1. Introduction

A system that answers questions automatically with just the right information will return answers of the correct type. The question itself often specifies the answer type it expects. When answering a question like “What flowers did Van Gogh paint?”, we prefer answers that are members of the class of flowers, e.g. tulips, begonias, or sunflowers, over other things Van Gogh might paint, like “watercolors”.

The answer type is also a poor search term: for the question “What countries produce coffee?”, answers may be of the form “Brazil produces coffee”, far from any mention of the term “country”.

2. Related Work

The first Answer Type Identification systems used named entity categories to require, for example, person or organization types for “Who” questions, and date or time types for “When” questions (Hirschman & Gaizauskas, 2001).

Today’s top systems use answer type taxonomies with hundreds to thousands of entries (Voorhees, 2004; Hovy et al., 2001; Hovy et al., 2002; Katz et al., 2003; Katz et al., 2004), allowing systems to answer questions that, for example, start with “Which countries” or “What president”.

Separately, others have explored automatic identification of hypernym–hyponym (henceforth “class”–“member”) relations in large bodies of text (Hearst, 1992; Carballo, 1999; Gruenstein, 2001; Snow et al., ; Pantel, 2005). Phrases in English like “corn, wheat, and other staple crops” identify corn and wheat as members of the class of “staple crops”. For the most part, these techniques have been used to augment WordNet, a machine-readable lexicon of English that encodes, among other things, these class–member relations.

Harabagiu et al. (Pasca & Harabagiu, 2001) observed that answer types can be identified with WordNet classes, and their members can be used as a set of possible answers. Fleischman (Fleischman et al., 2003) used automatic class identification to answer “Who is P?” with the automatically extracted classes of which P is a member. We are the first to put together these two ideas: to automatically identify classes in the same text that we will answer questions on, and to use those classes as answer types. The class information we extract also turns out to be instrumental in a number of related question–answering tasks.

3. Approach

We build on the observation that classes are answer types and their members are good candidate answers, and we use automatic class identification techniques to identify over a million classes and their members. These enable us to accurately answer questions with much more specific answer types than those previously available. Examples include: “Which non–OPEC countries ...” and “Which former Yugoslav President ...”.

3.1 Candidate Classes

Aaron Fernandes extended (Fernandes, 2004) Fleischman et al.’s work (Fleischman et al., 2003) on finding definitions from applying only to person names, to applying to most noun phrases. He generated candidate class–member pairs like those in Figure 1.

3.2 Aggregating Classes

Marton and Tellex then aggregated these class–member candidate pairs π, using a probability of correctness Freq(π) based on the number of times a pair occurred in the corpus, and on the precision p(z) of each pattern z: 1

\[
\text{Freq}(\pi) = \frac{\sum_z p(z) \cdot \text{count}(\pi)}{\sum_z (p(z) \cdot \sum_i \text{count}(i))}
\]

1The numerator is an expectation of times π was correctly seen; i in the denominator iterates over all pairs observed with the pattern z, making the denominator a normalizing constant. p(z) was estimated for each z from at least 100 examples.
Automatic Hypernym Extraction and Automatic Question Answering have been areas of intense study since the 1990s. Question answering is most often applied to a particular body of text. This work lets an automatic system read that text to learn the class information it needs to answer questions about the knowledge within.

References


