Automatic Question Answering

Jim Cowie, Evgeny Ludovik, Hugo Molina-Salgado,
Sergei Nirenburg & Svetlana Scheremetyeva
Computing Research Laboratory, NMSU
Dept 3CRL, Box 30001
Las Cruces, NM 88003, USA
{jcowie, eugene, hsalgado, sergei, lana}@crl.nmsu.edu

Abstract

We have developed a method for answering single answer questions automatically using a collection of documents or the Internet as a source of data for the production of the answer. Examples of such questions are 'What is the melting point of tin?', and 'Who wrote the novel Moby Dick?'. The approach we have adopted to the problem uses the Mikrokosmos ontology to represent knowledge about question and answer content. A specialized lexicon of English connects words, in English, to their ontological meanings. Analysis of texts (both questions and documents) is based on a statistical part-of speech tagger, and pattern-based proper name and fact classification and phrase recognition. The system assumes that all the information required to produce an answer exists in a single sentence and retrieval strategies (where possible) are geared to finding documents in which this is the case. In this paper we describe the overall structure of the system and the operation of the various components.

Introduction

Question answering (Q&A) for certain kinds of factual questions can be seen as purely an information retrieval task. The end users of a question answering system do not want the results of their question to be a set of documents. They need short, specific answers with possibly some supporting documents to confirm the answers' accuracy. Thus the problem seems to invite the use of both document retrieval, and information extraction, with both technologies being driven by the structure and content of the question. The borderline between document retrieval and information extraction is in fact a fuzzy one, see Cowie and Wilks (2000).

Q&A was adopted as an evaluation track for the eighth Text Retrieval Conference (TREC-8) (Harman, 1999). Each participating group was expected to supply a set of ten questions, whose answers were known to exist in the document corpora being used for the evaluation. These consisted of some 180 thousand documents from various news and government sources. This seemed a good opportunity to push forward on developing a question answering capability by integrating retrieval and extraction. In fact, software and data developed at CRL for several other tasks were integrated to produce our Q&A system: document retrieval, machine translation, summarization, and information extraction.

The Mikrokosmos ontology was initially created to support knowledge based machine translation, but recently we have been investigating its use as a control architecture for information extraction. To define an extraction task a static template consisting of named slots is created. Every slot contains one or more concepts from the Ontology, thus constraining possible slot fillers: any slot filler must be related to one of the slot concepts. For example:

```
ELECTION
   {"ELECT", "ELECT"}
   {"PERSON-ELECTED", "HUMAN}
   {"PLACE", "PLACE"}
   {"DATE", "TIME"}
   {"POSITION-ELECTEDTO","SOCIAL-ROLE"}
```

defines an election template. The first element in every line is the slot name (label), the second (there may be several of them) is the appropriate concept to which every possible filler must be related. Our idea for question answering is to use the question to dynamically define a similar template containing one slot per phrase plus the question target slot plus one slot used for information retrieval. Constraining slot concepts are defined by phrase head-words, and the phrase head word itself is added to the slot as an "artificial concept". Thus for any question a new template will be automatically produced, the first slot being the question target slot. The structured information in the template can be used to construct a set of Boolean queries to retrieve documents in which the key phrases, or equivalents occur. Information extraction is then carried out on the retrieved documents, with the slot(s) in the question template being filled. The filler of the first slot (newly found information) is the answer to the question.

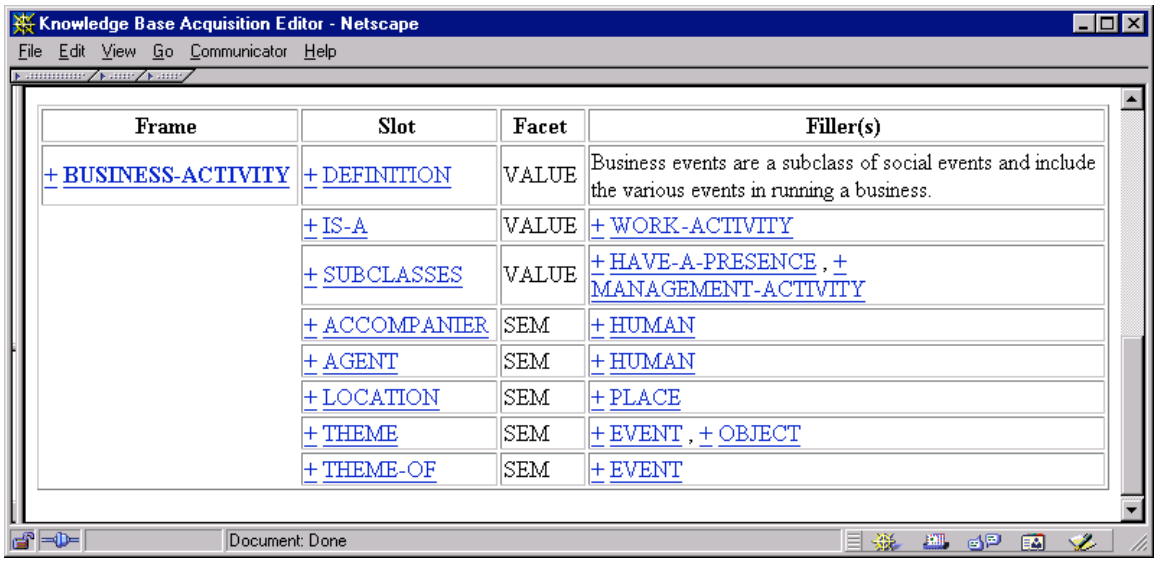
System Knowledge

For successful question answering two distinct types of expert knowledge are necessary: linguistic knowledge and world knowledge. Linguistic knowledge includes different kinds of lexicons with linguistic information for text parsing. It is a very important feature of our system that the lexicons are linked to the ontology, a language-neutral world model, that is used to explicate the meaning of lexical units and to 'fill gaps' in text meaning by making inferences based on the content of ontology conceptual knowledge. Thus our knowledge base consists of the Mikrokosmos ontology, a general lexicon, a format lexicon (for units describable by regular expressions), and rule-based recognition for dates, place, people, and organization names. These world-knowledge sources are used in combination with a part-of-speech tagger and a phrase grammar.

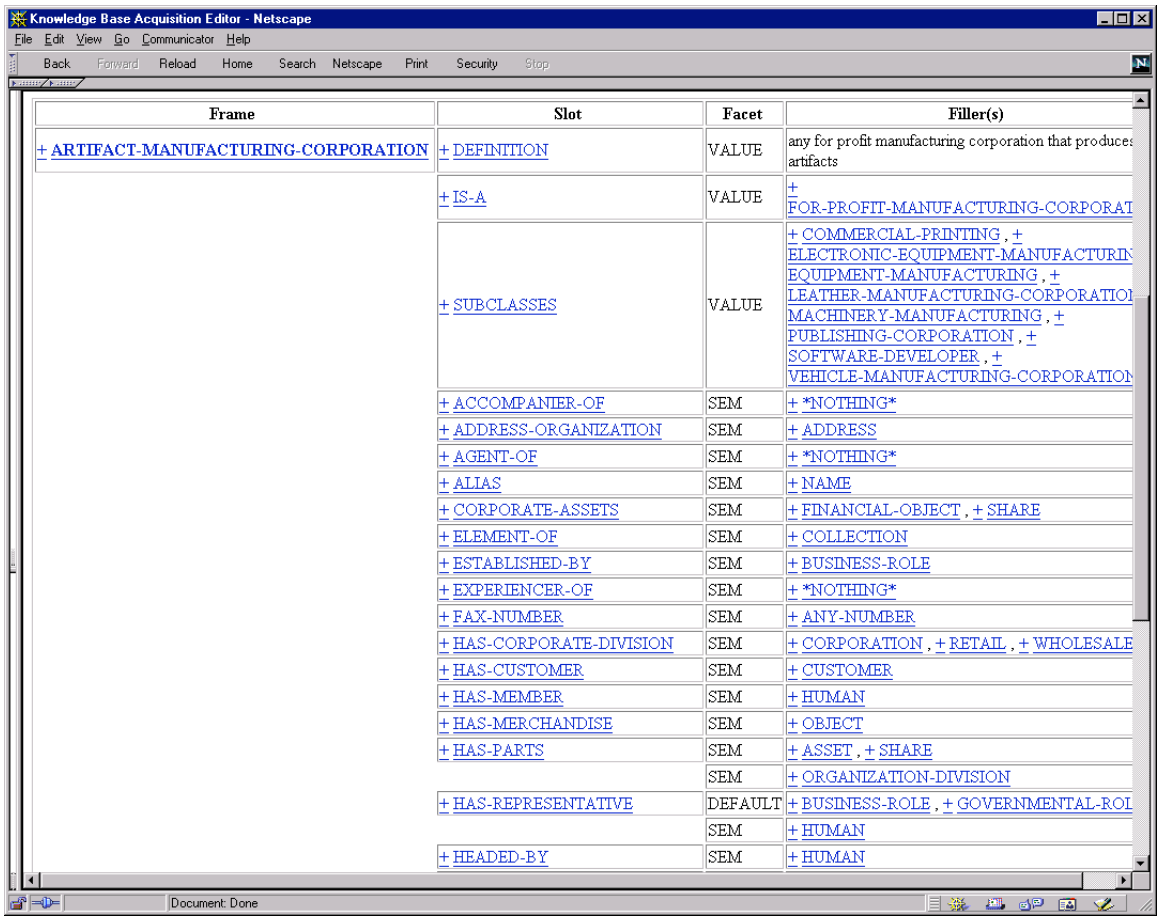
The Mikrokosmos ontology (http:// crl.nmsu.edu/Research/Projects/mikro/index.html) is a database with information about

- what categories (or concepts) exist in the world/domain,
- what properties they have, and
- how they relate to one another.

The ontology consists of around 5,000 concepts linked using 200 relationship types. Each concept is linked to other concepts through up to 16 different relationships. As described in Mahesh and Nirenburg (1995); and Mahesh (1996), the ontology includes a large collection of information about EVENTs (like BUSINESS-ACTIVITY), OBJECTs (like ARTIFACT-MANUFACTURING-CORPORATION) and PROPERTYs (like PRICE-ATTRIBUTE) in the world. In addition to the taxonomic multi-hierarchical organization, each concept has a number (currently averaging 14) of other local or inherited links to other concepts in the ontology, via relations (themselves defined in the PROPERTY sublattice). These links include case-role-like relations linking EVENTs to semantic constraints on the allowable fillers of those case-roles (i.e. selectional restrictions) and properties (see Figure 1).



Frame	Slot	Facet	Filler(s)
+ BUSINESS-ACTIVITY	+ DEFINITION	VALUE	Business events are a subclass of social events and include the various events in running a business.
	+ IS-A	VALUE	+ WORK-ACTIVITY
	+ SUBCLASSES	VALUE	+ HAVE-A-PRESENCE , + MANAGEMENT-ACTIVITY
	+ ACCOMPANIER	SEM	+ HUMAN
	+ AGENT	SEM	+ HUMAN
	+ LOCATION	SEM	+ PLACE
	+ THEME	SEM	+ EVENT , + OBJECT
	+ THEME-OF	SEM	+ EVENT



Frame	Slot	Facet	Filler(s)
+ ARTIFACT-MANUFACTURING-CORPORATION	+ DEFINITION	VALUE	any for profit manufacturing corporation that produces artifacts
	+ IS-A	VALUE	+ FOR-PROFIT-MANUFACTURING-CORPORAT
	+ SUBCLASSES	VALUE	+ COMMERCIAL-PRINTING , + ELECTRONIC-EQUIPMENT-MANUFACTURIN EQUIPMENT-MANUFACTURING , + LEATHER-MANUFACTURING-CORPORATIO MACHINERY-MANUFACTURING , + PUBLISHING-CORPORATION , + SOFTWARE-DEVELOPER , + VEHICLE-MANUFACTURING-CORPORATIO
	+ ACCOMPANIER-OF	SEM	+ *NOTHING*
	+ ADDRESS-ORGANIZATION	SEM	+ ADDRESS
	+ AGENT-OF	SEM	+ *NOTHING*
	+ ALIAS	SEM	+ NAME
	+ CORPORATE-ASSETS	SEM	+ FINANCIAL-OBJECT , + SHARE
	+ ELEMENT-OF	SEM	+ COLLECTION
	+ ESTABLISHED-BY	SEM	+ BUSINESS-ROLE
	+ EXPERIENCER-OF	SEM	+ *NOTHING*
	+ FAX-NUMBER	SEM	+ ANY-NUMBER
	+ HAS-CORPORATE-DIVISION	SEM	+ CORPORATION , + RETAIL , + WHOLESALE
	+ HAS-CUSTOMER	SEM	+ CUSTOMER
	+ HAS-MEMBER	SEM	+ HUMAN
	+ HAS-MERCHANDISE	SEM	+ OBJECT
	+ HAS-PARTS	SEM	+ ASSET , + SHARE
		SEM	+ ORGANIZATION-DIVISION
	+ HAS-REPRESENTATIVE	DEFAULT	+ BUSINESS-ROLE , + GOVERNMENTAL-ROL
		SEM	+ HUMAN
	+ HEADED-BY	SEM	+ HUMAN

Figure 1. Ontology frames for the concepts BUSINESS-ACTIVITYand ARTIFACT-MANUFACTURING-CORPORATION.

The information contained in the ontology allows for resolving semantic ambiguities and interpreting non-literal language by making inferences using the links in the ontology to measure the semantic affinity between meanings. It is also used to provide a grounding for representing text meaning in an interlingua and to enable lexicons for different languages to share knowledge.

The general lexicon has entries comprised of a number of zones (each possibly having multiple fields), integrating various levels of lexical information (morphological, syntactic and semantic).

The semantic zone of the lexicon is a focus of interest because it is the locus of interaction with the ontology, and thus the source of many of the building blocks of the eventual meaning representation. A partial sample entry of the general lexicon is shown in Table 1.

BUSINESS-ACTIVITY	business	N
BUSINESS-ACTIVITY	undertakings	N
BUSINESS-ACTIVITY	cease	V
BUSINESS-ACTIVITY	fulfill	V
BUSINESS-ACTIVITY	dealings	N
BUSINESS-ACTIVITY	matter	N
BUSINESS-ACTIVITY	affair	N
BUSINESS-ACTIVITY	relations	N
BUSINESS-ACTIVITY	cause-to-end	V
BUSINESS-ACTIVITY	interests	N
BUSINESS-ACTIVITY	bring-to-an-end	V
BUSINESS-ACTIVITY	do-business	V
BUSINESS-ACTIVITY	be-engaged-in	V
BUSINESS-ACTIVITY	fulfill	V
BUSINESS-ACTIVITY	concern	N
BUSINESS-ACTIVITY	stop	V
BUSINESS-ACTIVITY	end	V
BUSINESS-ACTIVITY	come-to-an-end	V
BUSINESS-ACTIVITY	conduct	
BUSINESS-ACTIVITY	accomplish	V
BUSINESS-ACTIVITY	close	V
BUSINESS-ACTIVITY	terminate	V
BUSINESS-ACTIVITY	activities	N
BUSINESS-ACTIVITY	drive	V
BUSINESS-ACTIVITY	halt	V
BUSINESS-ACTIVITY	finish	V

Table 1. A fragment of the general lexicon showing lexemes linked to the BUSINESS-ACTIVITY concept.

The format lexicon allows us to define useful sequences of concepts that are sought in the text. In the present system these mostly define "number MEASURING-UNIT" strings where measuring unit text representations are given in UK and US spellings, Plural and Singular, Full and Abbreviated forms. Combinations of different number-unit strings within every class are only given in some cases. In general they are supposed to be identified by the recognition program. The formats are classified according to the ontology concepts they are linked to:

AGE: = {NUMERIC-TYPE TEMPORAL-UNIT}
TEMPORAL-OBJECT(time period): = {NUMERIC-TYPE TEMPORAL-UNIT}
DATE
MONTH (month names)
DAY (weekday names)
YEAR
TIME-OBJECT (clock readings): = {NUMERIC-TYPE TIME-UNIT}
LINEAR-SIZE: = {NUMERIC-TYPE LINEAR-UNIT}
PLACE (area): = {NUMERIC-TYPE SQUARE-UNIT}
VOLUME: = {NUMERIC-TYPE CUBIC-UNIT}
LIQUID-VOLUME: = {NUMERIC-TYPE LIQUID-MEASURING-UNIT}
MASS: = {NUMERIC-TYPE MASS-WEIGHT-UNIT}
ELECTRICITY: = {NUMERIC-TYPE ELECTRICAL-POWER-UNIT}
ENERGY: = {NUMERIC-TYPE ENERGY-UNIT}
MONEY: = {NUMERIC-TYPE MONETARY-UNIT}

VELOCITY: = {NUMERIC-TYPE SPEED-UNIT}
ACCELERATION: = {NUMERIC-TYPE ACCELERATION-UNIT}
TEMPERATURE: = {NUMERIC-TYPE THERMOMETRIC-UNIT}
COMPUTER-MEMORY: = {NUMERIC-TYPE COMPUTER-MEMORY-UNIT}
RATE (rate of production): = {NUMERIC-TYPE RATE-UNIT}
PRESSURE: = {NUMERIC-TYPE PRESSURE-UNIT}
POPULATION-DENSITY: = {NUMERIC-TYPE POPULATION-DENSITY-UNIT}
REPRESENTATIONAL-OBJECT (other): = {NUMERIC-TYPE MEASURING-UNIT}

For example, the formats for the TEMPERATURE concept include text strings corresponding to the NUMERIC-TYPE (number) and THERMOMETRIC-UNIT concepts:

NUMERIC-TYPE degree C
NUMERIC-TYPE degrees C
NUMERIC-TYPE degree K
NUMERIC-TYPE degrees K
NUMERIC-TYPE degree F
NUMERIC-TYPE degrees F
NUMERIC-TYPE Kelvin
NUMERIC-TYPE Cent.
NUMERIC-TYPE Centigrade
NUMERIC-TYPE deg C
NUMERIC-TYPE deg
NUMERIC-TYPE deg K
NUMERIC-TYPE deg F
NUMERIC-TYPE C
NUMERIC-TYPE K
NUMERIC-TYPE F
NUMERIC-TYPE Fahrenheit
NUMERIC-TYPE Fahr
NUMERIC-TYPE Fr

System Operation

Our complete system consists of three main phases:

- Question Analysis - structure question content and recognize the question type
- Retrieval - build a structured query, retrieve and structure documents
- Answer Generation - information extraction and answer selection.

Figure 2 shows how these modules are linked.

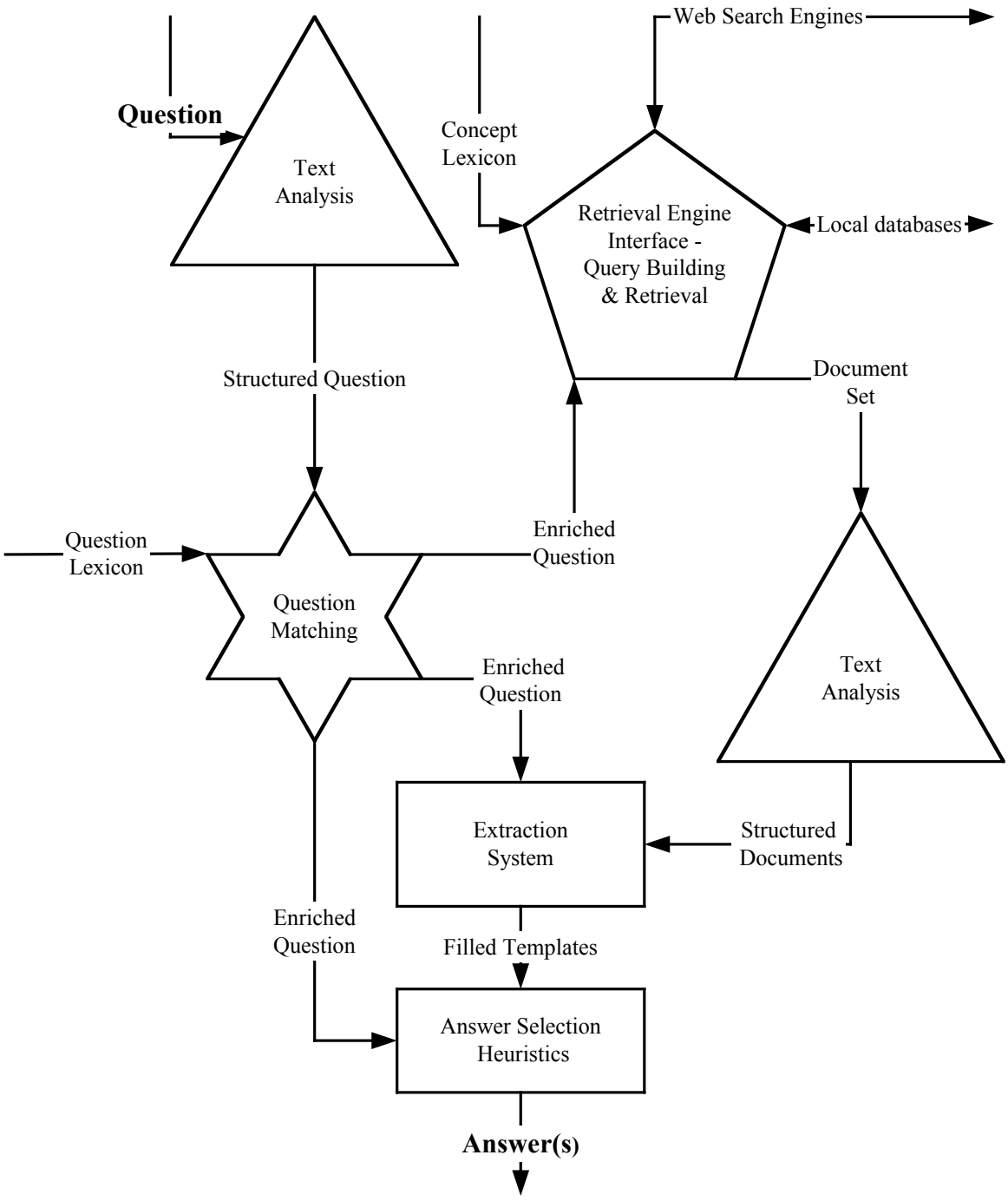


Figure 2: System Overview

Text Analysis

The basic processing undergone by the question and by sentences in the retrieved documents is the same. First, the document is processed by a part-of-speech tagger, this marks each word in the sentence with one part of speech. In our current system we use a statistical tagger from MITRE. The text is run independently through the CRL Diderot name recognition system (Cowie *et al.,* 1993). This recognizes names of organizations, places, people and a variety of other units of interest (dates, money, percentages etc.). These include the elements specified in the format lexicon. A complete list of the types handled is shown in Table 2 below. The labels are names of concepts from the Mikrokosmos ontology.

LINEAR-SIZE	ELECTRICITY	POPULATION-DENSITY	NATIONALITY
AREA	ENERGY	TEMPORAL-OBJECT	INHABITANT
VOLUME	VELOCITY	TIME-OBJECT	MATERIAL
LIQUID-VOLUME	ACCELERATION	AGE	EVENT-NAME
MASS	TEMPERATURE	NAME-HUMAN	PRODUCT-TYPE
RATE	COMPUTER-MEMORY	ORGANIZATION	NUMERIC-TYPE
PRESSURE		PLACE	DATE

Table 2: Types recognized for the Q&A task

The results of part-of-speech tagging and name and concept recognition are merged and the words are grouped into phrases, preference being given to the text units discovered by concept recognition. Verb and noun phrases and prepositional phrases are identified. A simple lexicon-based stemming algorithm is then applied to the heads of all phrases and provides the citation forms needed to support lookup in the English to Ontology Lexicon**.**

Patterns are then applied to recognize noun phrase and verb phrase; phrases recognized by the name and measure recognition phase are not merged into noun phrases. In every case, a head noun is identified. The head noun or verb is looked up in an English to ontology lexicon.

Question Matching

At this point we are ready to match the question against a set of skeletal question structures held in a 'question lexicon'. This allows the many ways that a question can be specified to all be mapped to a request for the same answer. Each entry consists of three parts:

<Type of Answer needed> <Additions to retrieval query> <Question pattern>

Where:

<Type of answer needed> specifies the ontological type of the answer needed

<Additions to retrieval query> specifies ontological concepts that should be mapped to lexical items to be used in the query process

<Question pattern> is a pattern containing strings that should be in the question, ontological types and Kleene stars, that allow matching any unit of question text. There is an implied '*' at the end of every question pattern. Currently there are some 500 question patterns in the system. Below we show the patterns used to handle questions on temperature.

```
TEMPERATURE    THERMOMETRIC-UNIT    * what * temperature
TEMPERATURE    THERMOMETRIC-UNIT    * how hot
TEMPERATURE    THERMOMETRIC-UNIT    * how cold
TEMPERATURE    THERMOMETRIC-UNIT    * how many degrees
TEMPERATURE    THERMOMETRIC-UNIT    * how high * temperature
TEMPERATURE    THERMOMETRIC-UNIT    * how low * temperature
TEMPERATURE    THERMOMETRIC-UNIT    * what * melting point
```

```
TEMPERATURE    THERMOMETRIC-UNIT     * what * boiling point
TEMPERATURE    THERMOMETRIC-UNIT     * what * freezing point
TEMPERATURE        THERMOMETRIC-UNIT          * how many *
THERMOMETRIC-UNIT
```

Temperature is a concept that is an object consisting of a NUMERIC-UNIT and a THERMOMETRIC_UNIT. The second element specifies that lexical entries attached to the concept THERMOMETRIC-UNIT should be included in the queries generated by the retrieval component of the system. The first pattern would recognize 'At what temperature does tin melt?'. The last pattern contains a concept in addition to strings, in lower case. This would match questions such as 'How many degrees centigrade is the melting point of tin?'.

The question recognition system uses dynamic programming to select the closest matching question pattern. Strings are matched with strings in the question, and concepts are matched with the head concepts found for each phrase. If a direct match is not found the concept's parent in the "IS-A" hierarchy will also be tried. This information is then passed both to the retrieval system query builder and to the answer extraction system.

Query Building and Retrieval

The goal is to find a text with a single sentence that specifies the answer in the context of all the constraints of the question. However, the constraints may need to be relaxed and synonyms generated to allow a matching sentence to be found. The query builder also expands the answer indicator concepts using the ontological lexicon. The THERMOMETRIC-UNIT will become 'centigrade OR fahrenheit OR kelvin OR c OR f OR k'. For local databases a boolean retrieval system is used and the initial query attempts to find all the phrases in a single sentence. The benefits of giving all the terms in a question equal weighting,, and of only performing stemming and term expansion in response to the initial query failure, are that texts are obtained where all the information specified is found in a close context. When a query is sent to a web search engine, which supports boolean queries, the documents returned have all the terms in one document. This places more reliance on the subsequent extraction step to select the document with the best concentration of information.

Document Structuring

The retrieved documents undergo the same language processing steps as was carried out on the query. Each sentence is part-of-speech tagged. Name recognition is run on whole documents, that allows much more accurate performance than processing single sentences. Phrases are recognized and heads of phrases are looked up in the English to ontology lexicon. A date analysis module detects a base date for the document and actual dates are computed for relative time phrases such as 'next week', and 'last Sunday'. The resulting structure for each document is then passed to the question answering phase.

Information Extraction and Answer Selection

The structured question is used as a template and matched against each sentence in the document. Each sentence receives a score for each string and each concept in the question that matches a text unit in the sentence. If no text unit matches the concept required then the sentence is rejected, otherwise the answer string is produced with a score for the number of question slots filled in producing this answer. Once all the documents have been processed, all the answers are sorted by score and the top five are picked.

The benefit of giving all the terms in a question equal weight in the query, and of only performing stemming and term expansion if the initial query fails, is that texts are obtained where all the information specified in the question is found in a close context in the document.

In this preliminary system the answer selection process only requires the answer concept and does not specifically check that the expected answer object is present. Thus *tall* would be an acceptable answer for a 'LINEAR-SIZE' question.

Sample Question and Answer

The following shows a question, the resulting structure, and the set of answers obtained, all from the same document from the *LA Times*.

```
% How tall is the Eiffel Tower?

answer-indicator LINEAR-UNIT
np "the Eiffel Tower" "tower"

LA061789-0071       2.0 CRL 1,000-foot
LA061789-0071       2.0 CRL short
LA061789-0071       0.95 CRL 76-foot
LA061789-0071       0.95 CRL 90-foot
LA061789-0071       0.95 CRL Too Tall
```

Future Work

Now that a complete working system is in place we particularly want to explore the differences between using a boolean retrieval system and a statistical system. It is our intuition that the boolean approach will provide both better precision and recall with a much smaller number of documents being processed.

More sophisticated heuristics for answer selection are also required. For example not producing as an answer something that was specified in the question. Some modicum of syntax will also help in matching sentences to the question template. Other more sophisticated language processing techniques such as co-reference resolution will be needed to handle cases where all the relevant information is not located in a single sentence.

The question lexicon needs to be expanded to cover more question types.

A web search version has been built to allow the demonstration of the process on non-TREC data and is currently undergoing testing. We intend to have the system cache answers to commonly asked questions. This will result in a fast growing fact database that will incorporate answers that are unlikely to be explicitly specified in documents (e.g. *What is the capital of France?*).

The method is not language independent, but the components used: part-of-speech tagging, phrase recognition, name recognition and an ontological lexicon are already available for Spanish and Chinese, so the development of question answering systems for these languages should be possible in a relatively short period of time (Cowie, 1996; Sheremetyeva *et al*., 1998).

Acknowledgements

We gratefully acknowledge the use of the MITRE part-of-speech tagger. We would also like to thank the organizers of the Q&A track of TREC-8 for stimulating our interest in this task.

References

Cowie, J. (1999) Collage: An NLP Toolset to Support Boolean Retrieval. In Tomek Strzalkowski (Ed.), *Natural Language Information Retrieval*. Kluwer Academic Publishers.

Cowie, J. & Wilks, Y. (2000). Information Extraction. In Robert Dale, Hermann Moisl, and Harold Somers (Eds.), *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker Inc.

Cowie, J., Ludovik, E., Molina-Salgado, H. (1998). Improving Robust Domain Independent Summarization. In *Proceedings of Natural Language Processing and Industrial Applications*. Moncton, Canada.

Cowie, J., Guthrie, L., Wakao, T., Jin, W., Pustejovsky, J., & Waterman, S. (1993). The Diderot Information Extraction System. In *Proceedings of the First Conference of the Pacific Association for Computational Linguistics, (PACLING 93)*. Vancouver, Canada.

Cowie, J. (1996). CRLs approach to MET (Multilingual Named Entity Recognition). In *Proceedings of the Tipster Text II 24 Month Workshop*. Morgan Kaufman.

Harman, D. K. (Ed.) (1999) The Eighth Text Retrieval Conference (TREC-8). NIST.

Mahesh, K. & Nirenburg, S. (1995). A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, International Joint Conference on Artificial Intelligence (IJCAI-95). Montreal, Canada.

Sheremetyeva S., Cowie, J., Nirenburg, S., & Zajac, R. (1998). A Multilingual Onomasticon as a Multipurpose NLP Resource. In *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain.