A Rule-Based Question Answering System for Reading Comprehension Tests

Ellen Riloff and Michael Thelen

Ramadan ALFARED
Computer-Master University of Henri Poincaré-Nancy 1
Oct 15 2007 ramadanalfared@neuf.fr
Outline of Talk

Introduction

Reading Comprehension Tests

A Rule-based System for Question Answer

Results

Conclusion
Introduction

- Quarc, that can read a short story and find the sentence in the story that best answers a given question.
- What used Quarc?
  - Quarc uses heuristic rules that look for lexical and semantics clues in the Question and story.
- System Quarc that takes reading comprehension tests.
- Given a story and a question, Quarc finds the sentence in the story that best answers the question.
Each test is followed by five **WH** question **WHO**, **WHAT**, **WHEN**, **WHERE** and **WHY**.

The answer to **WHEN** and **WHERE** question are also sometimes inferred from the dateline of story.
A Rule-based System for Question Answer

- Quarc QA for Reading Comprehension is a rule-based system that uses lexical and semantic heuristics to look for evidence that a sentence contains the answer to question.
- Each type of WH question looks for different type of answer.
- Quarc uses a separate set of rules for each question type WH*,...
- Given a question and a story $\Rightarrow$ Quarc parser the question and all of the sentences in the story.
- The rules are applied to each sentence in the story, as well as, the title of the story with the exception that the title is not considered for WHY question.
Each rule awards a certain number of points to a sentence. After all of the rules have been applied, the sentence or dateline that obtains the highest score is returned as the answer.

WordMatch function, which counts the number of words that appear in both the question and the sentence being considered.

Some rules also look for semantic classes which we will write in upper case e.g. HUMAN.
A Rule-based System for Question Answer

- Parser uses a dictionary and a semantic hierarchy, so that words can be defined with semantic classes.
- Parser recognizes two types of semantic entities:
  - A PROPER NOUN is defined as a noun phrase.
  - A NAME is defined as a PROPER NOUN that contains at least one HUMAN word.
- Each rule awards a specific number of points to a sentence, depending on how strongly the rule believes that is found the answer.
A rule can assign four possible point values:

- clue (+3), good-clue (+4), confident (+6), and slam-dunk (+20).
- These point values were based on the intuitions and worked.
- The main purpose of these values is to assess the relative importance of each clue.
- The next figure shows that the WHO rules, which use three fairly general heuristics as well as the Word-Match function rule 1.
A Rule-based System for Question Answer

1. Score(S) += WordMatch(Q,S)
2. if ¬ contains (Q,NAME) and contains (S,NAME)
   Then Score(S) += confident
3. if ¬ contains (Q,NAME) and contains (S,name)
   Then Score(S) += good clue
4. if contains(S,NAME,HUMAN)
   Then Score(S) += good clue

Figure 1: WHO Rules
A Rule-based System for Question Answer

- **WHAT** question were the most difficult to handle because they sought an amazing variety of answer.
- We noticed several what kind question, which looked for a description of an object.
- Rule 3 addresses these question by rewarding sentence that contain the word *call* or *from* e.g. *It is called...* or *It is made from ....*
- The rule set for **WHEN** question is the only rule set that does not apply the WordMatch function to every sentence in the story.
- almost always require a Time expression, so sentence that do not contain a Time expression are only considered in special cases.
A Rule-based System for Question Answer

- The **WHERE** question almost always look for specific location, so the **WHERE** rules are very focused.
- Quarc recognizes 21 preposition as being associated with location, such as in, at near, and inside.
- **WHY** question are handled differently than other question, the why rules are based on the observation that the answer to a why question often appears immediately before or immediately after the sentence that most closely matches the question.
- All sentences are assigned a score using the WordMatch function.
- Then the sentences with the top score are isolated.
- We will refer to these sentence as Best.
evaluated Quarc on the same data set that was used to evaluate the DeepRead reading comprehension system (Hirschman et al., 1999).

This data set contains 115 reading comprehension tests, 55 of which were used for development and 60 of which were reserved for testing purposes.

And also used the answer keys created by the DeepRead developers (Hirschman).
The *HumSent* answers are sentence that a human judged to be the best answer for each question.

The *AutSent* answer are generated automatically by determining which sentence contains the highest percentage of words in the published answer key.

Quarc achieved 40% Humsent accuracy overall, but the accuracy varied substantially across question types.

Quarc performed the best on WHEN question, achieving 55% accuracy, and performed the worst on WHAT and WHY questions, reaching only 28% accuracy.
### Results

<table>
<thead>
<tr>
<th>Question</th>
<th>HumSent</th>
<th>AutSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHO</td>
<td>0.41 (24/59)</td>
<td>0.49 (29/59)</td>
</tr>
<tr>
<td>WHAT</td>
<td>0.28 (17/61)</td>
<td>0.31 (19/59)</td>
</tr>
<tr>
<td>WHEN</td>
<td>0.55 (33/60)</td>
<td>0.28 (17/60)</td>
</tr>
<tr>
<td>WHERE</td>
<td>0.47 (28/60)</td>
<td>0.48 (29/60)</td>
</tr>
<tr>
<td>WHY</td>
<td>0.28 (17/60)</td>
<td>0.27 (16/60)</td>
</tr>
<tr>
<td>OVERALL</td>
<td>0.40 (119/300)</td>
<td>0.37 (110/300)</td>
</tr>
</tbody>
</table>

**Table 1: Overall Results**
we wanted to see how much effect the semantic classes had on performance,

added the rules that use semantic classes, only the WHO, WHEN, WHAT question types had such rules.

The final version of Quarc achieved 40% HumSent accuracy, which compares favorably with Deep-Read’s 36% HumSent accuracy.

Quarc is performance on WHAT, WHEN, and WHY questions improved by several percentage point, but performance on WHO and WHERE question was basically the same.
Conclusion

- Semantic classes were extremely useful for **WHO**, **WHEN** and **WHERE** questions because they look for descriptions of people, dates, and location.
- **WHY** questions are concerned with causal information.
- **WHAT** questions were the most difficult because the sought a staggering variety of answer.
Bibliographie

