Cross-Language French-English Question Answering using the DLT System at CLEF 2005

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

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

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; I.2 [Artificial Intelligence]: I.2.7 Natural Language Processing

General Terms

Question Answering

1. Introduction

This article outlines the participation of the Documents and Linguistic Technology (DLT) Group in the Cross Language French-English Question Answering Task of the Cross Language Evaluation Forum (CLEF).

2. Architecture of the CLEF 2005 DLT System

2.1 Outline

The basic architecture of our factoid system is standard in nature and comprises query type identification, query analysis and translation, retrieval query formulation, document retrieval, text file parsing, named entity recognition and answer entity selection.

2.2 Query Type Identification

As last year, simple keyword combinations and patterns are used to classify the query into a fixed number of types. Currently there are 69 categories plus the default 'unknown'.

2.3 Query Analysis and Translation

This stage is almost identical to last year. We start off by tagging the Query for part-of-speech using XeLDA (2004). We then carry out shallow parsing looking for various types of phrase. Each phrase is then translated using three different methods. Two translation engines and one dictionary are used. The engines are Reverso (2004) and WorldLingo (2004) which were chosen because we had found them to give the best overall performance in various experiments.

The dictionary used was the Grand Dictionnaire Terminologique (GDT, 2004) which is a very comprehensive terminological database for Canadian French with detailed data for a large number of different domains. The three candidate translations are then combined – if a GDT translation is found then the Reverso and WorldLingo translations are ignored. The reason for this is that if a phrase is in GDT, the translation for it is nearly always correct. In the case where words or phrases are not in GDT, then the Reverso and WorldLingo translations are simply combined.

Question Type	Example Question	Translation
who	0018 'Qui est le principal organisateur du concours international "Reine du futur" ?'	Who is the main organizer of the international contest "Queen of the Future"?
when	0190 'En quelle année le président de Chypres, Makarios III est-il décédé ?'	What year did the president of Cyprus, Makarios III, die?
how_many3	0043 'Combien de communautés Di Mambro a-t-il crée ?'	How many communities did Di Mambro found?
what_country	0102 'Dans quel pays l'euthanasie est-elle autorisée si le patient le souhaite et qu'il souffre de douleurs physiques et mentales insupportables ?'	In which country is euthanasia permitted if requested by a patient suffering intolerable physical or mental pain?
how_much_rate	0016 'Quel pourcentage de personnes touchées par le virus HIV vit en Afrique ?'	What percentage of people infected by HIV lives in Africa?
unknown	0048 'Quel contrat a cours de 1995 à 2004 ?'	Which contract runs from 1995 to 2004?

Table 1: Some of the Question Types used in the DLT system. The second column shows a sample question from this year for each type. Translations are listed in the third column.

The types of phrase recognised were determined after a study of the constructions used in French queries together with their English counterparts. The aim was to group words together into sufficiently large sequences to be independently meaningful but to avoid the problems of structural translation, split particles etc which tend to occur in the syntax of a question, and which the engines tend to analyse incorrectly.

The structures used were number, quote, cap_nou_prep_det_seq, all_cap_wd, cap_adj_cap_nou, cap_adj_low_nou, cap_nou_cap_adj, cap_nou_low_adj, low_nou_low_adj, low_nou_prep_low_nou, low_adj_low_nou, nou_seq and wd. These were based on our observations that (1) Proper names usually only start with a capital letter with subsequent words uncapitalised, unlike English; (2) Adjective-Noun combinations either capitalised or not can have the status of compounds in French and hence need special treatment; (3) Certain noun-preposition-noun phrases are also of significance.

As part of the translation and analysis process, weights are assigned to each phrase in an attempt to establish which parts are more important in the event of query simplification being necessary.

2.4 Retrieval Query Formulation

The starting point for this stage is a set of possible translations for each of the phrases recognised above. For each phrase, a boolean query is created comprising the various alternatives as disjunctions. In addition, alternation is added at this stage to take account of morphological inflections (e.g 'go'<->'went', 'company'<->'companies' etc) and European English vs. American English spelling ('neighbour'<->'neighbor', 'labelled'<->'labeled' etc). The list of the above components is then ordered by the weight assigned during the previous stage and the ordered components are then connected with AND operators to make the complete boolean query. This year we added a component which takes as input the query terms, performs Local Context Analysis (LCA) using the indexed document collection and returns a set of expansion terms. LCA can find terms which are related to a topic by association. For example if the input is 'Kurt Cobain' one output term could be 'Nirvana'. These terms are added to the search expresson in such a way that they boost the relevance of documents which contain them without their being required.

2.5 Document Retrieval

A major change this year was the adoption of the Lucene (2005) search engine instead of DTSearch (DTSearch, 2000). Lucene was used to index the LA Times and Glasgow Herald collections, with each sentence in the collection being considered as a separate document for indexing purposes. This followed our observation that in most cases the search keywords and the correct answer appear in the same sentence. We use the standard query language.

In the event that no documents are found, the conjunction in the query (corresponding to one phrase recognised in the query) with the lowest weight is eliminated and the search is repeated.

2.6 Text File Parsing

This stage is straightforward and simply involves retrieving the matching ’documents’ (i.e. sentences) from the corpus and extracting the text from the markup.

2.7 Named Entity Recognition

Named Entity (NE) recognition is carried out in the standard way using a mixture of grammars and lists. The number of NE types was increased to 75 by studying previous CLEF and TREC question sets.

2.8 Answer Entity Selection

Answer selection was updated this year so that the weight of a candidate answer is the sum of the weights of all search terms co-occurring with it. Because our system works by sentence, search terms must appear in the same sentence as the candidate answer. The contribution of a term reduces with the inverse of its distance from the candidate.

2.9 Temporally Restricted Questions

This year an additional question type was introduced, temporally restricted factoids. We did not have time to make a study of this interesting idea so instead we simply processed them as normal factoids. Effectively this means that any temporal restrictions are analysed as normal syntactic phrases within the query, are translated and hence become weighted query terms. As with all phases, therefore, the weight assigned depends on the syntactic form of the restriction and not on any estimate of its temporal restricting significance. This approach was in fact quite successful (see results table and discussion).

2.10 Definition Questions

This year 50 definition questions were included in the set of 200 queries with the remaining 150 being factoid (some temporally restricted, some not). At no stage have we made any study of these questions. For TREC we developed a very primitive component and so this was simply incorporated into the present system. Queries are first classified as def_organisation, def_person or def_unknown. The target is identified in the query (usually the name of an organisation or person). For an organisation query, a standard list of phrases is then added to the search expression, each suggesting that something of note is being said about the organisation. Example phrases are ‘was founded’ and ‘manufacturer of’. All sentences including the target term plus at least one significant phrase are returned. These are concatenated to yield the answer to the question. This approach does work on occasion but the result is rarely concise. For def_person queries the method is the same, but using a different set of phrases such as ‘brought up’, ‘founded’ etc. If the categoriser is unable to decide between def_organisation and def_person, it assigns def_unknown which results in both sets of patterns being used.

3. Runs and Results

3.1 Two Experiments

We submitted two runs which differed only in their use of LCA. Run 1 used it while Run 2 did not.

3.2 Results

Results are summarised by query type in Table 2. Concerning query classification it shows for each query type the number of queries assigned to that type which were correctly categorised along with the number incorrectly categorised. The overall rate of success was 84% which compares closely with the 85% achieved in the same task last year. This figure includes 33 queries which were ‘correctly’ classified as unknown. If these are not included then the figure becomes 67.5%. Effectively, answering these 33 queries (16.5% of the entire collection) lies outside the envisaged scope of the system.

Query Type	Classif.		Correct Classification								Incorrect Classification							
			Run 1				Run 2				Ru n 1				Run 2			
	C	**NC**	**R**	**X**	**U**	**W**	**R**	**X**	**U**	**W**	**R**	**X**	**U**	**W**	**R**	**X**	**U**	**W**
abbrev_expand	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
award	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
company	3	0	0	0	0	3	0	0	0	3	0	0	0	0	0	0	0	0
distance	2	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0
film	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
how_many3	10	3	3	0	0	7	4	0	0	6	1	0	0	2	1	0	0	2
hw_mch_mony	3	0	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0
hw_mch_rate	4	0	0	0	0	4	0	0	0	4	0	0	0	0	0	0	0	0
how_old	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
pol_party	3	0	0	0	0	3	0	0	0	3	0	0	0	0	0	0	0	0
population	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
profession	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
title	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
tv_network	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
what_capital	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
what_city	4	0	2	0	0	2	2	0	0	2	0	0	0	0	0	0	0	0
what_country	5	0	1	0	0	4	1	0	0	4	0	0	0	0	0	0	0	0
when	11	0	4	0	0	7	4	0	0	7	0	0	0	0	0	0	0	0
when_date	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
when_month	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
when_year	4	0	3	0	0	1	3	0	0	1	0	0	0	0	0	0	0	0
where	3	0	2	0	0	1	2	0	0	1	0	0	0	0	0	0	0	0
who	30	0	2	0	0	28	1	0	0	29	0	0	0	0	0	0	0	0
unknown	33	24	5	1	0	27	5	1	0	27	3	0	0	21	3	0	0	21
Subtotals	**123**	**27**	**26**	**2**	**0**	**95**	**26**	**3**	**0**	**94**	**4**	**0**	**0**	**23**	**4**	**0**	**0**	**23**
def_org	20	0	2	2	0	16	2	1	0	17	0	0	0	0	0	0	0	0
def_person	25	0	4	9	0	12	3	8	0	14	0	0	0	0	0	0	0	0
def_unknown	0	5	0	0	0	0	0	0	0	0	0	2	0	3	1	1	0	3
Subtotals	**45**	**5**	**6**	**11**	**0**	**28**	**5**	**9**	**0**	**31**	**0**	**2**	**0**	**3**	**1**	**1**	**0**	**3**
Totals	**168**	**32**	**32**	**13**	**0**	**123**	**31**	**12**	**0**	**125**	**4**	**2**	**0**	**26**	**5**	**1**	**0**	**26**

Table 2: Results by Query Type for 2005 Cross-Language French-English Task. 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 for each of the two runs Run 1 and Run 2. Finally, those classified incorrectly are broken down in the same way.

The performance in Run 1 can be summarised as follows. Taking all queries together (i.e. definitions and both types of factoid), 32 of the 168 queries classified properly were correctly answered. Of the 32 queries not classified properly, 4 were still answered correctly. Overall performance was thus 36 / 200, i.e. 18%. For Run 2, 31 of the 168 classified properly were answered correctly with an additional 5 of the 32 not classified properly still being right. This also gives a figure of 36 / 200, i.e. 18%. Our best figure for last year was in Run 1 where 19% was achieved. However, there were no definition questions in 2004 and this year we were able to devote little or no time to developing a component for these. If we consider just the factoid figures, performance in both runs is 26+4 / 150 i.e. 20%.

In terms of our overall position in the French-English task (see Table 6 in the QA summary paper) we are only in positions 5 and 6 out of 12 with the best performance being DFKI German-English at 25.50%. However, it turns out that the main difference between ourselves and high scoring competitors is in the definition questions where they score well and we do poorly. If we consider the performance in factoid questions, broken down into two types, non-temporally restricted and temporally restricted, our performance in the former is 20.66% in Run 1 and 19.83% in Run 2 while in the latter it is 17.24% in Run 1 and 20.69% in Run 2. This makes Run 1 the best system in the group for non-temporally restricted questions alone, and Run 2 the best equal system with LIRE French-English Run 2 for temporally restricted questions alone.

As mentioned above, we devoted very little time to factoids and hence our very poor result of 6 / 50 correct i.e. 12%. The judgement of definitions was quite strict (we were responsible for it) with any response containing both relevant and non-relevant information being judged as ineXact not Right. This probably explains why the scores assigned to systems in the English target task were lower than in some other tasks.

3.3 Platform

We used a Dell PC running Windows NT and having 256 Mb RAM. The majority of the system is written in SICStus Prolog 3.11.1 (SICStus, 2004) with Part-of-Speech tagging, Web translation and Local Context Analysis components being written in Java.

4. Conclusions

The overall performance was 18% which compares with 19% last year and 11.5% the year before. We were able to do very little work on the system this year and in addition there were 50 definition questions for which we only had a very primitive module inherited from our TREC system. If we exclude definitions, our performance compares more favourably with the other systems with Run 1 being the best system overall for normal factoids and Run 2 being equal best with LIRE for temporally restricted factoids.

Run 1 was our first experiment with Local Context Analysis for term expansion at the document retrieval stage. Informal observations have shown that this method provides very good expansion terms which are semantically related by topic and context. However, these experiments did not show any significant advantage for LCA compared to Run 2 which did not use it. Overall performance of the two runs was identical. Performance on non-temporal factoids was marginally better with the LCA (20.66% vs. 19.83%) but it was worse on temporal factoids (17.24% vs. 20.69%). Further analysis is necessariy to see why this was the case.

Definitions are an interesting category of question and we intend to devote much more time to them next year. We are hoping that the specification of a definition and the precise means by which it can be evaluated will be worked out in the mean time. A major defect of our approach is that it is imprecise. Under our strict scoring, accuracy was only 12%. However, we could easily have considered our inexact answers as correct. This would increase our score from 6 / 50 to 17 / 50, i.e. an improvement from 12% to 34%. To put this another way, if we were to select from the answer sentences more carefully, we could improve our algorithm considerably.

In CLEF generally, performance in the cross-lingual tasks is much lower than in the monolingual ones. One interesting experiment would be to eliminate the translation component from our system, thus making it monolingual, and then to try it on the English version of the same test collection. The level of performance would of course be higher and by measuring the difference we would be able to estimate how much information we are at present losing in the translation stage.

5. References

DTSearch (2000). www.dtsearch.com

GDT (2004) http://w3.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp

Lucene (2005). http://jakarta.apache.org/lucene/

Reverso (2004) http://grammaire.reverso.net/textonly/default.asp

SICStus (2004) http://www.sics.se/isl/sicstuswww/site/index.html

WorldLingo (2004) http://www.worldlingo.com/products_services/worldlingo_translator.html

XeLDA (2004) http://www.temis-group.com/temis/XeLDA.htm