Word proximity QA system

Philip Rennert
EC Wise, Inc. Machine Learning Center
408 Saybrooke View Dr., Gaithersburg, MD 20877
phil.rennert@ioip.com

Abstract

This is a question answering system with very little NLP, based on question-category-dependent selection and weighting of search terms, selecting answer strings centered around words most commonly found near search terms. Its performance was medium, with a 0.2 MRR. Certain category strategies may be of interest to other QAers, and are described.

1.0 Question categories

To save bandwidth, I won't repeat the description of the QA task or the box-and-line description of the system; see almost any paper here. Here are the question categories, with performance:

Category	N	Average run MRR	My MRR
???	20	0.151	0.0625
ABBREVFOR	1	0.037	0
ABBREVIATION	5	0.214	0.8
CANNED	11	0.296	0.523
CAPITAL	5	0.512	0.7
FIRST TO	14	0.173	0.107
HOW MANY	4	0.174	0.5
HOW MANY AREA	0		
HOW MANY DIST	8	0.300	0.619
HOW MANY DOSAGE	0		
HOW MANY MASS	2	0.177	0
HOW MANY MONEY	0		
HOW MANY SPEED	3	0.150	0.333
HOW MANY STATED	5	0.244	0.4
HOW MANY TEMP	0		
HOW MANY TIME	5	0.217	0.067
INVENTION	6	0.280	0.367
PLACE WHERE GEO	21	0.216	0.099
PLACE WHERE ORIG	0		
PLACE WHERE PHYS	0		
POPULATION	8	0.298	0.604
STAR OF	0		
SUPERLATIVES	20	0.219	0.2
WHAT EAT	1	0.560	1.0
WHAT IS DESCR	154	0.172	0.134
WHAT IS NAMED	114	0.242	0.110
WHAT PAPER	2	0.217	0.25
WHAT SHOW	1	0.198	0
WHAT SPORT	1	0.291	0.5
WHEN BORN	6	0.264	0.25
WHEN DID	30	0.302	0.262
WHERE BORN	0		
WHERE IS	16	0.445	0.349

WHO DID	22	0.392	0.447
WHO IS	3	0.338	0.167
WHY	4	0.154	0

Average MRR: 0.234
Number of questions: 492

Question categories were defined from the training set of previous TRECs;
some categories didn't occur in TREC 10.

2.0 Category strategies
The basic approach was to remove stopwords from the question and use
all remaining words as search terms, though some were upweighted in
some categories. The documents were viewed as an ordered list of words;
occurrence of a search term awarded points to all words within a certain
radius, typically five words. Points depended on the search term weight,
and slightly on the distance. Stopwords and most punctuation were usually
removed. Points were summed over all retrieved documents; answer strings
were centered on the highest scoring words.
This was the default strategy, used in categories ???, WHAT IS DESCRIBED
(a "What is" question with the object described but not named), and
FIRST TO (Who was the first to...), and it wasn't very successful.
In certain categories, predefined lists of search terms or answer patterns
produced better performance; these are described below.

ABBREVIATION - Look for the letters of the acronym in sequence at the
head of words in sequence. Usually one letter per word (e.g., laser),
but can be more (hazmat).

CANNED - Preloaded lists of US state and president information, and
winners of World Series and Superbowls, downloaded from the Web.
It's backwards to already know the answer and seek it in the test corpus;
I included this category only because it's valid in a real QA system.

HOW MANY - Answer must be a number followed by a unit. Elaborate
Perl regexp to identify a number; preloaded lists of units for
various physical quantities (mass, distance, speed,...); map from
question words to appropriate quantity. For some questions, unit
is stated in question (e.g., How many dogs pull a sled in the Iditarod?; unit is
"dogs"). Seek number-unit string near search terms.

INVENTION - Answer is a proper name, so will be Initcapped. Seek
search terms, word which stems to "invent" or "patent", and Initcap
string close together.

POPULATION - Answer is a number; seek search terms, "population" or "people",
and number close together. As a tiebreak, pick the larger
number; usually news stories describe a population and a subset.

WHAT IS NAMED (or Definition) - Unfortunate excess of these questions
in TREC 10 has been described elsewhere. Postfit strategy, after TREC 10
judging: if word has a WordNet gloss, extract the first two nouns before
the semicolon and add to search terms (if more than one sense, do it for all).
This improved MRR to .38 in this category. Another backwards
already-know-the-answer category; a mismatch with the news story corpus.

WHEN DID - Answer is a date; Perl regexp to define dates; seek them

near search terms. If question starts "When is...", append date from story timeline, to cover holidays and such.

WHERE IS - Preloaded list of proximity terms used with Initcap places, like "near", "neighboring", "border", etc.

WHO DID - Answer is a proper name, so Initcapped. Stem action verbs in question (who built, who killed, ...). In counting most common Initcaps near search terms, combine subset terms (e.g., "Fred Jones" and "Jones").

WHO IS - Preloaded list of occupations/activities to make this person famous (writer, leader, winner, etc.); seek the occupation most commonly found near the name. Stem the occupation words (writer = written = writing, etc.).

3.0 Document retrieval strategy comparison

I wrote my own search engine (mergesort-based), to be able to retrieve documents based on my search terms and weights. I submitted two runs, one with only documents from the PRISE top 50, one with those plus the top 50 from my strategy, interleaved to make a list of 100. The results were statistically no different. I conclude the PRISE top 50 document set is good enough to support the answer-finding strategies used here.

4.0 What's next?

With the advance to exact answers in future TRECs, answer patterns will become essential. I believe question parsing and question term expansion to generate them will also become essential, and proximity most-common strategies will not be effective. Some of the answer patterns used here may remain useful.