The Role of Text Analytics and Information Retrieval in the Accounting Domain

Ingrid E. Fisher
University at Albany, SUNY

Margaret R. Garnsey
Sunita Goel
Siena College

Kinsun Tam
University at Albany, SUNY

ABSTRACT: Textual documents proliferate throughout accounting practice, and a wide variety of groups make financial decisions based on written guidance. The Generally Accepted Accounting Principles (GAAP), along with annual corporate financial statements and other reporting narratives, are important sources of such guidance and information. This paper examines the literature in two major areas relevant to text analytics and information retrieval in the accounting domain: (1) the manual and computational content analysis of accounting narratives, accounting readability studies, and related text-mining work, and (2) the information retrieval literature stream that addresses the extraction of both text elements as well as quantities imbedded in text from accounting documents, and includes the impact of understanding the accounting lexicon upon retrieval from digital accounting documents. We use the goals in developing the GAAP Codification, as expressed by the Financial Accounting Standards Board (FASB) in their Notice to Constituents (FASB 2009), as a starting point for reviewing the literature. The paper concludes with a map for suggested future research in accounting text analytics and information retrieval.

INTRODUCTION

Background

As early as 1965, Goldberg (1965) claimed “it is scarcely an exaggeration to say that the problem of communication is the axial problem in accounting.” Textual documents including the Generally Accepted Accounting Principles (GAAP), annual corporate financial statements, along with other accounting narratives, are primary vehicles for communicating accounting information. These documents proliferate throughout accounting practice, and a wide variety of groups make financial decisions based on written guidance. However, the accounting language articulated in these sources is ever-changing, making it difficult for nonexperts to stay abreast of recent modifications in the use of accounting terminology. The chairperson of the
The Special Task Force on Standards Overload has stated, “Standard setters should stop inventing new terms every time a new standard is set” (Burke 1997).

The Financial Accounting Standards Board (FASB 2005) recognized the problem of understandability along with the problem of inefficient, ineffective information retrieval with regard to the financial accounting standards (FASs). In response, the FASB initiated “The FASB Codification and Retrieval Project” to codify the authoritative accounting literature (AAL). The FASB Accounting Standards Codification™ was launched on July 1, 2009, as the single source of authoritative nongovernmental U.S. generally accepted accounting principles (GAAP).

Research into the effective articulation of formal accounting communications, such as the financial accounting standards (FASs), is sparse, as is research that integrates the domains of linguistics and accounting. Haried (1972) observes, there has been a “conspicuous lack of empirical support” for the semantic problems in external accounting communications. He further identifies the designation of technical meanings to terms that express different meanings in common use or in other disciplines, and the lack of standardization of terms used in financial reports as particularly problematic. Despite these observations, there has been little subsequent research.

Objectives

This paper examines the literature in text analytics and information retrieval as relevant to the accounting domain. The literature in text analytics and information retrieval is subdivided into two streams. The first stream includes the manual and computational content analysis of accounting narratives, accounting readability studies, and related text-mining work. This is focused primarily on using the text elements in order to draw inferences (e.g., regarding future performance, stock price). The second stream comprises the information retrieval literature, which consists of the extraction of both text elements as well as quantities imbedded in text from accounting documents. This stream includes the significance of understanding the accounting lexicon and the role of an accounting thesaurus to retrieval from digital accounting documents. We exclude the literature that examines XBRL and focus on English language text. Figure 1 depicts the research streams reviewed.

In recognition of the significance of the authoritative accounting literature (AAL) as a text document, we use the goals in developing the GAAP Codification, as expressed by the FASB in their Notice to Constituents (FASB 2009), as one lens and a starting point for reviewing the literature. The FASB’s primary goal is to “simplify user access to all authoritative U.S. GAAP” (FASB 2009). The Codification was also designed to address stakeholder concerns that include: understandability, inefficient and ineffective information retrieval, excessive research effort, and inconsistent terminology. Inefficient and ineffective information retrieval, as well as excessive research effort, are specific aspects of accessibility, while the problem of inconsistent terminology impacts both accessibility and understandability. Therefore, accessibility and understandability are defined here as primary goals. In reviewing the literature we evaluate the degree to which these goals were met. This appears to be a logical starting point. The literature review is then extended more broadly to other accounting documents and across the research streams. The paper concludes with a map for suggested future research in accounting text analytics and information retrieval.

METHODOLOGY

This paper has reviewed 163 references, of which 127 are scholarly manuscripts in the accounting and business domains. We began our review by consulting the references of three recent dissertations in information science that employed computational linguistics in the accounting domain. We then expanded our search to the following databases: JSTOR, Science Direct, Emerald, the Digital Library of the American Accounting Association, and the Accounting Re-
search Network. Some of the frequently employed search terms include: content analysis, computerized content analysis, readability, information retrieval, text retrieval, accounting language, accounting documents. These terms were augmented by the term accounting to create domain-specific compound search terms. The reference lists of articles included were also examined for relevant references. We did not attempt to conduct an exhaustive review of content analysis research in the business domain. Our focus was primarily on accounting; however, the boundary between accounting and finance studies is not always clear-cut and, thus, we included relevant finance studies.

The chronological distribution of articles referenced in this paper reveals exponential growth of text analytics and information retrieval in business research in the past decades (see Figure 2). From only one paper in the 50s and two in the 60s, this number has grown to 32 in the 90s and 68 in the 00s. Advances in computer technologies (both hardware and software) and the increasing prominence of accounting narratives have fueled this explosion (Jones and Shoemaker 1994), and are expected to further encourage application of computational linguistic technologies in future text-based accounting and business research.

THE GAAP CODIFICATION GOALS: LESSONS FROM THE LITERATURE

Simplified User Access: Primary Goal Number One

Inefficient Information Retrieval

The inability to locate information quickly in FASs, as well as other textual accounting documents, translates directly into a more difficult and more time-consuming research effort. Fisher (2004) argues for consistency in the structure of FASs, including the use of section headings, sequencing and numbering of sections, along with a hierarchal structure as means to better enable both humans and machines to locate specific portions of text. Within the accounting
domain, Gangolly (1995) has also argued for the adoption of such hierarchical formalizations (that contain representations of meta-level information) in FASs. Gangolly (1995) maintains the separation of other meta-data contained in the standards (e.g., scope references, rule exceptions, justifications) can aid in the development of knowledge-based systems that employ this information to form and advance arguments.

The Codification project is the first undertaking by the FASB that explicitly addresses the importance of a uniform and consistent document structure for the AAL. Optimistically, the development and incorporation of a topical arrangement with corresponding subtopics provides the beginnings of a hierarchical structure (Fisher and McEwan 2009). The formal specification of prescribed ordered sequences of topics, subtopics, and sections provides the consistency in structure that can facilitate the development and implementation of indexing and searching procedures (Fisher 2004).

Fisher (2004) notes that FASs as historically drafted contain much redundant information. Such redundancy will negatively impact the precision of digital searches and creates particular difficulty when the documents are frequently modified (i.e., if information has been repeated across multiple locations, then those same locations, in turn, must all be modified). The Codification project focused, to the extent possible, on the elimination of redundant information and attempted to ensure that content is only presented at one location in the AAL. The elimination of redundant content may increase both the readability of the AAL and the precision of searches performed on the AAL (Fisher 2009). In addition, presenting information only once should facilitate maintenance of the knowledge base and reduce the likelihood of errors and inconsistencies in the update process (Fisher 2004).

Inconsistent Terminology

Thesauri aim to overcome the problem of language variability by categorizing terminology based upon meaning and use as well as by linking terms to terms with both broader and narrower meanings. Such classifications and linkages offer users multiple possible search terms. In 1974 Price Waterhouse & Co. published a print version of an accounting thesaurus for internal use (Price Waterhouse & Co. 1974). However, many terms in use today did not exist at the time it was compiled (e.g., securitization, comprehensive income). Thus, the thesaurus is woefully out of date and is also not available in electronic form.

The Codification search feature currently augments users’ search results with the option of narrowing their search by a “related” term or industry. However, the use of “related” term there is
not the same as in thesauri construction where a “related” term is one that has an associative relationship with the primary term. An associative relationship is defined by the International Organization for Standardization (1986) as one where a pair of terms is not equivalent, nor are they hierarchically related, yet there is a mental association strong enough to make the link explicit in the thesaurus. The related terms offered by the Codification are quite often either broader/narrower terms or other topics altogether. A better understanding of the accounting lexicon and an updated accounting thesaurus should enhance access to the AAL by facilitating a more effective selection of search terminology, which can serve as logical (keyword) entry points to the accounting standards literature.

**Understandability: Primary Goal Number Two**

Belkaoui (1980) determined that professional producers and users of accounting information differ in their understanding of selected accounting concepts, while Tanaka (1982) analyzed the structure of the accounting language as a means toward clarifying the basic accounting concepts, providing early signals of the potential problem of crafting an understandable AAL. The AICPA cited the “rapid change in and expansion of performance and reporting standards” in accounting as one factor contributing to their creation of a Special Task Force on Standards Overload (Burke 1997). The Special Task Force recommended the development of systems to “facilitate access to the professional literature and improve the understandability of that literature” (Burke 1997; emphasis added). Despite such urging, there has been an absence of academic studies that apply digital methods and technologies to facilitate the understandability of FASs.

Fisher (2004) asserts that the importance of carefully crafting official documents that prescribe professional practice(s) cannot be overstated. Careful construction of standards in accordance with appropriate criteria can produce a standard that is more understandable. The consistent use of a common hierarchical structure for the AAL, as described in the previous section, can facilitate users’ understanding of the AAL over time.

**EXTENSION OF THE REVIEW ACROSS THE RESEARCH STREAMS**

Accessibility and understandability continue to be major focuses of text analytics and information retrieval studies beyond the topic of the Codification of the AAL. Research on the retrieval of accounting text and accounting quantities from text-form financial reports assesses the accessibility issue from a programming perspective. Understandability of various accounting narratives such as annual reports, press releases, business news articles, and prospectus are examined in readability research. It is important to note that regulatory narratives (such as in the AAL) have different intended goals from reporting narratives (such as in the annual report); where regulatory narratives seek to only provide information, a goal of reporting narratives is to influence the reader through the information content (Henry 2008). This additional goal relates to our review of content analysis studies on accounting and business text documents.

**Content and Readability Analysis of Business Documents**

**Content and Readability Analysis Defined**

Content analysis and readability analysis are two related, but at times not clearly defined, areas of text-based study. Jones and Shoemaker (1994) classify empirical thematic analyses (studying specific trends, attitudes, or content categories from the text and then drawing inferences from them) and syntactic studies (study of text readability) as two separate streams of content analysis. Clatworthy and Jones (2001) divide academic research on accounting narratives into (1) content analysis studies, and (2) readability research. Clatworthy and Jones (2006) also distinguish between content analysis that is (1) meaning-oriented and focused on underlying themes from that
which is (2) form-oriented and focused on textual characteristics. This section separately discusses content analysis and readability analysis while acknowledging their common root.

**Content Analysis Studies in Accounting**

**Manual content analysis.** Prior text-based business research relies considerably on manual content analysis, where researchers read through text documents to draw inferences. In prior administrative science studies, researchers manually content-analyze annual reports to examine corporate strategies (Bowman 1984; Jarvenpaa and Ives 1990) and causal reasoning (Bettman and Weitz 1983; Staw et al. 1983; Salancik and Meindl 1984). Similarly, prior accounting and finance studies apply manual content analysis to draw inferences from public disclosures and press releases (Miller 2002; Razaei et al. 2003; Shon 2003), from management’s earnings forecasts and comments (Hoskin et al. 1986; Francis et al. 2002; Baginski et al. 2004), from journal articles (Previts and Brown 1993), from analyst reports (Govindarajan 1980; Previts et al. 1994; Bricker et al. 1995), from FASB comment letters (Yen et al. 2007), from professional accounting codes of ethics (Farrell and Cobbin 2000), from financial statement footnotes (Kelly-Newton 1980), from prospectus (Deumes 2008), and from shareholders’ letter to presidents (McConnell et al. 1986; Swales 1988; D’Aveni and MacMillan 1990; Clapham and Schwenk 1991), from management’s discussion and analysis (Cole 1990; Pava and Epstein 1993; Bryan 1997), and from other sections of the annual report (Singhvi 1968; Ingram and Frazier 1980; Steele 1982; Hooks and Moon 1993; Aerts 1994; Clarkson et al. 1994, 1999; Boo and Simnett 2002; Bhojraj et al. 2004). The conclusions that emerge from this work are that smaller and less profitable firms disclose less information; firms are more likely to disclose positive rather than negative information; management tends to attribute negative outcomes to uncontrollable factors while taking credit for positive outcomes; and firms disclose more information during periods of increased earnings.

Merkli-Davies and Brennan (2007) review and synthesize literature on discretionary disclosures in corporate documents and develop three taxonomies for the classification of the literature: the motivation (disclosures that constitute impression management and those that add incremental information), the research perspective (preparer versus user), and a group of seven disclosure strategies. Impression management is the selective communication of information so as to distort the reader’s understanding of corporate accomplishments. They conclude that it is difficult to arrive at compelling conclusions attributable to the fact that most studies investigate either impression management or incremental information, with only eight addressing both. Additionally, the overwhelming majority of studies adopt a preparer’s perspective. They suggest examining narratives from more than one perspective.

Brennan et al. (2009) review the literature in impression management as a foundation for the manual content analysis of 21 annual press release announcements in the U.K. They argue that impression management techniques are inherently subtle and sophisticated and, thus, require the detailed human attention afforded by manual content analysis. They also argue for a more holistic analysis of impression management through the use of multiple methods and factors.

**Partially computerized content analysis.** Manual content analysis could introduce researcher subjectivity (Rogers and Fogarty 1997; Krippendorff 2004). Moreover, high costs of manual processing limit sample size, and hence the power of tests and the generalizability of results. For instance, the sample size of accounting and finance studies based on manual content analysis summarized in the first paragraph of the preceding subsection ranges from 25 (Pava and Epstein 1993) to 278 documents (Yen et al. 2007). Miller (2002) admits that a larger sample can help strengthen the power of statistical tests and reduce the impact of noisy data, but hand-collecting disclosure data from text documents imposes high costs on large samples.

Using software to process electronic textual documents and count word frequencies, computerized content analysis overcomes the limitations introduced by manual content analysis. Drawing
inferences systematically from word frequencies is supported by the computational linguistics literature (Zipf 1929, 1949; Luhn 1957). Statistical analysis of word frequencies has been applied to resolve authorship disputes over historically significant writings. For instance, evidence from word frequencies ascribes to James Madison 12 essays written in 1788 to persuade New York State citizens to ratify the U.S. Constitution (Mosteller and Wallace 1964).

A few text-based accounting studies employ partially computerized content analysis. The studies conducted by Ingram and Frazier (1983), Frazier et al. (1984), and Tennyson et al. (1990) are based on small samples from an early period. They electronically analyze annual reports, but manually and subjectively interpret themes identified from factor analysis. Subsequently, Abrahamson and Park (1994) explore how corporate officers conceal negative outcomes from shareholders by conducting computer-assisted content analysis of over 1,000 presidents’ letters. Their computer program counts word frequencies, while human coders identify words connoting negative organizational outcome. Abrahamson and Amir (1996) use computer programs to identify high-frequency words with a negative connotation in 1,355 letters to shareholders from 1988, but again manually determine the context of such words.

Similarly, Clatworthy and Jones (2003) examine the narrative reporting practices (in the chairman’s statement, which is the U.K. equivalent of the president’s letter) of 50 very profitable and 50 very unprofitable companies in the U.K. using computerized methods to extract their data, but then using manual methods to interpret the meaning. They conclude that improving performers concentrate on good news. They also find that declining performers do not discuss the underlying causes of their poor performance. Consistent with other studies and with attribution theory, they find that both improving performers and declining performers take the credit for good news, but blame bad news on environmental factors. Using the same sample of companies and a similar methodology, Clatworthy and Jones (2006) again analyze the chairman’s statement and find that unprofitable, as compared to profitable, companies focus less on key financial indicators, use fewer quantitative results, use fewer personal references, use more passive sentences, and focus more on the future, thus affirming prior work in impression management.

Computerized content analysis. Accounting researchers have applied computerized content analysis since 2000 to large bodies of text, including various sections of the annual report, earnings press releases, and news articles. Smith and Taffler (2000) show that key words and themes (such as profit, dividends, turnover, and productivity) in discretionary narrative disclosures of the chairman’s statement contain important information and are closely associated with financial performance. Using concordance (OCP) and statistical software (SPSS-X), they compute keyword variables and theme variables based on frequency of occurrence, and find these variables to be systematically associated with subsequent firm failure or success.

Both Back et al. (2001) and Kloptchenko et al. (2004) convert words into numerical quantities before further processing. Back et al. (2001) introduce smart encoding to convert words into numerical values, and apply self-organizing maps and document histograms to analyze both quantitative financial and operations data and qualitative CEO report data. Their approach is computer-based and could be used to compare text documents and augment the text retrieval task of search engines. Kloptchenko et al. (2004) use data-mining and text-mining methods to retrieve indications of future financial performance from quantitative and qualitative parts of financial reports. They adopt a computerized approach that converts every word into a number, taking into account word length and the ASCII value of every character in a word. Histograms are accumulated for words and sentences. Euclidian distances between word histograms and sentence histograms are taken to indicate similarity in vocabulary and writing style.

Gerdes (2003) develops an EDGAR-analyzer based on keyword search and finds it to be an effective tool for retrieving text-blocks relevant to Y2K disclosures from 18,595 10-K filings between 1997 and 1999. Applying a standard text analysis software package to search the text of
electronic annual reports, Hussainey et al. (2003) devise a computerized methodology to automate computation of disclosure scores for large samples. Their disclosure scores based on word search are found to be comparable to those developed by the Association for Management Research-Financial Analysts Federation (AIMR-FAF). Beynon et al. (2005) use a data-mining technique known as rough set theory (specifically the variable-precision rough sets model) to successfully predict firm profitability based upon the textual characteristics (e.g., total words in the statement, percentage of good news words, percentage of bad news words) of the chairman’s statement.

Magnusson et al. (2005) combine visualization of quantitative data through self-organizing maps with the visualization of qualitative data using collocational networks to examine quarterly reports in the telecommunication industry. They conclude that changes in the textual data usually precede changes in the financial performance by approximately one fiscal quarter. Li (2006) electronically counts the frequency of words related to risk or uncertainty in 34,180 10-K filings between 1994 and 2005, and uses it as a risk sentiment proxy. He finds a statistical association between increase in risk sentiment and lower future earnings.

Li (2008) manually categorizes 30,000 sentences from the forward-looking statement section of MD&As along two dimensions: (1) positive versus negative tone, and (2) profitability, operations, and liquidity, etc. These manually coded sentences are then used as training data in a Naïve Bayesian machine-learning algorithm to classify 13 million forward-looking statements from more than 140,000 corporate 10-K and 10-Q MD&As between 1994 and 2007. The average tone of forward-looking statements is found to be positively associated with earnings and liquidity (Li 2010).

Using computer-based content analysis and a data-mining algorithm on earnings press releases, Henry (2006) develops verbal predictor variables including keyword (e.g., dividends, sales) counts and other measures intended to capture both content and style of firms’ performance commentaries. These variables and measures are found to enhance accuracy in prediction of market response. Tetlock et al. (2008) quantified the language used in financial news stories. Their main finding was that negative words in the financial press forecast low firm earnings. Henry (2008) conducts computer-based content analysis of 1,366 earnings press releases between 1998 and 2002 and finds that tone influences investors’ reactions. Demers and Vega (2008) use a textual-analysis program to measure managerial optimism expressed in the text of quarterly earnings press releases. Based on 20,000 earnings announcements, they find evidence that relates such optimism to stock returns. Engelberg (2008) employs the “typed dependency parsing” tool from natural language processing to count negative words (as defined by the Harvard Psychological Dictionary) in earnings announcements. He finds qualitative earnings information to have additional predictability for asset prices beyond that of quantitative information.

Davis et al. (2008) use textual-analysis software and word counts to measure levels of optimistic and pessimistic language in earnings press releases. Such language usage is found to be related to future firm performance. Sadique et al. (2008) use a computer content analyzer to analyze the tone of media news articles and earnings press releases. Their results suggest that positive tone is related to increases in a firm’s stock returns and to decreases in volatility, while negative tone is related to decreases in returns and increases in volatility.

Recently, content analysis studies have focused on drawing inferences with regard to fraud. Cecchini (2005) used support vector machines to examine the risk of financial events such as management fraud, bankruptcy, and restatements. The input vector of the fraud detection model included both quantitative financial variables (account ratios) and qualitative text variables (concept vector, which was preprocessed using accounting ontology) for fraudulent and nonfraudulent companies. Goel (2008) created a methodology to proactively detect fraud by examining the qualitative textual content of annual reports using natural language processing tools. The methodology employs machine-learning techniques to build an automated classifier that can predict the...
likelihood of fraud. In this study, a linguistic analysis was conducted to explore linguistic features (such as voice [active versus passive], uncertainty markers, readability index, tone, usage of proper nouns, type-token ratio, etc.) that could distinguish fraudulent annual reports from non-fraudulent reports. The results indicated that there are systematic differences in the communication and writing style of fraudulent annual reports, and the accuracy of fraud detection can be substantially improved by adding a linguistic dimension to fraud examination. The research in computerized content analysis demonstrates that automatic methods can successfully be used to obtain and analyze large samples, the results of which may offer greater generalizability.

**Readability Research in Accounting**

Another branch of research on accounting narratives is that of readability study. Readability study in accounting has a long history. Riahi-Belkaoui (1995) detailed the results of 42 studies performed from 1952 to 1993 on the readability of accounting documents, particularly financial statements and financial statement footnotes, along with eight studies focused on their understandability. The studies conclude that both readability and understandability by certified professional preparers (e.g., certified public accountants), certified professional nonpreparers (e.g., chartered financial analysts), and college-educated, nonprofessional users are problematic.

Early studies were primarily based on manual coding. Clatworthy and Jones (2001) review prior readability studies on annual reports (Pashalian and Crissy 1952; Soper and Dolphin 1964; Smith and Smith 1971; Still 1972; Dolphin and Wagley 1977; Healy 1977; Barnett and Leoffler 1979; Holley and Early 1980; Courtis 1986; Jones 1988; Schroeder and Gibson 1990; Smith and Taffler 1992a, 1992b; Baker and Kare 1992; Courtis 1995a, 1995b, 1998) and find them to have small sample sizes that range between 18 and 120 documents. Costly manual coding of hard copies apparently precludes large samples. Courtis (1998) and Clatworthy and Jones (2001) supplement their data compilation by using Microsoft Word’s Flesch package to compute the Flesch readability score from electronic annual reports.

Prior studies report variations in readability in multiple sections of the same annual report, as well as in the same section of multiple annual reports. For instance, footnotes are found to be more difficult to read than the chairman’s statement (Courtis 1986). Schroeder and Gibson (1990) examine the readability of 40 firms’ narrative disclosures and conclude that the MD&A and footnotes are more difficult to read than the president’s letter. Smith and Taffler (1992b) find poor readability of narratives to be strongly associated with poor financial performance, and easy readability associated with financial success. Courtis (1998) finds the first paragraph of annual reports to be the easiest to read and the second paragraph the most difficult, and suggests management would bury adverse information in the second paragraph.

Recent work on readability takes advantage of computational linguistic tools and computerized processing. Li (2008) uses a Perl software package to determine the readability of annual reports and finds readability to be related to earnings. In particular, annual reports of firms with lower earnings are less readable, and firms with readable annual reports have more persistent positive earnings. Goel and Gangolly (2009) use the GNU software STYLE to examine the readability of annual reports of fraudulent and nonfraudulent companies. They found that reports of fraudulent companies are much harder to read and comprehend. They show that readability index can be used as one of the features to distinguish fraudulent annual reports from nonfraudulent annual reports.

**Content and Readability Analysis in Finance**

Content and readability analysis in finance has gained momentum since 2001. This subsection discusses manual content analysis, computerized content analysis, and readability research together. Nelson and Pritchard (2007) manually collect and code safe harbor cautionary language
from the annual reports of 186 firms with a high risk of being sued in a securities fraud and 136 firms with low risk between 1996 and 2003. According to their results, firms with greater litigation risk disclose more cautionary language.

Finance researchers have successfully extracted market optimism or sentiment information from texts and found such information to be of value to the market. These texts include Federal Open Market Committee (FOMC) minutes, Internet message boards, news commentaries and articles, annual reports, prospectus, and disclosures by management, analysts, and financial press. Using an automated Latent Semantic Analysis (LSA) algorithm, Boukus and Rosenberg (2006) identify characteristic themes from 152 FOMC minutes from 1987–2005, and find these themes in the Federal Reserve’s communication to be correlated with current and future economic conditions.

Das and Chen (2001) use computerized natural language and statistical algorithms to automate classification of stock message board postings, and find investor sentiment extracted from such postings to be linked to market movements. Using software package and computational linguistics methods, Antweiler and Frank (2004) measure the bullishness of 1.5 million messages posted on Internet stock message boards. These messages are found to be predictive of market volatility, stock returns, and trading volume. Again using computational linguistics methods, Antweiler and Frank (2006) identify events from Wall Street Journal news stories from 1973 to 2001. Their event studies find news stories to be predictive of future returns.

Tetlock (2007) uses a quantitative content analysis program called General Inquirer to analyze electronic text of the Wall Street Journal’s “Abreast of the Market” column. His content analysis produces a numerical measure of the news media’s pessimism, which is found to be predictive of stock prices and trading volume. As in Tetlock (2007), Tetlock et al. (2008) use computerized content-analysis techniques to determine relative frequency of negative words in Wall Street Journal and Dow Jones News Service news stories, and find such relative frequency measure to be predictive of individual firms’ accounting earnings and stock returns.

Mohan (2006) examines the categories of words used in over 70,000 electronic 10-K filings from 1996–2005 with content analysis software, and finds that monitoring by independent boards, analysts, and institutional shareholders helps enhance the information value of corporate disclosure and reduce litigation risks. Feldman et al. (2008) construct a tone measure from the number of “positive” and “negative” words (as classified by the Harvard’s General Inquirer) in the MD&A of 10-K and 10-Q filings. They find significant association of the tone with stock returns around the filing dates. Hoberg and Phillips (2009) used computer-based algorithms to compute word vectors and cosine similarity measures from 49,408 product descriptions in 10-K filings between 1997 and 2005. Their findings indicate that firms merge and buy assets to exploit asset complementarities.

Hanley and Hoberg (2008) use computer algorithms to compute word vectors and document similarity measures based on 9,818 IPO filings and subsequent amendments, and find disclosure in the prospectus to be an important factor in IPO price formation. Using a similar methodology, Hanley and Hoberg (2009) download and parse 1,750 IPO prospectuses between 1996 and 2005. Their findings suggest that an informative prospectus results in more accurate offer prices and less underpricing.

Loughran and McDonald (2009a) develop computer programs to examine the readability of 42,357 10-K filings during 1994–2007. Their results significantly relate improved 10-K readability to increased small investor trading, a higher likelihood of seasoned equity issuance, and better corporate governance. Kothari et al.’s (2009) content analysis software counts target words in more than 100,000 disclosure reports by management, analysts, and news reporters. Their results link the tone of these texts to firms’ risk proxies, such as cost of capital, stock return volatility, and analyst forecast dispersion.
Retrieval of Accounting Information

Retrieval of Text: The Accounting Lexicon and the Role of Thesauri

Text retrieval entails the comparison of a user query to a set of free-text records in order to select appropriate matches. A key factor in successful retrieval is the choice of terminology in query formulation. In library science, thesauri, which address the variability in language, are used to improve retrieval. Fidel (1991) studied online searching behavior and the selection of descriptors or text words as search keys for free-text searching. He found, when using text words, synonyms were not used to compensate for the lack of vocabulary control. He suggested the need for easy, flexible, and inexpensive online access to thesauri and the construction of working switching languages and intermediary expert systems. Muddamalle (1998) noted that in domains where terminology changes frequently, a thesaurus needs to be updated regularly. For example, the Art & Architecture Thesaurus® Online is updated monthly (Art & Architecture Thesaurus® Online 2010). Jacob (1994) found different disciplines develop different terms for describing the same phenomena, which serves to isolate bodies of knowledge and prevent cross-disciplinary communication, including the “recognition of a commonality that exists across disciplinary boundaries.” He favored the development of more generalized structures that will facilitate the transfer of information across domains and encourage the dialog between them.

Thesaurus construction techniques are usually separated into manual, automatic, or semi-automatic approaches (Garnsey and Fisher 2008). Manual construction is complex, costly, and time-consuming. As a direct result of this, manual maintenance (update) can only be performed infrequently (Gangolly and Tam 2002). Gangolly and Tam (2002) argue for adopting computational linguistics methods and techniques for automatic generation of accounting dictionaries and thesauri from existing corpora. To demonstrate this automatic approach, they extracted high-frequency terms from 2000 EDGAR filings using the CMU-Cambridge Natural Language Modeling Toolkit. Manual approaches have also been criticized for the inconsistencies introduced when multiple people are almost certainly employed in the endeavor (Boguraev and Pustejovsky 1996; Gangolly and Tam 2002). Automatic approaches, on the other hand, while overcoming the time and cost limitations of manual approaches, have been criticized as resulting in thesauri that are weak in structure and contents and for their inability to recognize the relational complexities and semantic ambiguities intrinsic in languages (Schneider 2005). Thus, others (Schneider 2005; Anderson and Perez-Carballo 2001) assert that manual and automatic approaches (a hybrid approach) are complementary and that it is both beneficial and necessary to combine them.

Gangolly and Wu (2000) conduct preliminary work using term-document matrices to explore automatic classification of accounting concepts. Garnsey (2006) uses Latent Semantic Indexing (LSI) combined with hierarchical agglomerative clustering to automatically group related terms together and to successfully automatically classify accounting concepts. LSI has been shown to place terms used in the same context closer together in a term document space than when using a term-document matrix. Fisher and Garnsey (2006) use LSI combined with agglomerative clustering to identify related FASs amendments and to verify a taxonomy of changes in FASs. Financial statements may currently be prepared under either U.S. GAAP or the IASB standards. The preliminary research by Garnsey (2008) found numerous differences in terminology. One would expect that different terms in U.S. and International GAAP that represent the same concept could be grouped together using LSI. The method may also be helpful in developing an up-to-date accounting thesaurus.

Maintenance of an up-to-date thesaurus requires early and effective identification of relevant terminology. Garnsey and Fisher (2008) used computational linguistics on authoritative pronouncements and financial statements to identify new accounting terms in their work on how new
terms are incorporated into the accounting lexicon. They found that while many accounting terms first appeared in the financial statements, a significant number of new terms first appear in the finance literature.

**Fisher and Garnsey (2010)** employ a hybrid approach to thesaurus construction in accounting. They use the *Price Waterhouse & Co. (1974)* thesaurus as a starting point for the analysis of the hierarchical structure (tree) of terminology related to employee benefits in existence as of 1974, and then use automated methods to examine the AAL in existence today (e.g., GAAP Codification Section No. 715) in order to determine what new terminology should be added to the tree. The final result is a digital accounting thesaurus for a subset of the accounting lexicon, which could be used to augment the search mechanisms of online collections of accounting documents by offering users prompts or additional related terms and thereby improving access to and information retrieval from such collections.

**The Retrieval of Text-Query Formulation**

**Borthick et al. (2001)** investigated the effect of ambiguity on query development in relational databases. They took information requests and varied the semantic distance between the information request and the expression in the query language that would produce the desired information. In the context of their research, *semantic distance* is defined as “the distance between the information users want and the expression in the query interface language that will produce that information.” They found that accuracy and efficiency were inversely related to the ambiguity of information requests.

**Garnsey and Hotaling (2007)** found that when a group of accounting educators was given a list of accounting terms selected randomly from the AAL, at least one accounting educator could not identify any related terms for 13 of the 30 terms. They found that the number of terms with no response declined significantly when the educators were asked to identify related terms from a list. This agrees with research in psychology that has shown that recognition is generally better than recall (Gibson et al. 1964; Nelson 1978), and reinforces the role of “prompts” in facilitating query formulation. The use of a contemporary accounting thesaurus could also assist in overcoming such recall challenges.

The Examination Content section of the AICPA website for the Uniform CPA Examination identified one of the skills required of entry-level accountants as the ability to “perform research and analyze information” (AICPA 2002). FARS research requirements using simulations have been included in the computerized CPA exam since April 2004. Garnsey et al. (2009) demonstrated the usefulness of providing a list to students doing research in the AAL. Their suggested terms were taken directly from the FARS database and provided students with a means of reducing the semantic distance between their information need and their query to the database. When students were given a list of possible search terms, their ability to locate the correct pronunciation increased significantly. More generally, research in information retrieval has found that people have a higher success rate in finding targeted information objects as the number of index terms used for each object increases (Gomez et al. 1990).

**Retrieval of Accounting Quantities from Text**

The EDGAR database is a key public source of electronic financial reports. Since these financial reports are typically semi-structured or unstructured text documents, great variation exists in file structure and terminology. Extracting useful information from these reports is a challenge, especially for academic researchers and professional analysts who examine large numbers of firms. Various researchers (Tanaka 1982; Mui and McCarthy 1987; O'Leary and Munakata 1988; O'Leary and Eis 1991; O’Leary and Kandel 1992; Steier 1995; Ferguson 1997; Steier et
al. 1997; Nelson et al. 2000) have worked on building intelligent software to handle unstructured reports. Adding to these prior studies, accounting researchers have recently made new progress in this area.

Leinmann et al. (2001) introduced text-mining procedures to extract data from financial statements presented in ASCII table format in EDGAR filings. Bovee et al. (2005) work on automated access to financial information embedded in unstructured text documents such as EDGAR filings. They developed a prototype Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) system, which implements intelligent parsing to extract accounting numbers from financial statements by matching financial statement line-item labels to XBRL synonyms. Grant and Conlon (2006) extract employee stock options information from financial statement disclosure notes in the EDGAR database. Grant and Conlon’s (2006) EDGAR Extraction System makes use of KWIC, CMU Toolkit for Statistical Language Modeling, regular expressions, and pattern matching techniques, and is trained to recognize word phrases from Statement of Financial Accounting Standards (SFAS) No. 123.

Ding and Chen (2006) devise a text-mining system to automatically extract executive compensation data from Securities and Exchange Commission (SEC) proxy filings. They find individual retrieval technology to be deficient for the extraction task, but find a hybrid approach promising. Their hybrid approach integrates pattern matching with regular expressions, an ontology knowledge base for word classification, data range checking, and word context analysis. Cong et al. (2007) propose a template-based approach to retrieve the structure and contents from financial statements, and report promising performance with this approach. The template takes advantage of various contextual, semantic, geometric, and structural information, such as relative position of a given financial statement, title and header, percentage of nonwhite characters, and horizontal boundary.

The overall literature on retrieval of accounting quantities from text documents is quite sparse to date, although each prior study has reported moderate success. While more work is needed to improve the electronic accessibility of accounting quantities from text, the most recent studies suggest that regular expression-based pattern matching and template-based programming are promising tools.

SUGGESTED FUTURE RESEARCH PATHS

Overview

As technology continues to facilitate online access to information, we believe that the role of text analytics and information retrieval in the accounting domain will progress from simple applications (such as term recognition) to more sophisticated applications (such as thesaurus construction). More applications using natural language processing techniques for information retrieval and classification are required in order to successfully locate, extract, and organize information. The need for further research in text analytics and information retrieval falls into several areas: the success of the Codification Project in achieving its goals, identification of new terms entering the lexicon, standardization of accounting vocabulary, and thesaurus construction to facilitate retrieval. We briefly examine each of these areas.

The Codification

It is critical that research addresses how best to archive temporal versions of the AAL so as to enable effective access and retrieval. While changes in the AAL will be effected by directly amending the codified text, it is unclear how access to prior versions of the AAL will be provided, as well as at what level of temporal detail. The application of accounting principles occurs as of a specific point in time. Questions regarding the proper application of accounting principles often arise many years subsequent to the application date. For example, while the Securities and Ex-
change Commission investigation of Enron began in 2001, litigation of the case was still ongoing as of 2006. The events in the case pre-date both the investigation and the litigation, requiring the ability to review the AAL applicable as of the prior points in time. Thus, access to prior versions of the AAL is an important issue (Fisher 2007).

Fisher (2009) suggests that the Codification of the AAL will benefit from the incorporation of a formal XML structure. Such a structure can support improved search capabilities for the AAL by supplementing user-entered keyword searches with additional related terms that can be derived from markup. A clearer determination of the meaning (beyond simple word counts) of portions of the AAL can be described through markup tags enabling more effective user searches. Research into effective markup would be useful.

Ultimately the success of the Codification project will be determined by the users of the AAL. That is, will users find the Codification easier to navigate, more understandable, and will user searches of the AAL be more effective? Future academic research, which addresses how the codified text is used and by whom, should be performed. Research into how the technological specifications of document structure can better assist in facilitating the use of the AAL should also be conducted. For example, academic research into appropriate “keywords” that would best serve as indicators of the meaning and application of the various sections of the AAL should be valuable to the FASB as it maintains the Codification.

**Identification of New Terms Entering the Lexicon**

Often, individuals do not understand the meaning of new terms until they become “newsworthy.” Recently, the press has used the term “clawback” to refer to the return of some types of executive compensation. Prior to the recent financial crisis, many individuals were not familiar with the term. For individuals to understand current accounting news, they must be familiar with new terms as they enter the accounting lexicon. Additional research is needed to better determine where accounting terminology is first introduced: financial accounting or other authoritative pronouncements, corporate financial reports, or financial periodicals. Garnsey and Fisher (2008) concluded that the finance domain may be a substantial source of new terms. Further research is needed to verify this.

Many new terms entering the accounting lexicon are multiple-word phrases. Collocations are terms which occur next to each other and which taken together have meaning. Different statistical methods have been used to determine which consecutively occurring words in text collections are collocations (Manning and Schutze 1999). Garnsey and Fisher (2008) used the CMU-Cambridge Language Modeling Toolkit to identify terms which co-occur and to provide probabilities and frequencies of those bigrams. Bigram pruning was used to remove uncommon explicit bigram estimates, followed by manual pruning of nonaccounting-related bigrams. Future research will need to be more aggressive in excluding bigrams that do not contribute to accounting language from the language model. Different cutoffs should be examined, as well as the feasibility of pruning the list of bigrams identified at a particular cutoff point. Finally, research should explore the possibility of employing automatic methods to substantially expand the size of the corpus used by Garnsey and Fisher (2008). The expanded corpus should include additional financial statements, as well as additional sources of accounting information such as popular financial periodicals (e.g., Wall Street Journal), international accounting standards, and Statements on Auditing Standards.

**Standardization of Accounting Vocabulary**

One of the most important reasons for language standardization is that it allows the language to serve as a model for everyone in the community to agree upon in order to communicate
Standardization of language may refer to phonology, spelling, punctuation, grammar, and lexicon. In terms of accounting language standardization, we refer to spelling and lexicon.

The standardization of accounting vocabulary is a difficult problem because not only is the vocabulary of the accounting language large, but also, each term needs several features for its semantics to be adequately captured and represented. The main challenge is finding the right level of generalization. It will be useful to investigate how new words in the accounting domain are formed, i.e., are new words formed from a base word and, if so, are new (derived) words related to base words. For instance, if a new word entering the accounting lexicon is formed from a base word by means of suffixes or affixes and the meanings of the new word and base word are closely related, then it means that new word is both formally and semantically related to the base word (e.g., security becomes securitization by adding the suffix “ization” to denote the act of making/becoming something). This kind of analysis will help in formulating rules of word formation and contribute toward maintaining and extending thesauri in response to changes in the accounting domain.

The FASB’s recent Codification project instructed authors “to change as few words as possible when authoring their topics” (FASB 2008). However, the FASB Accounting Standards Codification™ Notice to Constituents specifically notes several efforts that were made to standardize language:

Over the years, standards have used numerous terms to describe an entity, such as “company,” “organization,” “enterprise,” “firm,” “preparer,” and so forth. To provide consistency, the FASB adopted the term “entity” as the appropriate term … Because the FASB adopted the term “entity” as the common term throughout the Codification, the FASB adopted the term “intra-entity” as the replacement for “intercompany.” The FASB has provided a definition for “intra-entity” and intends for the terms “intra-entity” and “intercompany” to be interchangeable … Similar to the multiple terms used to refer to an entity, standards have used various words or phrases to refer to a requirement, such as “should,” “shall,” “is required to,” “must,” and others. The FASB believes such terms represent the same concept—the requirement to apply a standard. To establish consistency, the Codification uses the term “shall” throughout the standards Sections. (FASB 2008)

Garnsey (2009) found that although the Codification is largely internally consistent, this is not universal. She also found instances where the Codification used different terms than the original pronouncements. More research is needed to determine if the inconsistencies and differences are significant and, if so, whether they are likely to impact information retrieval by users. It is also important to note that the Codification glossary does not include or define significant accounting elements (e.g., retained earnings, liability) that were previously included in the concept statements which were not incorporated into the Codification. This may negatively impact information retrieval, but investigation is required to determine the true impact of the omission.

Standardization across the international accounting lexicon may be impossible to achieve. Research in both information science (Bates 1986) and human-computer interaction (Furnas et al. 1987) has shown people use different terms to describe a concept based on their backgrounds. Accounting draws on concepts from finance and economics. Each field has its own vocabulary, further complicating the issue. In addition, preliminary work by Garnsey (2008) found numerous examples of language differences between international and U.S. GAAP. The differences included both the lexicon and spelling. The difficulties in standardization make research to develop a taxonomy and an up-to-date thesaurus for accounting more of an imperative.

Accounting Thesaurus Construction

The FASB’s focus throughout the Codification project on both standardizing the terminology used as well as reducing information redundancy makes it even more important that a user is able
to formulate a query that contains the language used in the database. This may be especially problematic for users who are inexperienced or whose information need is not well defined. Without sufficient access points, users will not be able to locate desired information. Thesaurus construction can also benefit from thematic analyses of accounting narratives to identify salient

### TABLE 1

**Selected Primary Lessons Derived from the Literature**

<table>
<thead>
<tr>
<th>Lesson Learned</th>
<th>Research Stream</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Codification provides a more consistent document structure and eliminates redundancies.</td>
<td>Retrieval of Accounting Text</td>
<td>Fisher 2009</td>
</tr>
<tr>
<td>There is an open question regarding temporal access to the AAL.</td>
<td>Retrieval of Accounting Text</td>
<td>Fisher 2009</td>
</tr>
<tr>
<td>Smaller and less profitable companies are less likely to disclose adequate information.</td>
<td>Drawing Inferences: Manual Content Analysis</td>
<td>Singhvi 1968; Miller 2002</td>
</tr>
<tr>
<td>Firms disclose more positive information than negative.</td>
<td>Drawing Inferences: Manual Content Analysis Partially Computerized Content Analysis</td>
<td>Pava and Epstein 1993; Boo and Simnett 2002 Abrahamson and Park 1994</td>
</tr>
<tr>
<td>Machine learning algorithms appear to be an effective means to classify text so as to predict incidents of fraud.</td>
<td>Drawing Inferences: Computerized Content Analysis</td>
<td>Cecchini 2005; Goel 2008</td>
</tr>
<tr>
<td>Poor readability of firm narratives is associated with poor financial performance.</td>
<td>Drawing Inferences: Readability Studies</td>
<td>Smith and Taffler 1992b Li 2008</td>
</tr>
<tr>
<td>Accurate information retrieval is increased by user knowledge of the query language and user prompts.</td>
<td>Retrieval of Accounting Text</td>
<td>Borthick et al. 2001 Garnsey and Hotaling 2007 Garnsey et al. 2009</td>
</tr>
<tr>
<td>Progress is being made in the development of software tools to extract useful information from unstructured accounting text.</td>
<td>Retrieval of Accounting Text Quantities from Text</td>
<td>Tanaka 1982; Mui and McCarthy 1987; O’Leary and Munakata 1988; O’Leary and Eis 1991; O’Leary and Kandelin 1992; Steier 1995; Ferguson 1997; Steier et al. 1997; Nelson et al. 2000</td>
</tr>
<tr>
<td>Computational linguistics can be successfully used to analyze accounting terminology.</td>
<td>Retrieval of Accounting Text</td>
<td>Gangolly and Wu 2000; Fisher and Garnsey 2006; Garnsey 2006; Grant and Conlon 2006; Garnsey and Fisher 2008</td>
</tr>
</tbody>
</table>
content categories. Future work in this area needs to focus on extracting new concepts and relations so that these could be assigned to categories that uniquely fit their description.

CONCLUDING OBSERVATIONS

Automated methods in text analytics and information retrieval are becoming important tools in accounting, finance, and business research.

Table 1 summarizes the lessons derived from the text analytics and information retrieval literature in accounting.

There is ample room for text analytics and information retrieval to grow. In accounting and finance research, such room exists in linguistic characteristics, the sources of documents, and the economic variables of interest. Current research focuses on a narrow set of computational linguistic characteristics (e.g., readability and tone). Other dimensions of linguistics awaiting future study include aspects of semantics, grammar, morphology, syntax, and phonology (especially those determinable by computational linguistics means). In addition, the Internet is now a significant source of electronic documents for computerized analysis. Every major company maintains a website which contains other documents besides the annual report, IPO prospectus, and press releases. These other documents (such as research and development narratives, employment opportunity postings, investor relation documents, etc.) could contain information of value to economic decision makers. Researchers of information economics could apply text analytics and information retrieval techniques to earnings and returns prediction studies, as well as to studies of economic events such as fraud, bankruptcy, and mergers and acquisitions. Other text documents of interest to accounting researchers that are available electronically for text analytics and information retrieval studies include: accounting pronouncements, SEC regulations, tax laws, tax court cases, comment letters to the FASB, accounting research articles and working papers, and accounting pedagogical materials (syllabus, program description, etc.).

Applying computational linguistic tools to accounting and finance research is not straightforward. Prior work (e.g., Tetlock 2007 and Tetlock et al. 2008) on pessimistic tone is based on counts of words identified as negative in Harvard’s Psychological Dictionary. Loughran and McDonald (2009a, 2009b) question the aptness of analyzing accounting and finance texts with a word list originally developed for psychology and sociology, and advocate the use of an alternative word list tailored for business research. In addition, simple word counts used in prior research ignore differences in relatedness of words. Pedersen et al.’s (2005) WordNet-Similarity project implements eight quantified measures of relatedness, which could be used to further fine-tune such a word list. Hanley and Hoberg’s (2008, 2009) and Hoberg and Phillips’ (2009) use of the “document similarity” measure to analyze IPO prospectuses are examples of applying Pedersen et al.’s (2005) work. Research on fine-tuning of computational linguistic methodologies for text-based accounting and finance research is likely to continue into the future.

REFERENCES


Art & Architecture Thesaurus® Online. 2010. About the AAT. Available at: http://www.getty.edu/research/conducting_research/vocabularies/aat/about.html.


———, ———, and ———. 1999. Evidence that management discussion and analysis (MD&A) is a part of a firm’s overall disclosure package. *Contemporary Accounting Research* 16 (1): 11–34.


———. 2009. A comparison of language in the Codification and the FASB and its predecessors pronouncements, paper presented at the AAA Information Systems Section Midyear Meeting, Charleston, SC.


Leinnemann, C., F. Schlottmann, D. Seese, and T. Stuempert. 2001. Automatic extraction and analysis of


——-. 2009b. When is a liability not a liability? Working paper, University of Notre Dame.


Pedersen, T., S. Patwardhan, and J. Michelizzi. 2005. WordNet:: Similarity—Measuring the relatedness of


Journal of Emerging Technologies in Accounting

Volume 7, 2010

American Accounting Association


