Heuristic and Syntactic Scoring for Cross-language Question Answering

Lisa A. Ballesteros and Xiaoyan Li

Computer Science Department, Mount Holyoke College
South Hadley, MA 01075
lballest@mtholyoke.edu, xli@mtholyoke.edu

Abstract

This paper describes the Marsha Cross-Language Question Answering System used by Mount Holyoke College in the English-Chinese, Chinese-Chinese, and English-English subtasks of the NTCIR Cross-Language Question answering task. The system was most effective in the Chinese and English monolingual tasks. However, improved translations and better query type identification remain challenges for more effective cross-language QA task performance.

Keywords: Question answering, cross-language information retrieval, syntactic information

1 Introduction

We use the Marsha question answering (QA) [5][6] system to perform the NTCIR 2006 monolingual English and Chinese QA tasks, and the English-Chinese cross-language QA task. Our approach is to combine syntactic information from questions and candidate sentences from top documents with heuristic ranking of candidate sentences. We also performed simple date resolution. Section 2 describes our system architecture, Section 3 presents our results, and Section 4 gives conclusions and future work.

2 Overview of MARSHA Architecture

The Marsha system has three main components: a query processing module, the Inquery search engine 0 or Hanquery search engine (for Chinese) [2], and an answer extraction module. The query processing module identifies particular question types and then generates a query for the search engine. The search engine retrieves a set of candidate documents from the test corpus. The answer extraction module infers passages most likely to contain question answers and extracts these answers when possible. Each module is described in more detail in Sections 2.1-2.3.

2.1 Query processing

Query processing proceeds in roughly the same way for both English and Chinese queries. The query processing module (QPM) processes queries as described below. When appropriate, differences in the way that English and Chinese queries are processed are explicitly noted.

1) Pre-defined question types that roughly correspond to standard named entity (NE) classes extracted by NE systems are recognized. Nine question types which include PERSON, LOCATION, ORGANIZATION, DATE, TIME, MONEY, PERCENT, NUMBER, and OTHER, are currently defined. Each question is first parsed using the BBN SIFT parser [3]. The question's type is then identified by matching syntactic information from the parse against syntactic pattern templates, of which there are currently 170 for Chinese. For example, the QPM would classify questions containing the strings "which city" or "which person" as LOCATION and PERSON, respectively.

2) Question words such as "which" and "what" are removed as they add no useful information.

3) BBN's Identifinder [4] is employed to identify named entities.

4) Dictionary translation of cross-language queries is performed from the source (English) to the target (Chinese) language. In a first pass, we do dictionary look-up for word bi-grams and entities identified in step 3. If no translation is found in the first pass, simple word-by-word translation is used for the remaining query terms.

5) Chinese queries are segmented to identify words, but NEs remain unsegmented. Note that English queries do not require segmentation.

6) Stop words are removed.

After processing, queries are submitted to the retrieval engine. This process is described more fully in the next section.

2.2 Retrieval Engine

Prior to submission to the retrieval engine, each query is reformulated with structured query operators. The structure that the operators impose makes more explicit the way in which the query words are being used. In other words, it emphasizes the importance of a word or group of words to the concept being conveyed. In this work, we select operators to yield a higher rank for those documents containing more of the query words. If there are five or more query words left after processing via the QPM, the query is wrapped in a probabilistic AND (*#and*). Here, documents containing a greater number of query words are ranked more highly. If the processed query contains fewer than 5 words, the terms are wrapped in a passage operator, *#Uwn*, where n is equivalent to two times the number of query terms remaining. In this case, documents in which query terms are found within an unordered window of n are ranked higher than those documents in which query words are found farther apart. English queries were submitted to the Inquery retrieval engine, while Chinese queries were submitted to the Hanquery retrieval engine. The top10 documents retrieved are then processed by the answer extraction module (AEM) as described in section 2.3.

2.3 Answer Extraction Module

The AEM is responsible for identifying potential answers and ranking them. Each of the top 10 documents retrieved in response to a question is analyzed in the following way to identify answer candidates.

1) Named Entities are extracted by Identifinder. The types of entities recognized include *person, organization, location, time expression, date, numeric expression, money, and percent.*

2) The document is partitioned into passages consisting of two adjacent sentences. The passages generated have a 1-sentence overlap.

3) Each passage is scored according to several heuristic rules including the number of query words occurring in the passage, whether matching words occur in the same sentence, the size of the best matching window, and the distance between an answer candidate and the center of the best matching window. More specifically, scoring proceeds as follows:

i. If no named entity is present, the passage receives a score of 0. If a named entity is present in the passage but it does not have the same type as that of the question, that NE is not considered. Additionally, a NE is removed from consideration if it is also present in the question.

ii. Calculate the number of matching query words, *count_m*, in the passage. If the number of matching words is less than a threshold, *t*, assign a score of 0. Otherwise, the passage score is *count_m*. Let *count_q* be the number of words in the query. The threshold *t* is defined as follows:

   a. If the number of query terms is less than 4, then let *t*=*count_q*. In other words, do not consider any passages that do not contain all of the query terms.

   b. If there are between 4 and 8 query terms, let the threshold $t = count\_q/2.0 + 1.0$.

   c. If there are greater than 8 query terms, $t = count\_q/3.0 + 2.0$.

   This step is employed to address query ambiguity. Short queries tend to be less well specified, so it is important that a retrieved passage contain more of the query terms to increase the likelihood that it will contain relevant information and thus an answer. As query length increases leading to better specification of the information need, a smaller proportion of the query terms need be present in a passage.

iii. We assume that word co-occurring in closer proximity are more closely related than those co-occurring in a larger text window. If all matching query words occur within one sentence, let *Sm* = 1, else *Sm* = 0. Add (0.5 * *Sm*) to the score.

iv. As word order can affect meaning, we give higher weight to passages in which matching words occur in the same order as they occur in the original question. Let *Ord* = 1 if all the matching words are in the same order as the original question, otherwise *Ord* = 0. Add (0.5 * *Ord*) to the score.

v. Passage score = score + (*count_m*/*W*), where *W* is the size of the best matching window. The best matching window is the one containing the greatest number of matching query terms in the smallest window. Like item (iii), this heuristic gives credit to passages in which matching query terms occur closer together.

4) The final score, *c_score*, for a Chinese passage is given by, *c_score = count_m + 0.5*Sm + 0.5*Ord +count_m/W*. The top scoring passage is selected from each document and they are placed in rank order.

5) Extract the answer candidates from the top ranked passage:

Recall that we consider as answer candidates only those named entities having the same type as that of the question. Furthermore, if an NE occurs in the original question, it is removed from consideration. The distance between each remaining answer candidate and the location of each matching query word in the passage is calculated and the one having the smallest distance is selected as the final answer. In the case of "Date" type answers, simple date resolution is performed. More specifically, we convert relative dates such as 'today' and 'tomorrow' to specific dates as determined by the date field of the document from which the answer was extracted. If no answer candidates are found, no answer is returned.

The final score of an English passage, *e_score*, is the sum of a heuristic score (*h_score*) and a syntactic score (*s_score)*. A combination of heuristic and syntactic scoring was shown to be more effective than that of heuristic scoring alone [SIGIR03].The heuristic score is a modified version of the scoring function for Chinese candidates and is given by, $h\_score = count\_m + 0.5*Sm + count\_m/W + 0.5/D$, where D is the distance between the answer candidate and the center of the best matching window.

The syntactic score is a weighted sum of six syntactic factors, F1-F6, where *s_score* = F1+0.5/F2+0.5*F3+F4+F5+F6. Each factor is defined as follows:

F1: Identify the sentence from the passage having the longest match between phrases or sub-phrases extracted from the question. Consider only the longest matching portion of any particular phrase. F1 = length of the total matched phrases/question length.

F2: the distance between the answer candidate and the main verb.

F3: For questions of type "Person", syntactic patterns were used to understand whether the relationship between the desired named entity and the main verb of the question is characterized as "passive" or "active". Check that the relationship between the answer candidate and main verb in the passage is consistent with this characterization. F3 = 1 if this factor is satisfied, 0 otherwise.

F4: For "Location" questions, check that the possessive formats of the answer and question match. F4 = 1 when formats match, otherwise it is 0.

F5: For questions of "Location" or "Date", determine whether the answer candidate is inside a prepositional phrase modifying the main verb. If so, F5 =1, else F5 = 0.

F6: For "Person" questions, determine whether the answer candidate and all query words are contained in an adjective noun phrase (NPA). If so, F6 = 1, else F6 = 0.

3 Results and Analysis

After answer candidates are identified as described in Section 2, final answers are selected from amongst answer candidates in two different ways. Approach one selects the answer candidate from the highest scoring passage. Approach two performs selection via 'majority vote'. In other words, the answer candidate identified most often amongst the top ranked passages is selected as the answer. If each answer candidate receives only 1 vote, then selection is via top score. Results for our runs using selection of 1 top answer via best heuristic score are given in Table 1.

Run	EE-01	CC-01	EC-01
Right	20	28	6
Unsupported	41	32	6
Accuracy	.1333	.1867	.04
MRR	.1333	.1867	.04
Top5	.1333	.1867	.04
Accuracy+U	.2733	.2133	.04
MRR+U	.2733	.2133	..04
Top5+U	.2733	.2133	.04

Table 1: Results of runs with selection of Top1 Answer via top score.

Table 2 gives results for runs using selection of 1 top answer via majority vote. Recall that this is the candidate answer that appeared most often in the top 10 answers.

Run	EE-02	CC-02	EC-02
Right	14	23	6
Unsupported	24	27	11
Accuracy	.0933	.1533	.0400
MRR	.0933	.1533	.0400
Top5	.0933	.1533	.0400
Accuracy+U	.1600	.1800	.0733
MRR+U	.1600	.1800	.0733
Top5+U	.1600	.1800	.0733

Table 2: Results of runs with selection of Top 1 Answer via majority vote.

Heuristic scoring (Table 1), is more effective than rank via majority vote (Table 2). The only exception to this is when unsupported answers are considered for cross-language QA, where majority vote ranking yields 5 more unsupported answers than does heuristic scoring.

In order of performance, our Chinese monolingual system does best, followed by the monolingual English system and then the cross-language system. None of our systems are able to return answers of type ARTIFACT because Identifinder, which we use to identify named entities, does not recognize this type of NE.

There are also some challenges specific to our Chinese system. First, the query processing module requires the grammatical structure of a question in order to identify question type. However, some of the Chinese questions are expressed more like statements than questions. For example, questions CLQA2-ZH-T3026-00, CLQA2-ZH-T3027-00, and CLQA2-ZH-T3037-00 do not contain phrases such as "What is" or "who is". Furthermore, our Chinese system is trained on simplified Chinese text, but the Chinese queries are written in traditional Chinese. Although the queries are automatically converted to simplified form prior to presenting them to the system, this character-by-character conversion produces slight variations in phrasing. The query processing module is less effective on these reformulations, although they are expressed in simplified Chinese.

Cross-language QA performance using either type of scoring is considerably lower than that for either English QA or Chinese QA. When using selection via heuristic scoring, our cross-language run correctly selects only 6 correct answers in the top 1 yielding an accuracy of 0.04, while the monolingual Chinese run selects 28 correct answers in the top 1. This is primarily due to translation error. For example, consider question CLQA2-EN-T3150-00: "Who was Soviet Premier in 1962?" The word "Soviet" is an adjective that should refer to the country "Soviet Union". However, our termlist has a small number of Chinese terms and yields the Chinese translation equivalent meaning "meeting".

When we consider the top 5 answers selected via heuristic scoring, the top 2 to 5 slots yield an additional 3 correct answers and an additional 5 unsupported correct answers for the cross-language run. Improvements are higher for the monolingual English and Chinese runs, with an additional 18 and 1 correct answers, and an additional 26 and 8 unsupported answers, respectively.

Run	EE-Top5	CC-Top5	EC-Top5
Right [1]	20	28	6
Right [2]	8	0	2
Right [3]	6	1	1
Right [4]	2	0	0
Right [5]	2	0	0
Unsupported [1]	41	32	6
Unsupported [2]	13	0	4
Unsupported [3]	7	1	3
Unsupported [4]	3	1	1
Unsupported [5]	3	6	0
Accuracy	.1333	.1867	.04
MRR	.1792	.1889	.0489
Top5	.2533	.1933	.06
Accuracy+U	.2733	.2133	.04
MRR+U	.3411	.2252	.0616
Top5+U	.4467	.2667	.0933

4 Conclusions

Our experiments show that it is possible to employ a combination of syntactic and heuristic scoring for answer candidate selection in cross-language QA. However, several improvements, such as a larger termlist, reducing translation ambiguity, and the of training our system on traditional Chinese text are needed to increase system accuracy in the cross-language environment.

References

[1] J. P. Callan, W. B. Croft, and S. M. Harding, "The INQUERY Retrieval System", in Proceedings of the 3rd International Conference on Database and Expert Systems. (1992).

[2] Broglio, J., Callan, J.P. and Croft, W.B. "Technical Issues in Building an Information Retrieval System for Chinese," CIIR Technical Report IR-86, Computer Science Department, University of Massachusetts, Amherst, (1996).

[3] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, and the Annotation Group, "Algorithms that learn to extract information--bbn: Description of the sift system as used for muc-7". Proceedings of the Seventh Message Understanding Conference (MUC-7). (1998).

[4] Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. Machine Learning, 34(1-3):211–231, 1999.

[5] X. Li and W.B. Croft, "Evaluating Question Answering Techniques in Chinese", Proc. HLT 01, 96-101, (2001).

[6] X. Li and W.B. Croft, "Syntactic Features in Question Answering", presented as a poster, Proceedings SIGIR03, 455-456, (2003).