Abstract. This article presents Ædipe, the question answering system that was used by the LIC2M for its participation in the CLEF-QA 2005 evaluation. The LIC2M participates more precisely in the monolingual track dedicated to the French language. The main characteristic of Ædipe is its simplicity: it mainly relies on the association of a linguistic pre-processor that normalizes words and recognizes named entities and the principles of the Vector Space model.

1 Introduction

Question Answering is at the edge of Information Retrieval and Information Extraction. This position has led to the development of both simple approaches, mainly based on Information Retrieval tools, and very sophisticated ones, such as [1] or [2] for instance, that heavily rely on Natural Language Processing tools. Previous evaluations in the Question Answering field have clearly shown that high results cannot be obtained with too simple systems. However, it still seems unclear, or at least it is not shared knowledge, what is actually necessary to build a question answering system that is comparable, in terms of results, to the best known systems. This is why we have decided to adopt an incremental method for building Ædipe, the question-answering system of the LIC2M, starting with a simple system that will be progressively enriched. Ædipe was first developed in 2004 for the EQUER evaluation [3] about question answering systems in French. It was designed mainly for finding passage answers and its overall design was not changed for its participation to the French monolingual track of CLEF-QA 2005. The main adaptation we made for CLEF-QA was the addition of a module that extracts short answers in passage answers for definition questions.

2 Overview of the Ædipe System

The architecture of the Ædipe system, as illustrated by Figure 1, is a classical one for a question answering system. Each question is first submitted to a search engine that returns a set of documents. These documents first go through a linguistic pre-processor to normalize their words and identify their named entities.
The same processing is applied to the question, followed by a specific analysis to determine the type of answer expected for this question. This search is performed through three levels of gisting: first, the passages that are the most strongly related to the content of the question are extracted from the documents returned by the search engine. Then, the sentences of these passages that are likely to contain an answer to the question are selected. These sentences can also be considered as passage answers. Finally, minimal-length answers are extracted from these sentences by locating their phrases that best correspond to the question features.

3 From Documents to Passages

3.1 LIMA

LIMA [4], which stands for LiC2m Multilingual Analyzer, is a modular linguistic analyzer that performs text processing from tokenization to syntactic analysis for 6 languages. More precisely, for CLEF-QA, the linguistic analysis of both documents and questions relied on the following modules:

- tokenizer
- morphological analysis

1 These languages are: French, English, Spanish, German, Arabic and Chinese. Full syntactic analysis is only available for French and English but the chunker module exists for the other languages.