Cross-Language French-English Question Answering using the DLT System at CLEF 2006

Richard F. E. Sutcliffe*[1], Kieran White
Darina Slattery, Igal Gabbay, Michael Mulcahy

Department of Computer Science*
University of Essex, Wivenhoe Park,
Colchester CO4 3SQ, UK

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

rsutcl@essex.ac.uk Kieran.White@ul.ie Darina.Slattery@ul.ie
Igal.Gabbay@ul.ie Michael.Mulcahy@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

Abstract

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. Factoid classification into 69 query types is carried out using keywords. Associated with each type is a set of one or more Named Entities. Xelda is used to tag the French query for part-of-speech and then shallow parsing is carried out over these in order to recognise thirteen different kinds of significant phrase. These were determined after a study of the constructions used in French queries together with their English counterparts. Our observations were that (1) Proper names usually only start with a capital letter with subsequent words un-capitalised, 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. The phrases are then translated into English by the engine WorldLingo and using the Grand Dictionnaire Terminologique, the results being combined. Each phrase has a weight assigned to it by the parser. A Boolean retrieval query is formulated consisting of an AND of all phrases in increasing order of weight. The corpus is indexed by sentence using Lucene. The Boolean query is submitted to the engine and if unsuccessful is re-submitted with the first (least significant) term removed. The process continues until the search succeeds. The documents (i.e. sentences) are retrieved and the NEs corresponding to the identified query type are marked. Significant terms from the query are also marked. Each NE is scored based on its distance from query terms and their individual weights. The answer returned is the highest-scoring NE. Temporarily Restricted Factoids are treated in the same way as Factoids. Definition questions are classified in three ways: organisation, person or unknown. This year Factoids had to be recognised automatically by an extension of the classifier. An IR query is formulated using the main term in the original question plus a disjunction of phrases depending on the identified type. All matching sentences are returned complete. Results this year were as follows: 32/150 (21%) of Factoids were R, 14/150 (9%) were X, 4/40 (10%) of Definitions were R and 2 List results were R (P@N = 0.2). Our ranking in Factoids relative to all thirteen runs was Fourth. However, scoring all systems over R&X together and including Definitions, our ranking would be Second Equal because we had more X scores than any other system. Last year our score on Factoids was 26/150 (17%) but the difference is probably the easier queries this year.

---

[1] On Sabbatical from University of Limerick.

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 2006 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'. In a major change this year, the queries were not tagged in the input file as Factoid, Definition or List. Instead this information had to be inferred. We altered the keyword classifier to recognise Factoids using last year's data for training. We made no attempt to recognise List questions and simply treated them as Factoids. This partly explains our low list score.

2.3 Query Analysis and Translation

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 two different methods. One translation engine and one dictionary are used. The engine is WorldLingo (2004). 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 two candidate translations are then combined – if a GDT translation is found then the WorldLingo translation is 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 WorldLingo translation is used.

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 un-capitalised, 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.

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 last year for each type. Translations are listed in the third column.

2.5 Document Retrieval

Lucene (2005) 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 conjunct 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

Temporally restricted factoids are processed in exactly the same way 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.

2.10 Definition Questions

Queries are classified as def_organisation, def_person or def_unknown during the query classification stage using keywords inferred from last year's data. This is necessary because Definitions are no longer tagged as such in the query file – a significant departure from last year. 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 and it can therefore result in inordinate number of answers being judged ineXact. 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 Runs

This year we submitted just one run.

3.2 Results

The performance can be summarised as follows: 32 out of 150 Factoids were Right (21%) and 14 out of 150 were ineXact (9%). 4 out of 40 Definitions were Right (10%). Unfortunately the count of ineXact answers is for Factoids and Definitions combined. For Lists, 2 Right answers were returned, P@N = 0.2. By comparison, last year our score on Factoids was 26/150 (17%) but the difference is probably that the queries were easier this year.

In terms of our overall position in the French-English task, there were thirteen runs in all and our ranking on Factoids is position 4, based on a simple count of correct answers. However we had a lot of X scores, more in fact than any other submitted run in this task. If we combine R and X and score these over Factoids and Definitions together our position would be Second Equal.

3.3 Platform

We used a Viglen PC running Windows XP and having 1 Gb 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 this year was similar to last. Unfortunately, we were able to do very little work on the system this year. The only real differences in the system were the automatic recognition of Factoids (quite successful), the non-recognition of Lists (which lowered our score for these significantly) and the use of just WorldLingo and GDT instead of these plus Reverso. The last change seemed to make very little difference although we have not yet quantified this.

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/

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