Chapter 6: Conclusions

6.1 Outline

This chapter first summarises the findings of our study and then suggests steps for further research.

6.2 Key findings

Our aim was to study the effectiveness of lexical patterns in capturing definitions in scientific text and to establish the characteristics of such definitions. The study involved running four quantitative and qualitative experiments using different sets of query terms and system configurations. Below are our key findings:

- Testing the system with query terms which were selected beforehand to match the theme of the document collection (salmon fish) achieved Average Binary Strict Responsiveness of 37% and Average Lenient Binary Responsiveness of 68%. These values was reduced considerably when we used a much larger set of terms from the glossary of the fish species database FishBase, many of which were general yet still in the fish domain;
- Improving the system focused on eliminating wrong answers which match the dominant patterns **`TERM, DEF`**, **`TERM (DEF),`** and **`TERM is DEF`**;
- The distribution of definition patterns in the SOK-i collection appears to be different from that in the news domain often used in question answering research;
- Elimination of further wrong answers should involve consideration of special features of scientific writing such as citations and tense;
- Adding patterns and simple elimination rules improved Average Strict Precision from 1.2% in the first experiment to 7.2% in the fourth experiment. Average Lenient Precision improved from 8.6% in the first experiment to 63.6% in the fourth experiment. Average Lenient Binary Responsiveness improved from 22%

to 60% over the three experiments using FishBase terms. Average Strict Binary Responsiveness remained low at under 10%;

- A lower `doc_so_far_term_count` was a better indicator of a sentence with a Vital or Okay answer than `doc_so_far_sentence_number` and `doc_so_far_sentence_proportion`;
- Definitions were not only found in the Introduction or Abstract of scientific papers. Some were even extracted from the References section;
- Definitions which were suitable for different levels of expertise of users were found in the collection.
- Our patterns captured hypernyms and hyponyms. Their usefulness may depend on the user's goal and expertise but even when too general, they may contribute to the future generation of a coherent response;
- Extraction of a definition in a sentence may not be necessary because users may prefer to see it in paragraph-sized context (Lin et al., 2003). When evaluating our extracts we often consulted the original document to judge the answer.
- Evaluation of definition answers is difficult even in a restricted domain. Expert users may find the same answer Vital or Okay in different circumstances.

6.3 Future Steps

Extracting definitions is a new area of research within question answering, and the scope for further research is wide. Our approach was basic, but the output of our system allowed us to study the topic and develop the following ideas for further research:

- Improving evaluation by adding more precise intermediate judgement categories so as to eliminate or reduce the need for Uncertain category;
- Testing definitions on real users by similar methods to the ones used in Klavans and Muresan (2001) and Blair-Goldensohn, McKeown and Schlaikjer (2003);
- Exploiting the rather uniform IMRD (Introduction, Methods, Results, Discussion) structure of scientific articles: typifying definitions in each section, using sentence position *within* each section to indicate the likelihood for definitions, determining the precision in each section and using the result to weight definitions from different sections;

- Correlating definition type (e.g., stipulative, lexical, operational, precising) with specific patterns and the IMRD sections. This could enable the users to specify their level of expertise, allowing the system to respond by only returning answers to the appropriate definition type;
- Classifying of terms according to part-of-speech or predefined categories. For example, a separate strategy or set of patterns may be developed for terms which are adjectives;
- Creating information extraction templates for certain categories of biological term and using auxiliary questions, similar to those used by Prager et al. (2003). For example, for the category Species we may extract its scientific name (the pattern **`TERM (DEF`** worked well in our experiments when the query term was a common name of a species), the taxonomic family or order, distribution, life cycle, synonym, and threats to the species;
- Adding syntactic patterns which detect defining verbal phrases and developing rules to filter false matching;
- Introducing patterns which match definition-by-cause and definition-by-exploration-of-origin;
- Ranking and combining answers to remove duplicates and generate a coherent response;
- Adopting simple anaphoric resolution in order to detect definitions containing words like 'it' or 'the fish'.

6.4 Summary

This chapter briefly described the findings related to the effectiveness of our lexical definition patterns, location of definitions in documents, improvement of performance and evaluation. We suggested areas for further work, including term classification, exploitation document structure, generation of responses, and use of an information extraction approach.