Extraction of Financial Information from Online Business Reports

Lakisha L. Simmons
Belmont University

Sumali J. Conlon
University of Mississippi

Abstract

CAINES, Content Analysis and INformation Extraction System, employs a semantic based information extraction (IE) methodology through a design science approach to extract unstructured text from the Web. Our system was knowledge-engineered and tested on an active business database by experts who use the database regularly to perform their job functions. We believe that by heavily involving business experts, we are able to advance our thinking about IS research. CAINES extracts information to meet three objectives that were deemed important by our experts: (1) understand what current market conditions impacted the growth of certain balance sheets (2) summarize management's discussion of potential risks and uncertainties (3) identify significant financial activities including mergers, acquisitions, and new business segments. These objectives were developed based on the advice of financial experts who regularly analyze financial reports.

A total of 21 online business reports from the EDGAR database, each averaging about 100 pages long, were used in this study. Based on financial expert opinions, extraction rules were created to extract information from financial reports. Using CAINES, one can extract information about global and domestic market conditions, market condition impacts, and information about the business outlook. User testing of CAINES resulted in recall of 85.91%, precision of 87.16%, and an F-measure of 86.46%. Speed with CAINES was also greater than manually extracting information. Users agreed that CAINES quickly and easily extracts unstructured information from financial reports on the EDGAR database. This study highlights the significance of creating a semantic based IE system that addresses practical business issues and solves a true business problem with the knowledge of business experts.

Keywords: Information extraction, Business intelligence, Sublanguage analysis, EDGAR.

ACM Categories: I.0, I.2, I.2.7.

General Terms: Documentation

Introduction

Useful information drives decision making, business solutions, and even competitive advantages. For example, the United States Securities and Exchange Commission (SEC) requires major financial businesses in the United States to submit financial reports to their electronic data gathering, analysis, and retrieval (EDGAR) online database. The reports dispersed throughout the EDGAR database are very useful for businesses searching for data for benchmarking and recent competitor moves. However, many reports on EDGAR are over 100 pages long and keyword searches are prone to irrelevant results.

This paper presents the idea of using sublanguage analysis and information extraction (IE) techniques to extract unstructured data from online text documents and make the semantic relationships within the text available for further analysis. Information extraction is automatic extraction of structured information such as entities, relationships between entities and attributes describing entities from unstructured sources (Sarawagi, 2008).

In this paper, CAINES extracts and analyzes information from reports over 100 pages long to meet three objectives that were deemed important by our experts: (1) understand what current market conditions impacted the growth of certain balance sheets (2) summarize management's discussion of potential risks and uncertainties with moving forward (3) identify significant financial activities including mergers, acquisitions, and new business segments. These objectives were built on the opinions of financial experts who regularly analyze financial reports. The experts were looking for a concise, easy, and useful way to retrieve information from lengthy EDGAR financial reports versus their manual retrieval method. The experts consisted of an Area Manager for SunTrust Bank, a Senior Audit Officer of a Tennessee based bank, and a Senior Cost Accountant for PepsiCo Beverages and Foods. We determined that conducting information extraction based on knowledge engineering and sublanguage analysis, was the most appropriate approach since the majority of the data is quite domain specific and each company had a tendency to provide reports with similar patterns and common terms.

Studies have been able to demonstrate the basic extraction of unstructured and semi-structured data from Web pages like theKnowItAll system (Etzioni et al. 2004; Etzioni, 2008) and financial databases like FIRST (Conlon, Hale, Lukose, & Strong, 2008). We go further by implementing a system with a knowledge engineering approach and test it on an active business database by experts who use the database regularly to perform their job functions. This research asks three questions:

1. Can an IE system (CAINES) be designed to assist business users in extracting information from online financial reports?
2. Can using CAINES result in performance greater than manually extracting information from financial reports?
3. Will users be satisfied with using CAINES as an information extraction system?

Analysis of these research questions can further the development of information extraction.

Over most extraction systems, CAINES was knowledge-engineered and tested on an active business database by experts who manually access the database to perform their job functions. We believe that by heavily involving business experts, we are able to advance our thinking about IS research. CAINES will handle lengthy documents with a wide range of information and will extract more semantic facts to better meet the needs of the target business users. CAINES will output semantic relationship phrases from the system to a user interface. It builds upon the use of IE techniques in the financial markets (e.g. Conlon, Hale, Lukose, & Strong, 2008) to advance more efficient and effective searches by using information extraction and sublanguage analysis. This study makes extensive use of information extraction from natural language for training to analyze content in text.

The next section begins with a discussion of information extraction, notable IE systems, and acceptance of systems. Next we describe the extraction and performance hypotheses. The semantic extraction methodology and testing procedures are then outlined before the results are revealed. The final section concludes with a summary of our research question results, a discussion of the limitations, and suggestions for future research.

Literature Review

Seminal IE Systems and Conferences

Information extraction (IE) is automatic extraction of structured information, such as entities, relationships between entities, and attributes describing entities, from unstructured sources, such as text corpora or text documents (Sarawagi, 2008). IE, like text summarization and machine translation, is a sub-field of natural language processing. IE is typically achieved by scanning a set of documents written in a natural language and populating a database with the extracted information. The origins of IE date back to the Cold War era of the 1960s. Several prominent IE systems were developed in the 1960s and 1970s. The ability to incorporate syntactic and semantic information in NLP techniques brought about new ideas in IE (Wilks, 1997). Naomi Sager and her colleagues first successfully applied IE technology to extract hospital discharge information from patient records in 1970. Dictionary lookup and pattern matching were used to extract relevant medical information in a limited context (Sager, et al. 1987).

In the early 1970s, Gerald DeJong developed an IE system named FRUMP. FRUMP used newswire articles to determine the relevant information using keywords and sentence analysis (Cowie & Lehnert, 1996). Even though it had weaknesses in performance, FRUMP became the basis of a later commercial IE system named ATRANS, the earliest IE system to be used for commercial purposes. ATRANS was able to extract bank money transfer information from telexes by using sentence analysis as FRUMP. The assumption of ATRANS was that the sentence structures of money transferring telex messages are predictable (Lytinen, 1993). A system named JASPER (Journalist's Assistant for Preparing Earnings Report), developed by Carnegie Group of Reuters Ltd., was designed to extract corporate news stories related to earnings, dividends, or income. JASPER requires manual evaluation using a set of test documents collected from PR newswire, which is a drawback (Andersen et al., 1992). JASPER employs knowledge representation, syntactic and semantic knowledge of sentences and domain dependent regularities of patterns, with higher precision of extraction than the previously developed systems. Its performance has been used as a baseline by TIPSTER and MUC (Message Understanding Conferences) systems.

CAINES (Content Analysis and INformation Extraction System) was built using a knowledge engineering approach. CAINES can analyze documents from several online sources including web blogs, customer reviews, business and government reports, and online news articles. In 2009, CAINES was used to analyze eWOM reviews from a sample of 18 action and adventure movies consisting of 20,677 individual reviews (Simmons, et al., 2011). It analyzes texts using syntactic and semantic techniques. Results from the system showed storyline is most important and consumers tend to leave more positive than negative reviews. Findings also reveal key sentiments of consumers' evaluations towards movies, something not found in many other studies.

Conlon, Hale, Lukose, and Strong (2008) created FIRST (Flexible Information extRaction SysTem) that is able to extract financial information from The Wall Street Journal (WSJ). Using a training set of documents from the WSJ, FIRST builds a knowledge base. The FIRST system relies on a service-oriented framework with information retrieval (IR) and IE components. The IR component retrieves source documents and the IE component analyzes the documents and converts news articles from The Wall Street Journal into a data template. It extracts specific information such as "sales rose 5%". FIRST demonstrated that it can be employed to extract information from unstructured Web documents and translate it into extensible markup language (XML). Recall was 85%, precision 90%, and F-measure at 87%.

Hypotheses Development

IE Performance

CAINES' performance is evaluated through speed, recall, precision, and F-measure. CAINES was developed to increase a user's efficiency in extracting information from online business reports. The average user processing time is defined as the time the participants spent answering questions with CAINES or by manually reading and extracting information from reports. It is expected that CAINES will allow a user to perform information extraction and answer questions faster than without CAINES.

> Hypothesis 1 (H1): Average processing time is faster when using CAINES than when manually extracting information from EDGAR 10-Q reports.

To test recall, a comparison will be made between the amount of relevant information a user is able to extract using CAINES versus manual extraction. Recall is measured by dividing the correct number of answers given by the total possible number of correct answers. Users of CAINES are expected to receive better recall results than from the manual extraction since it is so difficult to manually extract specific information from such lengthy reports.

> Hypothesis 2 (H2): Average recall is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.

Similarly, a comparison is made between the amount of information the users were able to extract using CAINES and manual extraction. Precision is measured by dividing the number of correct user answers by the total answers the user was able to produce for each method. Users of CAINES are expected to receive better precision results than from the manual extraction.

> Hypothesis 3 (H3): Average precision is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.

Accordingly, it is expected that the F-measure for CAINES will be greater than the F-measure for manual extraction. The F-measure is a combination of the equal weighted results of precision and recall.

Hypothesis 4 (H4): Average F-measure is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.

Technology Acceptance

For years researchers have studied why users accept certain systems and reject others (Goodwin, 1987; Venkatesh & Davis, 1996). The technology acceptance model (TAM) is one of the most widely used behavioral prediction models in the information systems (IS) field (Davis, 1986; Davis, 1989; Davis, Bagozzi, & Warshaw, 1989). The TAM is an information systems model to predict a behavioral outcome, the adoption and use of an information system. The two antecedents, perceived usefulness (PU) and perceived ease of use (PEOU), are related to behavior intentions (BI) in the TAM model. PU was first operationalized to refer to a user's subjective probability that using a specific system will increase job performance, and PEOU refers to the perception degree that the system will be free of effort (Davis, 1989). TAM is an application of Theory of Reasoned Action (TRA), a behavioral model of prediction of behavioral intention, developed by Fishbein & Ajzen (1975).

In this study, user satisfaction is the dependent variable hypothesized to be correlated with perceived ease of use and usefulness (Figure 1).

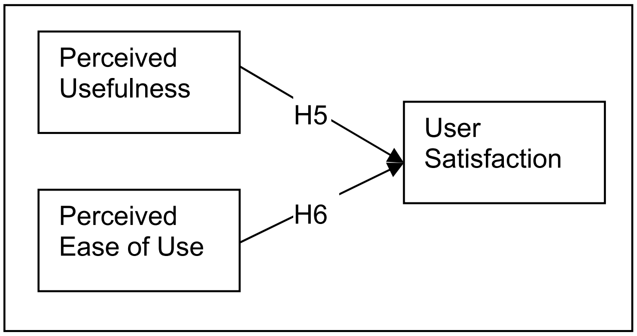


Figure 1. Conceptual model of CAINES' impact

CAINES was developed to be useful in quickly and efficiently finding information in lengthy online financial reports. Hence, perceived usefulness and satisfaction will positively correlate.

Hypothesis 5 (H5): There is a significant positive correlation between perceived usefulness of CAINES and user satisfaction.

CAINES was designed to be user friendly with a clean and sleek design. The extraction interface is very straightforward. Therefore, the perceived ease of use of CAINES is expected to positively correlate with user satisfaction.

Hypothesis 6 (H6): There is a significant positive correlation between perceived ease of use of CAINES and user satisfaction.

Research Methodology

Semantic Based Information Extraction from EDGAR using CAINES

The first step in developing CAINES for information extraction was to create a corpus from the 10-Q files of the financial statements. The corpus was inserted into a MySQL database using Perl regular expressions. Knowledge engineering techniques were applied to the corpus to determine the format of information and to detect patterns within the text. Various tools and techniques were used to analyze the patterns in the corpus. These include n-gram processing, Key Word In Context Index System (KWIC), Structured Query Language (SQL), WordNet (as a guide to help on finding more semantic relation terms), stemming, and knowledge engineering (for rule based extraction). A Web interface was designed to extract relationships and display the information. The overall structure of CAINES is shown in Figure 2.

Corpus Development

The use of corpora and knowledge engineering techniques to develop a domain-specific knowledge base has become increasingly popular since the 1990s (Chen, 2003). The corpus was first developed by selecting the sample of companies to be used in the study. The second stage prepared the text documents used in the corpus for analysis. To build the corpus the downloaded 10-Qs for the companies were converted to text format. The files containing the *Management's Discussion and Analysis of Financial Condition and Results of Operations* text portion of the 10-Q were used as the basis of the corpus.

Our senior level experts were professionals in the financial industry. They each heavily relied on the financial reports of competitors and other leading financial institutions to analyze the market and set appropriate goals for their organization. Based on the feedback of our experts, we chose to study U.S. companies that were outlined in the SEC's 2008 and 2009 reports or in news reports as affecting the overall US condition in some way. For the *information extraction*, six 10-Q Merrill Lynch reports were used for *training* to learn the patterns and create the rules: first and second quarter of 2007, third quarter of 2008, and first, second, and third quarter of 2009. *Information extraction testing* of CAINES was conducted using the following three files: third quarter of 2007, first and second quarter of 2008. The *training* corpus was comprised of Bank of America reports between 2007 and 2009. *Testing* occurred with Merrill Lynch 2007 to 2009 10-Q reports.

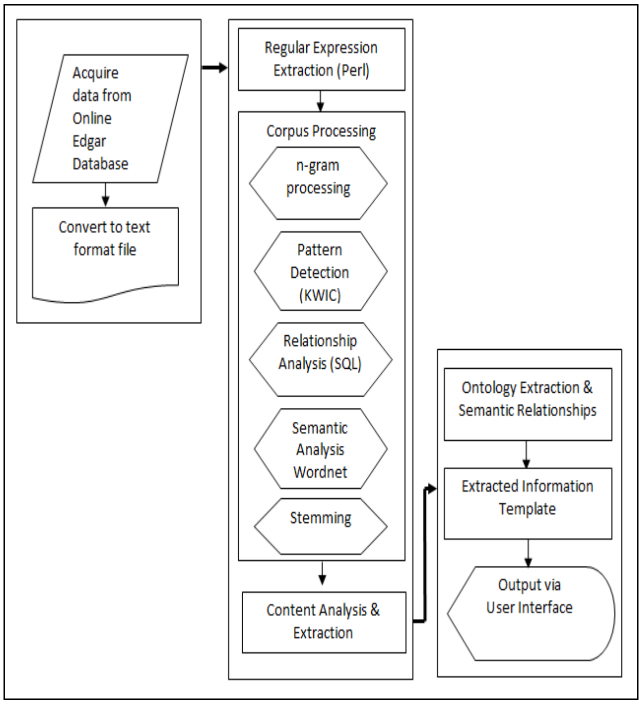

Acquire data from Online Edgar Database
Convert to text format file
Regular Expression Extraction (Perl)
Corpus Processing
n-gram processing
Pattern Detection (KWIC)
Relationship Analysis (SQL)
Semantic Analysis Wordnet
Stemming
Content Analysis & Extraction
Ontology Extraction & Semantic Relationships
Extracted Information Template
Output via User Interface

Figure 2. System Architecture of CAINES

IE Rule Creation

In line with the knowledge engineering approach, four major sets of rules were created to extract information. The rules were first created in a pseudo code fashion based on major categories of information that experts would be interested in. A set of programs were developed to extract information based on the four major rules. In order to ensure CAINES extracts useful information, financial experts who regularly analyze financial reports provided their opinions on what information the rules should extract. The experts consisted of an Area Manager for SunTrust Bank, a Senior Audit Officer of a Tennessee based bank, and a Senior Cost Accountant for PepsiCo Beverages and Foods. The three experts were shown a portion of the *Management's Discussion and Analysis of Financial Condition and Results of Operations* of a Bank of America 10-Q that was filed on 9/30/09. They were then interviewed about what they thought was important and were asked to point out specific examples from the report.

Results of the expert interviews revealed that the experts were interested in three major categories of information (1) understand what current market conditions impacted the growth of certain balance sheets (2) summarize management's discussion of potential risks and uncertainties with moving forward (3) identify significant financial activities including mergers, acquisitions, and new business segments. The three categories made up the four sets of information extraction rules. See Table 1 for the extraction rules.

N-gram Processing

The Lingua::En::Tagger and KWIC files were used to identify frequent n-grams and noun phrases for the extraction rules. CAINES has a part-of-speech tagger and noun phrase extractor that similarly assist with n-gram analysis. These subsystems assist with understanding the semantics of sentences in a corpus. The part-of-speech (POS) tagger assigns each word in a sentence a part of speech tag. The noun phrase extractor produces a list of noun phrases found in the corpus. These systems become useful when we aim to fill our subject-predicate-object format. We use Lingua::En::Tagger, available at http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.16/Tagger.pm to do these two tasks. Lingua::En::Tagger is a probability based, corpus-trained tagger. The tagger is based on a lookup dictionary and preset probability values. The tagger will try to assign a POS tag based on the known POS tags for a given word and the POS tag assigned to its predecessor. In this information extraction study, the terms that appear as noun phrases and verb phrases help us to identify the subjects and the objects of the key verbs that we are trying to find. A short list of relevant n-grams used in this study is found in Table 2.

Sublanguage Lexicon

Sublanguage theory explains that texts common to a specific subject area will share common vocabulary, specials symbols, abbreviations, and even sentence construction (Grishman and Kittredge, 1986; Harris, 1991; Liddy, Symonenko, and Roe, 2006). Sublanguage theory has assisted in analyzing documents in particular areas for years in biomedicine (Liddy et al., 1993; Friedman et al., 2002), weather forecasting (Somers, 2003), and others.

In our research, we developed a finance sublanguage lexicon to assist with analyzing the similar style, construction, and vocabulary of the texts. The items in our lexicon consisted of single terms (e.g., demand, interest, market, purchase, etc.) and phrases (e.g., commodity prices, market condition, and U.S. sub-prime residential mortgage). Abbreviations such as BOA (Bank of America), BAC (Bank of America Corporation), and NYMEX (New York Mercantile Exchange) are also included. The grammar rules in this domain are analyzed and used in our extraction algorithms.

Table 1. Financial Report Extraction Rule Sets

	IE Major Rule Sets
1. A rule to identify market conditions (globally and domestic)	***for*** each row in the portion of the report from which we want to extract information ***if*** the key word is a candidate denoting economic and market conditions (e.g., markets, economic conditions, market conditions, credit environment, indices, credit spreads, oil prices, commodity prices) ***then*** return the keyword noun phrase **and** return verb phrase immediately following the keyword and present them as the State of US or Global market (e.g. verb phrases-continued to, increased, slowed, declined,) **if** keyword is a verb phrase denoting cause (e.g. driven by) **then** return the verb phrase and following noun phrases as reasons ***end if*** ***end for***
2. A rule to identify specific economic and market conditions	**impacting growth and earnings to specific business assets** ***for*** each row in the portion of the report from which we want to extract information ***if*** the key terms are a verb phrase denoting impact (e.g. lower revenues , adversely impacted, resulted in) ***then*** return the noun phrase appearing prior to the key verb phrase as the market condition **and** return the noun phrases after the key verb phrase and present them as the business segment or asset affected ***end if*** ***end for***
3. A rule to identify forward looking statements	***for*** each row in the portion of the report from which we want to extract information ***if*** the key word is a candidate denoting forward looking statements (e.g., outlook, anticipate, demand) ***then*** return the noun phrases immediately following the keyword and present it as the Business segment outlook **and** return the verb phrase following the keyword and present it as the Outlook ***end if*** ***end for***
4. A rule to identify mergers, acquisitions, and new business segments	***for*** each row in the portion of the report from which we want to extract information ***if*** the verb phrase is a candidate denoting mergers, acquisitions, or new business segments ***then*** return the subject noun phrase as 'subject' **and** return verb phrase as 'predicate' (e.g. verb phrases- acquired, ceased, created, became, entered) **then** return the following noun phrase as 'object' ***end if*** ***end for***

Table 2. Relevant n-gram Output

Unigrams	Acquisition
	Demand
	Economic
	Indices
	Market
	Merger
Bi-grams	Business segment
	Business activity
	Commodity prices
	Credit environment
	Market condition
	Merger between
n-grams	Current portion of long-term borrowings
	Driven by cash equities
	Slowdown in U.S. economic growth
	Unprecedented credit market environment
	U.S. housing market downturn
	U.S. sub-prime residential mortgage

The next step is to implement subsystems to detect patterns in the text of the corpus based on the n-gram vocabulary list. For this process, the KWIC system, SQL, and stemming are used.

KWIC and SQL

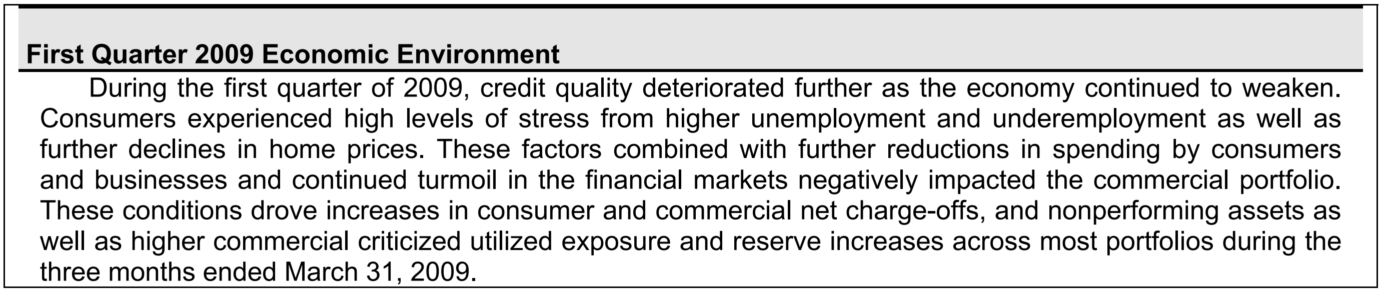
In 1958, Peter Luhn at IBM, developed the Key Word In Context (KWIC) system that uses automatic indexing to recognize word boundaries and frequencies (Luhn, 1960). For CAINES, KWIC is used to analyze word placement in relation to other words in a sentence to determine patterns in the text. Essentially, the KWIC system parses the text by paragraphs, denoted by a period followed by a new line. The results allow all sentences of the corpus to be included in the KWIC process.

The KWIC system then loads each word of the cleaned file into the first column of each row of a MySQL database table.

Table 3. Example Data output from The KWIC Index System

(For this example the database has been reduced to 7 columns rather than the 50 used in the actual database.)

Word1	Word2	Word3	Word4	Word5	Word6	Word7
During	the	first	quarter	of	2009,	Credit
the	first	quarter	of	2009,	credit	quality
first	quarter	of	2009,	credit	quality	deteriorated
quarter	of	2009,	credit	quality	deteriorated	further
of	2009,	credit	quality	deteriorated	further	as



First Quarter 2009 Economic Environment

During the first quarter of 2009, credit quality deteriorated further as the economy continued to weaken. Consumers experienced high levels of stress from higher unemployment and underemployment as well as further declines in home prices. These factors combined with further reductions in spending by consumers and businesses and continued turmoil in the financial markets negatively impacted the commercial portfolio. These conditions drove increases in consumer and commercial net charge-offs, and nonperforming assets as well as higher commercial criticized utilized exposure and reserve increases across most portfolios during the three months ended March 31, 2009.

Figure 3. Example text of Bank of America 10-Q

Each row in the database will consist of additional columns that contain words in the text that follow the word in the first column. What results is a shifting pattern that allows each word of the corpus to appear in each column of the database. Figure 3 shows a sample of text from a Bank of America 10-Q. Table 3 shows the sample text in the KWIC database format.

Semantic and Lexical Analysis

Developed in the early 1990s, WordNet has been used to classify information into hierarchical categories that can be adapted to develop a variety of IE systems (Bagga, J. Chai, A. & Biermann, 1996). The database recognizes and organizes parts of a sentence into machine-readable semantic relations (Miller, 1995). WordNet 3.0 was used in the CAINES corpus as well as other thesauri to determine synonyms for key words determined by the KWIC analysis. CAINES uses it to select synonyms for relationship terms. For example, *decline*, *loss*, *slow*, *weak*, and *lower* can be grouped and coded together in CAINES to extract phrases with predicates denoting some sort of decrease. As another example, *caused*, *driven by*, *resulted in*, *adverse impact*, and *impacted by* were used in rules where CAINES should extract information regarding causation.

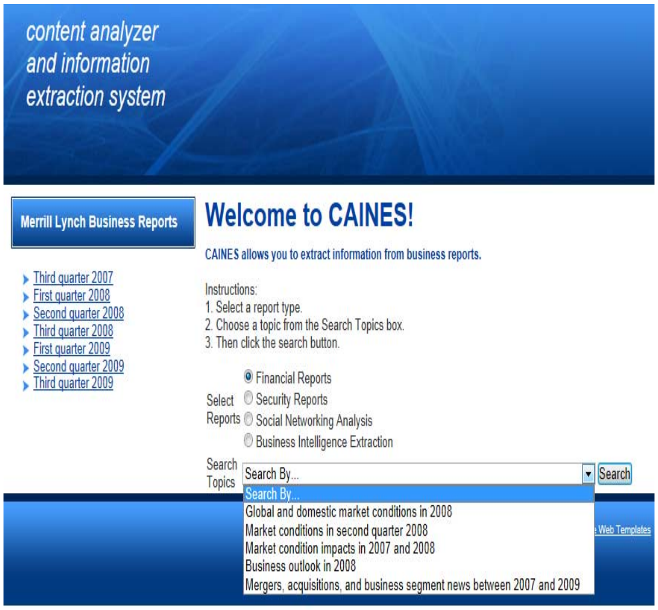
Stemming

Stemming is used in IE system development to improve recall. Stemming is the removal of the inflectional ending, such as –ed, -ing, and -s, from words to reduce word forms to its root. Root words are beneficial to CAINES because they can produce more effective results since words have different meanings in different contexts (Xu & Croft, 1998). CAINES incorporates stemming during the corpus development through SQL. Essentially, the roots of the words are combined with wildcard characters for SQL analysis. For example, in the SQL query for the word bank, the wildcard character "%" will be combined with the root word to form bank%. The query could return results such as– "bank", "banks" and "banking". Using wildcard characters in queries ensures all relevant references in the corpus are included.

CAINES Extraction Interface

Now that the extraction rules and subsystems are implemented within CAINES, users can extract information from long reports through the interface. The Web based interface was developed in HTML with PHP: Hypertext Preprocessor (PHP) scripting language. PHP was designed to allow web developers to create dynamic Web pages quickly (php.net). It was primarily chosen for this project because of its uncomplicated interfacing with

MySQL and overall ease of programming. See Figure 4 for a screen capture of CAINES.



content analyzer and information extraction system

Merrill Lynch Business Reports

- Third quarter 2007
- First quarter 2008
- Second quarter 2008
- Third quarter 2008
- First quarter 2009
- Second quarter 2009
- Third quarter 2009

Welcome to CAINES!

CAINES allows you to extract information from business reports.

Instructions:
1. Select a report type.
2. Choose a topic from the Search Topics box.
3. Then click the search button.

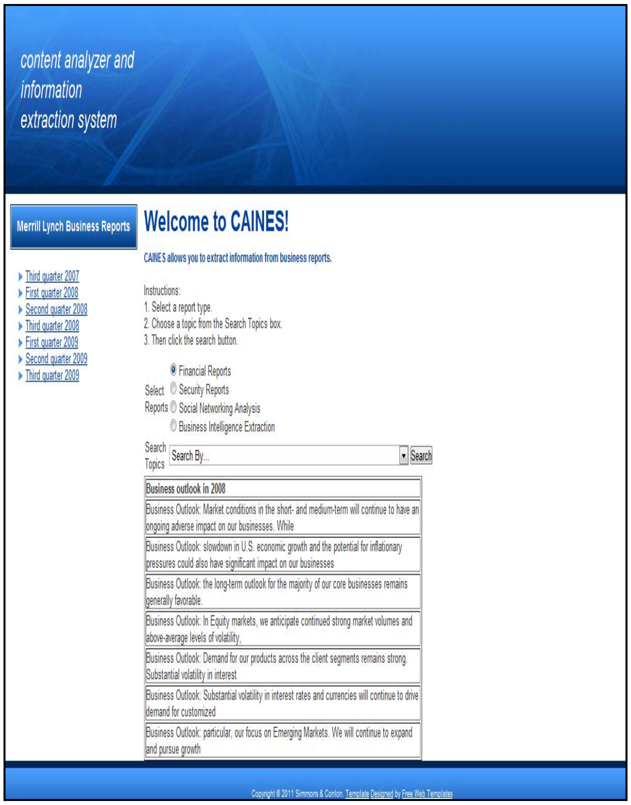
Select Reports
● Financial Reports
○ Security Reports
○ Social Networking Analysis
○ Business Intelligence Extraction

Search Topics: Search By... ▾ Search

Search By...
Global and domestic market conditions in 2008
Market conditions in second quarter 2008
Market condition impacts in 2007 and 2008
Business outlook in 2008
Mergers, acquisitions, and business segment news between 2007 and 2009

Figure 4. CAINES Extraction Interface

Figure 5 shows the business outlook output and Figure 6 shows the output of mergers, acquisitions, and business segment news between 2008 and 2009.



content analyzer and information extraction system

Merrill Lynch Business Reports

- Third quarter 2007
- First quarter 2008
- Second quarter 2008
- Third quarter 2008
- First quarter 2009
- Second quarter 2009
- Third quarter 2009

Welcome to CAINES!

CAINES allows you to extract information from business reports.

Instructions:
1. Select a report type.
2. Choose a topic from the Search Topics box.
3. Then click the search button.

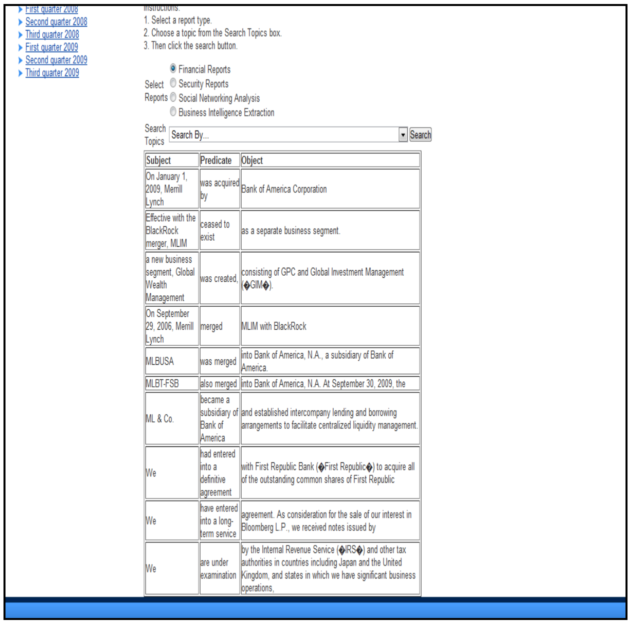
Select Reports
● Financial Reports
○ Security Reports
○ Social Networking Analysis
○ Business Intelligence Extraction

Search Topics: Search By... ▾ Search

Business outlook in 2008
Business Outlook: Market conditions in the short- and medium-term will continue to have an ongoing adverse impact on our businesses. While
Business Outlook: slowdown in U.S. economic growth and the potential for inflationary pressures could also have significant impact on our businesses
Business Outlook: the long-term outlook for the majority of our core businesses remains generally favorable.
Business Outlook: In Equity markets, we anticipate continued strong market volumes and above-average levels of volatility,
Business Outlook: Demand for our products across the client segments remains strong. Substantial volatility in interest
Business Outlook: Substantial volatility in interest rates and currencies will continue to drive demand for customized
Business Outlook: particular, our focus on Emerging Markets. We will continue to expand and pursue growth

Copyright © 2011 Simmons & Conlon. Template Designed by Free Web Templates

Figure 5. CAINES screen output of **"Business Outlook in 2008"**



- First quarter 2008
- Second quarter 2008
- Third quarter 2008
- First quarter 2009
- Second quarter 2009
- Third quarter 2009

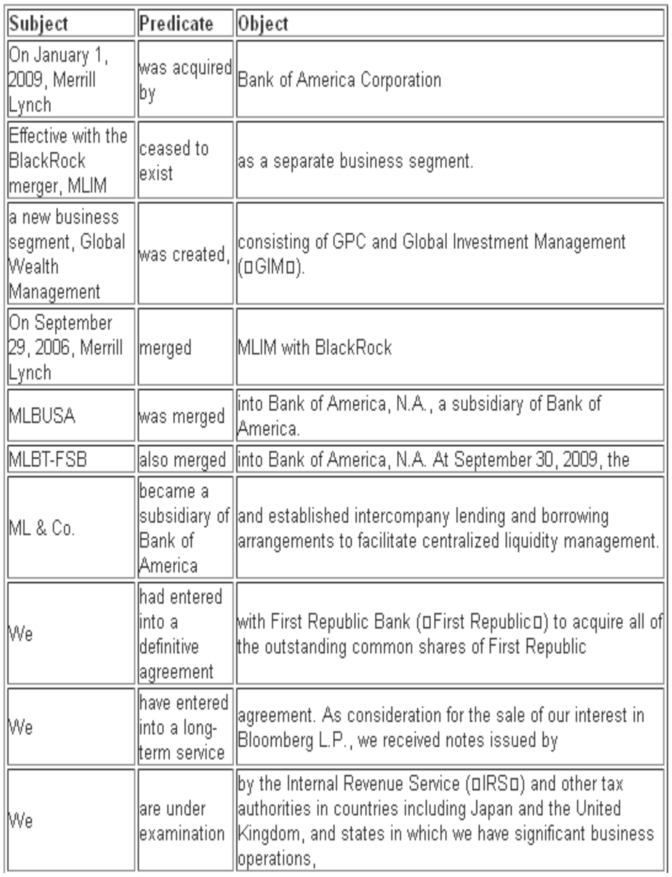
Instructions:
1. Select a report type.
2. Choose a topic from the Search Topics box.
3. Then click the search button.

Select Reports
● Financial Reports
○ Security Reports
○ Social Networking Analysis
○ Business Intelligence Extraction

Search Topics: Search By... ▾ Search

Subject	Predicate	Object
On January 1, 2009, Merrill Lynch	was acquired by	Bank of America Corporation
Effective with the BlackRock merger, MLIM	ceased to exist	as a separate business segment.
a new business segment, Global Wealth Management	was created,	consisting of GPC and Global Investment Management (�GIM�).
On September 29, 2006, Merrill Lynch	merged	MLIM with BlackRock
MLBUSA	was merged	into Bank of America, N.A., a subsidiary of Bank of America.
MLBT-FSB	also merged	into Bank of America, N.A. At September 30, 2009, the
ML & Co.	became a subsidiary of Bank of America	and established intercompany lending and borrowing arrangements to facilitate centralized liquidity management.
We	had entered into a definitive agreement	with First Republic Bank (�First Republic�) to acquire all of the outstanding common shares of First Republic
We	have entered into a long-term service	agreement. As consideration for the sale of our interest in Bloomberg L.P., we received notes issued by
We	are under examination	by the Internal Revenue Service (�IRS�) and other tax authorities in countries including Japan and the United Kingdom, and states in which we have significant business operations,

Figure 6a. CAINES screen output of "mergers, acquisitions, and business segment news between 2008 and 2009"



Subject	Predicate	Object
On January 1, 2009, Merrill Lynch	was acquired by	Bank of America Corporation
Effective with the BlackRock merger, MLIM	ceased to exist	as a separate business segment.
a new business segment, Global Wealth Management	was created,	consisting of GPC and Global Investment Management (□GIM□).
On September 29, 2006, Merrill Lynch	merged	MLIM with BlackRock
MLBUSA	was merged	into Bank of America, N.A., a subsidiary of Bank of America.
MLBT-FSB	also merged	into Bank of America, N.A. At September 30, 2009, the
ML & Co.	became a subsidiary of Bank of America	and established intercompany lending and borrowing arrangements to facilitate centralized liquidity management.
We	had entered into a definitive agreement	with First Republic Bank (□First Republic□) to acquire all of the outstanding common shares of First Republic
We	have entered into a long-term service	agreement. As consideration for the sale of our interest in Bloomberg L.P., we received notes issued by
We	are under examination	by the Internal Revenue Service (□IRS□) and other tax authorities in countries including Japan and the United Kingdom, and states in which we have significant business operations,

Figure 6b. Enlarged screen capture of the data output

Data Analysis

Pilot Testing

As a pilot test, we manually extracted semantic information from Bank of America reports. Recall, precision, and F-measure were the only tests for the pilot study. Twenty-one total relationships and 17 relationships regarding 'economic impact' were found. Then, using CAINES the author found 13 correct 'economic impact' relationships out of the total 17 semantic relationships that were found. Metrics were as follows: recall 76.4%, precision 92.8%, and F-measure 83.8%

In the pilot, only portions of the full system were used. Perl regular expression, KWIC, SQL, knowledge engineering techniques, and stemming procedures were used to receive the preliminary results. To increase accuracy and make the system more scalable, CAINES included n-gram, noun phrase, and semantic analysis in training and testing the full system.

Main Study Participants and Task

Business professionals, undergraduate and graduate business students participated in this study. Their task was to test the difference between manual and automated information extraction.

First, a pretest was conducted with six business professionals for one week. The professionals participated in the study and provided feedback on question wording, timing, and other flow issues. During pretesting, business professionals spent more than 30 minutes manually extracting information from the lengthy reports. Once all feedback was incorporated, the main study was conducted. A total of 54 complete sets of data from the main study were downloaded from Qualtrics survey suite. Four data sets were deleted because the reverse coded item was not answered appropriately. Six data sets were from the business professionals who pretested the study, and were therefore deleted. Forty-four samples were used in the data analysis.

The undergraduate business students were surveyed while attending a core MIS course. Graduate students participated in the study on their own time. The average age of the participants was 24. A little over half of the participants were male. About 35% of the students were majoring in accounting, finance, or economics. These three majors may be the most familiar with analyzing financial reports. However, the goal of the testing phase was to compare manual extraction of semantic information to that of extraction with CAINES. Therefore, a broad range of disciplines was acceptable for the sample.

During the undergraduate classroom administration of the study, the participants received a brief overview of the project. Then they were given access to the experiment which was housed online in Qualtrics. The first page of the Qualtrics survey was a description of the study. The second Web screen was the manual data extraction form noting the information to be extracted manually from the EDGAR database. This manual extraction page begins with instruction to answer five questions by accessing and reviewing specific financial reports on the EDGAR database. Web links to the specific reports were listed with the questions. The third Web screen displayed the questions to be answered using CAINES (see Figure 7).

The following questions ask about mergers, acquisitions, and business segment news between 2008 and 2009.

3. What acquisitions occurred in 2009?

Merrill Lynch was acquired by Bank of America Corporation

4. Were any new business segments created? If so, please state the name of the new business segment(s) below.

Global Wealth Management

5. Is Merrill Lynch ("we") under any examination? Circle one:

Yes – by the Securities and Exchange Commission (SEC)

No – not under any examination

Yes – by the IRS and other tax authorities

Figure 7. Questions and Answers for testing with CAINES

The participants used CAINES to query the database that was preloaded with the reports needed to answer the questions. The participants were given 10 minutes for each extraction method. There was a statement displayed on each screen that the page would advance after 10 minutes. This was to ensure that the study would not exceed a 30 minute time frame. Qualtrics was customized to capture the time spent on the manual (screen two) and the CAINES extraction pages (screen three). This data assists in our speed comparison calculation between manual extraction and CAINES extraction.

Survey Measurement Development

The user survey was introduced on the last screen in Qualtrics. Acceptance and satisfaction of CAINES were evaluated with a Likert scale survey. To analyze the perceived usefulness of CAINES, participants answered five questions. For example, “Using CAINES increased my productivity”. The next five survey items covered user feelings about the perceived ease of use of CAINES: “Learning to operate CAINES was easy for me”. Lastly, the survey asked questions about the user’s satisfaction with CAINES. For example, “I was satisfied with the time it took to search using CAINES” and “Using CAINES is a good way to search long financial reports”. User satisfaction was measured with three questions and the last question was reverse coded.

Main Study Analysis and Results

A total of 21 10-Q reports, averaging about 100 pages long, were used for information extraction, and ontology development. Five companies were used in this study: Countrywide Financial, HSBC, Merrill Lynch, Wachovia, Citigroup, and Bank of America.

For the semantic based information extraction, six 10-Q Merrill Lynch reports were used for *training.* Training is a technique used in CAINES to learn the patterns in the data and create the specific extraction rules. Merrill Lynch’s first and second quarters of 2007, third quarter of 2008, and first, second, and third quarters of 2009 reports were used for training. Information extraction testing of CAINES was conducted using the following three reports: third quarter of 2007, first and second quarters of 2008. From the three major extraction rules, specific rules were created to extract information from Merrill Lynch reports (Table 1). Using CAINES, one can extract information about *global and domestic market conditions in 2008*, *market conditions in second quarter 2008*, *market condition impacts in 2007 and 2008*, and information about the *business outlook in 2008*. Visit http://www.lakishasimmons.com/caines/index.php to extract information from reports.

The training corpus was comprised of Bank of America reports between 2007 and 2009. The testing was performed against Merrill Lynch 2007 to 2009 10-Q reports. CAINES extracted the mergers, acquisitions, and business information from the segment news between 2007 and 2009 using extraction rule 4 (Table 1).

Performance Results

Responses to the extraction questions, speed, and survey data were downloaded from Qualtrics. The speed comparison between using CAINES and manually extracting and processing information was conducted by having students answer questions based on information in the reports. They answered the questions by manually reviewing the reports and then by reviewing the information extracted by CAINES. Recall, precision, and F-measure were calculated based on correct and incorrect answers to the extraction questions.

The speed comparison between CAINES and manual extraction was conducted by having the participants answer questions based on information in the reports. They answered the questions by manually reviewing the reports and then by reviewing the information extracted by CAINES (Table 4). CAINES, the system itself, takes about 3 seconds to extract information from reports. The average user speed in the manual extraction process was a little over 9 minutes, compared to the average user speed with the help of CAINES of about 6 minutes. The 370 seconds consists of the CAINES extraction time *and* the user completing the questions. Potentially, the manual extraction time would have been longer if there was not a 10 minute time limit, which was created due to the feedback of the professionals during the pretesting phase. By limiting the time to 10 minutes, we were confident that we would receive the same end results – that CAINES is much more efficient and useful.

Recall of the participants for the manual extraction process was 10.45% compared to recall with CAINES of 85.91% (Table 4). Precision for the manual extraction was 16.86% compared to 87.16% for CAINES.

Lastly, the F-measure was used to combine the results of precision and recall. The F-measure for the manual process was 12.59% compared to the F-measure for CAINES of 86.46%.
A paired sample t-test was conducted to test the speed, recall, precision, and F-measure hypotheses (Table 5). All four hypotheses were supported as detailed in Table 6. H1 which suggested that CAINES would be faster, is strongly supported, t (43) = 8.861, p = <.01, α = .05.

Table 4. Speed, Recall, Precision, and F-measure results (n=44)

	User Processing Speed (seconds)		**Recall**		**Precision**		**F-measure**	
	Manual	**w/CAINES**	**Manual**	**w/CAINES**	**Manual**	**w/CAINES**	**Manual**	**w/CAINES**
Mean	556.39	370.61	.105	.859	.169	.872	.1259	.865
Min	163	182	.00	.40	.00	.40	.00	.40
Max	600	600	.60	1.0	1.0	1.0	.75	1.0
Std Dev	96.26	110.989	.180	.181	.309	.167	.220	.173

Table 5. Results of Paired T-Test

Paired Samples T-Test					
Paired Comparisons	Mean Difference	Standard Deviation	T-Value	Degrees of Freedom	P-Value
Time w/CAINES versus Manual	185.777	139.072	8.861	43	<.001
Recall w/CAINES versus Manual	-.755	.251	-19.940	43	<.001
Precision w/CAINES versus Manual	-.703	.351	-13.286	43	<.001
F-Measure w/CAINES versus Manual	-.739	.278	-17.645	43	<.001

Table 6. Performance Hypotheses Results

Hypothesis	T-Value	Supported
H1: Average processing time is faster when using CAINES than when manually extracting information from EDGAR 10-Q reports.	8.861*	Yes
H2: Average recall is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.	-19.940*	Yes
H3: Average precision is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.	-13.286*	Yes
H4: Average F-measure is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.	-17.645*	Yes

*Significant at the .05 level

Participants spent statistically significantly more time extracting semantic information manually (*M* = 556.39, *SD* = 96.26) than they did using CAINES (*M* = 370.61, *SD* = 110.989).

In regards to higher recall and precision with CAINES, both were supported, H2, *t* (43) = -19.940, *p* = <.01, α = .05 and H3, *t* (43) = -13.286, *p* = < .01, α = .05. Thus, there was evidence that a difference in recall existed between the manual extraction (*M* = .1045, *SD* = .1804) and CAINES (M = .8591, *SD* = .1809). There was also significant evidence that precision of the manual processing (*M* = .1686, *SD* = .3085) was less than for CAINES (*M* = .8716, *SD* = .1668). Similar results supported the F-measure hypothesis, H4, *t* (43) = -17.645, *p* = <.01, α = .05. The F-measure revealed significant differences between the manual extraction (*M* = .1259, *SD* = .2202) and CAINES extraction (*M* = .8646, *SD* = .1735).

Survey Measurement Analysis

Participants completed the survey after participating in the manual and CAINES extraction processes. Survey responses were downloaded from Qualtrics and analyzed in SPSS 17. Internal consistency was assessed by calculating Cronbach's alpha. The values were 0.82, 0.95, and 0.98 for satisfaction, perceived ease of use, and perceived usefulness respectively (see Appendix). The scales are deemed reliable since they were all greater than the accepted threshold of 0.70 (Nunnally, 1967).

Participants believed CAINES was useful for fast and effective searching of long reports and that it improved performance and productivity. Participants agreed that it was easy to learn to use and maneuver CAINES. They agreed that CAINES was clear and understandable and that they could become skillful with CAINES. They somewhat agreed that CAINES was flexible to interact with. Further, they were satisfied with using CAINES for long reports and the CAINES search time. They agreed that manually extracting was not the better search method (wording reverse coded).

To test the relationships between PU and satisfaction and PEOU and satisfaction, we analyzed the correlations between the constructs. PU and PEOU of CAINES were both positively correlated with satisfaction (Table 7). The relationship between PU and satisfaction was significantly correlated, r = .795, p < .01. PEOU and satisfaction were also significantly correlated, r = .831, p < .01. Thus, H5 and H6 are supported (Table 8).

Table 7. Correlations between Constructs

	PU	PEOU	SAT
PU	1	.795**	.776**
PEOU		1	.831**
SAT			1

**Correlation is significant at the 0.01 level (2-tailed).

Table 8. Satisfaction Hypotheses Results

Hypothesis	Path	Correlation (r-value)	Supported
H5: PU will positively influence satisfaction (US).	PU → US	.776**	Yes
H6: PEOU will positively influence satisfaction (US).	PEOU → US	.831**	Yes

**Correlation is significant at the 0.01 level (2-tailed).

Discussion

Overall, greater levels of recall, precision, and F-measure, were achieved with CAINES than with manual extraction of information from 10-Q reports. CAINES was also advantageous in terms of speed. Extraction with CAINES was faster than the manual extraction process. The time difference between CAINES and the manual was not very large probably because users were limited to 10 minutes for each extraction method. Thirty two out of the forty four participants spent the entire 10 minute period on the manual extraction questions. Only five participants required the entire 10 minute period using CAINES. In addition, one participant commented that they used the search feature in their Internet browser (Control + F) to find the answers to the manual questions, which was still less accurate. This could be another explanation for the small gap in processing time between the two methods. Lastly, the sample consisted of business students and therefore they may have answered some of the questions based on their common business knowledge. Thus, they may have answered some of the questions based on financial knowledge and not looked at the reports in detail.

Users extracted more relevant information with CAINES in its 6 minutes than they did during the manual process in 9 minutes. This time savings seemed to result in a larger amount of relevant information being processed, thus recall was 85.91% with CAINES versus 10.45% with the manual extraction process.

Regarding precision or accuracy of the manual process, three participants were able to achieve a 100% level. One participant answered three questions and those three were correct; two participants answered two questions and both were correct. The minimum precision for CAINES was 40%. Twenty-four participants were able to obtain 100% recall and precision levels using CAINES. CAINES clearly outperformed manual processing and users were satisfied with their experience with CAINES.

Thirty-five out of the forty-four participants left comments about their experience. Many participants alluded that the manual extraction process was "*arduous*", "*tedious*", "*cumbersome*", and "*frustrating*" and that CAINES "*saves a lot of time*", "*more efficient and a productive use of time*", and that it made the search "*more condensed*" .

Even with a few critiques, users were very satisfied with their experience with CAINES. CAINES brings a new level of performance and satisfaction because it is more advanced than similar systems. Like FIRST, CAINES extracts information from online business reports. FIRST extracted short articles that were about half a page long that consisted of structured financial data and converted it into XML for use in business applications. CAINES is more advanced than FIRST in that it can extract unstructured information from discussion text in lengthy reports. With high accuracy, CAINES can return relevant semantic phrases via a Web based user interface.

Conclusion

Extracting specific information that fulfills a nontrivial business need from online business reports is a challenging, but an important application for IE. We believe this implementation method of involving business experts and developing systems that solve specific business problems advances our thinking about IS research. Thus, we believe that this work, by following the design science approach, contributes to the growing body of literature on how to build an information system that helps business professionals find and manage information more effectively and efficiently. It will only become more important as the size of Web documents continue to grow with overlapping and conceptually equivalent facts. This work demonstrates that semantic based extraction can be conducted by knowledge engineered systems such as CAINES to enable business efficiency. CAINES can be useful to business managers, analysts, lenders, shareholders, and potential investors who want to quickly process online financial data for their specific needs.

In response to the research questions, CAINES was able to accurately extract information from online financial reports and allowed users to perform better than people extracting information manually. In addition, users overwhelmingly agreed that CAINES was useful and easy to use, and they were satisfied with the system.

It is important to note that it is appropriate to compare the performance of CAINES to the method that our experts typically used to extract information to meet their three specific business objectives, which happened to be a manual search*.*

Limitations and Future Work

This study found that CAINES was more accurate in retrieving relevant information and in less time than manually searching and extracting or keyword searching. The target population of CAINES in the business world is those who need to access financial reports. Although statistical power tests showed that the sample size was adequate, larger samples with a population of financial professionals could be used to replicate the results. However, this study shows that CAINES is great even for novice users. One could expect greater results with financial professionals who are well versed in the financial domain.

CAINES could be useful in extraction of other important information in domains such as health care. CAINES can assist the health care community with deep content analysis of treatment databases or extraction of specific health information. In sum, CAINES is a system that can be customized and scaled to accommodate extraction and analysis of many sources of online text.

References

Andersen, P., Hayes, P., Huettner, A., Schmandt, L., Nirenburg, I., & WeinStein, S. (1992). Automatic extraction of facts from press releases to generate news stories. *Processing of the 3rd Conference on Applied Natural Language Processing*, 170-177.

Bagga, A., Chai, J. & Biermann, A. (1996). The role of WordNet in the creation of a trainable message understanding system. *Proceedings of the 13th National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, 941-948.

Bardi, A., Calogero, R., & Mullen, B. (2008). A new archival approach to the study of values and value--behavior relations: Validation of the value lexicon. *Journal of Applied Psychology*, *93*(3), 483-497.

Chen, H. (2003). Web retrieval and mining, *Decision Support Systems, 35*(1), 1-5.

Conlon, S., Hale, J., Lukose, S., & Strong, J. (2008). Information extraction agents for service-oriented architecture using web service systems: A framework. *Journal of Computer Information Systems*, *48*(3), 74-83.

Cowie, J. & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80-91.

Cunningham, H. (1999). *Information extraction - A user guide* (second edition). Retrieved from: http://www.dcs.shef.ac.uk/~hamish/IE/userguide/main.html.

Davis, F. D. (1986). A technology acceptance model for empirically testing new end-user information systems: Theory and results. (*Doctoral dissertation, Sloan School of Management, Massachusetts Institute of Technology*).

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly*, 13(3), 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003.

Etzioni, O., Banko, M., Soderland, S., & Weld, D. (2008). Open Information Extraction from the Web. *Communications of the ACM, 51*(12), 68-74.

Etzioni, O., Cafarella, M., Downey, D., Popescue, A.M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2004). Methods for domain-independent information extraction from the web: An experimental comparison, *Aaai Conference On Artificial Intelligence*, 391-398.

Fishbein, M. & Ajzen, I. (1975). Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley Publishing Company, Reading, MA.

Friedman, C., Kraa, P., & Rzhetskya, A. (2002). Two Biomedical Sublanguages: a Description Based on the Theories of Zellig Harris. Journal of Biomedical Informatics, 35(4): 222-235.

Goodwin, N.C. (1987). Functionality and usability.

Communications of the ACM, 30, 229-233.

Grishman, R. & Kittredge, R. I. (Eds.). (1986). Analyzing Language in Restricted Domains: Sublanguage Description & Processing: Lawrence Erlbaum Assoc.

Harris, Z. (1991). *A theory of language and information: a mathematical approach*. Oxford: Clarendon Press.

Hendler, J. & Berners-Lee, T. (2010). From the semantic web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence*, *174*(2), 156-161.

Lee, Younghwa, Kozar, K. A. & Larsen, K. R.T. (2003). The Technology Acceptance Model: Past, Present, and Future, *Communication of the AIS,* (12), 725 -780

Liddy, E. D., Jorgensen, C. L., Sibert, E. E. & Yu, E. S. (1993). A Sublanguage Approach to Natural Language Processing for an Expert System. *Information Processing & Management*, 29(5): 633-645.

Liddy, E. D., Symonenko, S., Rowe, S. (2006). Sublanguage analysis applied to trouble tickets. *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference,* 752-757.

Luhn, H. (1960). Keyword-in-context index for technical literature (KWIC Index), *American Documentation, 11,* 228-295.

Miller, G. (1995). Wordnet: A lexical database for English. *Communication of the ACM*, *38* (11) 39-41.

Sager, N., Friedman C., & Lyman, M. (1987). *Medical Language Processing: Computer Management of Limited Data*. Addison Wisely, Reading, MA.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases*, *1*(3).

Simmons, L. L., Conlon, S. J. Mukhopadhyay, S. & Yang, J. (2011). A computer aided content analysis of online reviews. *Journal of Computer Information Systems* 52, 1, 43-55.

Somers, H. (2003). Sublanguage. In H. Somers (Ed.), *Computers and Translation: A translator's guide.*

U.S. Securities and Exchange Commission. (2008). *2008 Performance and accountability report*. Retrieved from http://www.sec.gov/about/secpar/secpar2008.pdf#sec1

Venkatesh, V. & Davis, F.D. (1996). A model of the antecedents of perceived ease of use: Development and test, *Decision Sciences, 27*(3), 1996, 451-481.

Wilks, Y. (1997). *Information Extraction as a Core Language Technology*. M-T. Pazienza (ed.): Springer, Berlin.

Xu, J. & Croft, W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems, 16*(1) 61-81.

About the Authors

Lakisha L. Simmons is an Assistant Professor of Information Systems Management at Belmont University. She holds a Ph.D. in Management Information Systems from the University of Mississippi. Her research interests include sentiment and semantic extraction, virtual and e-commerce issues, and cross-cultural issues in technology and education. Her work has appeared in *MISQ Executive*, *Journal of Computer Information Systems,* and *the Decisions Sciences Journal of Innovative Education,* among others. Prior to pursuing a Ph.D., Lakisha held positions as a 6 Sigma Black Belt and Business Analyst for Caterpillar Financial.

Sumali J. Conlon is an Associate Professor of Management Information Systems at the University of Mississippi. She received her B.A. in Statistics with Economics minor from Thammasat University, Thailand and Ph.D. from the Illinois Institute of Technology. Her teaching and research interests include Sentiment Analysis, Semantic Web, Web Services, Web Mining, Natural Language Processing, Information Retrieval, Knowledge Management, and Database Systems. Her work has appeared in Decision Support Systems, Omega, Information Processing & Management, Journal of the American Society for Information Science, and Journal of Computer Information Systems, among others.

Appendix

Survey Response Summary (strongly disagree (1) to strongly agree (7))

Construct	Items	Mean	StdDev
PU (.98)	PU1- faster search	6.43	1.065
	PU2-improved search performance	6.39	1.017
	PU3-increased productivity	6.36	1.080
	PU4-enhanced search effectiveness	6.41	1.085
	PU5- would be useful again	6.45	1.066
	Grand Mean	6.41	
PEOU (.95)	PEOU1-learning to use was easy	6.14	1.112
	PEOU2-easy to maneuver	5.98	1.191
	PEOU3-clear and understandable	6.07	1.129
	PEOU4-flexible to interact with	5.71	1.374
	PEOU5-can be become skillful with	6.16	1.219
	Grand Mean	6.01	
SAT (.82)	SAT1-good for long reports	6.23	1.118
	SAT2-satisfied with search time	6.36	1.036
	SAT3-manual search was not better	6.18	1.040
	Grand Mean	6.30	