Using Structured, Knowledge-Rich Corpora in Question Answering

Federico Mora  
Jesse Louis-Rosenberg  
Gregory Marton  
Boris Katz  

FEDERICO@CSAIL.MIT.EDU  
JESSEL@CSAIL.MIT.EDU  
GREMIO@CSAIL.MIT.EDU  
BORIS@CSAIL.MIT.EDU  

MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge MA, 02139 USA

1. Introduction

“Question answering (QA) is a type of information retrieval. Given a collection of documents (such as the World Wide Web or a local collection) the system should be able to retrieve answers to questions posed in natural language. QA is regarded as requiring more complex natural language processing techniques than other types of information retrieval such as document retrieval, and it is sometimes regarded as the next step beyond search engines” (Wikipedia, 2005).

To answer domain-independent questions precisely, one requires an extensive and accurate corpus from which to extract answers. These two aspects are often at odds with each other: large corpora (like the Web) often contain much incorrect data, while smaller corpora such as encyclopedias and dictionaries are generally too limited in scope.

Wikipedia, an online encyclopedia created by volunteers across the web, provides a well-structured, wide-coverage corpus available for free. It contains over 600,000 articles, which are updated daily. Because it provides such a large domain of knowledge-rich articles, it serves as an excellent corpus for question answering.

2. Approach

We incorporate Wikipedia as a source in our question answering system for the TREC 2005 QA track using various orthogonal methods (Katz et al., 2004). The TREC 2005 QA track divides questions into three categories: list, definition and factoid. These questions are clustered into small groups, each of which is about the same topic (Voorhees, 2004). The answers to these questions must be found with AQUAINT, a corpus of newspaper articles from 1998 to 2000. We integrated Wikipedia into our list and definition answering system. Our factoid system was unaltered, but there has been previous work using Wikipedia for factoid questions (Ahn et al., 2004).

2.1 Finding a Wikipedia Article

The first step in employing the Wikipedia for a question is finding the relevant article for a topic. Topics in previous years were restricted to simple noun phrases, and over 90% of them appeared in the Wikipedia in some form. Of the 75 topics this year, 10 were “event” noun phrases like “1998 Nagano Olympics” and 4 were headline-like events, like “Liberty Bell 7 space capsule recovered from ocean”. Fewer of these appear in the Wikipedia. We found the correct Wikipedia article for 87% of the noun phrase topics, for 70% of the noun phrase event topics, and none of the headline topics — 81% accuracy overall.

A topic was often listed as a Wikipedia title. If not, we tried several variants: removing pluralization and allowing non-capitalized words in the body instead of the title. If all of those, and some combinations, failed to find an article, then we took the top Google result when searching Wikipedia, and took the top available article from the main namespace. Sometimes this best Google article had only a few paragraphs on the desired topic. If fewer than 25% of the paragraphs contained words from the topic, then only those paragraphs that did contain some topic word were selected.

We used Google to find 39 of the Wikipedia topics; all 14 misidentifications were among these.

2.2 Lists

List questions ask for a list of entities that satisfy some condition, for example, “What diseases are prions associated with?”. We must return members of the class of diseases, and ideally only those that involve prions. Wikipedia helps by providing members of many classes, and by offering synonyms for key terms. Having class-member relations is crucial also because answers often do not have the class name (“disease”) near the answer (“Prions are known to cause Creutzfeld-Jacob syndrome”).

Wikipedia provides class-member relations in three different forms: First there are entire articles that are just lists, whose titles are the class name (the article, “prizes,
medals and awards”) (48,412 class-member pairs); second, some articles have lists within them, (“Nirvana (band)” has “discography”) (166,263 pairs); third, articles may mention a Category, for example the “Line Feed” article is a member of the category “Control characters”, but we deemed these too dirty for immediate use.

Our baseline system used only about 3000 lists — in comparison, over 200,000 manually-compiled lists made many more answer types available.

Wikipedia also provides synonyms in its Redirect structure and subtitles. For instance, we can expand “disease” to “medical condition”, “Woodrow Wilson Guthrie” to “Woody Guthrie” and “Woodie Guthrie”, and “Nagano” to “Nagano, Japan”. These manually constructed synonyms can significantly aid recall.

2.3 Definitions

Definition questions ask for any additional interesting information that can be found on a topic, without repeating any previous answers. Encyclopedia entries aim to have the definitional content we want, so we can recast the problem as trying to find the Wikipedia content in the newspaper articles.

<table>
<thead>
<tr>
<th>Run</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure ($\beta = 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>0.125</td>
<td>0.185</td>
<td>0.132</td>
</tr>
<tr>
<td>2005</td>
<td>0.315</td>
<td>0.123</td>
<td>0.155</td>
</tr>
<tr>
<td>2005+wiki</td>
<td>0.237</td>
<td>0.163</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Table 2: 2004—last year’s results; 2005—improved system, same lists; 2005+wiki—improved system, adding lists from Wikipedia

Why then is the difference so small (non-significant), given such a huge gain in knowledge of answer types? The lists we knew best weren’t the lists being asked about. Over the 75 questions, only 83 Wikipedia lists were used at all, and most of these were either augmenting existing lists (“countries by continent” duplicates “countries”) or incorrectly used. Only 3 lists contributed meaningfully to the system.

Definition results improved qualitatively with Wikipedia strategies, but reliable quantitative evaluation is elusive.

Our preliminary results indicate that Wikipedia has improved QA recall through synonymy, and QA precision by guiding our search for answers in the corpus.

4. Future Work

The Wikipedia has untapped structure that promises to be useful:

- Use structured information from tables and fields.
- Build class-member relations from Wikipedia categories.
- Use class-member relations from Wikipedia for other tasks like anaphora resolution, duplicate detection, disambiguation, and query expansion.
- Incorporate Wikipedia into our factoid system.

References


1 official judgements will not be available until November