



Figure 1: An architecture for knowledge-based Question/Answering

uments containing the required information, based on *keywords techniques*. Q/A systems are presented with natural language questions, far richer in semantics than a set of keywords eventually structured around some operators. Furthermore, the output of Q/A systems is either the actual answer identified in a text or small text fragments containing the answer. This eliminates the user's trouble of finding the required information in sometimes large sets of retrieved documents. Open-domain Q/A systems must also be distinguished, on the other hand, from IE systems that model the information need through database templates, thus less naturally than a textual answer. Moreover, open-domain IE is still difficult to achieve, because its linguistic pattern recognition relies on domain-dependent lexico-semantic knowledge.

To be able to satisfy the open-domain constraints, textual Q/A systems replace the linguistic pattern matching capabilities of IE systems with methods that rely on the recognition of the *question type* and of the *expected answer type*. Generally, this information is available by accessing a classification based on the *question stem* (i.e. *what*, *how much*, *who*) and the head of the first noun phrase of the question. Question processing also includes the identification of the *question keywords*. Empirical methods, based on a set of ordered heuristics operating on the phrasal parse of the question, extract keywords that are passed to the search engine. The overall precision of the Q/A system depends also on the recognition of the *question focus*, since the answer extraction, succeeding the IR phase, is centered around the question focus. Unfortunately, empirical meth-

ods for focus recognition are hard to develop without the availability of richer semantic knowledge.

Special requirements are set on the document processing component of a Q/A system. To speed-up the answer extraction, the search engine returns only those paragraphs from a document that contain all queried keywords. The paragraphs are ordered to promote the cases when the keywords not only are as close as possible, but also preserve the syntactic dependencies recognized in the question. Answers are extracted whenever the question topic and the answer type are recognized in a paragraph. Thereafter the answers are scored based on several bag-of-words heuristics. Throughout all this processing, the NLP techniques are limited to (a) named entity recognition; (b) semantic classification of the question type, based on information provided by an off-line question taxonomy and semantic class information available from WordNet (Fellbaum 1998); and (c) phrasal parsing produced by enhancing Brill's part-of-speech tagger with some rules for phrase formation.

However simple, this technology surpasses 75% precision on trivia questions, as posed in the TREC-8 competition (cf. (Moldovan et al.1999)). An impressive improvement of 14% is achieved when more knowledge-intensive NLP techniques are applied at both question and answer processing level. Figure 1 illustrates the architecture of a system that has enhanced Q/A performance.

As represented in Figure 1, all three modules of the Q/A system preserve the shallow processing components that determine good performance. In the *Question Processing* module, the Question Class recognizer, working against a taxonomy of questions,