# Issues in Extraction and Categorization for Question Answering

**David Eichmann**

School of Library and Information Science
The University of Iowa
Iowa City, IA 52242
david-eichmann@uiowa.edu

### Abstract

We describe our approach in constructing question answering systems, which involves natural language parsing enhanced with named entity extraction and part-of-speech tagging. Performance of our question categorizer is discussed, particularly with respect to misclassifications and how they can be challenging to correct. Finally, recent work in connecting question answering to image and video retrieval components is discussed.

## Introduction

Our work in question answering (QA) systems has focussed on a simple, modular approach to the problem. Our research hypothesis is that by careful introduction of a limited number of techniques into an extensible architectural framework, we will be able to construct a system with respectable performance that the same time is amenable to incremental evaluation of the contribution that those individual components make towards the overall efficacy of the system.

Participation in the Question Answering Track in TREC-[8-10] and TREC-2002 has yielded valuable insight into the value of lightweight implementation of components, relying strongly on reuse of existing subsystems. Our TREC-8 system was effectively a sentence-level retrieval engine, capitalizing on the extended size of the allowed response string to achieve modest performance. By TREC-10, we had settled on a core architecture involving sentence recognition, named entity recognition, syntactic parsing of sentences (using the CMU link grammar parser (Grinberg 1995)) and scoring based upon syntax-based clause extraction with lexical relationships (using WordNet). The transition to the exact single answer rules for TREC-2002 hence proved to be straight-forward, since we were already targeting that as our definition of answer.

In this paper we discuss the rationale for our attention to architecture and our choice of entity extraction as a key supporting technology. We then consider two key aspects of QA performance, the categorization of user questions and the scoring of potential answers. Finally we present preliminary work on multimedia question answering.



## Information Retrieval Architectures

The basic nature of information retrieval systems has failed to keep up with software engineering principles. In spite of government initiatives such as TIPSTER ((Grishman 1996), the majority of information retrieval researchers are using a system architecture first developed in the early 70's, that used by SMART (Salton 1971). While SMART was designed for extensibility of algorithm, its core architecture is inherently batch, coloring the view of researchers using it. More recent systems such as GATE (Cunningham et.al 1996a, 1996b), Calypso (Zajac 1998, Zajac et.al. 1997) and our own work (Eichmann 1994, Eichmann 1998, Eichmann 1999) have built upon the lessons learned in SMART, but capitalize upon the developments in software engineering and architectures, producing systems capable of operating both interactively and incrementally. TIPSTER itself was meant to interplay retrieval and extraction, but few systems actually coupled the two into an integrated environment (Cowie and Lehnert 1996).

We intend to demonstrate the power and flexibility of such an integrated architecture designed and constructed using standard software engineering practices such as domain analysis (Prieto-Diaz 1991) of current 'best-of-breed' approaches to information retrieval, extraction, detection and summarization. As this analysis has progressed, we have adapted our existing systems to conform to the resulting domain architecture, forming the core of an extensible family of systems that are composable on demand for a variety of retrieval and analysis tasks. This approach has proven fruitful in related areas such as database systems (Batory 1988) and $C^3$ (Dowle 1989).

### Extraction Mechanisms

Information extraction is typically viewed as having two major phases, local text analysis and discourse analysis (Grishman 1997). Local text analysis is comprised further of lexical analysis, syntactic analysis and some form of recognition / pattern matching. Discourse analysis involves the matching of various forms of reference to specific entities. A broad variety of techniques are used in these phases (finite state transducers (Hobbs et.al 1993), machine learning (Neri and Saitta 1997), conceptual graph matching (Jacobs and Rau 1990), etc.), Zechner's survey (Zechner

1997) serves as an excellent starting point for current techniques.

(Gaizauskas and Robertson 1997) explored the use of a Web search engine (Excite) as a filter for information extraction. One of the limits to this approach is that it is dependent upon the coverage of the search engines for recall. They have no leverage on the major engines with respect to crawling policies or amount of data available in response to a given query.

Our work in entity extraction to date has followed a path similar to that used in the HUB evaluation (Burger et.al 1998, Robinson et.al. 1999), concentrating on the general categories of persons, organizations, places and events, with medical terminology included for special situations (such as TREC-9). Table 1 shows the top entities by occurrence frequency for the TREC AP newswire corpus 1988-1990, as generated by our named entity recognizer. Note the double appearance of George Bush, due to the recognizer finding last name only occurrences in some documents and full name occurrences in other documents.[1] Mis-classifications in persons include Britain, due to the CIA Fact Book not using it as a place name, and Fed, a slang term for the Federal Reserve. Each of these is correctable by the addition of a new rule in the respective category in the recognizer. Our recognizer defaults to organization for acronyms – the run generating this data did not include MeSH terms, and so AIDS appears as an organization. Note that stop entities exist as readily as stop words. As you might expect, all high frequency events are days of the week or months.

Lower frequency entries have proven to be fertile ground for the identification of new entity categories. Table 2 shows a sample of entities drawn from middle of the frequency range (but having more than one occurrence). Note the two sports figures (including team names) and a television program under persons; a warship, a personal computer and two documents under organizations; a variety of regions and mid-level geography under place names; and weather and sports occurrences under events. Some of these are instances of refinements of the general category (river valleys for place names, storms for events) and some are mis-classifications that belong in new categories (frigate as a kind of ship, IBM-AT as a model of personal computer).

The mis-classifications do little to damage similarity calculations for questions regarding other entities, since they appear with very low frequency. Since question categorization shares the same entity recognition support with the document analyzer, there is actually little impact on answering a question relating to a mis-classified entity, since the mis-classification is reflected both in the question representation and the corpus representation.

Even more intriguing are mis-classifications that can lead directly to the instantiation of attributes of and relationships between entities. Consider the following:

- 74-year history of Wrigley Field
- 75-year-old tree-cutting expert Casimier Wojakowski of Bark River
- 800-mile Rhine River
- associate director of Illinois Council 31 of the American Federation
- Corn Insects Research Unit of the U.S. Department of Agriculture
- New Zealand-born entrepreneur Bruce Judge
- outspoken critic of the new three-year National Master Freight Agreement
- Trojan nuclear plant operator Portland General Electric Co.

all single occurrence 'poor' classifications from early versions of the extraction component that carry a rich collection of properties and relationships. Cowie and Lehnert note in (Cowie and Lehnert 1996) that: "[r]eal text is rich in proper names and expressions for dates, values and measurements. These phrasal units are used productively, usually posing no problem to human readers." Information extraction lies then on the boundary between 'traditional' information retrieval and natural language processing. Our goal here is to move beyond simple named entities as the unit of recognition to (potentially fact attributed) entities or structures of entities as the unit of recognition. The 'mis-classifications' discussed above will form our starting point for this work. We will then move into fusion techniques to combine separate occurrences and then add relationships and properties derived from grammatic structures.

## Semi-Structured Data

Extraction of information from even highly structured databases can be problematic when connections between information are blurred through the noise of misspellings, missing information, etc. As early as the late 1950's, the U.S. Census Bureau was struggling to achieve record linkage(Fellegi and Sunter 1969, Newcombe et.al 1959), the merging and unduplication of lists that may be used as survey frames or in conjunction with administrative files. As today's Web continues its dramatic growth, increasing numbers of databases and other structured sources of data are being made available as what the database community refers to as semi-structured data (Atzeni et.al. 1997, Hammer et.al. 1997), highly regular information whose actual structure must be inferred from its embedding HTML markup.

There are a number of approaches to accessing and manipulating semi-structured data. Wrapping involves the construction, either by hand or through induction (Kuskmerick), of a translator for an information source into a common integrating model (Widom 1995). Mediation is similar, addressing issues of semantic mismatch between heterogeneous data representations (Qian 1993). Also popular as a perspective is to view the Web as a connected graph of document 'objects' and to then impose query semantics on the traversal of that graph, e.g., (Abiteboul 1997). The common thread in all of these approaches is the viewing of the task being one of structure navigation and predicate assertion, in a manner similar to that applied to (usually object-oriented) database query evaluation. (Cluet 1997) describes a particularly useful

1.We also recognize titles of persons, maintaining a frequency list of them with the entity, but have not shown them here.

**Table 1:** Top 10 Entities by Occurrence Frequency, AP 1988-1990

| Persons | Organizations | Places | Events |
|---|---|---|---|
| • Bush | • Congress | • United States | • Friday |
| • Reagan | • Senate | • Soviet Union | • Tuesday |
| • George Bush | • House | • Japan | • Monday |
| • Britain | • White House | • New York | • Wednesday |
| • Mikhail S. Gorbachev | • AIDS | • United Kingdom | • Thursday |
| • Michael Dukakis | • Police | • Europe | • Sunday |
| • Treasury | • New York Stock Exchange | • Germany | • Saturday |
| • Republican | • Democrats | • Israel | • December |
| • Fed | • City | • Washington | • January |
| • Saddam Hussein | • Securities and Exchange Commission | • Iraq | • June |

**Table 2:** Midrange Frequency Entities (# > 1), AP 1988-1990

| Persons | Organizations | Places | Events |
|---|---|---|---|
| • Illinois strong safety Quintin Parker | • U.S. Navy frigate Constitution | • Atlantic coast of South America | • Typhoon Pat |
| • P. C. Leung | • 96-member MIT Corporation | • Toutle River Valley | • Typhoon Ruby weeks |
| • Ann Marie Cullen | • Experimental Aviation Association | • Stamford American | • television broadcast of the All-Star Game |
| • James McIntyre Jr. | • IBM-AT | • Atlantic remnants of Tropical Storm Klaus | • throes of the Cultural Revolution |
| • Kenneth J. O'Donnell | • AHMSA | • Tennessee Jersey | • Martin Luther King Jr. International Peace Award |
| • Antiques Roadshow | • U.S. Nike | • Gulf Canada Res | • UCLA Invitational |
| • Hugo Rivero Villavicencia | • Israeli International Institute for Applied Economic Policy Review | • Atlantic town of Puerto Cabezas | • U.S. Senior Sports Classic |
| • Jude Harmon | • U.S. Observers | • Stan DeMaggio of Capistrano Valley | • UC Riverside Tournament |
| • Kenneth J. Pinkes | • Island Properties Report | • Gulf Coast-Southwestern | • throes of the ultra-leftist Cultural Revolution |
| • Angel newcomer Luis Polonia | • ABN-Amro | • Standiford Field | |
| • Michael Lovejoy | • Fairchild Fastener Group | • CS Los Angeles | |
| • Japan Advertiser | • Fifth U.S. Circuit Court of Appeals | • Stan Rivers | |
| • Freddie Fender | • UN Security Council Resolution | • Gulf Coast States | |
| • Illinois surgeon James Chow | | • Stand Hill | |
| | | • Bhootahi Balan River | |
| | | • Ventura River Valley | |

scheme with the Object Extraction Model (OEM), which is used to represent both document structure and queries.

We view support for semi-structured data analysis as a natural and indeed critical component in our architecture, given the ability to use this information in sense disambiguation and named entity identification and extraction. Sulla uses a mediation approach to Web meta-search, where the mediators use an open template definition standard (Apple's Sherlock specification) to extract search hits from result pages (Eichmann 1996). We plan to extend this mechanism with a graph-based query language to support connectivity related semantics. We will also incorporate current work in record linkage (Winkler 1994, Winkler).

## The Web as Open Corpus

Using a topological analysis approach, (Albert, et. al.) used Lawrence and Giles original numbers (Lawrence and Giles 1996) to estimate the diameter of the Web (shortest distance between any two documents) at 18 links and that a tenfold increase in size would only increase the diameter to 20 links. They then observe that intelligent agents can therefore easily identify and acquire relevant information, but that crawlers operating solely on matching strings would need to index fully 10% of the Web to find all information relevant to a target query. We have found this argument strongly supported in our work in Web retrieval agents, where results from interrogation of the major search engines are used to seed a topic-specific crawl of a conceptually-connected Web space.

(Gaizauskas and Robertson 1997) observes that "IE systems in MUC-6 can perform text-filtering to a high degree of accuracy… However, the proportion of relevant documents is much higher in MUC than is the case in a typical IR task and very much higher than is the case in a set of documents retrieved from the WWW." We see this as the basic challenge in focusing on the Web as a major information source for our open-domain QA work. Utility of this approach is easily established from considering, for example, the number TREC QA participants that enrich their candidate document selection with Web-derived data. We choose not to do this, preferring to separate issues of closed- and open-corpus question answering by employing only structured data sources as system domain knowledge sources. WordNet, dis-

cussed in the next section, is an excellent example of such a source - non-specific to any given question domain.

### Digital Thesauri – WordNet

WordNet is an electronic thesaurus under development at Princeton University (Miller). WordNet words are classified into four lexical categories: nouns, verbs, adjectives, and adverbs. Essentially, only relations between meanings in the same category are considered. Basic relations include: synonymy (similarity of meaning. The basic units in the WordNet database is a synset which is a group of words that have similar or identical meanings), hyponymy/hypernymy ("is a" relation. A hyponym has all of the features of its hypernym and adds at least one distinguishing feature. The hypo/hypernym relation produces a hierarchical organization of synsets), meronymy/holonymy: ("has a" relation. A meronym is a part, a member or a substance of its holonym), antonymy: (an opposite. This relation is defined between word forms, not between synsets.) These relations form a web of semantic interconnection between the words of English that can be used to extract semantic meaning from text. WordNet contains 168,000 synsets with 126,000 different word forms.

WordNet and similar semantic databases have been used for concept-based text retrieval in two ways. One way is to use synset relations to expand a query to include synonyms of the relevant words. A second is to use wordnet relations to measure the semantic similarity of query and document sets. Word sense disambiguation is crucial to the success of both techniques. While recall is improved even in the case of bad disambiguation, precision can decrease dramatically because documents are found that are related to another meaning of a search term. As long as the disambiguation algorithm can correctly determine 70% of word senses, searching documents by synset is more effective than searching by word matches (Gonzalo 1988). It is not yet known if practical algorithms for word sense disambiguation can reach this level. However, in the case of short documents, WordNet-based techniques can show significant improvements even in the absence of good word-sense disambiguation methods and accurate conceptual distance measures (Smeaton and Quigley 1996).

We take a significantly different approach, based upon the notion of synset as the basis of entity category identification and matching. This allows us to support question evaluation approaches that avoid missing documents not retrieved by current retrieval engines due to variations in word choice. We are also working towards support of query expansion and refinement through the use of meronymy (part-of relationships), particularly when exploring neighborhood semantics relating to an information need.

## Question Categorization

When are there too many categories? Consider the following question: *Who sued Exxon?* Over specificity in question categorization can force a QA system to choose between Person and Organization as the desired response. In fact, the answer might be instance of either or both. Similarly, there is advantage to discrimination between various quantitative categories (e.g., between duration and area), but there is also difficulty in something as straightforward as this - consider the question: *How big is the Amazon?* Even when a system can establish that there are two distinct alternatives, the rain forest and the river, there is still the issue of whether it is the area of the rain forest or the length of the river that is the desired response.

Given the degree of potential ambiguity in a posed question, we have chosen to keep the number of question categories used by our system to a very small number: who (including both Person and Organization), what, when, where, how_much and how. Table 3 shows how this scheme categorizes three years worth of TREC questions (TREC-[8-10]).

**Table 3:** Category Distribution of 1393 TREC QA Questions

| Category | # of Questions | % of Questions |
|---|---|---|
| how | 12 | 00.86 |
| how_much | 185 | 13.28 |
| what | 486 | 34.88 |
| when | 146 | 10.48 |
| where | 254 | 18.23 |
| who | 310 | 22.25 |

Development of our question categorizer has involved taking a development set of questions and known categories and developing a rule set using prefix trigger words (e.g., "When did...") and WordNet syntactic hierarchies (to categorize target-referencing question vocabulary) which was then tested against a separate set of questions. We did this in successive iterations by using TREC-9 to test the rules derived for TREC-8 and TREC-10 to test the rules for TREC-[8-9]. As shown in Table 4, the distribution of questions varies broadly over the data, but by TREC-10 questions we achieve good recognition levels. Table 5 breaks down the errors by type and year..

**Table 4:** Actual Question Categories vs. Test Results

| Category | TREC-8 | | TREC-9 | | TREC-10 | |
|---|---|---|---|---|---|---|
| | actual | test | actual | test | actual | test |
| how | 3 | 3 | 4 | 3 | 5 | 5 |
| how_much | 39 | 42 | 75 | 91 | 71 | 74 |
| what | 31 | 35 | 230 | 198 | 225 | 214 |
| when | 24 | 23 | 73 | 72 | 49 | 49 |
| where | 41 | 41 | 133 | 135 | 80 | 88 |
| who | 62 | 61 | 178 | 194 | 70 | 70 |

The largest areas of performance issues with the system involve *what* questions miscategorized as *how much* questions and *what* questions miscategorized as *who* questions. Much of this is attributable to linguistic ambiguity of words

**Table 5:** Categorization Performance by Miscategorization

| Test Category | Actual Category | TREC-8 | TREC-9 | TREC-10 | Overall | Overall % |
|---|---|---|---|---|---|---|
| how | how_much | | | | | |
| how | what | | 1 | | 1 | < 0.1% |
| how | when | | | | | |
| how | where | | | | | |
| how | who | | | | | |
| how_much | how | | 2 | | 2 | < 0.1% |
| how_much | what | 2 | 16 | 9 | 27 | 0.2% |
| how_much | when | | | | | |
| how_much | where | | | | | |
| how_much | who | | 1 | | 1 | < 0.1% |
| what | how | | | | | |
| what | how_much | | 2 | 5 | 7 | < 0.1% |
| what | when | 1 | 1 | | 2 | < 0.1% |
| what | where | 1 | 1 | 4 | 6 | < 0.1% |
| what | who | 3 | 5 | 3 | 11 | < 0.1% |
| when | how | | | | | |
| when | how_much | | 1 | | 1 | < 0.1% |
| when | what | | | 1 | 1 | < 0.1% |
| when | where | | | | | |
| when | who | | | | | |
| where | how | | | | | |
| where | how_much | 1 | | 1 | 2 | < 0.1% |
| where | what | | 3 | 11 | 14 | 0.1% |
| where | when | | | | | |
| where | who | | 3 | 2 | 5 | < 0.1% |
| who | how | | | | | |
| who | how_much | | | | | |
| who | what | | 21 | 2 | 23 | 0.2% |
| who | when | | 1 | 1 | 2 | < 0.1% |
| who | where | | 3 | 2 | 5 | < 0.1% |
| | Total | 8 | 61 | 41 | 110 | |
| | % | 4.0% | 8.8% | 8.2% | 7.9% | |

**Table 6:** Scoring Effects on Correct Answers by Entity Category, TREC-2002

| Entity Type | Best Single Score | Summed Scores | Response Frequency |
|---|---|---|---|
| Person | 29.2% | 19.2% | 4.2% |
| Organization | 4.2% | 3.8% | 0.0% |
| Place Name | 16.7% | 11.5% | 12.5% |
| Event | 20.8% | 38.5% | 58.3% |
| Amount | 20.8% | 23.0% | 20.8% |
| What | 8.4% | 3.8% | 4.2% |

having both quantity senses and general senses (e.g., term). Even then, we're achieving sufficiently high levels of categorization (91-92%) to concentrate on other aspects of the system for additional work. It is also interesting to note that this level of categorization was achieved with approximately one person-week of effort. An additional person-week of effort devoted to improving performance failed to result in any gain, yielding only 'jitter' in WordNet sense assignments.

## Response Scoring

Given adequate levels of question categorization (as discussed in the previous section) and the ability to extract good candidates, as generated by our entity recognition and syn-

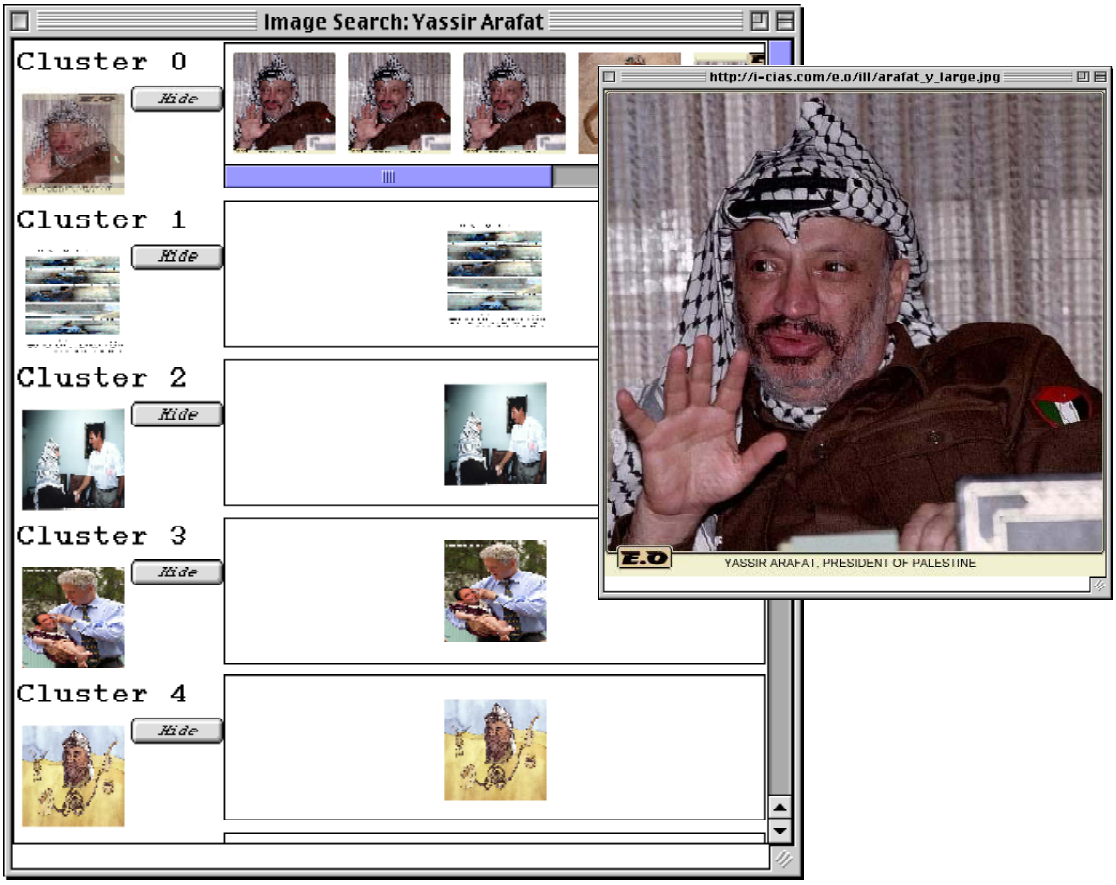


**Figure 1:** Results of a hybrid text / image search on Yassir Arafat

tactic extractors, the remaining challenge lies in scoring the generated candidates. Our initial notions of scoring (in TREC-8) revolved around simple scoring of the entire sentence or some pruned chunk of it based upon the level of term match between sentence and question (term vector similarity). This approach becomes increasingly problematic as the size of the response is reduced, with the extreme being the exact answer requirements of TREC-2002. This approach to scoring weights source substructure of a sentence more heavily than the target substructure of a sentence. We have subsequently been dependent upon use of the syntactic structure of a sentence to identify the appropriate clauses in a sentence, usually based upon their position relative to that portion of the sentence 'responsive' to the question vocabulary.

Hence, for TREC-2002, we submitted three distinct scoring schemes. The first, termed *best single score,* is the technique described in the previous paragraph. The second, termed *summed scores*, considers the aggregated scores for a given answer across all candidate sentences evaluated for a given question. The third, termed *response frequency,* uses the number of times a given answer appears in the answer candidate pool. Table 6 shows the relative performance of each scheme

These results seem to indicate that persons are reasonably easy to identify in isolation, without support from other sentences. Events on the other hand are better handled using the frequency with which they appear as candidates – in raw numbers, response frequency generated more than twice the number of correct answers for events than best single score generated for persons. Even more interesting, this version of the system had only persons and organizations as a development focus, with no particular attention towards events or other entity types. Whether comparable improvements for events can be achieved in our best single score approach will require additional experimentation.

## Multimedia Question Answering

Our initial experimentation with media-enhancement of question answering has involved Web meta-search (Eichmann 1996). As an example, we have extended our clustering scheme to include images and video. Figure 1 shows the result of clustering a set of images returned on a Web search for Yassir Arafat. This is a hybrid system that performs a meta-search using a textual query and retrieves images present in the result Web pages. These images are then clustered based upon a range of similarity measures and the clus-

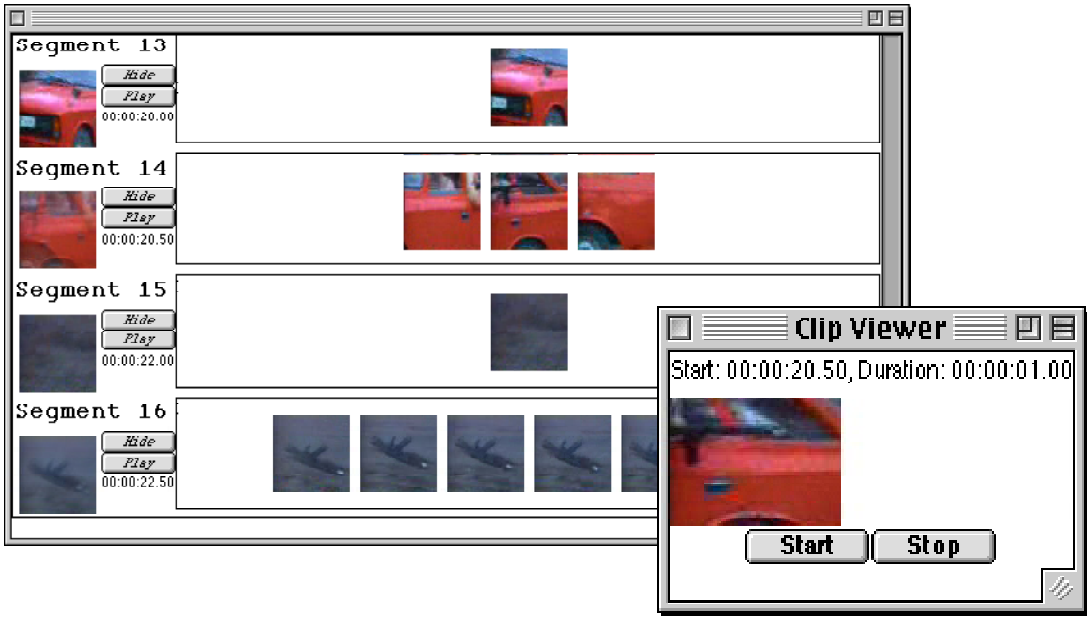


**Figure 3:** Segmentation of a video of a gunman discarding an AK-74 from a car.

ters displayed for user browsing. The Web is a rich source of imagery, but identification of relevant images can be daunting. Our work in hybrid retrieval schemes has proven useful in exploiting the correlation between the text of a Web page and the images linked into that page (see Figure 1). We use this 'additional channel' for retrieval and classification of a variety of image types – general photographs such as those in Figure 1, diagrams, images with substantial metadata such as those available from the Microsoft TerraServer project[1], etc. Figure 2 shows an image retrieved from TerraServer showing the south entrance to the building containing the author's office. We can currently recognize author/address blocks from Web server acquired PostScript and PDF files and we can, given a set of lat-long coordinates, extract available imagery. However, the mapping from an address to coordinates currently requires user intervention. More complete integration of mapping data from structured flat files, gazetteer data from relational databases (e.g., Tiger census data) and external resources such as TerraServer is one of the major goals of our current research

**Figure 2:** TerraServer query result,
41N 39′34″ 91W 32′22″

## Video imagery drawn from the Web

Video is increasingly available on the Web as bandwidth makes on demand delivery a viable option. Much of the material is from unofficial sources (e.g., the clip segmented in Figure 3 was retrieved from http://club.guns.ru, a site devoted to Russian firearms) and of widely varying quality. It provides an excellent source of noisy video data that has connected written text.

Figure 3 shows the result of using inter-frame similarity thresholds to segment a similarly retrieved video clip involving a gunman throwing an AK-74 from a car as it passes the camera The clusters in this case comprise the sequences of frames sampled from the video. Each frame is individually viewable as was done in Figure 1 and as shown in the inset window in Figure 3, the generated segment can be played in isolation. Both of these interfaces share a common architecture with our QA system..

1.The Microsoft TerraServer project, http://terraserver.homeadvisor.msn.com provides a public Web interface to digital orthophoto quadrangle (DOQ) images maintained by the USGS. Querying TerraServer provides a demonstration of image extraction from an external resource.

## Conclusions and Future Work

The QA system presented here is definitely a work in progress. Particular components, e.g., question categorization and feature extraction, are performing at a level that we believe to be adequate for overall good performance. Current performance focusses on 'factoid' type answers occurring in a single candidate sentence, however. Moving beyond a single sentence being suitable for answer generation will involve scoring schemes that propagate match weights across multiple syntactic structures. Our currently plans for this are similar to the ternary expressions used by (Lin 2001) or (Litowski 1999). Our use of relational database representations of lexical features are easily extended along these lines.

Expanding from pure text as response to additional forms of media shows excellent promise. Experimentation with Web documents as the granularity associated with search-media correspondence yields sufficient levels of precision for most simple user requirements. We expect significant improvements in precision by reducing scoring granularity down to the sentence and/or hypertext anchor level. Some form of image analysis and feature extraction is likely to be required for further improvements.

## References


Abiteboul, S. 1997. "Querying Semi-Structured Data," *ICDT '97*, Delphi, Greece.

Albert, R., J. Jeong and A.-L. Barabási, "The diameter of the world wide web," submitted for publication. Available at http://xxx.lanl.gov/abs/cond-mat/9907038.

Atzeni, P., G. Mecca and P. Merialdo. 1997. "Semistructured and Structured Data in the Web: Going Back and Forth," *Proc. Workshop on Management of Semistructured Data (in conjunction with PODS/SIGMOD)*, Tucson, AZ, May 16, 1997.

Batory, D. S. 1988. "Concepts for a Database System Compiler," *Proc. ACM Principles of Database Systems Conf.*

Burger, J, D. Palmer, and L. Hirschman. 1998. "Named Entity Scoring for Speech Input," *COLING-98*, Montreal.

Cluet, S., "Modeling and Querying Semi-structured Data," in (Pazienza 1997).

Cowie, J. and W. Lehnert. 1996. "Information Extraction," *Communications of the ACM*, v. 39, no. 1, January 1996, pp. 80-91.

Cunningham, H., Y. Wilks and R. Gaizauskas. 1996a. "Software Infrastructure for Language Engineering," *Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition*, University of Sussex.

Cunningham, H., Y. Wilks and R. Gaizauskas. 1996b. "New Methods, Current Trends and Software Infrastructure for NLP," *Proc. 2nd Conf on New Methods in Natural Language Processing*, Bilkent University.

Dowle, S. W. 1989. "Domain Modelling for Requirements Specification," *Proc. 3rd Int. Conf. on Command, Control, Communications and Management Information Systems*, London, UK, May 1989, pp. 1-7.

Eichmann, D. 1994. "The RBSE Spider – Balancing Effective Search Against Web Load," *First International Conference on the World Wide Web*, Geneva, Switzerland, May 25-27, 1994, p. 113-120.

Eichmann, D. 1995. "Semantic Levels of Web Index Interaction," *Web-wide Indexing and Semantic Header Workshop* at *Third International World-Wide Web Conference*, Darmstadt, Germany, April 10, 1995.

Eichmann, D. 1996. "Sulla - A User Agent for the Web," poster paper, *Fifth International WWW Conference*, Paris, France, May 6-10, 1996, poster proceedings p. 1-9.

Eichmann, D. 1998. "Ontology-Based Information Fusion," *Workshop on Real-Time Intelligent User Interfaces for Decision Support and Information Visualization, 1998 International Conference on Intelligent User Interfaces*, San Francisco, CA, January 6-9, 1998.

Eichmann, D., M. Ruiz, P. Srinivasan, N. Street, C. Culy, F. Menczer. 1999. "A Cluster-Based Approach to Tracking, Detection and Segmentation of Broadcast News," *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA, February 28 - March 3, 1999, pp. 69-75.

Fellegi, I. P and A. B. Sunter. 1969. "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1969, pp. 1183-1210.

Gaizauskas, R. and A. M. Robertson. 1997. "Coupling Information Retrieval and Information Extraction: A New Text Technology for Gathering Information from the Web," *Proc. of RIAO '97: Computer-Assisted Information Searching on the Internet*, Montreal, Canada, 1997, pp. 356-370.

Gonzalo, J., F. Verdejo, I. Chugur and J. Cigarrán. 1988. "Indexing with WordNet synsets can improve text retrieval" in *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal 1988.

Grinberg, D., J. Lafferty, and D. Sleator. 1995. *A robust parsing algorithm for link grammars.* In Proceedings of the Fourth International Workshop on Parsing Technologies, Prague/Karlovy Vary, Czech Republic.

Grishman, R. 1996. *TIPSTER architecture design document version 2.2*, http://www.tipster.org/

Grishman, R. 1997. "Information Extraction: Techniques and Challenges," in (Pazienza 1997).

Hammer, J., H. Garcia-Molina, J. Cho, R. Araha and A. Crespo. 1997. "Extracting Semistructured Information from the Web," *Proc. Workshop on Management of Semistructured Data (in conjunction with PODS/SIGMOD)*, Tucson, AZ, May 16, 1997.

Hobbs, J.R., Appelt, D., Bear, J. et al. 1993. "FASTUS: A System for Extracting Information from Text," *Human Language Technology: Proceedings of a Workshop*, Plainsboro, San Francisco, CA Morgan Kaufman Publishers.

Jacobs, P. S. and L. F. Rau. 1990. "SCISOR" Extracting Information from On-line News," *Communications of the ACM*, v. 33, no. 11, November 1990, pp. 88-97.

Kushmerick, N. "Wrapper Induction: Efficiency and Expressiveness," submitted for publication.

Lawrence, S. and C. L. Giles. 1998. "Searching the World Wide Web," *Science*, 280, April 3, 1998, p. 98-100.

Lin, J. J. 2001. *Indexing and Retrieving Natural Language Using Ternary Expressions*, Master's Thesis, Massachusetts Instituteof Technology, February 2001.

Litowski, K. C. 1999. *Question-answering using semantic relation triples,* in Proc.of the 8th Text Retrieval Conference (TREC-8).

Miller, G., *Five Papers on Wordnet,* Cognitive Science Laboratory, Princeton University. Available at http://www.cogsci.princeton.edu/.

Neri, F. and L. Saitta. 1997"Machine Learning for Information Extraction," in (Pazienza 1997).

Newcombe, H. B., J. M. Kennedy, S. J. Axford and A. P. James. 1959. "Automatic Linkage of Vital Records," *Science*, 130, pp. 954-959.

Pazienza, M. T. (ed.). 1997. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology,* International Summer School, SCIE-97, Frascati, Italy, July 14-18, 1997, Lecture Notes in Artificial Intelligence, vol. 1299, Springer-Verlag.

Prieto-Díaz, R. and G. Arango. 1991. *Domain Analysis and Software Systems Modeling,* IEEE Computer Society Press, Los Alamitos, CA.

Qian, X. 1993. "Semantic Interoperation Via Intelligent Mediation," *Proc. Int. Workshop on Research Issues in Data Engineering*, p. 228-231.

Robinson, P, E. Brown, J. Burger, N. Chinchor, A. Douthat, L. Ferro and J. Hirschman. 1999. "Overview: Information Extraction from Broadcast News," *Proc. DARPA Broadcast News Workshop*, p. 27-30.

Salton, G. (ed.). 1971. *The Smart Retrieval System – Experiments in Automatic Document Processing,* Prentice-Hall.

Smeaton, A. F. and I. Quigley. 1996, "Experiments on Using Semantic Distances Between Words in Image Caption Retrieval," in *Proc. 19th International Conf. on Research and Development in IR*, Zurich, August 1996.

Widom, J. 1995. "Research Problems in Data Warehousing," *Proc. 4th Int. Conf. on Information and Knowledge Management (CIKM)*, November 1995.

Winkler, W. E. 1994. "Advanced Methods of Record Linkage," *American Statistical Association, Proc. of the Section of Survey Research Methods*, pp. 467-472.

Winkler, W. E., "Matching and Record Linkage," in B. G. Cox, et. al. (eds.) *Business Survey Methods*, J. Wiley, New York, pp. 355-384.

Zajac, R. 1998. "Reuse and Integration of NLP Components in the Calypso Architecture," *Distributing and Accessing Language Resources Workshop, 1st Int. Conf. on Language Resources and Evaluation,* University of Grenada, May 26-30, 1998.

Zajac, R., M. Casper and N. Sharples. 1997. "An Open Distributed Architecture for Reuse and Integration of Heterogeneous NLP Components," *Proc. 5th Conf. on Applied Natural Language Processing*, March 31 - April 3, 1997, Washington, D.C.

Zechner, K. 1997. *A Literature Survey on Information Extraction and Text Summarization,* Carnegie Mellon University, April 14, 1997.