Applying Link Analysis on Automatically Extracted Information from Texts Within KNOW-IT[1] (KNOWledge Base Information Tools)

Woojin Paik, Elizabeth Liddy, Eileen Allen, Eric Brown, Andrew Farris,
Robert Irwin, Jennifer Liddy, and Ian Niles

TextWise Inc.
2-212 Center for Science and Technology
Syracuse, NY 13244

{woojin, liz, eileen, eric, drew, rjirwin, jennifer, ihniles}@textwise.com

Introduction

A robust information extraction system, which can accurately and rapidly convert vast amounts of textual data into a semantic network type knowledge representation scheme, will be a useful preprocessor for various link analysis applications. In this paper, we describe an end-to-end knowledge discovery system which takes raw texts as input and generates a knowledge base which can be utilized for a variety of information access systems, such as a visual knowledge base browser and a question-answering system.

We wish to explore the possibility of using our visual knowledge base browser as a link analyzer, and show that it is possible to apply link analysis techniques to vast amounts of textual data through an automatic information extraction system.

First, we will explain the KNOW-IT system. Then, underlying natural language processing techniques which comprise the information extraction system will be described. Finally, browsing will be discussed in the context of information retrieval, and the visual knowledge base browser will be described.

KNOW-IT Overview

KNOW-IT is an integrated suite of robust tools which transform vast, unmanageable bodies of raw textual data into accessible and surveyable intelligence. The initial component of the system is an information extraction module which maps open source documents into structured, unambiguous representations of their content. Unlike most other information extraction systems, the input is not restricted to newswire text - it will encompass everything from State Department travel advisories, to online battlespace manuals, to WWW home pages, to mention just a few examples. Another distinctive aspect of the extraction component of KNOW-IT is that it constitutes a unique compromise between deep, domain-dependent information extraction and shallow, domain-independent extraction. KNOW-IT's pattern-matching heuristics and sense-disambiguation module provide a representation of texts of any subject area. However, political content (particularly that of international significance) is provided with an enriched representation by means of handcrafted templates for verbs and nouns relating, respectively, to political events and actors.

Once the extractions are available, they are loaded into a database where they can be accessed by the browser. The key innovation of the browser is that it seamlessly incorporates the organizational principles of both an ontology and a flat semantic network. One window of the browser allows the user to navigate through the hierarchy of meaning relationships between concepts, while a second window permits one to expand any node of this hierarchy into a flat network representing all of the fact-based relationships involving this node. Thus, the metaphor behind the browser is to structure information in the form of a three-dimensional cone, whose depth corresponds to the semantic relationships between concepts and whose breadth represents the factual information extracted from text. Since this browser is a graphically navigable, automatically constructed index of all of the fine-grained content expressed by input text, it will enable the analyst in a very short period of time to target precisely that information which is relevant to a critical decision.

Aside from being a stand-alone system, KNOW-IT can also be used to enhance the performance of other knowledge-based systems. Since the KNOW-IT extractions are sets of triples which can be easily mapped

[1] This research and development project has been supported in part by Rome Laboratory, USAF under contract F30602-96-C-0164 and DARPA under the High Performance Knowledge Base (HPKB) project.

to both frame-based and logic-based knowledge representations, these extractions can be used by any knowledge-based system to complement its own specialized base of common-sense or domain knowledge. Once the extractions have been loaded into a database, they can be called by a knowledge-based system whenever it requires a repository of general, factual information to answer a query, perform an inference, or locate instances of a concept or relation.

Information Extraction

In recent years, there has been increased interest in textual information extraction research using natural language processing techniques. The most common medium of storing knowledge is texts; textual information extraction is an approach to acquire knowledge from texts.

Many noticeable research efforts have been in Message Understanding Conferences (MUC). The goal of MUC (Chincor, 1992) was to automatically extract information from news texts to populate databases. Participants of MUC were given the task of extracting information about clearly defined event types (or domains) such as "terrorism in South America." For each event type, the MUC participants were given pre-determined categories of information that their systems were required to extract.

The approaches which were employed in MUC depend on the careful analysis of common terminologies which are used in each event type. Thus, every participating system has to be reworked either to capture the typical roles of the exhaustive list of entities which have potential to occur in the designated event or to identify all possible verbs which can be used to describe the event and the associated roles of the syntactic arguments of the verbs. These processes can take long periods of time, varying from a few weeks to several months.

Most participating systems were successful in extracting relevant information; however, given that there are almost infinite number of event types or subject domains, it does not seem feasible to build a domain-independent textual information extraction system by following MUC's one domain at a time approach.

Thus, we have taken an alternative approach to extract domain independent, time-stamped information from various types of texts including news texts by extracting information about named entities without recognizing domain events (Paik, 1994.) This particular approach takes advantages of the common practice among writers of including predictable information-rich linguistic constructions in close proximity to related proper names. By recognizing proper names in a texts, then locating and recording the associated linguistic constructions, it is possible to construct complex, in-depth histographies of events and their specific relation to a given proper name that can span several decades.

In addition, our approach utilizes more general semantic relations to link concepts which occur in texts in contrast to the relations which are used in MUC. For example, MUC used semantic relations such as 'weapons used' or 'victim' in the 'terrorism in South America' domain. However, we limited the semantic relations to fairly generic ones such as affiliation, agent, duration, location, or point-in-time. These differences allowed simplification of the developed information extraction algorithm and the application of the information extraction system to domain-independent data.

The application of natural language processing techniques for knowledge discovery in the KNOW-IT system can be summarized as conversion from 1) 'text' to the 'extraction' output; 2) 'extraction' output to 'semantic representation'; and 3) 'semantic representation' to 'knowledge product'.

The Figure 1 shows input text to the KNOW-IT information extraction system. The pseudo-Standard Generalized Markup Language (SGML) tags, which are enclosed between angle brackets, such as <HL> signals the beginning of the headline and <SO,DATE> shows that the source and date information of the document will follow the code.

<*HL*> Annan to stay in Baghdad Friday-Sunday
<*SO,DATE*> AFX-Europe, 2/18/98

During a meeting in Tehran with Iraqi Foreign Minister Mohammed Said al-Sahhaf, Kharazi called for cooperation with the UN and denounced the foreign military presence in the Gulf.

Iran, which opposes the growing US military might in the Gulf, says that a U.S. attack on Iraq would serve Israel's interests.

Figure 1. Text

The document processing module, which consists of various natural language processing programs, identifies various fields, clauses, parts-of-speech (POS), and punctuation in a text, and annotates a document with identifying tags for these units. The identification process occurs at the sentence, paragraph, and discourse levels and is a fundamental precursor to the concept-relation-concept (CRC) extraction step. The extracted CRCs form semantic networks which can be used for link analysis or other knowledge discovery processes.

<PN>
0 30 Kamal Kharazi
31 Foreign Minister
7 Iran
1 1 Tehran
7 Iran
2 30 Mohammed Said al-Sahhaf
31 Foreign Minister
7 Iran
3 50 United Nations
4 6 Persian Gulf
5 7 Iran
6 7 United States
...
<FA> During|IN a|DT meeting|NN in|IN Tehran|NP|1 with
Iraqi_Foreign_Minister_Mohammed_Said_al-Sahhaf|NP|2 ,|, Kharazi|NP|0 called|VBD
for|IN cooperation|NN with|IN the|DT UN|NP|3 and|CC denounced|VBD the|DT
<cn> foreign|JJ military|JJ presence|NN </cn> in|IN the|DT Gulf|NP|4 *</FA>*

<FA> Iran|NP|5 ,|, which|WDT opposes|VBZ the|DT growing|VBG US|NP|6
<cn> military|NN might|NN </cn> in|IN the|DT Gulf|NP|4 ,|, says|VBZ *</FA>*
<AN/FU> that|WDT a|DT <cn> U.S.|NP|6 attack|NN </cn> on|IN Iraq|NP|7 would|MD
serve|VB <cn> Israel|NP|8 's|POS interests|NNS </cn> .|.
</AN/FU>

Figure 2. 'extraction' output

Figure 2 shows the annotated result from the POS tagger which assigns one of 48 possible grammatical forms such as preposition (IN), determiner (DT), or singular noun (NN) to words. Each word in the text is followed by a '|' symbol and a POS tag.

Certain phrases such as 'foreign military presence' or 'military might' are enclosed within the '<cn>' and '</cn>' tags, which represents the boundaries of a type of noun phrase called complex nominals.

Proper names (PN) including multi-part names are identified and categorized according to a previously defined PN classification scheme (Paik, et al, 1996.) 'NP' and 'NPS' are the POS tags for proper names and each tag is followed by a numeric id. For example, 'Tehran' is followed by singular PN POS tag, 'NP', and the id, '1'. This id '1' is indication that the information about the PN identified as 1 is shown in record number 1 in the PN table at the beginning of the annotated document. The SGML tag, '<PN>' signals the beginning of the PN table. The first column of the table is the record number. The second column is the PN category id. For example, PN category 30 stands for the name of a person, 31 stands for a title, 7 stands for a name of a country, and 1 stands for a name of a city. Thus, 'Tehran' is categorized as a name of a city (1) which is located in Iran which is a country (7). Multi-part names such as 'Iraqi Foreign Minister Mohammed Said al-Sahhaf' are identified as one unit and categorized by each part.

Finally, the KNOW-IT document processor delineates the discourse-level organization of a document's content by assigning discourse component tags, such as '<FA>' which stands for factual information and '<AN/FU>' which represents for analysis and future, to each clause in the text.

Based on the automatically annotated syntactic, semantic, and discourse level information about the text (shown in Figure 2 in conjunction with the output from semantic parsing), CRC triples which form the semantic representation of the document content are extracted from the text. CRC triples consist of semantic relations, which are widely used in Conceptual Graph related studies (Sowa, 1984), and concepts, which are linked by the semantic relations.

For example, if 'A opposes B' then 'A' is the agent of 'opposing' and 'B' is the object of 'opposing'. This semantic information is represented as the following:

AGENT (oppose, A)
OBJECT(oppose, B)

During a meeting in Tehran with Iraqi Foreign Minister Mohammed Said al-Sahhaf, Kharazi called for cooperation with the UN and denounced the foreign military presence in the Gulf.
...
AGENT (call for, Kamal Kharazi|person)
OBJECT (call for, cooperation)
ASSOCIATION (cooperation, United Nations|organization)

Iran, which opposes the growing US military might in the Gulf, says that a U.S. attack on Iraq would serve Israel's interests.

AGENT (oppose, Iran|country)
OBJECT (oppose, military might)
AFFILIATION (military might, United States|country)
LOCATION (military might, Persian Gulf|region)
...

Figure 3. Semantic Representation

Figure 3 shows partial output of the CRC extraction and the resulting semantic representation of the text.

Currently, the accuracy of CRC extraction, based on testing data which are not used for training, for texts

such as reports, briefings, and area studies, is 90% and the coverage of the extraction is 75%. Accuracy is defined as a number of correctly extracted CRCs divided by the number of all extracted CRCs. Coverage is defined as the number of correctly extracted CRCs divided by the number of all correct CRCs in the testing data.

Browsing is described as a type of searching where the initial criterion is only partially defined or known, and is characterized by the presence of a search goal but not a strategy.

Finally, Figure 4 shows a schematic view of a semantic network which is a collection of extracted CRCs.

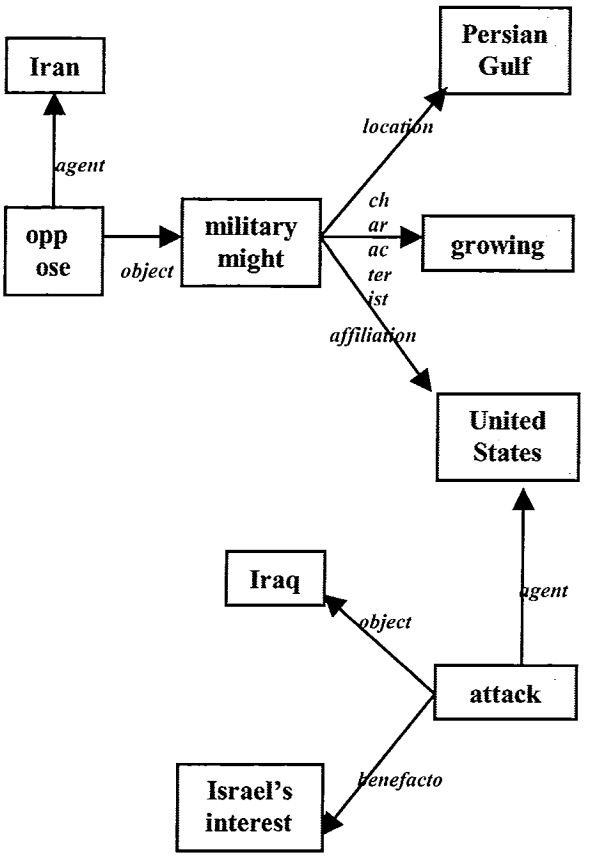

Figure 4. Knowledge Product

Browsing of Semantic Networks

Traditional task-oriented, information retrieval systems have query-based, command-driven user interfaces. Users are required to express and submit their information needs in explicitly formed Boolean or natural language queries to the information retrieval systems. The systems then bring back a list of documents which are possibly relevant to users' queries. The users usually go through the list to find what is relevant by reading documents in the list one at a time. Browsing can be viewed as an alternative approach to the traditional information retrieval tasks.

According to Cove & Walsh (1987), browsing can be divided into three categories:

- *Search browsing*, which is directed, specific, or goal-oriented;
- *General purpose browsing*, which is semi-directed predictive, or purposive; and
- *Serendipity browsing*, which is undirected and not goal oriented

Originally we developed the visualization capability as a knowledge base inspector/editor for the KNOW-IT system. The knowledge base inspector/editor was designed to be used by domain experts to inspect or verify the integrity or the correctness of an automatically constructed knowledge base, as well as to edit the knowledge base by adding or deleting entries. Subsequently, we have determined that KNOW-IT's knowledge base editor/inspector functions well as a semantic network visualizer, and is able to support all three browsing activities described above. Thus, it is useful as a browser for information retrieval tasks.

Figure 5 is a snap shot of the browser. This snapshot utilizes a knowledge base concerning Iran. The knowledge base was constructed by extracting information from documents which contain the country name, Iran, from New York Times documents published between July 1994 and December 1996. There were 1,458 documents in this data set.

A window in the browser, labeled as the Hyperbolic Browser, located on the upper left side of Figure 5, is the ontology navigation window described in the KNOW-IT Overview section. This example shows the instances from the Religion portion of the ontology which occurred in the Iran knowledge base. To explore these instances further, users can select one concept/node from the window. Any concept/node in the left window can be selected in turn to further explore or drill down and retrieve relevant information about a specific concept.

The window with the label, Concept Visualzer, on the right side of the screen, is the flat semantic network representing all of the fact-based relationships based on the selected concept/node from the ontology window. Currently, the window shows the state when a user has selected 'Muslim' from the left window and selected 'Party of God' successively from the initial right window view about 'Muslim.' The Concept Visualizer window displays related information about a guerilla group called, 'Party of God', including the facts that Iran supports the group, and the group is a radical anti-Israel movement.

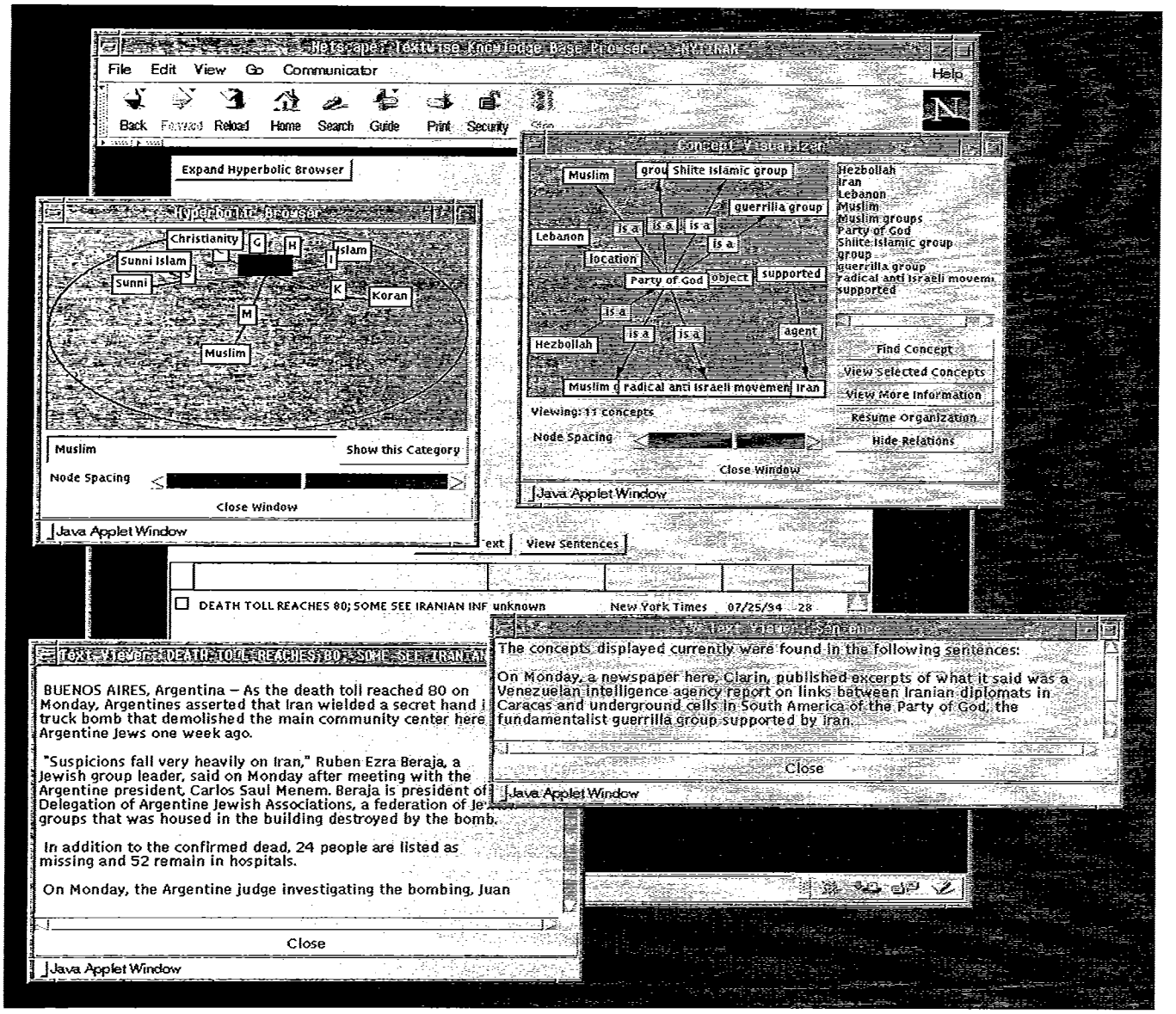

Figure 5 Visual Knowledge Base Browser

The bottom left window shows the full text view of the document which is the source of the extracted information and the bottom right window shows a specific sentence from which the visualized information was extracted.

We believe that the browsing which begins with the selection of a concept in the ontology window supports the previously mentioned *General browsing* and *Serendipity browsing* types. In addition, the browser has the functionality which allows users to type in a concept they wish to display, and this feature supports *Search browsing*.

Conclusion

We have discussed automatic information extraction system, which is a part of the larger knowledge discovery system, and the browser, which visualizes the semantic network representation of the extracted information.

Recently, we have constructed a knowledge base from one year's worth of a newswire which consists of 410,000 documents. From this document set, 4.2 million CRCs were extracted and formed a knowledge base. There were about 1.2 million unique concepts in the extracted CRCs.

We believe our domain-independent information extraction system can be used to extend the potential of link analysis to information which resides in a vast number of textual documents. In addition, our unique combination of ontology and semantic network based visual browsing strategy might serve as an interesting method to conduct link analysis.

References

Chincor, N. "MUC-4 Evaluation Metrics," Proceedings of the Fourth Message Understanding Conference (MUC-4), McLean, VA June 16-18, 1992.

Cove, J.F. & Walsh, B.C. "Browsing as a Means of Online Text Retrieval," Information Services and Use, 1987.

Paik, W., Liddy, E.D., Yu, E. & McKenna, M. "Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval," Corpus Processing for Lexical Acquisition, Boguraev, B. & Pustejovsky, J. (eds.), Cambridge, MA: MIT Press, pp. 62-73, 1996.

Paik, W. "Chronological Information Extraction System (CIES)," Dagstuhl-Seminar-Report: 79 on Summarizing Text for Intelligent Communication, Endres-Niggermeyer, B., Hobbs, J. & Jones, K.S. (eds.), Wadern, Germany: IBFI, 1994.

Sowa, J., Conceptual Structures: Information Processing in Mind and Machine, Reading, MA: Addison-Wesley, 1984