Question Answering using the DLT System at TREC 2005

Richard F. E. Sutcliffe[1]

Natural Language Engineering and
Web Applications Group
Department of Computer Science
University of Essex, Wivenhoe Park
Colchester CO4 3SQ, UK

rsutcl@essex.ac.uk

Michael Mulcahy, Kieran White
Igal Gabbay, Aoife O'Gorman

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

Michael.Mulcahy@ul.ie Kieran.White@ul.ie
Igal.Gabbay@ul.ie Aoife.OGorman@ul.ie

1. Introduction

This article summarises our participation in the Question Answering (QA) Track at TREC. Section 2 outlines the architecture of our system. Section 3 describes the changes made for this year. Section 4 summarises the results of our submitted runs while Section 5 presents conclusions and proposes further steps.

2. Outline of System

2.1 Overall Strategy

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.

These steps are very typical of first generation QA systems.

2.2 Factoids

Factoids were the first type of question to appear at TREC and they are still the most frequent, with 362 appearing in the current test collection. Each asks for a single piece of information such as a name or a date. The key to our strategy in processing factoids (in common with most other participants) is to predict the expected type of answer (e.g. a date) from the type of question (e.g. 'When...').

[1] On Sabbatical from University of Limerick.

Construct	Weight	Example
quote	80	82.1 "Howdy Doody Show"
all_cap_wd	60	85.1 NCL
cap_dot_wd	1	108.2 U.S.
cap_nou_prep_det_seq	40	113.4 Hole in the Wall Foundation
cap_wd_seq	40	68.1 Port Arthur
number	20	77.3 first
low_adj_low_nou	40	72.2 foreign city
non_cap_nou_seq	40	67.4 contest
wd (vrb)	20	66.4 lost
wd (adj)	6	85.3 private
wd (adv)	5	* quickly

Table 1: 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 the 2005 queries except the example adverb – none were encountered this year.

2.3 Lists

List questions are flagged as such in the test collection and ask for a series of pieces of information all of the same type. They are essentially factoids where multiple answers are expected and this is the way they are treated in our system. We have so far devoted very little time to lists and our method of approaching them has remained the same: all results are returned whose match score exceeds a fixed threshold.

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

Sadly we have also been able to devote very little time to definition questions. We first retrieve documents which contain the phrase specified in the question. We then search for instances of phrasal patterns which are intended to indicate the presence of important information about the target. Any such phrases are returned in a group as the answer.

2.5 Question Groups

In common with last year, the questions are grouped by topic as expressed by a short phrase such as '1998 Baseball World Series'. There are 140 such topics this year. Each contains one or more factoids, lists and definitions. The markup makes it clear which of these three types each question in the group falls into.

The use of question groups is intended to stimulate research concerning the relationship between answers to different questions on the same topic. However, our approach to them is identical to the one we adopted when questions were presented in a simple list. While anaphors do occur in the questions (e.g. 71.2 (F16): 'How fast can **it** fly?) we do not attempt to resolve this but instead simply add the target (F16 in this case) to the end of the query before processing it.

The next section outlines particular aspects of our system.

3. DLT System Components

3.1 Summary of Enhancements

Due to shortage of time, very little was altered in our system. However, there were two major changes. Firstly, we experimented with query term expansion algorithms in the document retrieval stage. Secondly, we improved the technique used for answer selection.

3.2 Query Types and their Identification

Over the years, we have added more query types to the system. In 2003 there were 47 types plus 'unknown'. In 2004 this rose to 79 plus unknown. This year there are 82 plus unknown, a very small change. However, in 2003, 29 out of the 47 different query types were recognised in the test queries. In 2004, 28 out of the 79 were recognised. This year, 33 out of the 82 were used (see Table 2). Thus, less than half of our query types are actually being used and the majority in fact occur very infrequently.

3.3 Query Analysis

The following steps are carried out on the query:

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

The eleven constructs are shown in Table 1 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

Searches of the document collection use boolean queries. 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.

We tried an experiment in term expansion using the Local Context Analysis (LCA) algorithm of Xu and Croft (2000). The input terms were the phrases identified in the question. The top 200 documents (i.e. sentences) returned from the collection when these phrases were used as a non-Boolean query were used. The parameter θ was set to 0.01. The ten best scoring terms identified in the collection by the LCA algorithm were selected and used to boost the boolean query used in document retrieval.

3.5 Document Retrieval

The entire corpus is split into individual sentences each of which is indexed separately using the Lucene system. Each 'document' retrieved by the system is thus a sentence. The complete query is submitted and the first *n* results found are returned (Lucene orders documents even for boolean queries). *n* is set to 30. If no document is found, the query is relaxed by removing the least significant term and then re-submited. The process continues until results are returned or no further simplification is possible.

Query Type	Classif.		Correct Classification				Incorrect Classification				Total
	C	NC	R	X	U	W	R	X	U	W	
award	1	0	0	1	0	0	0	0	0	0	1
colour	1	0	0	0	1	0	0	0	0	0	1
company	7	3	2	0	0	5	0	0	0	3	10
distance	4	3	2	0	0	2	0	0	0	3	7
film	1	0	0	0	0	1	0	0	0	0	1
how_did_die	3	0	2	0	0	1	0	0	0	0	3
how_many3	39	0	9	0	1	29	0	0	0	0	39
how_much_money	1	0	0	0	0	1	0	0	0	0	1
how_much_rate	2	0	0	0	0	2	0	0	0	0	2
how_old	7	0	0	0	0	7	0	0	0	0	7
language	1	0	0	0	0	1	0	0	0	0	1
name_part	1	0	0	0	0	1	0	0	0	0	1
organisation	2	0	0	0	0	2	0	0	0	0	2
pol_party	1	0	1	0	0	0	0	0	0	0	1
population	1	0	0	0	0	1	0	0	0	0	1
profession	5	0	1	1	0	3	0	0	0	0	5
sci_name	1	0	0	0	0	1	0	0	0	0	1
speed	3	0	1	0	0	2	0	0	0	0	3
team	1	0	0	0	0	1	0	0	0	0	1
title	3	0	2	0	0	1	0	0	0	0	3
tv_network	1	0	0	0	0	1	0	0	0	0	1
what_city	2	0	0	0	0	2	0	0	0	0	2
what_country	8	0	1	0	0	7	0	0	0	0	8
what_mountain	2	0	0	0	0	2	0	0	0	0	2
what_mountain_range	1	0	0	0	0	1	0	0	0	0	1
what_sea	1	0	0	0	0	1	0	0	0	0	1
what_state_us	1	0	1	0	0	0	0	0	0	0	1
when	53	1	18	4	0	31	0	0	0	1	54
when_date	3	1	2	0	0	1	0	0	0	1	4
when_year	7	0	0	0	0	7	0	0	0	0	7
where	39	1	9	2	1	27	0	0	0	1	40
where_school	2	0	1	0	0	1	0	0	0	0	2
who	53	9	9	1	1	42	0	0	0	9	62
unknown	42	44	0	0	0	42	3	0	0	41	86
Total	300	62	61	9	4	226	3	0	0	59	362

Table 2: Results by Query Type for Run 1. 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. Next, those classified incorrectly are also broken down. The final column shows the total number of queries for each type.

3.6 Named Entity Recognition

As previously, NE recognition uses our own module which is based on grammars together with some exhaustive lists. Following Clarke et al. (2003), queries of unknown type are answered by searching for general names.

3.7 Answer Selection

During this stage, each candidate NE found within a returned document is scored and the highest scoring NE is returned as the answer to the question. Scoring is done using a measure which incorporates the number of co-occurring key phrases, their assigned weights and their distance from the NE. The distance between a candidate NE and a key phrase is measured in words, e.g. if the phrase is adjacent to the NE its distance is 1, if one word separates them it is 2 and so on. Certain stop words such as prepositions do not contribute to this distance. The reciprocal of the distance is taken and this is multiplied by the weight assigned to the phrase. The sum of all such values is taken to provide an intermediate score for the NE. The final score is this intermediate score multiplied by the Lucene score assigned to the containing document. Following this process, the highest scoring NE is returned.

4. Runs and Results

Two runs were submited. The first did not use the LCA expansion while the second did. *n* (see earlier) was set to 30 throughout. Run 1 was better and the analysis for it is shown in Table 2. Only those query types which were actually encountered in the test collection are shown. Thus we see that 3 types were used plus unknown. Performance in query classification can be discerned from the first two columns of the table. Of the 362 factoid queries, 300 were classified correctly, i.e. 82.87%. This is a good result given our simple methods and is consistent with previous years.

Note that the types of query are not at all evenly distributed in the test collection. On the contrary, the vast majority are who (62 or 17.13%), when (54 or 14.92%), where (40 or 11.05%), how_many3 (39 or 10.77%). These four categories account for 53.87% of the queries. Following this we drop down to company (10 or 2.76%), what_country (8 or 2.21%), distance, how_old and when_year (each 7 or 1.93%), profession (5 or 1.38%) and when_date (4 or 1.10%). The remaining types each occur less than four times.

Classification accuracy on the top four categories is 85.48% for who, 98.15% for when, 97.5% for where and 100.00% for how_many3. Thus performance on who queries seems markedly worse than the others. Concerning unknown queries, we consider a query to be correctly classified as unknown if should not in fact be classified as any other type in the system. Conversely, classification is incorrect if an existing type should have been used. 42/86 i.e. 48.84% of unknown queries were correctly classified. This means that 42/362 i.e. 11.60% of queries lie outside the designed scope of the system, a surprisingly small figure.

Turning to the 44 queries incorrectly classified as unknown, an analysis can be seen in Table 3. The 44 queries should have been distributed among twenty different query types with between one and four unknown queries being assigned to each type. In other words, mis-classified unknown queries are fairly evenly distributed among the query types of the system.

Qids	Correct Classification of Unknown Question & Examples
	abbrev_expand (4)
111.5	What is the name "AMWAY" short for?
	area (2)
115.3	How large is it?
	award (2)
72.4	What is Bollywood's equivalent of the Oscars?
	company (2)
85.2	What cruise line attempted to take over NCL in December 1999?
	distance (1)
130.3	What is its maximum height?
	how_did_die (2)
99.6	What caused Guthrie's death?
	how_many3 (3)
127.3	What is the enrollment?
	how_much_money (2)
102.3	What was the original estimated cost of the Big Dig?
	length_of_time (1)
140.5	What is the average waiting time for this organization to determine benefits?
	nick_name (3)
78.5	What was his English nickname?
	profession (3)
113.2	What is his second successful career?
	team (3)
100.2	What was Sosa's team?
	title (1)
116.3	What was it originally called?
	element + unknown (1)
87.6	Give the name and symbol for the chemical element named after Enrico Fermi.
	what_city (2)
72.2	From what foreign city did Bollywood derive its name?
	what_country (1)
137.6	Of the two governments involved over Kinmen, which has air superiority?
	when (1)
86.1	Give the month and year that General Abacha had a successful coup in Nigeria.
	where (2)
122.5	From where did he begin his famous ride?
	where_school (1)
76.5	He is an alumnus of which university?
	who (4)
106.5	What is the name of the winning manager?

Table 3: Breakdown of Incorrectly Classified Unknown Questions. The number in brackets following each query type is the count of unknown queries which should have been assigned to that type. Following each type is an example which was mis-classified.

Qids	Type of Difficulty & Examples
	Unclear answer type (17)
70.4	What was the affiliation of the plane?
80.1	Where in the Atlantic Ocean did Flight 990 crash?
83.1	What was the Louvre Museum before it was a museum?
87.5	What is Enrico Fermi most known for?
110.1	What is the mission of the Lions Club?
136.3	What was this person's relationship to the Prophet Mohammad?
	Unusual phrasing (4)
81.5	What was the track attendance for the 1998 Preakness?
82.5	The main puppet character was based on what person?
89.2	On what street are the fields where the Little League World Series is played?
114.2	What is his birth name?
	Complex query (4)
72.3	What is the Bollywood equivalent of Beverly Hills?
85.4	How does NCL rank in size with other cruise lines?
	Definition question (1)
134.1	What is a genome?
	List question (1)
96.1	What materials was the 1998 Olympic torch made of?
	Why question (1)
85.5	Why did the Grand Cayman turn away a NCL ship?
	Fine-grained question (1)
113.2	What is his second successful career?
	General question (1)
117.5	Why is it a problem?

Table 4: Examples of Difficult Questions at TREC 2005. Eight kinds of difficulty are shown with the number found shown in brackets. Examples of each are then listed.

Question answering performance is also summarised for factoids in Table 2. Columns 4-7 show the number of queries of each type which were rated Right, ineXact, Unsupported and Wrong following correct type classification. Columns 8-11 show the same information for queries following incorrect type classification. Overall performance is 61+3/362 i.e. 17.68% as compared to 16.96% last year.

Performance in Run 2 using LCA (not shown in the tables) was 55/362 i.e. 15.19%. We carried out an analysis of the results to see where the differences lay. In particular, were there any queries where LCA returned the correct answer in Run 2 where the answer in Run 1 was incorrect? It turns out that there are only three such queries: 68.1, 81.3 and 123.3. Conversely there are twelve queries for which LCA causes a correct answer to be lost: 82.5, 95.2, 100.2, 101.1, 111.2, 116.5, 119.2, 124.4, 127.1, 129.2, 131.3 and 134.4. Overall then, Run 2 is better by three and worse by twelve, i.e. worse by nine overall.

A final analysis we carried out was of 'difficult' questions which are either impossible or very difficult for a system to answer. A breakdown of these is shown in Table 4. The most frequent category is those queries where the type of the answer is unclear from the question. 17 of these were observed in the test collection. A typical example is 110.1 'What is the mission of the Lions Club?'. In this case we have no way of knowing what a mission is before inspecting the text. Of course, it could be a motto, an idea a theme and so forth, but these could be expressed in many ways and are not really amenable to an NE-based approach. Four queries used unusual phrasing, e.g. 81.5 'What was the *track attendance* for the 1998 Preakness?'. Here, track attendance is presumably the number of horses which took place but this is a very rare way of expressing this information. Four queries expressed complex notions, e.g. 72.3 'What is the Bollywood equivalent of Beverly Hills?'. Equivalence is a nebulous concept and once again it would be necessary to read and comprehend the text before understanding it. There was a definition question, a list questio and a why question. These are not strictly factoids. There was a fine-grained question 113.2 'What is his second successful career?'. This is very complex to answer since several careers for the one person are being discussed simultaneously. Finally, there was a why question which was also very general: 117.5 'Why is it a problem?'.

5. Conclusions

This year, very little time was available for our experiments and in consequence the system we used was very similar to last year. Results were also similar with an improvement on factoids of less than 1%. One major addition was the use of query expansion using Local Context Analysis (LCA). While this method is definitely very good at returning contextually related terms which moreover are guaranteed to be within the vocabulary of the corpus, our experiments did not show an improvement in QA performance, at least for the particular way in which we used it to boost certain results in Boolean searches. We need to carry out further analysis to find out exactly why this is so.

Two general themes running through our results over the years are, firstly that classification accuracy using simple keyword method is surprisingly good at 80% or more (this year 82.87%). This means that classification is not limiting the performance of our system. Secondly, queries are not evenly distributed over query types. We noticed this particularly at NTCIR this year where our performance was still 14.00% in a system which only had twelve types plus unknown. It follows from this that with limited time available, we should devote our attention to the small subset of the query types which occur frequently.

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.

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

Xu, J. and Croft, W. B. (2000). Improving the Effectiveness of Information Retrieval with Local Context Analysis, *ACM Transactions on Information Systems* **18**(1): 79-112.

Acknowledgement

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