

Chinese QA and CLQA: NTCIR-5 QA Experiments at UNT

Jiangping Chen, Rowena Li, Ping Yu, He Ge, Pok Chin, Fei Li, Cong Xuan
School of Library and Information Sciences
University of North Texas, P.O. Box 311068, Denton, TX 76203, USA
{jpchen, rll0099, pingyu, hg0022, pwc0007, fl0030, cx0005}@unt.edu

Abstract

This paper describes our participation in the NTCIR-5 CLQA task. Three runs were officially submitted for three subtasks: Chinese Question Answering, English-Chinese Question Answering, and Chinese-English Question Answering. We expanded our TREC experimental QA system EagleQA this year to include Chinese QA and Cross-Language QA capabilities. Various information retrieval and natural language processing tools were incorporated with our home-built programs such as Answer Type Identification, Sentence Extraction, and Answer Finding to find answers to the test questions. Future development will focus on investigating effective question translation and answer finding solutions.

Keywords: Chinese Question Answering, Cross Language Question Answering, natural language processing, system development.

1 Introduction

Question Answering (QA) systems identify answers from a large document collection or online information resources to users' natural language questions. Such systems can release the users from digesting huge amounts of text in order to locate particular facts or numbers. Current research on QA is mainly conducted in English. However, today's information sources are becoming more and more multi-linguistic, especially on the Internet. A survey of distribution of languages on the Internet (<http://www.netz-tipp.de/sprachen.html>) shows that in 2002, only 56.4% of Web pages were in English. Statistics on a Web site listing "Internet Users by Language" (<http://www.internetworldstats.com/stats7.htm>) updated on March 24, 2005, indicate that English language usage is further reduced to only 32.8% of Internet users. As a result, research is needed to explore solutions for QA in languages other than English and for Cross-Language Question Answering (CLQA). CLQA research explores effective and efficient solutions to find answers for users' questions from documents written in languages different from the questions. It is a more challenging

task than monolingual question answering because it involves translation among different languages. Among non-English languages, Chinese, Japanese, and Spanish are the top three languages used on the Web. Therefore, Chinese QA and CLQA research is needed in order to allow users to find answers from a collection of resources in multiple languages.

This year, NTCIR-5 initiated the evaluation of Chinese Question Answering, English-Chinese Question Answering, and Chinese-English Question Answering, along with other QA tasks. Chinese Question Answering (C-C) aims to find answers to Chinese questions in Chinese documents; English-Chinese Question Answering (E-C) finds answers to questions written in English among documents written in Chinese; the purpose of Chinese-English Question Answering (C-E) is to find answers to Chinese questions in English documents. The Chinese document set used this year is a collection of 901,446 news articles spanning from 2000 to 2001 taken from UDN.COM, and the question files are in BIG 5 encoding. The English document set contains news articles from the Daily Yomiuri in 2000 and 2001, and the question files are encoded with ASCII. For formal runs, there were 200 testing questions for each subtask. Answers to all questions were restricted to named entities to create a simpler question target for this pilot task, and all answers were judged with three scores – correct (S), unsupported (A), and incorrect (C). Performance of a run is measured by *accuracy* – the percentage of questions which are correctly answered.

We expanded our TREC experimental QA system EagleQA this year and participated in the CLQA task. Our purposes for participation include: 1) investigate and evaluate a Chinese QA and CLQA solution; 2) evaluate several software tools for certain tasks such as document retrieval and text annotation; and 3) understand the challenges of the CLQA tasks for future improvement. This paper describes our efforts on three subtasks: Chinese Question Answering, English-Chinese Question Answering, and Chinese-English Question Answering. It is arranged as follows: Section 2 briefly describes current Chinese monolingual QA approaches and CLQA approaches. A general overview of our EagleQA system is

provided in Section 3. Section 4 lists all linguistic resources and tools we have used for NTCIR-5 experiments. Section 5 describes our strategies specific to the three subtasks that we carried out for the NTCIR-5 Workshop. Section 6 reports our submissions and results. Section 7 reports our analysis of some processes including Question Translation and Answer Type Identification. The paper concludes with future directions for research.

2 Current research on monolingual Chinese QA and CLQA

Research on Monolingual Chinese Question Answering systems is still at its developing stage. Li and Croft [8] built a Chinese QA system utilizing similar approaches to those of English systems. Huang and Yao [6] used the Web as their search engine and knowledge base for Chinese QA combining with natural language parsing and an Entity-Relation-Entity relational model to boost performance. Peng, Weischedel, Licuanan, and Xu [12] explored QA strategies for Chinese definitional questions by combining deep linguistic analysis with surface pattern learning. Meanwhile, Zhang and Zhang [15] employed a rule-based logic form representation algorithm and lexical knowledge extracted from HowNet for logic proving.

Research on Cross-Language Question Answering systems was initiated at the Cross Language Evaluation Forum (CLEF) in 2003. CLEF focuses mainly on evaluating and encouraging CLQA systems for European languages. In its first campaign year, five languages (Italian, Spanish, Dutch, French, and German) were tested and searched against an English corpus [13]. How-questions and definition questions were introduced as testing queries in 2004. At CLEF 2005, nine target languages and ten source languages were explored for 73 cross-language tasks [14]. Various approaches were utilized to bridge the language barriers including shallow linguistic analysis, statistical analysis, and question translation using bilingual dictionaries or machine translation systems [1, 5, 9, 10, 11].

3 EagleQA architecture

In general, our current QA system EagleQA contains six modules as illustrated in Figure 1. They are Question Processing, Document Retrieval, Text Annotation, Sentence Extraction, Answer Finding, and Submission, as described below. These modules have integrated several freely available NLP software tools for the purposes of Chinese QA and/or Cross-Language QA. Those NLP tools will be introduced in the next section.

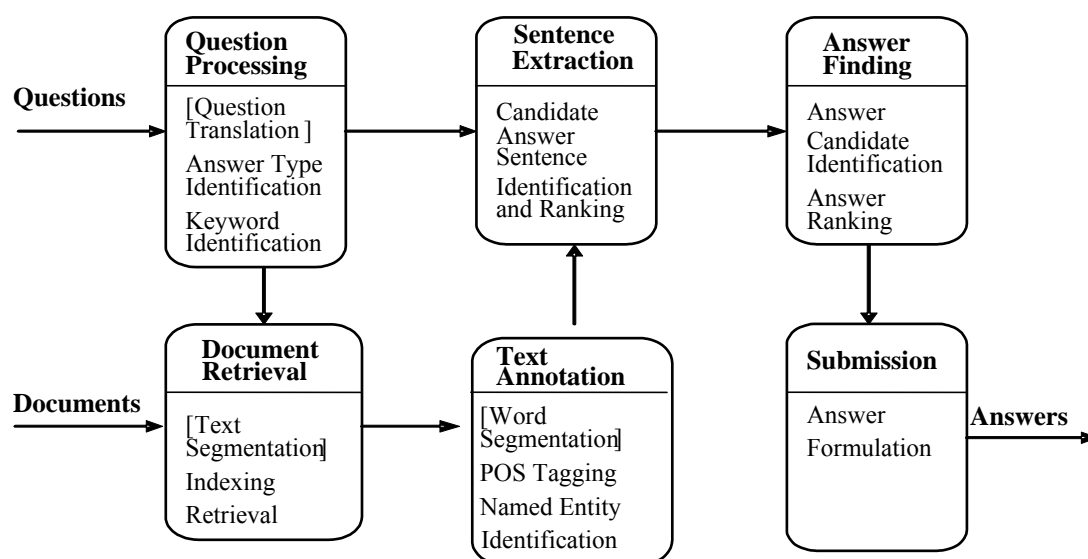


Figure 1. EagleQA architecture

3.1 Question Processing

The Question Processing module performs three processes: Question Translation, Answer Type Identification, and Keyword Identification.

Question Translation is applicable only to Cross-Language subtasks such as Chinese-English

and English-Chinese QA. Current implementation includes submitting Chinese or English queries to Babel Fish, an online machine translation system, for translation. The ongoing development in Question Translation includes constructing a lexical knowledge base from the document collection in combination with Babel Fish or other MT system for query translation. However, the development could not be

completed when NTCIR-5 CLQA experiments were conducted.

Answer Type Identification is the second process performed by this module. *Answer type* refers to the category in which the answer to a question should belong. For example, "PERSON" is the answer type for question "Who is the first astronaut in the world?" EagleQA automatically assigns an answer type to an incoming question by comparing the question with a manually developed answer type pattern file. The pattern file is extracted from 2393 TREC sample questions (for English), and 200 Chinese sample questions distributed by NTCIR-5 CLQA organizer (for Chinese). The most common answer types are: PERSON, LOCATION, ORGANIZATION, NUMBER, TIME, MONEY, and NAME (ARTIFACT).

Keyword Identification extracts important words or phrases from the annotated question. Each question is annotated applying the Text Annotation process described below in Section 3.3. A word or a phrase is regarded as important if it is not included in the stopword list of the system. The stopword list was prepared from training questions (2393 previous TREC questions for English stopword list, and 200 sample Chinese questions for the Chinese stopword list) by taking into account each word's Part-of-Speech (POS) and its frequency. For English questions, word expansion is also performed. Nouns and verbs were expanded by adding their synonyms and derivation forms to the keyword list based on WordNet 2.0 (www.princeton.edu).

3.2 Document Retrieval

We used Lemur to retrieve relevant documents from the provided document collections for both Chinese and English. Our NTCIR-5 Chinese Information Retrieval paper [4] describes in detail the evaluation we conducted on Chinese Information Retrieval experiments using Lemur. Based on our previous experimental results, we consider that Lemur's performance is acceptable. Prior to document retrieval, Chinese texts were segmented into bi-grams. Then we used Lemur to index the document collections. For both Chinese and English text retrieval, we chose to use Lemur's Okapi BM25 retrieval module with relevance feedback. The document number for relevance feedback is 5. The maximum number of new terms that were added to the original questions was 20.

3.3 Text Annotation

The retrieved documents obtained from the Document Retrieval module were annotated before the system performed sentence extraction.

3.3.1 English Text Annotation. We used LingPipe and Minipar together to perform Part-of-Speech tagging, named entity categorization, and noun phrase detection for English text annotation. The method is described in our TREC paper [3]. In general, LingPipe is used first to detect sentence boundaries, then the identified sentences are sent to Minipar for Part-of-Speech tagging and named entity categorization. We also keep the named entity categorization from LingPipe and combine the annotation results from the two systems.

3.3.2 Chinese Text Annotation. Chinese Text Annotation is our new development this year. Due to time constraints we performed preliminary annotation including following steps: a) Chinese segmentation dictionary construction, which combined multiple lexical resources [4]; b) Chinese word segmentation using forward maximum matching approach [4]; c) Part-of-Speech (POS) tagging. Since the segmentation dictionary also contains the POS for each word, it was also used to assign Parts-of-Speech to the collection. If a word had more than one POS, the most frequent used POS was selected. Also, simple rules were applied to identify number, time, and English words. The segmentation and POS tagging approaches were originally employed to automatically develop a draft annotation for human annotators to create training data, which will be used to develop statistical solutions to Chinese word segmentation and POS tagging. However, we could not complete the whole development process in time for the NTCIR-5 evaluation.

3.4 Sentence Extraction

The Sentence Extraction module identifies a certain number of non-duplicate sentences (500 sentences maximum for this year) from the annotated documents as sentence candidates which may contain an answer to each test question from the retrieved documents. The keyword lists and answer type information obtained in Question Processing are utilized to weigh extracted sentences for each question. The top 500 sentences were returned to the Answer Finding module to find the answers.

3.5 Answer Finding

The Answer Finding module applies multiple evidences to find answers for test questions. First, answer candidates were identified based on their annotation, and/or Part-of-Speech tagging. Then, each candidate was weighed based on certain factors including: 1) answer type: whether the candidate was annotated a category which is the same as the answer type of the question; 2) weight of the sentence, which is inherited from the sentence extraction module; 3)

