Question Answering using the DLT System at TREC 2003

Richard F. E. Sutcliffe, Igal Gabbay, Michael Mulcahy, Kieran White

Documents and Linguistic Technology Group
Department of Computer Science
and Information Systems
University of Limerick
Limerick, Ireland

+353 61 202706 Tel
+353 61 202734 Fax
Richard.Sutcliffe@ul.ie Email
www.csis.ul.ie/staff/richard.sutcliffe URL

1. Introduction

This article outlines our participation in the Question Answering Track of the Text REtrieval Conference organised by the National Institute of Standards and Technology. This was our second year in the track and we hoped to improve our performance relative to 2002. In the next section we outline the general strategy we adopted, the changes relative to last year and the approaches taken to the three question types, namely factoid, list and definition. Following this the individual system components are described in more detail. Thirdly, the runs we submitted are presented together with the results obtained. Finally, conclusions are drawn based on our findings.

2. Outline of System

2.1 Overall Strategy

In common with most other QA systems we use the following general strategy to answer questions:

- **Question analysis**: Process the input query attempting to find its type (e.g. who or colour).
- **Document retrieval**: Formulate a search query based on the results of the previous stage. Use this together with a search engine indexed on the document collection to produce a list of candidate documents which are likely to contain answers to the question.
- **Named entity recognition**: Based on the query type identified in the first stage, search for appropriate named entities (NEs) in the candidate documents which co-occur with terms derived from the query.
- **Answer selection**: Decide which NE (or NEs) should be chosen as the answer.

We now outline how this strategy was adapted to handle the three types of question.

2.2 Factoids

Factoid questions formed the majority (413) of the total of 500 in the collection. They are intended to ask about straightforward pieces of information which can be extracted from free text fairly readily. Our approach to these was based entirely on traditional NEs, i.e. numbers, places, person names and so on.

Question Type	Example Question	Candidate Answer
abbrev	’What does ACLU stand for?’	American Civil Liberties Union
airport_name	’What is the name of the airport in Dallas Ft. Worth?’	Dallas Fort Worth International Airport
colour	’What color is the top stripe on the U.S. flag?’	Red
company	What company manufactures Sinemet?’	Hangzhou MSD Pharmaceutical
currency	’What is the currency of Denmark?’	Kroner
distance	’How far is it from Earth to Mars?’	249 million miles
gen_name	’What is the name of the chart that tells you the symbol for all chemical elements?’	Periodic Table
how_did_die	’How did Cleopatra die?’	asp bite
how_many3	’How many time zones are there in the world?’	24
how_much	’What percent of the nation’’s cheese does Wisconsin produce?	28 percent
length_of_time	' How long is a quarter in an NBA game?’	12 minutes
name_part	' What is Britney Spears' ' middle name?'	Jean
nickname_state	' What is the Bluegrass state?'	Kentucky
population	' What is the population of Iceland?'	275000
sci_name	' What is the scientific name for red ants?'	Solenopsis invicta
speed	' How fast does light travel through space?'	186,000 miles per second
temp	' How hot is the sun?'	two million degrees centigrade
title	' What book did Rachel Carson write in 1962?'	Silent Springs
translat	' How do you say "cat" in the French language?'	chat
unknown	' What passage has the Ten Commandments?'	Exodus
what_city	' What city is Disneyland in?'	Anaheim
what_continent	' What continent is the world' ' s largest dessert on?'	Africa
what_county	' What county is San Antonio Texas in?'	Bexar
what_country	' What country is Aswan High Dam located in?'	Egypt
what_state	' What state was Amelia Earhart born in?'	Kansas
when	' When was "Cold Mountain" written?'	1997
when_date	' What date did the U.S. civil war start?'	April 12th 1861
where	' Where is Mount Olympus?'	Greece
who	' Who created the literary character Phineas Fogg?'	Jules Verne

Table 1: Question Types used in the DLT system. The second column shows a sample question for each type. The third column shows sample answers. These are all of appropriate types for the question but are not necessarily correct. Fifteen question types handled by the system did were not used. They are anthem, atomic_number, atomic_weight, country_religion, element, planet, profession, state_motto, state_nickname (the opposite of nickname_state), what_airport_code, when_interval, when_month, when_week_day, when_year and where_airport.

The type of the question as identified in the first stage was directly mapped onto one or more named entities which were then searched for in the text. For example if the question was identified as being of type what_city then the NEs used were nea_x_us_city and nea_x_non_us_city. The answer to the question was defined to be the NE which scored best by one of two measures. The highest_scoring

method looked at how many query-derived terms co-occurred with the NE. The highest_google method used the World Wide Web (WWW) to predict the NE most likely to be the correct answer by adopting an algorithm similar to that of Magnini et al. (2002).

2.3 Lists

The approach to answering the X list questions was identical to that for factoids except in the answer selection stage. Here, multiple answers were selected based on their exceeding a simple threshold.

2.4 Definitions

In order to answer the Y definition questions, the NE recognition stage was adapted to find instances of simple phrasal patterns based around terms from the query. These had been developed in another project concerned with scientific definitions. All such phrases were extracted during answer selection. The first two stages of processing were the same as for factoids and lists.

In the next section we describe the various components of the system in more detail.

3. DLT System Components

3.1 Summary of Enhancements

Our 2002 TREC system was constructed in a short time frame and was thus very basic. A number of significant extensions were made to the system for this year which we summarise here. Firstly additional query types were added, taking the total from 19 plus 'unknown' up to 43 plus 'unknown'. Examples of the 29 types actually used this year are shown in Table 1 with the remaining fifteen types listed in the caption. Secondly the process of query analysis was completely changed in order to allow a strategy for formulating and subsequently simplifying search expressions to be implemented. Thirdly documents in the Aquaint collection were indexed and searched using a commercial engine (DTSearch, 2000) rather than relying on the NIST TOPDOCS files. Fourthly recognisers for various NEs were added, including one for recognising general names. Finally, an answer re-ranking component using WWW hit counts was developed. These enhancements are described below.

3.1 Query Type Identification

As before this was accomplished using simple keyword combinations and without carrying out syntactic parsing of the query. 24 additional types were added relative to last year. However, many of these did not in fact come up in the test questions (see Table 1). Once again an 'unknown' type was adopted to allow a default strategy for queries not falling into other categories.

3.2 Query Analysis

Following type identification the query is subjected to a detailed analysis to assist in the process of search expression formulation as follows:

- Initial words and phrases are removed;
- Capitalised word sequences and expressions within quotation marks are recognised;
- Alternatives for all-capitals words and initial sequences are computed (e.g. ACLU could also be A.C.L.U. and T.E. could be TE);

Query Type	C	NC	R	X	U	W	Total
abbrev	2	0	0	0	0	2	2
airport_name	1	0	0	0	0	1	1
colour	3	0	1	0	0	2	3
company	3	0	0	0	0	3	3
currency	1	0	0	0	0	1	1
distance	17	3	2	0	0	18	20
gen_name	5	2	0	0	0	7	7
how_did_die	23	0	4	0	0	19	23
how_many3	39	6	4	0	0	41	45
how_much	6	1	1	0	0	6	7
length_of_time	3	0	0	0	0	3	3
name_part	2	0	0	0	0	2	2
nickname_state	1	1	0	0	0	2	2
population	2	1	0	0	0	3	3
sci_name	1	0	0	0	0	1	1
speed	4	1	0	0	0	5	5
temp	5	0	0	0	0	5	5
title	3	1	0	0	0	4	4
translat	3	0	1	0	0	2	3
unknown	51	97	6	1	0	141	148
what_city	16	0	2	0	1	13	16
what_continent	4	0	3	0	0	1	4
what_county	1	0	0	0	0	1	1
what_country	22	2	1	0	0	23	24
what_state	0	2	0	0	0	2	2
when	39	0	3	0	0	36	39
when_date	6	0	0	2	0	4	6
where	1	0	0	0	0	1	1
who	23	9	5	0	0	27	32
	287	126	33	3	1	376	413

Table 2: Results by Query Type. The columns C and NC show the numbers of queries of a particular type which were classified correctly and not correctly. Those classified correctly are then broken down into Right, ineXact, Unsupported and Wrong.

- Alternative formulations for numbers are computed (e.g. Apollo-11 could be Apollo-Eleven);
- Stopwords are removed;
- The tense of any remaining verbs is changed to simple past.

3.3 Search Expression Formulation

Based on the results of query analysis a search expression is composed. All searches are boolean. Remaining query terms (i.e. quotations, capitalised word sequences or individual words, each possibly in disjunction with alternatives) are assigned an importance score using a scheme reminiscent of Magnini et al. (2002). Quotations score 10, capitalised word sequences 9, numbers 8, pure nouns and verbs 7, superlative adjectives 6, pure adjectives 2, pure adverbs 1 and every other term 5. A 'pure' noun is one which can not have any other part-of-speech and similarly for pure verbs etc. Terms are then ordered by increasing score and then joined with AND operators to make a single boolean query. This is then used to search for documents.

3.4 Document Retrieval

Before the runs, the Aquaint collection was indexed by treating each paragraph (marked by a <p> tag) as a separate document. During retrieval a boolean query as formulated in the previous stage is submitted to the search engine and the first *n* matching documents (i.e. paragraphs) are returned. *n* was set to 30 throughout. If no documents are returned the least significant term (i.e. the first) is removed and the search is repeated. This process continues until at least one document is returned or no terms remain.

3.5 Named Entity Recognition

NE recognition is similar to last year and uses our own module which is based on grammars. Some extra types were added including nea_x_title and nea_x_general_name. The former recognises quotes expressions while the latter can recognise capitalised expressions with spurious matches being eliminated using simple heuristics. Following Clarke et al. (2003) queries of unknown type are answered by searching for general names.

3.6 Answer Selection

In order to decide which NE candidate (or candidates in the case of list questions) should be returned, two strategies for answer selection were used. The first is highest_scoring, where we return the NE occurring in a context which matches terms in the query best. The second is highest_google which uses a similar algorithm to Magnini et al. (2002). Specifically for candidate answer we

- Submit the candidate answer to Google with search terms and record the number of hits;
- Submit the candidate answer to Google alone and record the number of hits (it will be many more);
- Divide the first value by the second.

During answer selection for factoid questions the candidate with the highest score is chosen. For list questions a threshold of 0.03 was used.

4. Runs and Results

Two runs were submitted. The first used highest_scoring for answer selection while the second used highest_google. The results of Run 2 are shown in Table 2. The first two columns show the numbers of queries which were classified correctly (C) and incorrectly (NC) broken down by query type. 287 of the 413 factoid questions were classified correctly i.e. 69%. However if unknown and gen_name queries (which are effectively the same) are disregarded the rate of success in recognising known query types is 231 out of 258 which is 90%. As noted before, fifteen query types did not come up at all.

Query Num	Query Text	Answer	Supporting Doc	Text Extract
1925	What did Ozzy Osbourne bite the head off of?	bat	NYT19981111.0479	Maybe this is sacrilege (or at least the devil-metal equivalent), but I've always preferred bat-biting rocker Ozzy Osbourne's solo material to his work with Black Sabbath.
1925	What did Ozzy Osbourne bite the head off of?	bat	APW20000823.0047	One of his most notorious stunts was biting off a bat's head during a 1982 concert in Des Moines, Iowa.
1939	How did Einstein die?'	ruptured abdominal aortic aneurysms'	20000314_NYT	Some 15,000 die from ruptured abdominal aortic aneurysms each year. <new para> Albert Einstein died from one in 1955 when he was 76.
2004	What do the opposite sides of a die add up to?	seven	NYT19981202.0181	For instance, the opposite sides of a die always add to seven.
2037	How did Iowa get its name?	ouaouia	NYT19991026.0143	Iowa (from the Siouan ''ouaouia,'' meaning ''one who puts to sleep''),
2059	How did Chicago get its name?	''she-kag-ong,''	NYT19981008.0164	have gotten its name from the Ojibwa words ''she-kag-ong,'' meaning ''wild onion place.''
2259	How did Hawaii become a state?	NO ANSWER		
2262	How did Cincinnati get its name?	given its name by Arthur St. Clair	NYT19991028.0282	Founded in 1788, it was given its name by Arthur St. Clair, governor of the Northwest Territory.
2262	How did Cincinnati get its name?	Cincinnatus	NYT19991028.0282	The society was named for Cincinnatus, a Roman patriot who returned to his farm after saving his city in battle, and this city, in turn, was named for the society.
2287	What divides Haiti from the Dominican Republic?	NO ANSWER		
2311	How did the Lindy Hop get its name?	Charles Lindbergh	NYT19990820.0230	''Shorty'' Snowden invents the Lindy Hop, a dance craze named for Charles Lindbergh's trans-Atlantic flight.
2313	What does an English stone equal?	14 pounds	NYT20000203.0201	And nearly all English people weigh themselves in ''stones'' (14 pounds) rather than pounds or ... shudder, kilograms.
2334	What did Peter Minuit buy for the equivalent of $24.00?	Manhattan	NYT19991120.0177	New York recalls how Peter Minuit bought Manhattan Island from local Indians for a Dutch song. <new para> Legend has it that Minuit paid $24, possibly in beads and trinkets.
2334	What did Peter Minuit buy for the equivalent of $24.00?	Manhattan	NYT19980723.0209	Everyone knows that, back in 1626, Peter Minuit bought Manhattan from the Indians for $24 worth of trinkets.

Table 3: Examples of Hard Questions at TREC 2003.

The columns marked R, X, U and W show the numbers of answers judged Right, ineXact, Unsupported and Wrong by the NIST assessors. The overall rate of success was thus 33 out of 413 (8%) or 36 out of 413 (9%) including inexact answers.

5. Conclusions

After carrying out a significant number of enhancements to our system relative to last year, the results are still no better. This is a disappointing result. There are several reasons for this. Firstly, the questions appear to be much harder this year. A significant number of the factoid questions could not even in principle be answered by our type of system based mainly on NEs. A selection of hard queries can be seen in Table 3. Secondly, certain queries are conventional in form but the answer is not stated in a way which is easy to find. An example of this is numbers with unexpected spacing as in the answer to Q-1980 of '250 , 000 miles'. Another example is anaphoric reference as in Q-1939. Thirdly, we are missing a lot of categories for 'normal' forms of TREC query such as what things made of, animals, rock bands and pop groups, baseball teams, musical instruments, how often, how late etc.

References

Clarke, C. L. A., Cormack, G. V., Kemkes, G., Laszlo, M., Lynam, T. R., Terra, E. L., & Tilker P.L. (2003). Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002). In E. M. Voorhees and L. P. Buckland (Eds) *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002), Gaithersburg, Maryland, November 19-22, 2002*. NIST Special Publication 500-251. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

DTSearch (2000). www.dtsearch.com .

Magnini, B., Negri, M., Prevete, R., & Tanev H. (2002). Is it the Right Answer? Exploiting Web Redundancy for Answer Validation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia*, 425-432.

Acknowledgement

Many thanks to Ken Litkowski for providing answer patterns and supporting documents for the question collection.