Question Answering using the DLT System at TREC 2004

Richard F. E. Sutcliffe, Igal Gabbay, Kieran White
Aoife O'Gorman, Michael Mulcahy

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. We first provide an outline of the system. We then describe the changes made relative to last year. After this we summarise our results before drawing conclusions and identifying some next steps.

2. Outline of System

2.1 Overall Strategy

In common with most other QA systems, the following stages are at the core of our approach:

- **Question analysis**: Process the input query attempting to find its type (e.g. who or colour) and to identify significant phrases.
- **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 corresponding 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.

2.2 Factoids

Most of the questions in the TREC task are factoid. They ask for straightforward pieces of information which can be extracted from free text fairly readily. Our approach to these is conventional and relies on two simple ideas: Firstly, the required NE type can be predicted from the question itself, and secondly NEs can be identified in the text by simple means which are determined in advance.

Question Type	Number	Example
organisation	26	Rat Pack
person	25	James Dean
man_made_object	5	USS Constitution
animal	2	agouti
substance	2	prions
art_work	1	Tale of Genji
medical_condition	1	cataract
natural phenomenon	1	Hale Bopp comet
political_agreement	1	Good Friday Agreement
scandal	1	Teapot Dome scandal
Total	65	

Table 1: Breakdown of Question Groups by Target Types. This shows an analysis of the 65 targets in TREC 2004 into a representative set of types. As can be seen, the vast majority of targets are organisations (interpreted broadly) and persons.

Two key stages in this process are the formulation of the search query during Document Retrieval, and the choice of correct answer from a list of candidates during Answer Selection.

2.3 Lists

We have not so far devoted much attention to list questions so the approach adopted is simply to treat them as factoids and to return all answers which exceed a threshold.

2.4 Definitions (i.e. type 'Other')

Definitions are approached differently from factoids and lists. Having returned documents which contain the appropriate target phrase, we search for instances of phrasal patterns which can indicate the presence of important information about the target. Each sentence containing such a phrase is then returned. The phrases used were adapted from another project (Gabbay and Sutcliffe, 2004).

2.5 Question Groups

For this year the overall task changed. Previously the object was to answer 500 individual questions which could each be of factoid, list or definition type. No link existed between questions. However, for this year the task was based around groups of questions each concerning an overall topic which was identfied by a marked-up target. There were 65 targets in total and their breakdown by topic can be seen in Table 1. The vast majority (51) concern organisations and persons. Within each group, a question is explicitly marked-up as being of type Factoid, List or Other (i.e. Definition).

While the grouping of questions allows interesting experiments concerning interactions between answers, we did not attempt this. Instead we adopted the simplistic approach of adding the target terms to the end of each query within a group and then proceeding to process the individual questions using the same basic method as last year. As it turned out, this was not a bad strategy.

In the next section we describe the major additions which were made to the system relative to the last TREC.

Question Type	Example Question (Target)	Correct Answer
abbrev_expand	'What does AARP stand for?' (AARP)	American Association of Retired Persons
colour	'What is their gang color?' (Crips)	blue
company	'What record company is he with?' (Fred Durst)	Interscope Records
film	'What film introduced Jar Jar Binks?' (Jar Jar Binks)	Star Wars : Episode I _ The Phantom Menace
how_did_die	'How did he die?' (James Dean)	car crash
how_many	'How many seats are in the cabin of Concords?' (Concorde)	100
how_much_money	'How much is the Sacajawea coin worth?' (Sacajawea)	$1
how_often	'How often does it approach the earth?' (Hale Bopp)	every 3,000 years
how_old	'How old was he when he won the title?' (Floyd Patterson)	21
length_of_time	'How long does one study as a Rhodes scholar?' (Rhodes scholars)	two or three years
nationality	'What is Vilar's nationality?' (philanthropist Alberto Vilar)	American
nickname	'What is her nickname?' (USS Constitution)	Old Ironsides
pol_party	'What is his party affiliation?' (Bashar Asaad)	Baath Party
profession	'What is her occupation?' (Eileen Marie Collins)	space shuttle commander
religion	'What is the religious affiliation of the Kurds?' (Kurds)	Sunni sect
speed	'How fast does the Concorde fly?' (Concorde)	twice the speed the sound
sport	'What sport do they play?' (Harlem Globe Trotters)	basketball
title	'What is their biggest hit?' (the band Nirvana)	Smells Like Teen Spirit
unknown	'What are prions made of?' (prions)	mutated proteins
what_city	'What town was he native of?' (Chester Nimitz)	Fredericksburg
what_country	'What country is he associated with?' (Horus)	Egypt
what_us_state	'What state does he represent?' (senator Jim Inhofe)	Oklahoma
when	'When was he born?' (Franz Kafka)	1883
when_date	'On what date was Bashar Assad inaugurated as the Syrian president?' (Bashar Assad)	July 17, 2000
when_year	'In what year did the PLO condemn him to death?' (Abu Nidal)	1974
where	'Where is its headquarters?' (AARP)	Washington
where_school	'Where do Rhodes scholars study?' (Rhodes scholars)	Oxford
who	'Who is the president or chief executive of Amtrak?' (Amtrak)	George Warrington

Table 2: Question Types used in the DLT system. The second column shows a sample question for each type. The third column shows the correct answer, not necessarily that produced by our system. 27 question types are listed here plus 'unknown'. A further 52 question types handled by the system were not used.

3. DLT System Components

3.1 Summary of Enhancements

Three main changes were made this year. Firstly, new query types were added together with a method for recognising them. This necessitated the development of some new NEs. Secondly, the question analysis stage was refined and this led to changes in document retrieval. Thirdly, we indexed the document collection by sentence rather than by paragraph.

3.2 Query Types and their Identification

Last year's system could recognise 47 question types plus the default type 'unknown'. Instances of 29 types occurred in the test queries. This year, a further 32 types were added, bringing the total up to 79 plus 'unknown'. These types are shown in Table 2, together with examples of each, taken from the 2004 test questions. It is interesting to observe that while last year 29 out of 47 types in the system were actually used in the runs, this year only 28 out of 79 types were used, 14 of which were new. Perhaps the new style of grouping questions makes them less diverse.

3.3 Query Analysis

Following type identification the query is subjected to a detailed analysis to assist in the process of search expression formulation. Last year we attempted to identify capitalised sequences and other simple constructs in order to use these as search phrases. Following work on CLEF this year (Sutcliffe et al., 2004) and a study we undertook concerning the relationship between query terms and effective search terms (White and Sutcliffe, 2004) we decided adopt the CLEF strategy which uses the following steps:

- Tag the query for part-of-speech using Xelda (2003);
- Recognise instances of 9 different constructions;
- Weight these according to their importance;
- Order them according to weight;
- Use the conjunction of these as the initial search expression.

The nine constructions are shown in Table 3 together with their weights and an example of each. Weights are assigned using a scheme reminiscent of Magnini et al. (2002).

3.4 Search Expression Formulation

Based on the results of query analysis a search expression is composed. All searches are boolean. Constructs as identified in the previous stage are ordered by increasing score and then joined with AND operators to make a single boolean query. This is then used as the starting point of a search for documents.

3.5 Document Retrieval

Last year, we indexed Aquaint by <p> tag, i.e. treating each paragraph as a separate document. The use of <p> tags is uneven in the collection so we decided instead to split the entire collection into separate sentences and then to index by these instead. Thus each sentence was considered a separate document.

Construct	Weight	Example
quote	80	"Cold Mountain" *
all_cap_wd	60	AARP
cap_dot_wd	1	U.S. *
cap_nou_prep_det_seq	40	President of the United States
cap_wd_seq	40	Black Panthers
number	20	first
low_adj_low_nou	40	annual revenue
non_cap_nou_seq	40	cataract surgery
wd	20	year

Table 3: Construct Types used in Query Analysis. The second column is the integer weight assigned to the construct and the third shows a sample phrase for the type. All come from this year's queries except those marked with * which did not occur and are thus taken from last year.

This effectively ensures that all search terms must occur in close proximity to each other. Of course, it also means that where a concept is mentioned explicitly and then referred to anaphorically in a local context containing other search terms (and indeed a candidate answer) that such a sentence will not be retrieved and hence the correct answer may be lost. However the White and Sutcliffe (2004) study investigated this point among a number of other related issues for a sample set of TREC queries, and found that it is too infrequent to have a significant impact on our existing system.

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. sentences) are returned. *n* was set to 30 for Run 1 and 100 for Runs 2 and 3. (Results reported here are Run 2.) If no documents are returned in the search, 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.

3.6 Named Entity Recognition

NE recognition is similar to last year and uses our own module which is based on grammars together with some exhaustive lists. Some new NEs were added this year for basic entities such as museums. Following Clarke et al. (2003), queries of unknown type are answered by searching for general names. We attempted to restrict these names relative to last year by disallowing instances of places, person names, mountains, mountain ranges, bodies of water and companies.

3.7 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).

During answer selection for factoid questions the candidate with the best score is chosen, this being either highest_scoring or highest_google. For list questions, up to 20 answers which are of the correct NE type(s) and which co-occur with at least one other query term are returned. Definition answers are not subjected to any selection – the answer is the concatenation of all sentences matching a definition pattern, however many there happen to be.

Query Type	C	NC	R	X	U	W	Total
abbrev_expand	0	1	0	0	0	1	1
company	2	0	0	0	0	2	2
distance	0	2	0	0	0	2	2
film	2	0	1	0	0	1	2
how_did_die	1	0	1	0	0	0	1
how_many	20	0	7	0	0	13	20
how_much_money	2	0	0	0	0	2	2
how_often	1	0	0	0	0	1	1
how_old	3	0	0	0	0	3	3
length_of_time	1	0	0	0	0	1	1
nationality	1	0	0	0	0	1	1
pol_party	1	0	0	0	0	1	1
profession	3	0	1	0	0	2	3
speed	1	0	0	0	0	1	1
sport	2	0	0	0	0	2	2
title	2	0	0	0	0	2	2
unknown	43	20	5	1	0	57	63
what_city	1	0	0	0	0	1	1
what_country	3	0	2	0	1	0	3
what_us_state	1	0	0	0	0	1	1
when	48	0	11	0	0	37	48
when_date	1	0	0	0	0	1	1
when_year	7	0	3	0	0	4	7
where	26	0	4	1	0	21	26
where_school	2	0	0	0	0	2	2
who	31	2	4	0	0	29	33
	205	25	39	2	1	188	230

Table 4: Results by Query Type for Run 2. 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.

4. Runs and Results

Three runs were submitted. The first and second used highest_scoring for answer selection while the third used highest_google. Run 1 used *n*=30 for the (maximum) number of documents analysed while Runs 2 and 3 used *n*=100. Run 2 was the best and the results for it are shown in Table 4 for queries of factoid type only. The first two columns show the numbers of queries which were classified correctly (C) and incorrectly (NC) broken down by query type. A significant number of factoid queries i.e. 43 out of 230 or 19% are in fact unclassifiable by the system because no appropriate category exists for them. If the system assigns 'unknown' to a query appropriately it is technically correct behaviour but of course is unlikely to result in the correct answer being returned. Considering all such assignments as wrong, therefore, the classifier accuracy is 70%. Considering them right when they are technically correct, the classifier accuracy is 89%. As noted above, no fewer than 52 out of the system's 79 query types did not

Query Num	Query Text	Answer	Supporting Doc	Text Extract
12.1	What industry is Rohm and Haas in?	crop protection	XIE19981215.0341	<P> The Crop Protection Association of China (CPAC), which was recently set up in Beijing… </P> <P> The members of CPAC are AgroEvo, American Cyanamid, BASF, Bayer, Dow, DuPont, Elf Atochem, FMC, Monsanto, Novartis, Rhone-Poulenc, Rohm and Haas…
13.3	To what alien race does he belong? (Jar Jar Binks)	Gungan	NYT19990517.0227	<P> JAR JAR BINKS (voice of Ahmed Best): A Gungan, (a species of amphibians on Naboo), he's the cutest of the new computer-generated creatures. Big floppy basset-hound-like ears. Long neck. Kind eyes set well above the rest of his, um, head. Buddies with Qui-Gon and Obi-Wan. </P>
17.2	What kind of cases does it try? (International Criminal Court)	war criminals and instigators of genocide	XIE19970916.0107	<P> JOHANNESBURG, September 16 (Xinhua) -- The Southern African Development Community (SADC) is supporting the establishment of an international criminal court to try war criminals and instigators of genocide, a SADC official said here today. </P>
29.1	Why is the 'Tale of Genji' famous?	the world' s first novel and Japan' s greatest literary achievement	NYT19990527.0336	He was the title character in ` ` Tale of Genji,' ' the nearly 1,000-year-old love story that is sometimes described as the world's first novel and is usually considered Japan' s greatest literary achievement.
46.6	Why did they commit suicide? (Heaven's Gate)	appeared to believe that the Hale-Bopp Comet now streaking across the sky was their ticket to heaven.	XIE19970328.0083	<P> -- The 39 men and women who committed suicide were members of a cult known as Heaven's Gate, the authorities said. They left behind detailed videotapes describing their intention and appeared to believe that the Hale-Bopp Comet now streaking across the sky was their ticket to heaven. </P>

Table 5: Examples of Hard Questions at TREC 2004. All these are classified as Unknown by our system which currently cannot answer 'why' questions or resolve anaphors.

come up at all. Naturally these have all occurred in previous TRECs, suggesting that there may be fewer query types in the new 'grouped question' task of TREC 2004 than were found in previous years.

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 39 out of 230 (17%) or 41 out of 230 (18%) including inexact answers. This is approximately twice the performance of our system last year (8% or 9% including inexact answers). Regarding the list and definition questions, the mean results were F=0.101 and F=0.171 respectively. Last year the scores were 0.034 and 0.133 so this is also a small improvement.

5. Conclusions

Our performance is still poor but doubled since last year. This can be attributed mainly to indexing the document collection by sentence rather than by paragraph, and identifying appropriate search phrases within the query prior to document retrieval. In addition, many extra query types were added to the system and there were a few small refinements such as the tightening of the definition of a general name.

The main weakness of the system is the search component. Next steps should include improvements to the term expansion and query relaxation strategies together with detailed evaluation experiments concerning these. In addition we should take greater advantage of the target field which effectively states the topic of a group of queries. Another area for further work is to reduce the number of unknown queries from their present level of 63 out of 230 i.e. 27%. Some of these could be handled by new question types together with refinements to the query categorisation module. Examples of others which remain very difficult can be seen in Table 5.

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 .

Gabbay, I., & Sutcliffe, R. F. E. (2004). Comparing Scientific and Journalistic Texts from the Perspective of Extracting Definitions. In *Proceedings of the Workshop on Question Answering in Restricted Domains, 25 July, 2004, at the 42nd Meeting of the Association for Computational Linguistics, Forum Convention Centre Barcelona 21-26 July, 2004*, 16-22.

Sutcliffe, R. F. E., Gabbay, I., Mulcahy, M., & O'Gorman, A. (2004). Cross-Language French-English Question Answering using the DLT System at CLEF 2004. In *Proceedings of the Cross Language Evaluation Forum, CLEF 2004, Bath, UK, 16-17 September, 2004*, 305-309.

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.

White, K., & Sutcliffe, R. F. E. (2004). Seeking an Upper Bound to Sentence Level Retrieval in in Question Answering. In *Proceedings of the Workshop IR4QA: Information Retrieval for Question Answering, 29 July 2004, at the 27th Annual International ACM SIGIR Conference, University of Sheffield, UK, 25-29 July 25, 2004*, 59-63.

Xelda (2003). www.temis-group.com .

Acknowledgement

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