# Question Answering using the DLT System at TREC 2006

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

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

Richard.Sutcliffe@ul.ie Kieran.White@ul.ie
Igal.Gabbay@ul.ie Michael.Mulcahy@ul.ie

## 1. Introduction

This article summarises our participation in the Question Answering (QA) Track at TREC 2006. 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

As in previous years, 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.

## 3. DLT System Components

### 3.1 Summary of Enhancements

Relative to last year there were only two changes. Firstly, instead of using the Xelda tagger we used the Connexor (2006) parser and extracted part-of-speech tags from the parse trees. This was because Xelda was not available at Essex. Secondly, instead of including the Topic in addition to the Question for both document retrieval and answer identification we experimented with the use of just the Topic information

| Type | Count |
|---|---|
| person | 20 |
| event | 18 |
| organisation | 13 |
| miscellaneous | 11 |
| monument | 6 |
| company | 5 |
| tv_show | 2 |
| **Total** | 75 |

**Table 1: Breakdown of 75 Question Groups in TREC 2006.** This is an approximate classification.

for the first stage and just the Answer information for the second. This was in Run 1, with Run 2 using both texts in both stages as was previously the case.

## 3.2 Query Types and their Identification

Our system can recognise 82 different query types plus unknown, but few of these are used in practice, especially with the more homogenous style of questions in the new question groups. This year, 34 question types were actually used in answering questions (see table) exactly the same number as last year. In addition a further seven of the question types should have been used (mostly where the system incorrectly assigned the type 'unknown'.

## 3.3 Query Analysis

Exactly as last year (except use of Connexor), the following steps are carried out on the query:

- Parse the query using the Connexor parser;
- Extract part-of-speech information and convert to Xelda tag set;
- 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.

## 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.

|  | Classif. |  | Correct Classification |  |  |  | Incorrect Classification |  |  |  | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Query Type | C | NC | R/L | X | U | W | R/L | X | U | W |  |
| unknown | 36 | 47 | 3 | 1 | 0 | 33 | 0 | 0 | 0 | 47 | 83 |
| who | 56 | 5 | 8 | 2 | 2 | 44 | 0 | 0 | 0 | 5 | 61 |
| how_many3 | 56 | 0 | 8+2 | 0 | 0 | 46 | 0 | 0 | 0 | 0 | 56 |
| when_year | 32 | 3 | 4+1 | 0 | 1 | 26 | 0 | 0 | 0 | 3 | 35 |
| when | 27 | 3 | 5 | 2 | 0 | 20 | 0 | 0 | 0 | 3 | 30 |
| where | 24 | 0 | 2 | 4 | 1 | 17 | 0 | 0 | 0 | 0 | 24 |
| when_date | 18 | 0 | 1 | 2 | 0 | 15 | 0 | 0 | 0 | 0 | 18 |
| what_city | 17 | 0 | 2 | 0 | 0 | 15 | 0 | 0 | 0 | 0 | 17 |
| what_country | 12 | 1 | 2 | 0 | 0 | 10 | 0 | 0 | 0 | 1 | 13 |
| distance | 7 | 3 | 2 | 0 | 0 | 5 | 0 | 0 | 0 | 3 | 10 |
| how_old | 9 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 9 |
| title | 5 | 3 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 3 | 8 |
| film | 5 | 0 | 1 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 5 |
| how_much_money | 4 | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 4 |
| when_month | 3 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 |
| company | 3 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 |
| what_state_us | 3 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 |
| how_much_rate | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| name_part | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| pol_party | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| tv_network | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| anatomy | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| animal | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| colour | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| how_did_die | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| how_often | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| organisation | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| profession | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| sport | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| what_continent | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| what_county_us | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| what_island | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| what_mountain_range | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| where_school | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| **Totals** | 337 | 66 | 47 | 12 | 4 | 275 | 0 | 0 | 0 | 66 | 403 |

**Table 2: 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+Logically Correct, ineXact, Unsupported and Wrong. Next, those classified incorrectly are also broken down. The final column shows the total number of queries for each type.

### 3.5 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.

### 3.6 Named Entity Recognition

We have our own module for NE recognition which uses a mixture of simple grammars and lists. Because of the changes in question content and form in recent years, few of the lists are in fact used. 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 submitted. In Run 1, we used the Target information to search for documents and then used the Query to identify terms in retrieved documents (sentences) in the vicinity of candidate NEs. In Run 2, both Target and Query were used in both stages. The results are shown in Table 2. Concerning classification accuracy, 337 out of 403 queries were classified correctly, i.e. 83.62%. The figure for last year was very similar. However in considering such figures we need to bear in mind that this includes the 'correct' classification of 36 queries as 'unknown'. This is effectively rewarding the system for saying that it can not classify a query. The breakdown of query types in Table 2 is illuminating. The first two columns show correct and incorrect classification and the rows are by decreasing frequency of query type. Essentially a very small number of types (for example nine) account for a very high proportion of the queries (337 out of 403 factoids). This means that very few query types need really be processed by a QA system because the TREC query collections are much more homogenous than they used to be in the days of state mottos and baseball scores. As 36 of the 403 queries are correctly classified as 'unknown', we can interpret this as meaning that 8.93% of queries are totally outside the scope of the system. However, the true figure is probably much higher than that.

The performance of Run 2 was better than Run 1 so only the former is summarised in Table 2. Overall QA performance was 44 Right out of 403 i.e. 10.92%. This is much worse than the figure of 17.68% achieved last year. Even the lenient figure this year of 59 out of 403 (44R+3L+12X) is only 14.64% compared to 20.17% last year. This year there were 12 ineXact answers compared to 9 in Run 1 last year. This is an increase but it is unlikely to be significant. By contrast, other groups have reported a large increase in the number of ineXact judgements.

| |
|---|
| **141.1:** What position did Moon play in professional football? ... **141.2**: WHERE did Moon play in college?<br>*Ellipsis in second question relative to first* |
| **147.4:** Where was Edward in line for the throne at the time of the wedding?<br>*Question-convolution of a rare idiom* |
| **151.6:** What is considered the minor league for the Winston Cup series?<br>*Unclear question* |
| **156.6:** Who holds the record of career victories in NASCAR?<br>*Unclear* |
| **162.3:** What other form of treatment has been used for multiple myeloma?<br>*Very open-ended* |
| **163.4:** What is the size of the Hermitage Museum collection?<br>*How do you quantify the answer?* |
| **166.3:** How may chickens were slaughtered to stop further spread of the disease to humans?<br>*Typographic error!* |
| **168.4:** What is his (Prince Charles') usual painting medium?<br>*Hard to understand without specialised knowledge* |
| **168.5:** What are his usual subjects?<br>*Ditto* |
| **169.6:** What is the oldest stone circle in the UK?<br>*Hard to predict answer type* |
| **188.2:** What is the FAT CONTENT of an avocado?<br>*Must understand the concept* |
| **198.4:** What is the claimed primary purpose of this facility?<br>*When is something a purpose and how do you discern it?* |
| **206.3:** How much water fell on Johnstown?<br>*Hard to predict the answer – inches of rainfall perhaps?* |

**Table 3: Examples of Difficult Questions at TREC 2006.**

We provide a couple of tables giving further information on the queries this year. Table 1 shows an approximate breakdown of the 75 question groups into seven categories. Not surprisingly, most of the groups concern persons, events and organisations. Our processing of the main NEs concerning these is very primitive by today's standards.

Table 3 shows examples of some of the ’difficult’ questions this year. The problem in most cases is that the precise question is hard even for a person to understand and the form of the answer is difficult to predict.

## 5. Conclusions

Very little time was available for our experiments this year and the results were poor even by comparison with last year. We plan to devote much more time to next year's system.

## 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.

Connexor (2006). www.connexor.com .

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.