An Open-domain Question Answering System for NTCIR-8 C-C Task

Guiping Zhang, Wubei Zhang, Yu Bai, Shengqiao Kang, Peiyan Wang
Knowledge Engineering Research Center, Shenyang Institute of Aeronautical Engineering
Email:nlpxiaobai@yahoo.com

Abstract

In this paper, we described our CCLQA system and the evaluation results for the C-C task at NTCIR-8 ACLIA. The system consists of a Question Analysis module, IR module and Answer Extraction module. The Question Analysis module was developed for NTCIR-7 CCLQA, which is based on the Question pattern library and HowNet. The IR module was developed for NTCIR-8 IR4QA task, and the results of KECIR-CS-CS-01-T were used as question related documents. For answer extraction, a Surface-Based Multi-Strategy approach was used. It deals retrieval results with different Strategies. The evaluation results show that our system achieves 0.3450 average F-score (beta = 3).

Keywords: Question Answering, Surface-Based, Question Analysis, Answer Extraction

1. Introduction

Current research in QA is moving beyond factoid questions, so there is significant motivation to evaluate more complex questions in order to move the research forward.[1] We participate in simplified Chinese-Chinese(C-C) QA subtask in NTCIR-8, whose goal is to develop effective CCLQA(Complex Cross-lingual Question Answering) evaluations for complex questions as well as factoid questions.

According to the requirement of the subtask, we implement a surface-based QA system that can handle questions with different types. We focused on the processing of five types of complex questions: DEFINITION, BIOGRAPHY, RELATIONSHIP, EVENT, and WHY; four types of factoid questions: PERSON, ORGANIZATION, LOCATION, and DATE.

In this paper, we described our CCLQA system and the evaluation results for the C-C task at NTCIR-8 ACLIA. The system consists of a Question Analysis module, IR module and Answer Extraction module. The Question Analysis module was developed for NTCIR-7 CCLQA, which is based on the Question pattern library and HowNet. The IR module was developed for NTCIR-8 IR4QA task, and the results of KECIR-CS-CS-01-T were used as question related documents. For answer extraction, a Surface-Based Multi-Strategy approach was used. It deals retrieval results with different Strategies. The evaluation results show that our system achieves 0.3450 average F-score (beta = 3).

The rest of this paper is organized as follows: In section 2, we present related studies. Section 3 provides an overview of the system architecture. In section 4 and 5, we introduce the question analysis methods and answer extraction strategies respectively. Section 6 gives the evaluation results and error analysis. Finally, the conclusion is given in section 7.

2. Related Studies

2.1 Question Pattern Built

We focused on the evaluation of five types of complex questions and four types of factoid questions. Examples are shown in Table 1 below.

Table 1. Example of Questions

Type	#	Example of Question
DEFINITION	10	What is the Human Genome Project?
BIOGRAPHY	10	Who is Howard Dean?
RELATIONSHIP	20	What is the relationship between Saddam Hussein and Jacques Chirac?
EVENT	20	What are the major conflicts between India and China on border issues?
WHY	20	Why doesn't U.S. ratify the Kyoto Protocol?
PERSON	5	Who is the Finland's first woman president?
ORGANIZATION	5	What is the name of the company that produced the first Fair-trade coffee?
LOCATION	5	What is the name of the river that separates North Korea from China?
DATE	5	When did Queen Victoria die?

The Question pattern library is used to the Question Analysis module in our QA system, which is built manually according to the different question types.

To build the question pattern, the Xinhua corpus (1998-2001) and evaluation questions of NTCIR-7 were used. For a question, each named entity was replaced by the named entity type correspondingly. The Table 2 gives some sample of the question pattern library. For details of building patterns, see [2].

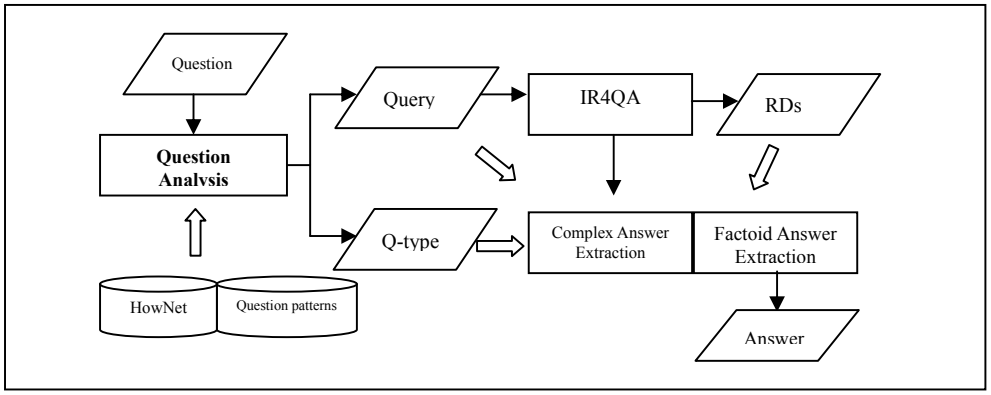

Figure 1. C-C QA architectural module

Table 2. Examples of Chinese question pattern with Query type

Q-type	Question patterns
Eve.	[列举/举出/罗列/历数] <EVENT>[有/是]［哪些/什么]
Bio.	［介绍/简介］谁是<PERSON>; ［介绍/简介］<PERSON>是谁
Rel.	［介绍/说明/描述］<Entity1> 和<Entity2> [有/是][什么/哪些][关系/联系/区别]
Def.	［什么是/何谓]< term >; < term >的[定义/解释/含义]是什么
Why	[是]为什么；什么原因；原因是什么
Pers.	是谁; 谁是; 哪个<def:=Person>
Loc.	[在/位于]哪[里]；哪个 <def:=Location>
Date	什么时候；何时；哪[个]<def:=time>
Org.	哪[个/家/支]<def:=org>

2.2 Name Entity Recognition

The NE recognition is one of the most important parts for answering factoid questions. We employ a NE recognition system based on conditional random fields [3], and its F-score in close test is 0.8801. In order to improve the effect of NE recognition, a classified dictionary of entities was built mainly by using the open database of the Chinese Wikipedia, and about 300 thousand entities were content in it.

2.3 Wiki-based Abbreviation Expansion

We can get some keywords in question analysis step, but they only appear in the question, and it's possible for them to appear in corpus in another way, especially for abbreviations. Thereby, for C-C task we must find out the synonyms of those abbreviations in Chinese, so as to improve recall of our system. The online-Wiki (http://baike.baidu.com) was used as abbreviation expanding sources.

For example, the explanation of the term "SARS病毒" given by online-Wiki contains "SARS是'非典'学名的英文缩写", and then, by using rules matching approach, we can get "非典" as the expansion of "SARS病毒".

3. System Description

As in NTCIR-7, we follow a modular architectural approach to QA represented in Figure.1.

The system contains three modules: Question Analysis module, IR module and Answer Extraction module. It carries out the following steps: For a given question, we did question classification first, the question was classified into its expected answer category by using pattern matching with predefined templates in the Question pattern library, and then, a query form was generated by the original question, it could be used to retrieve relevant documents from the target corpus. A few sentences were extracted from those relevant documents to form a candidate pool. Here, a scoring approach was used to rank candidate answers in this period.

Since the ACLIA task also contains IR4QA (Information Retrieval for Question Answering) subtask, During the evaluation, the question text and QA system question analysis results were provided as input to the IR4QA system, which produced retrieval results that were subsequently fed back into the end-to-end QA systems[4]. Our group also participated in the IR4QA task at NTCIR8. Therefore, we use this IR4QA system as the document retrieval module. For details on it, we refer the reader to the KECIR IR4QA system description paper [5]. Details of other modules will be introduced in the following sections.

4. Question Analysis

For a given question, the entire passage retrieval and answer extraction process relies on having the correct question (or answer) type and question forces. Meanwhile, the purpose of question analysis is to identify the question (or answer) type and detect question forces of the inputted question. Therefore, Question Analysis is an important task in most Question Answering systems [5, 6, and 7]. It contains Question classification and Keyword extraction.

The purpose of question classification is to obtain the question type and the answer type. Question classification can be formulated as:

$$f : Q \rightarrow T$$

Here, Q represents the question set; T is the set of question types; and f means to assign a particular class to a question based on the type of answer entity the question represents.

In NTCIR-8 Question Analysis Track**:** Question Analysis results contain key terms and answer types extracted from the input question. The released formal run topics were described in the xml-structure as shown in Figure.2.

```
<TOPIC ID="ACLIA2-CS-0001">
    <QUESTION
LANG="EN"><![CDATA[Who is the best
actor in the 76th Oscar's?]]></QUESTION>
    <QUESTION
LANG="CS"><![CDATA[第76界奥斯卡最佳
男主角是谁？]]></QUESTION>
    <NARRATIVE
LANG="EN"><![CDATA[The user would like
to know the name of the best actor in the 76th
Oscar's.]]></NARRATIVE>
    <NARRATIVE
LANG="CS"><![CDATA[使用者想知道第76
届奥斯卡最佳男主角是
```

Figure2. A released formal run topic

For C-C Subtask, only Chinese QUESTION field is used. Usually, in a question, there are dominant words and the indicate words. dominant words like “谁(who)”, “为什么(why)”, “何时(when)”, “何地(where)” and other words such as “何人(which person)”, “哪个国家(which country)”, etc. These can be regarded as dominant words that show the question type directly. Take a Chinese question sentence “电影《无极》什么时候公映？(When is "The Promise" screened?)” as an example, we can easily identify the question type as DATE type because of the dominant word “什么时候(When)”. Since the traditional regular expression model performs well with those questions, which containing dominant words, using the templates and rules, we build a Question pattern library.

Table 3. The coverage of Chinese question type classification with patterns

Type	Patterns matched / All questions
DEFINITION	10/10
BIOGRAPHY	10/10
RELATIONSHIP	18/20
EVENT	10/20
WHY	19/20
PERSON	5/5
ORGANIZATION	3/5
LOCATION	4/5
DATE	5/5

As shown in Table 3, Because of the limitation of the questions set that was used to construct patterns, we can not design all the patterns that could match every question. For instance, some interrogative words like “什么(what)” and “哪个(which)”, they could make a pair with many different words and represent different types. For example, the question type of “在克什米尔地区有哪个**国家**的边界？(Which **countries** have borders in the Kashimir region?)” is LOCATION, however, the question “王建民是哪个**球队**的？(Which **baseball team** does Chien-Ming Wang belong to?)” is an ORGANIZATION type question. Therefore, we consider the indicate words which always follow the interrogative words in a question to identify the query type.

HowNet is an on-line extra linguistics knowledge system for computation of meaning in Human Language Technology (HLT) [9]. HowNet unveils inter-concept relations and inter-attribute relations of the concepts as connoting in lexicons of the Chinese and their English equivalents. The concept definition (DEF) in HowNet is not written in natural language as in LDOCE or WordNet, but in a mark-up language, whose basic units are sememes such as “human|人”, “InstitutePlace|场所”, “Occupation|职位”, “medical|医”, “doctor|医治”, and semantic roles and features such as “HostOf”, “agent”, “modifier” and “domain”. Therefore, the definition of “doctor” can be literally paraphrased as “A doctor is a human being, who has the attribute of occupation; he doctors (gives medical treatment to); he belongs to the domain of medicine”.

```
NO=127941
W_C=医生
G_C=N[yi sheng]
E_C=
W_E=doctor
G_E=N
E_E=
DEF={human|人:HostOf={Occupation|职业}.
        Domain={medicl|医}.{doctor|医治:
        agent={~}}}
```

Figure 3. The description of “doctor” in HowNet

To find the indicate words we use the strategy proposed in [8], following the algorithm in figure 4.

```
Step1.Get interrogative word in question Q
Step2.Select the first noun behind
      interrogative word as the indicate word
      of the question.
Step3.If there is no noun exist behind
      interrogative word, then, choose the
      nearest one left the interrogative word
      as the indicate word of the question.
```

Figure 4. Approach of question indicate word selection

Then, we gain the concept definition of the indicate word of the question. The first part of the concept definition in HowNet was used to identify the type of the question.

5. Answer Extraction

5.1 Number of Snippets

The experiment [8] results has already shown that when varied the number of snippets from 1 to 1000, performance improves sharply as the number of snippets increases from 1 to 50 (0.243 MRR for 1 snippet, 0.370 MRR for 5, 0.423 MRR for 10, and 0.501 for 50), somewhat more slowly after that, and then falling off somewhat after that as more than 200 snippets are included. Since that, the top 100 results of KECIR-CS-CS-01-T were used as question related documents for answer extraction. Then we utilize different strategies to deal with different type of questions.

5.2 Answer factoid Question

For answering question in 'PERSON', 'LOCATION' and 'ORGANIZATION' type, we first use the NER tool to recognize the NE which is correspond to the question type as the candidate answers, and then A Abbreviations was used to normalize similar candidate answers. At last we used the voting approach to ranking them by the following formula:

$$Score(Ai) = \sum_{j=0}^{n} \frac{Count(Ai)}{Count(Aj)} (f5.2.1)$$

Here, the Score(X) is the povo (percentage of votes obtained) of candidate answer X; the Count(X) is the occurrence number of candidate answer X in relate snippets; and the n is the number of different candidate answers.

If some candidate answers got the same score, then the rank of them will be decided by the minimal mean distance weighting method:

$$W(Ai) = \left(1 + \frac{\sum_{j=1}^{m} \min(Ai, Qj)}{m} \right)^{-1} \quad (f5.2.2)$$

The W(Ai) is the weight of the candidate answer Ai, the m is number of keywords contained in the relate snippet. min(Ai,Qj) is the minimal number of words between candidate answer Ai and keyword Qj.

For answering question in 'DATE' type, the document ID was used for time reasoning. Rules sample is shown in Table 4.

Table4. Rules sample of date expanding

Document ID	Time Entity	Answer
XIN_CMN_ 20050620.0000	今年	2005年
	年底*	2005年12月
	本月	2005年6月
	今天*	2005年6月20日
	X日*	2005年6月X日
	X月*	2005年X月

* If there is another time entity before this in the same sentence, then don't follow the rules.

And then, we use formula f5.2.1 and f5.2.2 to calculate the score of the candidate answer. Here, the Count (Ai) is the weighted score of time expressions in candidate answers.

Here, we consider a whole time expression will be including three items, day, month and year. Define the time expression which is including all of three as T-expression, including two of them as T2-expression, and just including one of them as T1-expression, and then Count (Ai) will be the following formula:

$$Count(Ai) = \begin{cases} \text{if Ai is an T-expression} \\ \text{then:} \\ C(Ai) + 0.5 \times C(T2) \\ +0.3 \times C(T1) \\ \text{if Ai is an T2-expression} \\ \text{then:} \\ C(Ai) + 0.3 \times C(T1) \\ \text{if Ai is an T1-expression} \\ then: \\ C(Ai) \end{cases} (f5.2.3)$$

C (Ai) is the occurrence number of candidate answer Ai in relates snippets; C (T2) and C (T1) will be the occurrence number of candidate answer in T2-expression and T1-expression respectively, and all of them were included in Ai.

5.3 Answer complex Question

The complex questions of CCLQA task in NTCIR-8 is composed of the types of 'DEFINITION', 'BIOGRAPHY', 'EVENT', 'RELATION' and 'WHY'. Answer types for these questions are not explicit. On the other hand, nuggets, which represent the minimum unit of correct information in answer sentences, are required as answer candidates in the task. For answer extraction, we utilize different strategies to deal with different type of questions.

5.3.1. 'DEFINITION' type

To a definition question, we first split the document based on punctuation. If a sentence is an interrogative sentence, we filter the sentence. Some hint words were collected to identify whether it's a candidate answer or not. The list of the hint words for "DEFINITION" is shown in Table5.

Table5. Hint words of 'DEFINITION' type

sets	words
definition_verbs	是、就是、叫、叫做、称为、称作、称(之)为、命名(为) 、简称、又称 、俗称、(被) 认为
property_verbs	构成、组成、成分、属于、功用

We also summarized the patterns to extract the

answer of these questions. The patterns are shown as follow:

Table6.Patterns to extract the answers of the 'DEFINITION' type of questions

pattern	remark
keyword + V+ <ANSWER>	V∈definition_verbs or V∈proterty_verbs
keyword + 由 +<ANSWER> +V	V∈proterty_verbs

5.3.2. 'BIOGRAPHY' type

To a biography question, the answer actually refers to the description of a person. In this case, we utilize some attributes of a person. Like '国籍'(nationality) , '出生年月'(birthday) , '职业'(career) , '成就'(achievements) , etc. We believe these kinds of information are helpful to extract the answer of 'BIOGRAPHY' type. Answer patterns are shown in Table7.

Table7. Examples of Answer Patterns for "BIOGRAPHY" Type

attribute	pattern
Place	<PERSON>生于<PLACE> <PLACE>[Career]<PERSON>
Career	演员、科学家、领导人、局长、总统
Date	<PERSON>生于<TIME>, <PERSON>[于]<TIME>[出生/生人] <PERSON>[现年]<NUM>岁
Achievements	<PERSON>[获得/荣获/得到] <PERSON>[创办/领导/组织/参加]<OGANIZATION> <OGANIZATION>创始人<PERSON>

5.3.3. 'RELATION' type

For this type of questions, answers mostly refer to the relationship between two name entities. The heuristic rule to identify the potential targets is : if a sentence contains two named entities which exist in the question then the items between them are most likely to be an answer. We classify the relationship of two name entities as: person-person, person-organization, person-event, event-event , and others. The answer extraction method of every relationship pair are shown in Table8.

5.3.4. 'EVENT' type

We consider 'EVENT' type as compounded by 'List' question and 'Description' question. For the 'List' type, some works, places and persons are frequently mentioned. For the 'Description' type, the answer should indicate relate facts about the event. 'List' type and 'Description' type are classified by use words shown in table9.

5.3.5. 'WHY' type

For this type of questions, the answer actually refers to the reason of phenomena. We use some rules and patterns to identify the potential targets. Patterns and methods to extract are shown in table10.

Table8. The description of "relationship" answer extraction method

Relationship type	methods
person-person	1) if a sentence contains two named entities which exist in a question and has a segment between two name entities, the segment will be an answer. 2) NE 和 NE + <ANSWER>, NE与NE + <ANSWER>， NE + <ANSWER> + NE. If a sentence matches the above pattern, extract the segment between the first NE and verb.
person-organization	extract the sentences that contain all keywords
person-event	extract the sentences that contain all keywords
event-event	extract the sentences that contain two events
other	extract the sentences that contain all keywords

Table9. Hint word of classification

event type	hint words
List	列举、列出、相关、哪些、举出
Description	Other

Table10.The methods to extract the answers

id	pattern
1	因为+ <ANSWER> +所以+ keyword ..
2	<ANSWER>.，因此 + .keyword
3	<ANSWER> ，所以+ .keyword
4	是由于+ <ANSWER>
5	是为了+ <ANSWER>
6	A的原因是+ <ANSWER>
7	由于+ <ANSWER> + 将（会）+keyword
8	<ANSWER> +将对+ keywords
9	<ANSWER> + .造成+ keywords
10	<ANSWER> + 引起+ keywords

For a 'WHY' type question, if we can't use patterns to identify any sentence as an answer, instead, we extract the sentences which contain keywords. If the sentence in candidates is an interrogative sentence simile to the question, we just take the next sentence as an answer.

5.3.6. Answer Ranking

For the "List" type of the "EVENT" question, the ranking method is the same to factoid ones. For other complex types, two strategies were used separately, as shown in table11.

Table11.The methods of answer ranking

CS-CS Runs	methods
KECIR-CS-CS-01-T	With minimal mean distance of keywords.
KECIR-CS-CS-02-T	Re-rank answers which got same score in Run-01 with answer length.

6. Evaluation

To evaluate system responses, Official per-topic F-score definition based on nugget pyramid method [1] was used.

Table12.The official human evaluation

CS-CS Runs	all
KECIR-CS-CS-01-T	0.3450
KECIR-CS-CS-02-T	0.3354

The average F-scores of every type of question are shown in figure5, it shows run-01 and run-02 respectively.

7. Conclusion

In this paper, we described our CCLQA system and the evaluation results for the C-C task at NTCIR-8 ACLIA. For answer extraction, a Surface-Based Multi-Strategy approach was used. It deals retrieval results with different Strategies. The evaluation results show that our system achieves 0.3450 average F-score (beta = 3).

Reference

[1]. Teruko Mitamura, Hideki Shima, Tetsuya Sakai, etc. Overview of the NTCIR-8 ACLIA Tasks:Advanced Cross-Lingual Information Access. The proceedings of the 8th NTCIR workshop meeting. 2010.

[2]. Yu Bai, Li Guo, Lei Liu, etc. KECIR Question Answering System at NTCIR7 CCLQA. The proceedings of the 7th NTCIR workshop meeting. 2008.

[3]. Zhou Bo, Cai Dongfeng. Recognition of Chinese organization name based on conditional random fields. Journal of Shenyang Institute of Aeronautical Engineering. 2009.

[4]. Tetsuya Sakai etc. Overview of NTCIR-7 ACLIA IR4QA Task. The proceedings of the 7th NTCIR workshop. 2008.

[5]. Thamar Solorio, Manuel Pérez-Coutiño, Manuel Montes-y-Gémez, et al. A language independent method for question classification. In: Proceedings of the 20th international conference on Computational Linguistics, Switzerland, pp.1374-es. 2004.

[6]. CUI Huan, CAI Dong-feng, MIAO Xue-lei. Research on Question Comprehension for Question Answering System [C] (in Chinese). In: Process for Artificial Intelligence in China [M]. Beijing University of Posts and Telecommunications (BUPT) Publishing House. 2003.

[7]. Dongfeng Cai, Yu Bai etc. Chinese Question Classification Using Combination Approach. The proceeding of the International conference on semantics, knowledge and Grid. 2007.

[8]. Susan Dumais, Michele Banko. Web Question Answering: Is More Always Better?. The proceedings of the 25th Annual International ACM SIGIR conference on Research and development in Information Retrieval. 2002.

[9]. DONG Zhen-dong, DONG Qiang. HOWNET and the Computation of Meaning. World Scientific Publishing Co. Pte. Ltd. 2006.

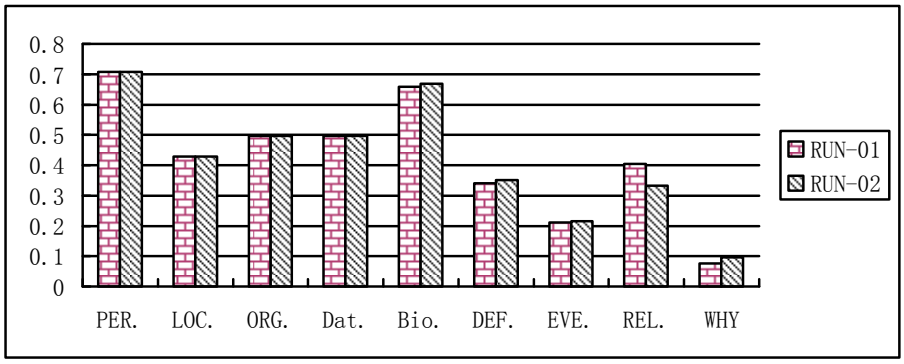

Figure 5.The average F-scores of every type of questions