmax_{\overline{y}' \in \overline{Y}}\{\mathbf{w}^T \Psi(\overline{x}, \overline{y}')\} \quad (1)$$

by solving the following optimization problem.

$$min_{\mathbf{w},\xi \geq 0} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \quad (2)$$

$$s.t.: \ \forall \overline{y}' \in \overline{Y} \backslash \overline{y} : \mathbf{w}^T[\Psi(\overline{x}, \overline{y}) - \Psi(\overline{x}, \overline{y}')] \geq \Delta(\overline{y}', \overline{y}) - \xi \quad (3)$$

where, $\mathbf{w}$ is a parameter vector, Ψ is a function that returns a feature vector describing the match between $(\mathbf{x_1}, ..., \mathbf{x_n})$ and $(y'_1, ..., y'_n)$, Δ denotes types of multivariate loss functions, and ξ is a slack variable.

4 Experiments

The NTCIR 2008 test data set (Mitamura et al., 2008) contains 30 complex questions[4] we discussed here. However, a small number of test questions are included for some question types, e.g.; it contains only 1 hazard-type, 1 scale-type, and 3 significance-type questions. To form a more complete test set, we create another 65 test questions[5]. Therefore, the test data used in this paper includes 95 complex questions.

For each test question we also provide a list of weighted nuggets, which are used as the gold standard answers for evaluation. The evaluation is conducted by employing Pourpre v1.0c (Lin and Demner-Fushman, 2006), which uses the standard scoring methodology for TREC other questions (Voorhees, 2003), i.e., answer nugget recall NR, nugget precision NP, and a combination score F_3 of NR and NP. For better understanding, we evaluate the systems when outputting the top N sentences as answers.

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

[5]The approach of creating test data is same as that in the NTCIR 2008.

	F_3 (%)			NR (%)			NP (%)		
	N = 1	N = 5	N = 10	N = 1	N = 5	N = 10	N = 1	N = 5	N = 10
Baseline	9.82	18.18	21.95	9.44	19.85	27.64	34.35	25.32	18.96
TransM	9.76	20.47	24.76	9.44	19.85	33.10	31.96	21.73	13.57
Ours$_{lin}$	10.92	22.61	25.74	10.49	25.95	34.70	34.98	23.40	15.11
Ours$_{errorrate}$	12.37	23.10	27.74	12.05	26.98	37.03	33.22	26.48	18.67
Ours$_{pre@k}$	8.96	22.85	29.85	8.72	25.67	38.78	26.28	28.82	20.45

Table 3: Overall performance for the test data

4.1 Overall Results

Table 3 summarizes the evaluation results for several N values. The baseline refers to the conventional method introduced in Section 3, which does not employ question-type-specific classifiers before the Answer Extraction. The baseline can be expressed by the formula:

$$sim(q,s) = \frac{\langle V_q \cdot V_s \rangle}{\|V_q\| \times \|V_s\|} \quad (4)$$

where, V_q and V_s are the vectors of the question and candidate answer. The TransM denotes a translation model for QA (Xue, et al., 2008) (Bernhard et al., 2009), which uses Q&A pairs as the parallel corpus, with questions to the "source" language and answers corresponding to the "target" language. This model can be expressed by:

$$P(q|S) = \prod_{w \in q}((1-\gamma)P_{mx}(w|S) + \gamma P_{ml}(w|C))$$
$$P_{mx}(w|S) = (1-\zeta)P_{ml}(w|S) + \zeta \sum_{t \in S} P(w|t)P_{ml}(t|S) \quad (5)$$

where, q is the question, S the sentence, $P(w|t)$ the probability of translating a sentence term t to the question term w, which is obtained by using the GIZA++ toolkit (Och and Ney, 2003). We use six million Q&A pairs to train IBM model 1 for obtaining word-to-word probability $P(w|t)$. Ours$_{errorrate}$ and Ours$_{pre@k}$ denote our models that are based on classifiers optimizing performance measure error-rate and prec@k, respectively. Ours$_{lin}$, a linear interpolation model, that combines scores of classifiers and the baseline, which is similar to (Mori et al., 2008) and can be expressed by the equation:

$$sim(q,s)' = sim(q,s) + \alpha \times \phi(s) \quad (6)$$

where, $\phi(s)$ is the score calculated by classifiers (Thorsten Joachims, 2005) and α denotes the weight of the score.

This experiment shows that: (1) Question-type-specific classifiers can greatly outperform the baseline; for example, the F_3 improvements of Ours$_{errorrate}$ and Ours$_{pre@k}$ over the baseline in terms of N=10 are 5.8% and 7.9%, respectively. (2) Ours$_{errorrate}$ is better than Ours$_{pre@k}$ when $N < 10$. The average numbers of sentences retained in Ours$_{errorrate}$ and Ours$_{pre@k}$ are 130, and 217, respectively. That means the precision of the classifier optimizing errorrate is superior to the classifier optimizing prec@k, while the recall is relatively inferior. (3) Ours$_{lin}$ is worse than Ours$_{errorrate}$ and Ours$_{pre@k}$, which indicates that using question-type-specific classifiers by classification is better than using it by interpolation like (Mori et al., 2008). (4) Our models also outperform TransM, e.g.; the F_3 improvement is 5.1% when N is set to 10. TransM exploits the social Q&A collection without consideration of question types, while our models select and exploit the social Q&A pairs of the same question types. Thereby, this experiment also indicates that it is better to exploit social Q&A pairs by type of question. The performance ranking of these models when N=10 is: Ours$_{prec@k}$ > Ours$_{errorrate}$ > Ours$_{lin}$ > TransM > Baseline.

4.2 Impact of Features

In order to evaluate the contributions of individual features to our models, this experiment

is conducted by gradually adding them. Table 4 summarizes the performance of Our$_{prec@k}$ on different set of features, L and P represent lexical and PoS-based features, respectively. This table demonstrates that all the lexical and PoS features can positively impact Our$_{prec@k}$, especially, the contribution of the PoS-based features is largest.

Features	F_3	NR	NP
Lunigram	23.44	31.23	17.32
+Lbigram +Ltrigram	25.34	33.15	18.87
+Punigram	28.24	36.27	20.18
+Pbigram	29.85	38.78	20.45

Table 4: Impact of features on Our$_{prec@k}$.

4.3 Improvement

As discussed in Section 2, the writing style of social Q&A collections slightly differs from that of our complex QA system, which is an unfavorable circumstance in utilizing social Q&A collections. For better understanding we randomly select 100 Q&A training pairs of each type of question acquired in Section 3, and manually classify each Q&A pair into `NON-NOISE` and `NOISE`[6] categories. Figure 1 reports the percentage of `NON-NOISE`. This figure indicates that 71% of the training pairs of the scale-type questions are noises, which may lead to a small improvement.

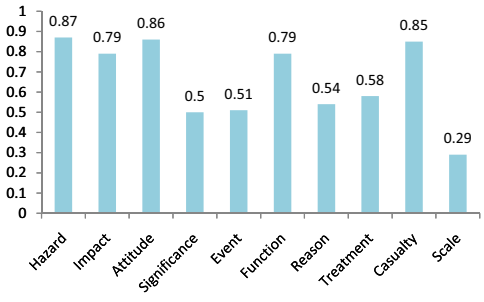

Figure 1: Percentage of `NON-NOISE` pairs by type of questions.

To further improve the performance, we employ k-fold cross validation to remove noises from the collected training data in Section 3.1. Specifically, the collected training data are first divided into k (= 5) sets. Secondly, k-1 sets are used to train classifiers that are applied to classify the Q&A pairs in the remaining set. Finally, part of the Q&A pairs classified as negative pairs are removed[7]. According to Figure 1, we remove 20% of the training data from the negative pairs for the hazard-type, impact-type, and function-type questions, and 40% of the training data for significance-type, event-type, and reason-type questions. Because the sizes of the training pairs of the other four types of questions are small, we do not use this approach on them. Table 5 shows the results of Ours$_{pre@k}$ on the above six types of questions. The numbers in brackets indicate absolute improvements over the system based on the data without removing noises. *N* is the number of answer sentences to a question. The experiment shows that the performance is generally improved by removing noise in the training Q&A pairs using k-fold cross-validation.

	F_3 (%)	NR (%)	NP (%)
N = 1	$9.6_{+2.1}$	$9.3_{+2.0}$	$30.8_{+7.4}$
N = 5	$21.6_{+0.7}$	$24.9_{+1.2}$	$26.0_{-1.3}$
N = 10	$28.6_{+0.9}$	$37.9_{+1.7}$	$19.2_{-0.2}$

Table 5: Performance of Ours$_{pre@k}$ after removing noises in the training Q&A pairs.

4.4 Subjective evaluation

Pourpre v1.0c evaluation is based on n-gram overlap between the automatically produced answers and the human generated reference answers. Thus, it is not able to measure conceptual equivalent. In subjective evaluation, the answer sentences returned by systems are labeled by a native Chinese assessor. Figure 2 shows the distribution of the ranks of the first correct answers for all questions. This figure demonstrates that the Ours$_{pre@k}$ answers 57 questions which

[6] `NOISE` means that the Q&A pair is not useful in our study.

[7] We do not remove all negative Q&A pairs to ensure the coverage of training data because the classifiers have relatively lower recall, as mentioned in Section 3.3.

first answers are ranked in top 3, which is larger than that of the baseline, i.e., 49. Moreover, the Ours$_{pre@k}$ contains only 11.5% of questions which answers are ranked after top 10, while this number of the baseline is 20.7%.

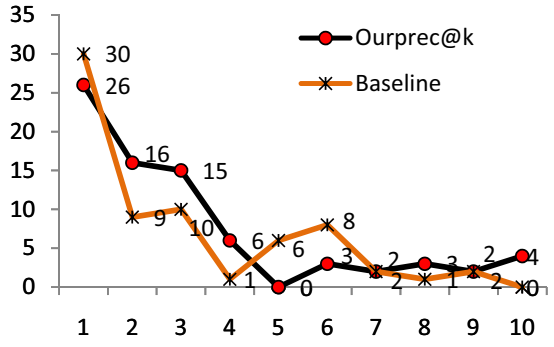

Figure 2: Distribution of the ranks of first answers.

5 Related Work

Recently, some pioneering studies on the social Q&A collection have been conducted. Among them, much of the research aims to retrieve answers to queried questions from the 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) proposed searching questions semantically equivalent or close to the queried question for a question recommendation system. (Agichtein et al., 2008) investigated techniques of finding high-quality content in the social Q&A collection, and indicated that 94% of answers to questions with high quality 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 answer part.

Another category of study regards the 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 is very limited work reported on this aspect. This paper is similar to (Mori et al., 2008), but different from it as follows. (1) (Mori et al., 2008) collects training data for each test question using 7-grams for which centers are interrogatives, while this paper collects training data for each type of question using answer type informers. (2) About the knowledge learned, we extract lexical/class-based, PoS-based unigrams, bigrams, and trigrams. (Mori et al., 2008) only extracts lexical bigrams. (3) They incorporated knowledge learned by interpolating with the baseline. However, we utilize the learned knowledge to train a binary classifier, which can remove noise sentences before answer selection.

6 Conclusion

This paper investigated a technique for mining knowledge from social Q&A websites for improving a sentence-based complex QA system. More specifically, it explored a social Q&A collection to automatically construct training data, and created question-type-specific classifier for each type of question to filter out noise sentences before answer selection.

The experiments on 10 types of complex Chinese questions show that the proposed approach is effective; e.g., the improvement in F_3 reaches 7.9%. In the future, we will endeavor to reduce `NOISE` pairs in the training data, and to extract type-of-question dependent features. Future research tasks also include adapting the QA system to a topic-based summarization system, which, for example, summarizes accidents according to "casualty", "reason", and summarizes events according to "reason", "measure," "impact", etc.

Appendix A. Examples of 10 Types of Questions.

References

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

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*, Singapore, pp728-736.

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

Qtype	Examples
危 害/Hazard-type	全球气候变暖的**危害**是什么？What are the **hazards** of global warming?
作用/Function-type	联合国的**作用**是什么？What are the **functions** of the United Nations?
影 响/Impact-type	列举911事件对美国的**影响**。List the **impact** of the 911 attacks on the United States.
意义/ Significance-type	列举中国加入WTO的**意义**。 List the **significance** of China's accession to the WTO.
态 度/Attitude-type	列举各国对巴以冲突的**态度**。List the **attitudes** of other countries toward the Israeli-Palestinian conflict.
措施/Measure-type	日本在节能减排方面采取了哪些**措施**？What **measures** have been taken for energy-saving and emissions-reduction in Japan?
原 因/Reason-type	全球气候变暖的**原因**是什么？What are the **reasons** for global warming?
伤亡/Casualty-type	列举洛克比空难的**伤亡**。List the **casualties** of the Lockerbie Air Disaster.
事 件/Event-type	列举北爱尔兰和平和平谈判**事件**。List the **events** in the Northern Ireland peace process.
规 模/Scale-type	介绍一下昆明世界园艺博览会的**规模**。Give information about the **scale** of the Kunming World Horticulture Exposition.

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

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 Yong Yu. 2008. Searching Questions by Identifying Question Topic and Question Focus. In *Proc. of ACL 2008*, Canada, pp 156-164.

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

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In *Proc. of ACL 2008*, Ohio, USA, pp 719-727.

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

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.

Sanda Harabagiu, Finley Lacatusu, Andrew Hickl. 2006. Answering Complex Questions with Random Walk Models. In *Proc. of the 29th SIGIR*, pp 220-227, ACM.

Ves Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-Perspective Question Answering Using the OpQA Corpus. In *Proc. of HLT/EMNLP 2005*, Canada, pp 923-930.

Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, Tetsuya Sakai, Donghong Ji and Noriko Kando. 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.

Vladimir Vapnik 1998. Statistical learning theory. John Wiley.

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

Yutaka Sasaki. 2005. Question Answering as Question-biased Term Extraction: A New Approach toward Multilingual QA. In *Proc. of ACL 2005*, pp 215-222.

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*, Canada, pp 427-434.

Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, Yong Yu. 2008. Understanding and Summarizing Answers in Community-Based Question Answering Services. In *Proc. of COLING 2008*, Manchester, pp 497-504.