View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DCU Online Research Access Service

ICT-DCU Question Answering Task at NTCIR-6

Zhang Sen
IR group, ICT, CAS
Beijing, 100080, China
zhangsen@ict.ac.cn

Bin Wang
IR group, ICT, CAS
Beijing, 100080, China
wangbing@ict.ac.cn

Gareth J.F. Jones
School of Computing
Dublin City University
Dublin 9, Ireland
Gareth.Jones@computing.dcu.ie

Abstract

This paper describes details of our participation in the NTCIR-6 Chinese-to-Chinese Question Answering task. We use the "retrieval plus extraction approach" to get answers for questions. We first split the documents into short passages, and then retrieve potentially relevant passages for a question, and finally extract named entity answers from the most relevant passages. For question type identification, we use simple heuristic rules which cover most questions. The Lemur toolkit was used with the okapi model for document retrieval. Results of our task submission are given and some preliminary conclusions drawn.

Keywords: *NTCIR, Chinese-to-Chinese Question Answering, Information Retrieval, Information Extraction*

1 Introduction

This paper describes details of our participation in the Chinese-to-Chinese (C-C) Question Answering (QA) subtask for NTCIR-6. We use a standard QA strategy of information retrieval (IR) plus answer extraction to obtain answers to questions. There are two basic steps: first, we retrieve short document passages that are potentially relevant to a question and may contain answers to that question; then, named entities identification methods are used to mark and obtain the most likely answer from the retrieved passages.

The NTCIR-6 C-C QA task contained 150 questions to be answered from 901,446 news article documents spanning two years (2000-2001). Both the questions and documents are encoded with BIG5 which is widely used in Taiwan Province of China, and formatted according to standard TREC conventions. Because in our work we mainly use the GBK encoding in the mainland of China, we had to convert the questions and documents from BIG5 encoding into GBK encoding to enable processing using our text processing tools trained on Simplified Chinese corpora encoded with the GBK encoding.

We used some heuristic methods and pattern matching rules for question type analysis and classification, based on word splitting and part-of-speech (POS) tagging of questions. The Lemur toolkit package was used for the retrieval of relevant document passages (which were produced by splitting documents). We made use of POS tagging and statistical information of tagged Chinese words for the answer acquisition process.

The remaining parts of this report are organized as follows: Section 2 describes the architecture of our C-C answering system; Section 3 describes the components of our C-C answering system; Section 4 gives our results and makes some analysis on the results; and finally, conclusions and some closing thoughts are given in Section 5.

2 System Architecture

The architecture of our C-C QA system is shown in Figure 1. The work flow of our QA system is as follows:

1. All the data (including both the document collection and the questions) are converted from the provided BIG5 encoded dataset into GBK encoding.

2. The GBK-encoded documents are split into short passages.

3. These short passages and the GBK-encoded questions are split into Chinese words, and tagged with POS.

4. A search index is built of the short passages.

5. Relevant passages for each question are retrieved using Lemur.

6. Possible answers are obtained from the retrieval result, tagged document passages and tagged question. The possible answers are those that are most likely according to some statistical possibility rank.

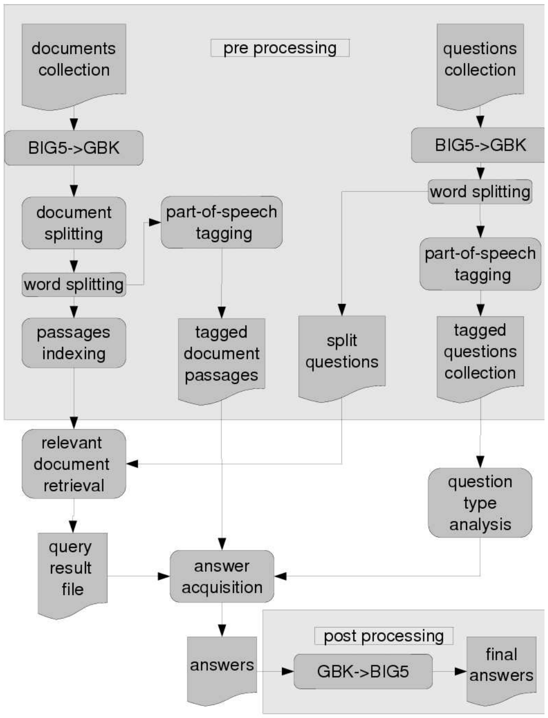

Figure 1. QA System Architecture

7. The answers are converted back to BIG5-encoding.

3 System Components

3.1 Pre-processing and Post-processing

As pointed out in Section 1, all the documents and questions are encoded in BIG5 encoding, because we are not familiar with BIG5-encoded Chinese processing and the tools available to us are trained on Simplified Chinese corpora, we have to convert the documents and questions to GBK-encoding. After character conversion we split each document into short passages, where the short passages are originally passages in each document. This is followed by word splitting and POS tagging which are applied separately to the passages and questions. Next we use Lemur to build an index for the passages which have been split into words.

We use Textpro [1] to convert the documents and questions to GBK-encoding. Textpro is generally highly effective for this task, but there are still some illegal sequences of character codes in the converted documents and questions. So the GNU "iconv" [2] program with the "-c" switch is used to first convert the GBK-encoded documents and questions to UTF-8 encoding to get rid of the illegal code sequences. Illegal code sequences of a certern character encoding are these sequences which can't be recognized correctly by that encoding(in our work, GBK).Later "iconv" is used again to convert back to GBK-encoding.

After all the data has been converted to GBK-encoding, we segment it into Chinese words, assign POS tags and recognize the named entities. ICTCLAS, a Chinese segmentation, POS tagging and named entity identification tool is used for this task. The questions were split into words and also POS tagged using the same method. Although ICTCLAS is trained based on Simplified Chinese corpora news articles from the Chinese People's Daily, we found that ICTCLAS can generate reasonable segmentation results and named entity labels for the transformed CLQA documents and questions set.

In the post-processing stage, the obtained answers are converted back to BIG5-encoding to form the final answers for submission.

3.2 Relevant Passages Retrieval

We used the Lemur toolkit to perform passage retrieval. Lemur was developed to facilitate language modeling IR research by CMU & UMASS. However, it also includes traditional IR methods such as the vector-space model (VSM) and some probabilistic models such as okapi.

In our experiments, we tried a simple TFIDF VSM model and the BM25 okapi model to retrieve relevant passages. Although these two methods retrieve different results (at least the orders of retrieved passages are different), the final answer results are actually very similar. So, finally we chose the BM25 okapi model for retrieval in our C-C QA experiments.

Some stop items such as interrogatives and other common stop words were eliminated from questions. We also explored to giving different weights to some words (e.g., proper nouns, entity names) that seem to be more important, but the results were not clearly improved.

3.3 Question Type Analysis and Classification

For questions asking about different entities, the returned answer should be the corresponding proper entity type, so we classify questions by their required answer types. There are nine types of questions:

1. PERSON
2. LOCATION
3. ORGANIZATION
4. DATE

[1] See http://www.fodian.net/tools/TextPro5.zip

[2] See http://www.gnu.org/software/libiconv/

5. TIME

6. NUMEX

7. MONEY

8. PERCENT

9. ARTIFACT

In our work, we used pattern matching heuristic rules to classify each question as one of the nine question types. We built a table of rules for these nine question types. When a question is entered, pattern matching based on keywords is performed to assign the question to one of the available question types. The following table gives some examples to illustrate assignment of the answer type:

Table 1. Examples of special words

type of named entities	special word examples
PERSON	是谁 谁是 由谁 由何人 什么名字
DATE	哪一天 何时 哪一年 几岁 几月
ARTIFACT	什么奖 哪一款 叫什么 哪部
MONEY	多少钱 几元
LOCATION	哪个城市 哪边 哪个地方 哪一州 哪一国

3.4 Answer Acquisition

Candidate answers for each question are extracted from the retrieved passages based on the question type using different strategies. Different types of questions require different types of answers.

We fetched the first twenty passages in the order of descending relative score between the question and the passages. Heuristic rules together with ICTCLAS are used to find the candidate answers from these twenty passages. A disadvantage of ICTCLAS currently is that it can only recognize proper nouns (including person name, location name and organization name) and other types of words such as temporal words (including date and time), numeral words (money, numex, percent) and ordinary artificial nouns, but it cannot discriminate between numercial types money, numex and percent against each other, although it can tag proper nouns further as names of person, organization or location. This is not sufficient for the C-C QA task, where we in fact need nine types of answer entities for all the questions, which are person, location, organization, date, time, numex, percent, money and artifact. Because ICTCLAS performs well in recognizing the names of a person, a location and an organization respectively, we expect to get more correct answers for questions in these three types of questions (which include PERSON, LOCATION, ORGANIZATION). But for other types of questions, this will not necessarily be the case. For example, ICTLAS can only recognize and tag all the numex, percent and money as numeral words, this may possibly decrease the reliability of the answer seeking program.

There are some obviously impossible words presented in the candidate answers, so we build a list of impossible words to get rid of these noisy words. For example, the Chinese characters including ”后”, ”期”, ”末” and ”夜” which appears too frequently are noisy for these questions asking for date.

Then, the number of the occurrences of each candidate answer is counted. The candidate answer with the largest number of times of occurrences is chosen as the most likely answer to the question.

4 Results and Analysis

Using this simple approach we obtained correct answers to 51 questions, representing % of the 150 questions. Some improvement is observed over the result using a simple system for the NTCIR-5 C-C QA tasks. The MRR for our QA system is 0.340, which is higher than the result we obtained previously.

We obtained different outcomes for different question types, as illustrated in the histogram shown in Figure 2.

As stated in subsection 3.4, we get higher correctness for those questions asking about names of person, location and organization and no correct answers were found for numex, percent and money. The high correctness for date and time are perhaps due to the fact that there are no significant difference between these two concepts.

Further it is possible that tools trained on Simplified Chinese corpora are not good at processing the Traditional Chinese documents and queries. Improvement can be made on the classification of questions using some learning methods. The coarse grain question types decreased the overall

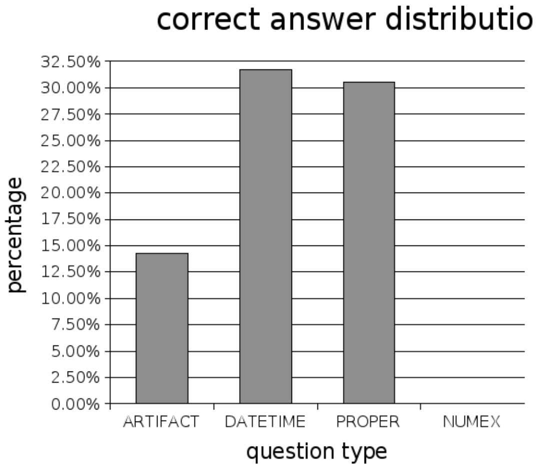

Figure 2. Answers Distribution

system performance, and the answer extraction method was simple and naive with much room for improvement.

5 Conclusions and Future Work

There are several aspects where we can improve the performance of the system:

- Combine both rule-based pattern matching, statistical methods and learning methods to assign each question to a question type more accurately. Some questions are ambiguous when using only rule-based pattern matching based on keywords.

- Make different and more detailed policies for each question type to find the correct answer. We made only four different kinds of coarse grain polices: (PERSON, LOCATION, ORGANIZATION), (NUMEX), (ARTIFACT), (DATE, TIME). This obviously decreases the potential accuracy and performance of our QA system.

- The ICTCLAS segmentation and POS tool can recognize only limited types of named entities. We need to improve this tool or find another alternative tool to be able to identify more accurately the types of named entities.

6 Acknowledgement

This work is supported by a China-Ireland Science and Technology Collaboration Research Fund award under Grant No. CI-2004-12 and China High Technology 973 project under Grant No. 2004CB318109. Thanks to Huaping Zhang for assistance with the ICTCLAS tools.

References

[1] Huaping Zhang Hong-Kui Yu, De-Yi Xiong, Qun LIU; HMM-based Chinese Lexical Analyzer ICTCLAS, *In Proceedings of the second SIGHAN Workshop affiliated with 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo Japan, 2003

[2] Dell Zhang, Wee Sun Lee, Question Classification using Support Vector Machines, *in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 26-32, Toronto, Canada, 2003.

[3] Seung-Hoon Na, In-Su Kang, Sang-Yool Lee, Jong-Hyeok Lee, Question Answering Approach Using a WordNet-based Answer Type Taxonomy, *in Proceedings of the 11th Text REtrieval Conference (TREC2002)*, NIST, 2003.

[4] U. Hermjakob, Parsing and Question Classificationfor Question Answering, *in Proceedings of the Workshop on Open-Domain Question Answering at ACL-2001*, 2001.

[5] Mei, Jia-Ju, Yi-Ming Zhu, Yun-Qi Gao, Hong-Xiang Yin, Tongyici CiLin (Chinese Synonym Forest), *Shanghai Press of Lexicon and Books*, 1983.

[6] S.E.Robertson, S.Walker, S.Jones, M. Hancock-Beaulieu and M.Gatford, Okapi at TREC-3, *In Proceedings of the Third Text REtrieval Conference (TREC-3)*. NIST, 1995.

[7] In-Ho Kang, GilChang Kim, *Query Type Classification for Web Document Retrieval*, *in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp64-71 , Toronto, Canada, 2003.

[8] Bin Wang, Gareth J.F. Jones, *LCC-DCU C-C Question Answering Task at NTCIR-5*, it in Proceedings of the Fifth NTCIR Workshop on Research in Information Access Technologies - Information Retrieval, Question Answering and Cross-Lingual Information Access, Tokyo, Japan, pp262-267, 2005.