Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2004 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2004

Extracting Financial Information from Text Documents

Susan Lukose
University of Mississippi

Frank Mathew
University of Mississippi

Sumali Conlon
University of Mississippi

Pamela Lawhead
University of Mississippi

Recommended Citation

Lukose, Susan; Mathew, Frank; Conlon, Sumali; and Lawhead, Pamela, "Extracting Financial Information from Text Documents" (2004). *AMCIS 2004 Proceedings*. 232.
http://aisel.aisnet.org/amcis2004/232

Extracting Financial Information from Text Documents

Susan Lukose
University of Mississippi
svlukose@olemiss.edu

Frank Mathew
University of Mississippi
fmathew@bus.olemiss.edu

Sumali Conlon
University of Mississippi
sconlon@bus.olemiss.edu

Pamela Lawhead
University of Mississippi
lawhead@cs.olemiss.edu

ABSTRACT

The majority of electronic data today is in textual form. Financial data such as articles in the Wall Street Journal are written as texts. These electronic documents contain a wealth of information but require human interpretation. For financial analysis, rapid up-to-date information is critical. Most software tools currently require data which are better structured than text (such as data in relational databases). Thus, our research goal is to build a system, "FIRST" (Flexible Information extRaction SysTem), that will extract data from financial articles and store the output in an explicit format. FIRST uses natural language processing techniques and resources such as the lexical database WordNet and collocation information to extract information. We hope to be able to extract data such as an organization's name, its profit/loss status, and sales status, from financial articles to input into a database. The data will come from international corporate reports which appear in the Wall Street Journal.

Keywords

Information Extraction, Natural Language Processing, Report Generation

INTRODUCTION

Textual information available in electronic form has grown exponentially with the growth of the Internet. In financial applications, current financial decision making processes use outputs from financial analysis tools. Most of these tools require a huge set of data in the database. However, since most of the original data is in textual form, manually deciphering information is practically impossible when the data is spread over a huge number of documents. There is therefore a need to efficiently automate information extraction whenever possible.

There have been quite a few attempts at automating information extraction, most of which have been domain specific. Domain specific systems have limited application; this makes them more accurate and easier to build. The scope of such systems, however, is very limited. Another approach has been to develop systems that, once developed, have only a portion which is domain dependent. For each new application, only this portion must be rebuilt, while the rest of the system remains unchanged (Bagga, Chai and Biermann 1997). FIRST (Flexible Information extRaction SysTem) is such a system, that we are in the process of building. It will perform information extraction using methods that involve semantic and syntactic text analysis and natural language processing techniques.

Although FIRST can potentially be used for information extraction from text in any area, the application on which we focus is the generation of reports on financial data that appear in the international corporate reports of the Wall Street Journal. FIRST uses a two step approach to extract information. Initially, in the domain training phase, a knowledge base is built. This knowledge base will consist of a keyword lexicon, which will enable the system to determine the different items of interest, and a set of rules to be used for extracting information. In the second phase the documents are analyzed using natural language processing techniques to output the extracted information in a form (say tabular) suitable for entering into a relational database.

In the rest of the paper, we discuss related work, present a system overview and the system architecture, explain how we will evaluate the system, and conclude.

RELATED WORK

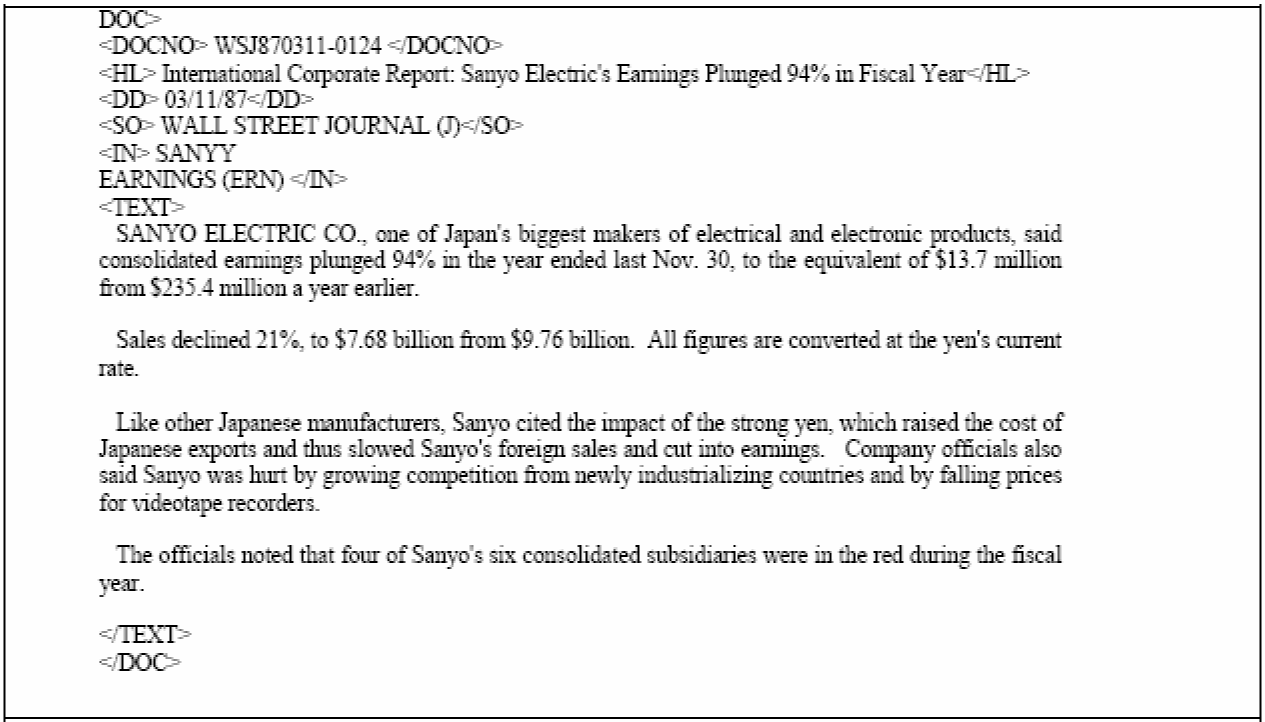
The use of Natural Language Processing techniques to extract information from text has been explored for at least the last 15 years. The DARPA-sponsored Message Understanding Conference (MUC) was one of the earliest conferences to feature systems that use natural language processing techniques for information extraction (Cardie, 1997). In 1990, Jacobs and Rau, developed SCISOR, a system that searches financial news to find and summarize corporate merger stories (Jacobs and Rau 1990). Later, in 1997, Holowczak and Adam built a system that automatically classifies legal documents (Holowczak and Adam, 1997). A trainable domain independent system that used WordNet was built by Bagga et al. at Duke University (Bagga, Chai, and Biermann, 1997). Gerdes built an information extraction system called "EDGAR-Analyzer." The system analyzes corporate data contained in the SEC's EDGAR database (Gerdes, 2003). Leroy et al. extract relations between noun phrases automatically from a collection of biomedical text (Leroy et al.. 2003).

The main drawback of domain-dependent systems is that they are not portable, while the accuracy of a domain independent trainable system depends on an aptly chosen set of training data. Therefore the accuracy, robustness and portability of information extraction systems can still be improved (Cardie, 1997). This can be done by making the system flexible enough to adapt to a new domain with minimal changes while maintaining accuracy.

SYSTEM OVERVIEW

The system is trained for a specific domain during a Building Phase. During this Building Phase, the system uses a training set of documents which, in our application, will be financial articles from the Wall Street Journal. The system uses this training set to build a knowledge base consisting of a keyword lexicon and a set of extraction rules. This knowledge base is used in the "Functional Phase" to perform information extraction using text analysis and natural language processing techniques.

Figure 1 shows a sample article on a corporate report, published in 1988 in the Wall Street Journal.



DOC>
<DOCNO> WSJ870311-0124 </DOCNO>
<HL> International Corporate Report: Sanyo Electric's Earnings Plunged 94% in Fiscal Year</HL>
<DD> 03/11/87</DD>
<SO> WALL STREET JOURNAL (J)</SO>
<IN> SANYY
EARNINGS (ERN) </IN>
<TEXT>

SANYO ELECTRIC CO., one of Japan's biggest makers of electrical and electronic products, said consolidated earnings plunged 94% in the year ended last Nov. 30, to the equivalent of $13.7 million from $235.4 million a year earlier.

Sales declined 21%, to $7.68 billion from $9.76 billion. All figures are converted at the yen's current rate.

Like other Japanese manufacturers, Sanyo cited the impact of the strong yen, which raised the cost of Japanese exports and thus slowed Sanyo's foreign sales and cut into earnings. Company officials also said Sanyo was hurt by growing competition from newly industrializing countries and by falling prices for videotape recorders.

The officials noted that four of Sanyo's six consolidated subsidiaries were in the red during the fiscal year.

</TEXT>
</DOC>

Figure 1. Sample article on a Corporate Report, Article Published in 1988 in the Wall Street Journal.

The article is coded in the SGML format and is about the earnings of the Sanyo Electric Co. It describes what happened to sales over the previous year and how earnings were affected. FIRST is initially configured to extract desired information. Therefore when it receives this article as an input it should be able to generate a report by extracting information of interest which could be seen in a tabular form as shown in Table 1 (note that, at this stage in our research, this table is created by hand, for illustrative purposes, and is not yet generated automatically by our system). From the table we see that the contents of the article have been sieved to leave behind just the information of interest as instructed during configuration. This information explains what has happened to the subject which in this case is the firm's earnings. It also presents the firm's sales information, if it has been provided in the article.

Organization Name:	SANYO ELECTRIC CO
Organization Description:	One of Japan's biggest makers of electrical and electronic products
Fact / Prediction	Fact (Has Happened)
Financial Item:	Earnings
Financial Item Status:	Fell
Financial Item % Change:	94%
Financial Item Change Description:	From $235.4 million to $13.7 million
Sales Status:	Fell
Sales % Change:	21%
Sales Change Description:	from $9.76 billion to $7.68 billion

Table 1. Sample Corporate Report Table

SYSTEM ARCHITECTURE

FIRST has five major sub-processes. The Keyword Lexicon Builder and the Extraction Rule Generator form the Building Phase, while Document Filter, Tokenization process, and the Extraction Engine are the major components of the Functional Phase. The general architecture of FIRST is shown in Figure 2.

Building Phase

The Building Phase is the domain training portion of FIRST. The system is provided with a subset of domain documents for use in training. Using these training documents, knowledge is generated which will be used later to extract information. This knowledge consists of keywords and extraction rules, and is maintained in the Knowledge Base to be used later in the Functional Phase. The keywords are generated by the Keyword Lexicon Builder, which is fed with an initial set of keywords by a domain expert, while the extraction rules are generated by the Extraction Rules Generator. The Keyword Lexicon Builder consists of three main components: WordNet, the KWIC Index builder, and the Semantic Classifier.

WordNet

WordNet is a lexical database developed at the Princeton University Cognitive Center (Miller et al., 1990, Miller, 1995). It is widely used for natural language processing research, as it provides a means to identify semantically similar words. In contrast to several information extraction systems that rely on a hand-coded Knowledge Base, FIRST uses WordNet in the Keyword Lexicon Builder for this process. WordNet helps to build upon the domain expert's initial set of keywords by providing synonyms. For example, the keyword "decline," which appears as a verb for the word "sales" in the sample article above, has the synonyms shown in Table 2.

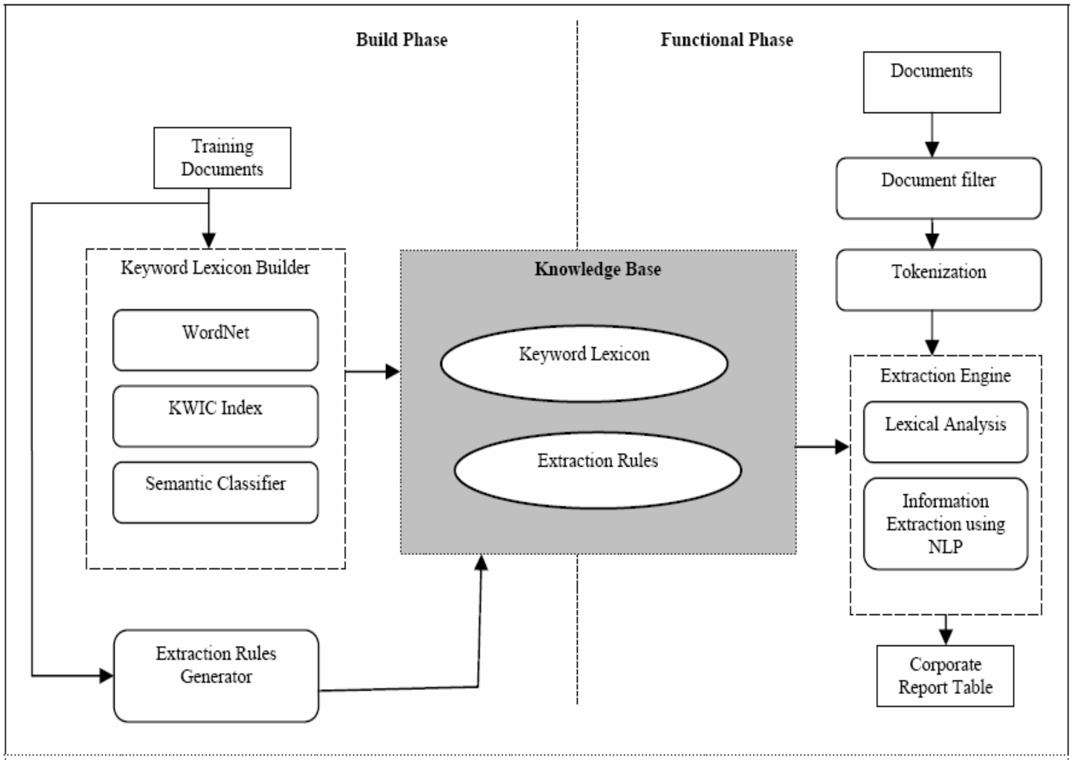


Figure 2. System Architecture of FIRST

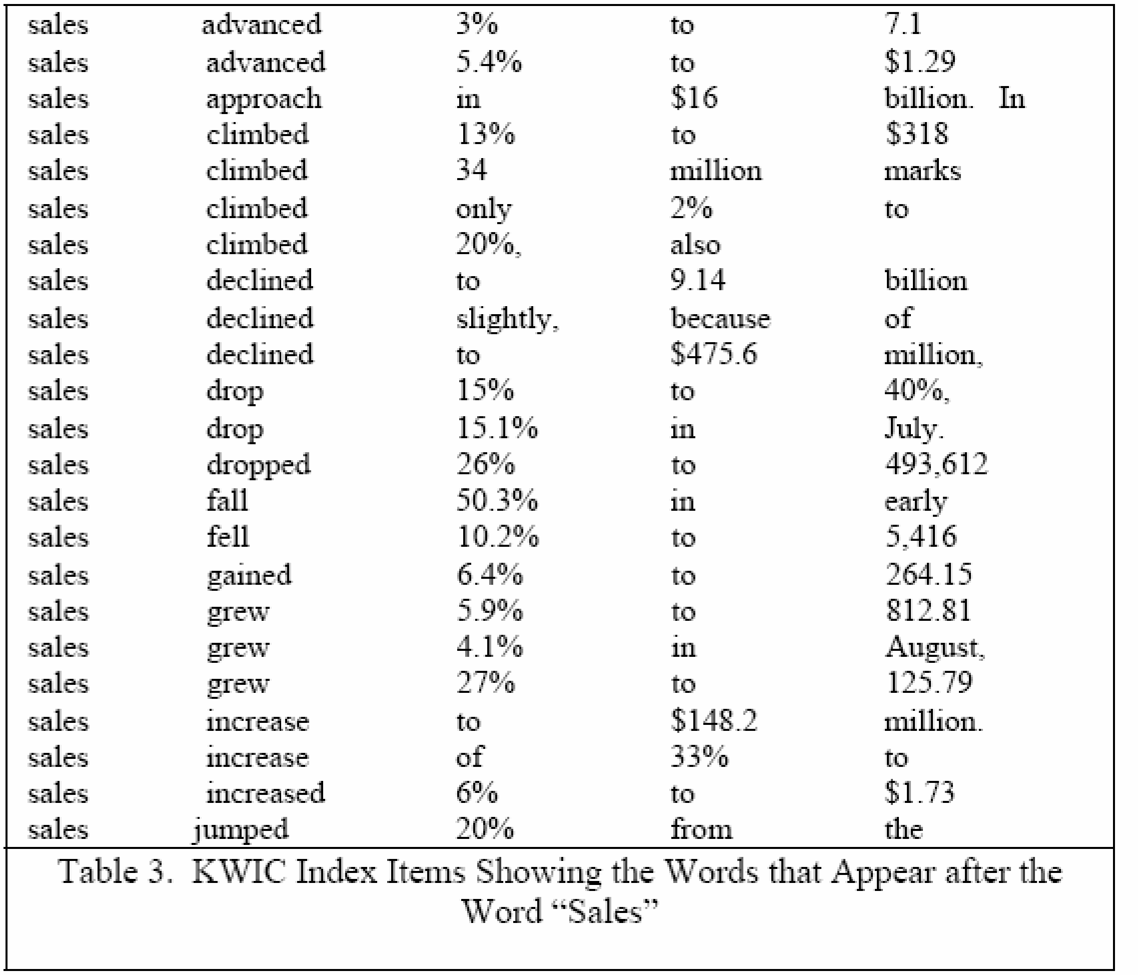
Change state	incline
condition	inflect
decrease	lessen
decrement	side
diminish	slope
drop	status
fall	turn

Table 2. WordNet Synonyms for "Decline"

Unfortunately, WordNet provides too many words that express the same meaning as "decline." FIRST therefore uses a KWIC Index to identify the keywords that are relevant to the given context.

KWIC Index

A KWIC (Key Word In Context) Index is an index of all the words in a text, showing every context in which each word occurs (Luhn, 1960). For each occurrence of each word, it shows the words that appear before and after the word in question. Table 3 shows some sample KWIC index items, specifically, the words that appear after the word "sales" from Wall Street Journal text. In the training stage, the Keyword Lexicon Builder feeds each word obtained from WordNet to a KWIC Index obtained from the set of training documents. The output of the KWIC Index is analyzed to predict if this word should be a keyword, based on the contexts in which it occurs. If this word is predicted to be a keyword, it is passed on to the semantic classifier. The word "fall," for example, which is returned by WordNet as a synonym for "decline," also appears as a verb for the word "sales," so the system will accept the word "fall" as a synonym for "decline" in this domain.



sales	advanced	3%	to	7.1
sales	advanced	5.4%	to	$1.29
sales	approach	in	$16	billion. In
sales	climbed	13%	to	$318
sales	climbed	34	million	marks
sales	climbed	only	2%	to
sales	climbed	20%,	also	
sales	declined	to	9.14	billion
sales	declined	slightly,	because	of
sales	declined	to	$475.6	million,
sales	drop	15%	to	40%,
sales	drop	15.1%	in	July.
sales	dropped	26%	to	493,612
sales	fall	50.3%	in	early
sales	fell	10.2%	to	5,416
sales	gained	6.4%	to	264.15
sales	grew	5.9%	to	812.81
sales	grew	4.1%	in	August,
sales	grew	27%	to	125.79
sales	increase	to	$148.2	million.
sales	increase	of	33%	to
sales	increased	6%	to	$1.73
sales	jumped	20%	from	the

Table 3. KWIC Index Items Showing the Words that Appear after the Word “Sales”

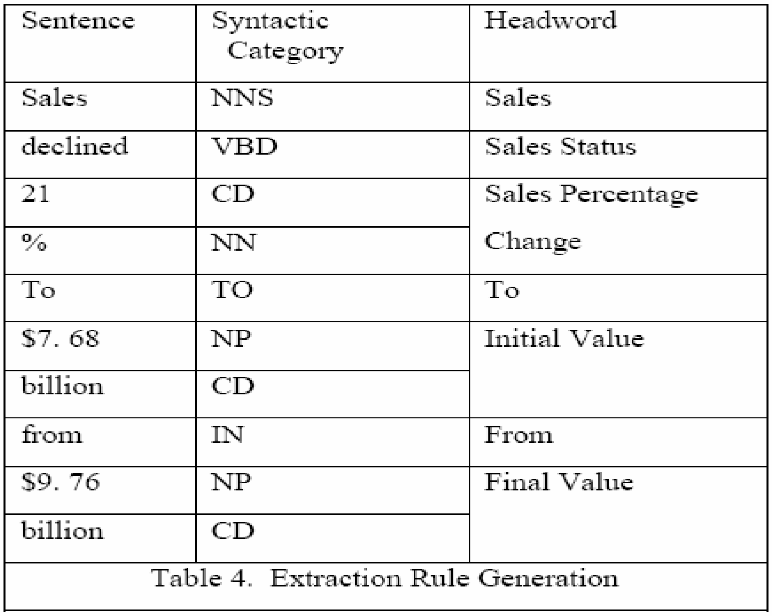
Semantic Classifier

For the current Financial Data application, FIRST will maintain a separate set of keywords to extract values for the attributes: Financial Items, Financial Status, Financial Cause and Sales Status. Once a word is identified as a keyword, the Semantic Classifier decides to which of these sets the new keyword should be added.

Extraction Rules Generator

The training set of documents is analyzed to identify patterns in the text’s presentation of the data containing the information of interest. Based on these patterns, extraction rules are generated that will enable the Extraction Engine to extract information. For example, Table 4 illustrates the generation of a rule based on the following sentence that is parsed and split into parts of speech:

Sales declined 21%, to $7.68 billion from $9.76 billion.



Sentence	Syntactic Category	Headword
Sales	NNS	Sales
declined	VBD	Sales Status
21	CD	Sales Percentage Change
%	NN	
To	TO	To
$7. 68	NP	Initial Value
billion	CD	
from	IN	From
$9. 76	NP	Final Value
billion	CD	

Table 4. Extraction Rule Generation

Where NNS – Plural Noun; VBD – Past Tense Verb, CD – Cardinal Number, NP – Noun Phrase,

NN -- Singular Noun, TO – Infinitive Marker, IN – Preposition (from Manning and Schutze, 1999, pp142).

The rule generated based on the sentence in Table 4 is: When a sentence with the headword "Sales" is observed, the verb in the past tense next to it is the Sales Status. This is followed by the sales percentage change and then by the actual change in sales "to" a final value "from" an initial value. Using the same technique as above, a comprehensive set of extraction rules will be generated.

Functional Phase

In the Functional Phase, from the input documents, the relevant articles are filtered out. In this application, these are the articles concerning corporate reports. Each of these articles is then split into sentences and then passed on to the Extraction Engine.

Extraction Engine

The sentences from an article received by the extraction engine are subjected to part of speech tagging by a POS (Part Of Speech) Tagger (Schmid, 1994; Brill 1994). The lexical analyzer uses the Keyword Lexicon in the Knowledge Base to determine the sentences containing information to extract. Extraction Rules together with the Keyword Lexicon are then used to extract information and generate the report table. Optionally, the corporate report table that is generated can also be stored in a relational database like Oracle.

SYSTEM EVALUATION

FIRST is in its developmental stage. The Keyword Lexicon has been built using the Keyword Lexicon Builder and the process of filtering and tokenizing corporate report articles has also been completed. We are now in the process of generating extraction rules which will then be followed by the Extraction Engine. Articles that appeared in the Wall Street Journal will be used to train and evaluate the system (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93T3B). Articles that appeared in 1987 and 1988 are used as training documents. Once the system is ready it will be evaluated with documents from 1989. The resulting reports table will be evaluated by a domain expert, who will determine the accuracy of the system.

CONCLUSION

Financial data are important for business decision makers and government officials. These decision makers need accurate and up-to-date data. Currently, people have to manually convert the data implicit in article format to explicit forms. Our ongoing research project, "FIRST," is intended to be a flexible information extraction system that has a domain training portion. Its unique architecture uses WordNet and a KWIC Index to generate keywords, and its Extraction Engine uses Natural Language Processing (NLP) techniques. The system is being developed in Perl while WordNet and the KWIC Index are accessed from an Oracle Database. We intend to have FIRST developed by the summer of 2004 and expect that it will be more accurate and portable when compared to existing systems.

REFERENCES

1. Bagga, A. , Chai, J. , and Biermann, A. (1997) The Role of WordNet in the Creation of a Trainable Message Understanding System, *Proceedings of Ninth Conference on Innovative Applications of Artificial Intelligence (IAAI-97)*.
2. Brill, E. (1994) Some advances in rule based part-of-speech tagging, *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 722--727, Seattle, WA.
3. Cardie, C. (1997) Empirical methods in information extraction, *AI Magazine*, 18(4):65-80.
4. Gerdes, John (2003) EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database, *Decision Support Systems*, Vol. 35: 7-29. Elsevier Science Publishers.
5. Holowczak, R. D. and Adam, N. R. (1997) Information Extraction based Multiple-Category Document Classification for the Global Legal Information Network, *Proceedings of the Ninth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-97),* Providence, Rhode Island.
6. Jacobs, P. S. and Rau, L. (1990) SCISOR: Extracting information from On-Line News. *Communications of the ACM*, 33(11):88-97.
7. Leroy, A, H. Chen, and Martinezb J. (2003) A shallow parser based on closed-class words to capture relations in biomedical text. *Decision Support Systems*, Vol. 36: 145-158. Elsevier Science Publishers.
8. Luhn, H. P. (1960). Keyword-in-context index for technical literature (KWIC index), *American Documentation* 11:288-295.
9. Manning, C., and Schutze H. (1999) Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge.
10. Miller, G. A. , Beckwith, R. , Fellbaum, C. , Gross, D. and Miller, K. J. (1990) Introduction to WordNet: an on-line lexical database, *International Journal of Lexicography*, Vol. 3, No. 4: 235 - 244.
11. Miller, G. A. (1995) WordNet: a Lexical Database for English, *Communication of the ACM,* Vol . 38, No 11: 39-41.
12. Schmid, H. (1994) Probabilistic part-of-speech tagging using decision trees, *Proceedings of International Conference on New Methods in Language Processing*.