Chapter 1: Introduction

1.1 Outline

This chapter first states the objectives of the project. The next section describes the subject domain and is followed by an outline of the system's architecture. The experiments, their results and key conclusions are then summarised and followed by a guide to other chapters.

1.2 Objectives

An information retrieval system 'informs on the existence (or non-existence) and whereabouts of documents' relating to the request of a user (Lancaster, 1968, cited in van Risjbergen, 1979). On the other hand, a question answering system attempts to allow a user to ask a question in natural language and receive a concise answer, possibly with a validating context (Hirschman and Gaizauskas, 2001).

Questions asking about definitions of terms (i.e., 'What is X?') occur frequently in the query logs of search engines (Voorhees, 2003). However, due to their complexity, recent work in the field of question answering has largely neglected them and concentrated instead on answering factoid questions for which the answer is a single word or short phrase (Blair-Goldensohn, McKeown and Schlaikjer, 2003). Much of this work has been motivated by the question answering track of the Text REtrieval Conference (TREC), which evaluates systems by providing them with a common challenge.

This project was born as a result of the participation of the Documents and Linguistic Technology Group at the University of Limerick in TRECs 11 (2002) and 12 (2003). The idea of focusing on answering definition questions emerged after the organisers of TREC had announced that such questions would be included in the main question answering task of 2003. However, for this thesis, we decided to select a domain other than news reports, which is the one used not only in TREC but also in most other recent research in question answering. The objectives of the project were

- To test the effectiveness of lexical patterns without deep linguistic knowledge in capturing definitions in scientific papers;
- To discover simple features which indicate sentences containing definitions;
- To study qualitatively definitions which match lexical patterns in scientific text;
- To try to improve performance of the system by iterative extension and refinement of the lexical pattern set.

1.3 Subject Domain

We chose the terminology-rich field of salmon fish biology as the research domain. A collection of papers on salmon was created by downloading 1,000 documents from ScienceDirect (ScienceDirect, 2003) and named SOK-i (see Section 3.3 for an explanation of the name which is pronounced Sockeye). We used it as the source of definitions. The test query terms for the first experiment were suggested by salmon researchers. In the subsequent three experiments we extracted a much larger set of terms from the glossary of a fish database on the Web called FishBase (FishBase, 2003).

1.4 System Architecture

We pre-processed the documents in SOK-i by splitting them into sentences using heuristics. We inserted tags to mark a sentence boundary and assigned two numbers to each sentence: the ordinal number of the sentence within the document and the total number of sentences in the document. We used the search engine dtSearch (dtSearch, 2003) to index each sentence as a separate document by means of segmentation rules.

A system was built which operated in the following manner: The input was a term as described in Section 3.4. The term was formed into a Boolean query within a batch script, which was then submitted to dtSearch. The system was set to mark the terms in the top 1,000 documents (sentences) which were retrieved. Next, a set of syntactic rules matched all instances of definition patterns in the marked sentences. The output was a section of a sentence that matched the definition part of a pattern.

We extended the set of lexical definition patterns which we originally created for the definition question subtask in TREC-12. The patterns consisted of the term, a lexical phrase and the definition segment, usually extracted text up to a sentence boundary. For example, when matching the pattern **`TERM is the term for DEF`**, the definition (DEF) returned was the text after the word 'for' and up to the end of the sentence.

In the fourth experiment (see below) we used the XeLDA tagger (XeLDA, 2003) to recognise parts-of-speech, so that we could refer to this in patterns. For example, to match the pattern **`TERM, DEF, VERB`** we had to recognise verbs.

1.5 Experiments

We judged each pattern instance to be either Vital, Okay or Wrong. After the first experiment we added the category Uncertain. When evaluating the answers, we kept in mind a range of hypothetical sophisticated users. In the first and fourth experiments we calculated Average Precision of two types: Strict and Lenient. Average Strict Precision was the proportion of Vital pattern instances in each experiment; Average Lenient Precision was the proportion of Vital Okay and Uncertain pattern instances in each experiment (only Vital and Okay in the first experiment). In all the experiments we computed Average Strict Binary Responsiveness—the proportion of queries for which at least one Vital pattern instance was found—and Average Lenient Binary Responsiveness—the proportion of queries for which at least one Vital, Okay or Uncertain pattern instance was found.

To discover contextual characteristics associated with the usefulness of sentences and documents we gathered the following statistics in the first experiment:

- the ordinal number of the instance of the term in the document (`doc_so_far_term_count`);
- the total number of instances of the term in the document (`doc_total_term_count`);
- the ratio of the above two numbers (`doc_so_far_term_proportion`);
- the length of the sentence containing a pattern (`doc_pattern_instance_sen_length`);

- the total number of sentences in the document (`doc_sentence_number`);
- the position of the sentence containing a pattern (`doc_so_far_sentence_number`);
- the relative sentence position (`doc_so_far_sentence_proportion`) calculated by dividing `doc_so_far_sentence_number` by `doc_sentence_number`.

In the second, third and fourth experiments we tested the system with a larger and broader selection of fish-related terms. In each of these experiments we extended and refined the set of lexical definition patterns based on observations in the preceding experiment. In the fourth experiment we exploited shallow syntactic information to restrict some of the patterns. In this last experiment we associated different sets of patterns with common and uncommon terms in the document.

1.4 Results

In the first experiment, Average Strict Binary Responsiveness was 37.1%, whereas Average Lenient Binary Responsiveness was 68.6% (Lenient Precision). Average Strict Precision and Average Lenient Precision were 1.2% and 18.6% respectively.

The patterns **`TERM, DEF`**, **`TERM (DEF)`**, and **`TERM is DEF`** accounted for about 90% of the answers.

Statistically significant differences were found between the judgement categories in `doc_so_far_term_count` and `doc_total_term_count` but not in `doc_so_far_term_proportion`, `doc_so_far_sentence_number`, `doc_pattern_instance_sen_length`, `doc_sentence_number`, and `doc_so_far_sentence_proportion`.

In the second, third and fourth experiments Average Binary Responsiveness (both Lenient and Strict) was lower than in the first experiment, probably because the terms in the first experiment were more specific to salmon biology. The lower responsiveness could also be due to stricter judgement against the definitions in the FishBase glossary.

However, we improved Average Lenient Binary Responsiveness from 22% in the second experiment to 65% in the fourth experiment.

The greatest improvement was in the Average Precision which was measured in the fourth experiment: Average Strict Precision was 7.2% (compared to 1.2% in the first experiment) and Average Lenient Precision was 63.6% (18.6% in the first experiment).

Definitions were not limited to the Introduction and Abstract sections of the documents. Some definitions were even found in the References section (e.g., `Canthaxanthin: a pigmenter for salmonids`).

Despite the homogeneity and small size of the document collection, the definitions that were retrieved might satisfy users who possess different levels of expertise.

1.5 Key Conclusions

- Improvements to the system should focus on eliminating wrong answers which matched the most common terms;
- Elimination of wrong answers should take into account features of scientific writing such as frequent citations and use of tense;
- The distribution of patterns in the answers which were retrieved from the SOK-i collection suggests it was different from the distribution in the news domain frequently used in question answering;
- `doc_so_far_term_count` was a better indication of a sentence with a Vital or Okay answer than `doc_so_far_sentence_number`, which is a feature used by others working in the news domain (Joho and Sanderson, 2000; Blair-Goldensohn et al., 2003).
- Adding patterns and simple elimination rules can improve Lenient Precision, but Strict Precision is likely to remain low when retrieving definitions from a small document collection;
- Evaluation is a difficult task and requires more discerning judgement categories to reduce the number of answers classified as Uncertain.

1.6 Guide to Other Chapters

Chapter 2: Literature Review presents a general background to the research field of question answering and the associated track in TREC. It summarises previous work on answering definition questions, and provides some theoretical background on definitions.

Chapter 3: Domain of Application justifies the choice of fish-related terminology and salmon in particular as the research domain. It also describes the creation of the test document collection (SOK-i) and explains how the query terms were obtained.

Chapter 4: Implementation of the System introduces the DLT question answering system that served as the starting point for this project and then reports on the pre-processing of the documents in SOK-i and the modification of the system for the purpose of our experiments.

Chapter 5: Experiments and Results describes how answers were evaluated and presents the four experiments and their results. The findings of each experiment are discussed.

Chapter 6: Conclusions summarises the project and suggests steps for further research.