TASK- AND DOMAIN-DEPENDENT NATURAL LANGUAGE PROCESSING FOR SUPPORTING SPOKEN DOCUMENT USE

by

Siavash Kazemian

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Computer Science
University of Toronto

© Copyright by Siavash Kazemian 2018
TASK- AND DOMAIN-DEPENDENT NATURAL LANGUAGE PROCESSING FOR SUPPORTING SPOKEN DOCUMENT USE

Siavash Kazemian
Doctor of Philosophy
Department of Computer Science
University of Toronto
2018

Abstract

This dissertation investigates how Natural Language Processing (NLP) can improve users’ access to spoken content in multimedia documents. Information seeking in electronic environments is a highly variable process shaped by information-seeking factors. It is shown that incorporating knowledge about these factors into the design process—task- and domain-dependent (TDD) development—leads to effective Spoken Document Support System (SDSS) prototypes. The presented field studies identify Automatic Speech Recognition (ASR) and Sentiment Analysis (SA) as two requisite NLP technologies for supporting information seekers in the studied domain. The dissertation also demonstrates how SA models can be optimized with TDD feedback taken into account. The described optimization techniques rely on task-based evaluation, which is not commonly used in SA and ASR research. The experiments reported here identify important scenarios in which non-task-based evaluation is not correlated with the outcome of task-based appraisal. Although some task-based evaluation of SA, including the one advocated here, is no more resource-intensive than calculating intrinsic measures, others can be more expensive if they involve human-subject experimentation. This work also introduces a semi-automatic TDD measure of ASR quality that simulates human-subject experimentation, minimizing resource consumption without sacrificing empirical grounding.
Acknowledgments

I am greatly indebted to my advisor, Gerald Penn, for his wisdom, support, and patience over the years. Thanks to him, my graduate studies have been a tremendous learning opportunity. I owe a great deal of this work to his intellectual guidance and technical contributions. I am also deeply grateful to my committee members, Allan Borodin and Mark Chignell, as well as my external examiner, Gary Marchionini, for their guidance and insightful comments. I would also like to show my gratitude to Graeme Hirst, for his teachings, valuable feedback, and leadership in the Computational Linguists lab. I am also deeply grateful to my collaborators, Ani Nenkova, Benoit Favre, Clare Voss, Cosmin Munteanu, Frauke Zeller, and Yang Liu, for playing such an important role in my education.

I have been very fortunate to be surrounded by remarkable mentors and whose friendship has brought so much meaning to my life. Special thanks to Adam Lee, Aditya Bhargava, Afra Alishahi, Afsaneh Fazly, Arvid Frydenlund, Aida Nematzadeh, Anthony McCallum, Amin Tootoonchian, Barend Beekhuizen, Chloe Pou-Prom, Chris Parisien, Dennis Ochei, Ehsan Variani, Elizabeth Patitsas, Eric Corlett, Frank Rudzicz, Gagandeep Singh, George Dahl, Ilya Sutskever, Ioan Stefanovicì, Jackie Cheung, Jasper Snoek, Julian Brooke, Katie Fraser, Kinfe Tadesse, Kyla Cheung, Laurent Charlin, Libby Barak, Michael Chiu, Michael Guerzhoy, Michael Tao, Michael Reimer, Muuo Wambua, Navdeep Jaitly, Nicole Sultanum, Nona Naderi, Patricia Thaine, Paul Cook, Rouzbeh Farahmand, Sajad Shirali-Shahreza, Sean Robertson, Sepand Mavandadi, Serena Jeeblee, Shunan Zhao, Stefan Tartz, Suzanne Stevenson, Sven Dickinson, Timothy Fowler, Tong Wang, Ulrich Germann, Vanessa Wei Feng, Varada Kolhatkar, Xiaodan Zhu, Yevgen Matusevych, and Yuval Filmus. I owe special thanks to Patricia Thaine for proof reading my dissertation. I would also like to thank Darren Deering, John Hancock, Josheph Raghubar, Kevin McCormick, Lynda Barnes, Philip Chan, Relu Patrascu, Luna Keshwah, Marina Haloulos, Lisa DeCaro, Sara Burns, and the rest of DCS staff for their great support.

I am also grateful to all of my other teachers, especially, the late Mr. Vegh, a concerned, wise, and tireless English teacher who gave me the confidence to speak up. I would also like to thank Mr. Sorgente, another compassionate high school teacher, who helped me buy my first computer and taught me how to use the DOS command prompt.
Finally, I would like to dedicate this dissertation to my sister, Golbahar Kazemian, a strong woman, whose support I counted on throughout, as well as my parents, Aghdas Memarzadeh and Iraj Kazemian, who immigrated to Canada for our sake.
Table of Contents

Acknowledgments iii

Table of Contents v

List of Tables viii

List of Figures ix

Chapter 1 Introduction 1

Chapter 2 Background and Literature Review 8

2.1. Automatic Speech Recognition 8

2.1.1. Measuring ASR transcript accuracy 10

2.1.2. Using ASR to gain access to spoken content: Speech retrieval 12

2.1.3. Measuring speech retrieval success 14

2.2. Sentiment Analysis 17

2.2.1. Sentiment analysis workflow 18

2.2.2. Using SA to support information seeking in the financial domain 22

2.3. Information Seeking Process 25

2.3.1. Examining information-seeking factors affecting the use of spoken documents 29

2.3.2. Task-based evaluations of SDSSs 32

2.3.3. Interacting with spoken documents 34

2.3.4. Representing a collection of spoken documents 39

2.3.5. Grounding experiments in ISP theory 42

Chapter 3 Investigating Spoken Document Use by Investment Management Professionals 44

3.1. Background 45

3.2. Methodology and data 46

3.3. Analysis 50

3.3.1. Q1: Setting 50

3.3.2. Q2: Tasks 51
5.4. Conclusion .................................................................................................................. 102

Chapter 6 Evaluating Automatic Speech Recognition.................................................. 104


6.2. Method....................................................................................................................... 108

6.2.1. Task....................................................................................................................... 109

6.2.2. Data....................................................................................................................... 109

6.2.3. Tools and experimental design ................................................................. 109

6.3. Automatic Evaluation Metric.............................................................................. 111

6.4. APP Features......................................................................................................... 113

6.5. Experiments and Discussion.............................................................................. 114

6.5.1. Predicting Auditor Performance ..................................................................... 114

6.5.2. Decision-Detection ......................................................................................... 117

6.6. Related Work ........................................................................................................ 118

6.7. Conclusion............................................................................................................. 120

Chapter 7 Conclusions and Future Work ................................................................. 121

Bibliography .............................................................................................................. 125
List of Tables

Table 3.1: Examples of discussions of actions and outcomes ................................................ 58

Table 3.2: Sought-after knowledge for predicting future actions and outcomes .................. 55

Table 4.1: Participants’ professional history in the participatory design (PD) study ......... 67

Table 5.1: Average accuracy of the sentiment classifier using different feature sets .......... 95

Table 5.2: Returns of the experimental, baseline, and topline trading strategies ............... 97

Table 5.3: Sentiment classification accuracy and trade returns of different feature sets ....... 102
List of Figures

Figure 2.1: Overview of a typical Automatic Speech Recognition system ........................................ 11
Figure 2.2: Collecting lexical features from a text document .......................................................... 17
Figure 2.3: SVM’s optimal decision hyperplane ................................................................................. 21
Figure 2.4: The SCAN user interface (Whittaker et al., 1999) ............................................................ 35
Figure 2.5: SCANMAIL user interface (Whittaker et al., 2002) .......................................................... 36
Figure 2.6: A sample ASR transcript of a meeting with a WER of 38.9% ............................................ 37
Figure 2.7: The result list produced by Buchenwald portal for the query “bezetting” ......................... 41
Figure 3.1: Information-task hierarchy of institutional financial analysts ............................................. 51
Figure 3.2: A Bloomberg-terminal chart of historic EUR-USD index prices ........................................ 59
Figure 3.3: Information subtasks performed by financial analysts when using spoken records ........ 64
Figure 4.1: Cashflow overview ........................................................................................................... 71
Figure 4.2: The Sensitivity Analysis widget .......................................................................................... 72
Figure 4.3: Sentiment comparison widget ............................................................................................. 72
Figure 4.4: The taxonomy of spoken document support tasks .............................................................. 78
Figure 4.5: The proposed SDSS design ................................................................................................. 80
Figure 4.6: Interacting with the natural language search box in the dynamic pane ............................ 81
Figure 4.7: Interacting with the tagged production figures in the transcript pane .............................. 82
Figure 4.8: Microsoft’s historic and forecasted commercial cloud revenue run rate ......................... 88
Figure 4.9: Line chart comparing a company’s actual and projected total revenue .......................... 88
Figure 4.10: Analyst profile .................................................................................................................. 89
Figure 5.1: Market capitalization of each trade with one-day holding period
Figure 5.2: Percent change of trade returns vs. SVM scores for different holding periods
Figure 5.3: Returns for different SVM distance thresholds
Figure 6.1: JFerret interface as used in the human-subject experiment
Figure 6.2: Meeting-level WER vs average H-score for all transcript conditions
Figure 6.3: Precision/recall curve for each of the leave-one-out trials
Figure 6.4: Effects of holding out the test meeting (during training) on APP
Figure 6.5: APP feature ablation experiment
Figure 6.6: Effects of training dataset WER on models' decision-detection performance
Figure 6.7: Decision-detection performance on held-out meetings
Chapter 1
Introduction

In recent years, multimedia content has become ubiquitous, accounting for 64% of all Internet data transfers in 2014, and projected to account for 80% to 90% by 2019, at which point approximately one million minutes of multimedia data are expected to cross the network every minute (Cisco, 2015). Much of this growth is associated with the increasing use of multimedia documents in corporate settings and in news media transmissions (Cisco, 2015; Levy, Newman, & Nielsen, 2015). This dissertation is concerned with making information stored in these documents more accessible to information seekers. A significant portion of this information is in spoken form (Chelba, Silva, & Acero, 2007), motivating this dissertation to focus on improving access to spoken content by using Natural Language Processing (NLP). For the sake of simplicity here, multimedia documents containing spoken content are referred to as spoken documents.

NLP research is concerned with the processing of human (natural) language and has led to the development of technologies that could potentially improve access to spoken documents. These technologies include Automatic Speech Recognition (ASR), the automatic transcription of spoken documents; and Sentiment Analysis (SA), the automatic interpretation of authors’ subjective attitudes towards discussed topics (e.g., negative vs. positive). ASR and other NLP technologies have enabled users to interact more naturally with computational devices through Voice User Interfaces (Pearl, 2016), available in products such as Google Assistant (Google Assistant, 2018). However, these technologies have not been as effective in supporting information seekers in accessing spoken content, with the notable exception of assisting people with hearing impairments. Increasing the accessibility of multimedia content in other settings has proved to be more challenging, despite numerous attempts (Fiscus, Ajot, Garofolo, & Doddingtion, 2007; Garofolo, Auzanne, & Voorhees, 1999; Cui, et al., 2013).

At the same time, speech and NLP research have enjoyed remarkable progress. ASR accuracies of 93.1% or more are now achievable in challenging genres such as conversational speech (Xiong, et al., 2017). Similar progress has made SA accuracies of more than 90% attainable (Paltoglou & Thelwall, 2010). Moreover, feats achievable by speech and NLP technologies, such as the ability to transcribe spoken documents, have already been found to be useful in related
field research and user studies (Whittaker, Tucker, Swampillai, & Laban, 2008; Whittaker, et al., 2002). As will be discussed in Chapter 3, at least for some prominent user groups, namely, investment management professionals (IMPs), NLP technologies still do not play a significant supporting role in making spoken content more accessible. This dissertation sheds light on how such support can be provided.

We focus on supporting IMPs’ information-seeking which is central to their day-to-day activities. IMPs’ information-seeking notably involves the use of spoken records covering public events, such as earnings calls and central bank news conferences, or private internal investment meetings. As will be discussed in Chapters 3 and 4, manual transcripts are commonly used to cover earnings calls and news conferences. Further automatic processing of these transcripts can lead to significant time-savings for IMPs. Utilizing NLP in this context will not only give rise to more comprehensive support systems for IMPs, but will also provide a valuable opportunity for NLP researchers to evaluate their models outside research settings. Records of private internal investment meetings commonly include personal notes, meeting minutes, and shared artifacts. However, as will be discussed in Chapters 2 and 6, these records lack essential contextual information. Moreover, taking notes or minutes can limit participants’ engagement in meetings. As will be discussed in Chapter 2, raw audio/video recordings are also of limited use due to the absence of usable navigation tools. NLP technologies, as well as ASR, can enable the development of such tools, allowing information systems to automate the transcription of meetings, alleviating the need to take manual notes, and providing more comprehensive support to information seekers in this domain.

This dissertation examines the benefits of restricting information system design and NLP development to supporting IMPs’ spoken document use. This restriction can be implemented by incorporating domain knowledge about IMPs’ information seeking activities into the design of spoken document support systems (SDSSs) and optimizing the necessary NLP technologies. This restrictive paradigm is referred to as here as Task- and Domain-Dependent (TDD) development. “Task” in TDD development refers to the SDSS’s mission. Such tasks include, for instance, presenting spoken records of earnings calls and news conferences. “Domain” refers to the contextual information about the system’s usage scenario. The SDSS’s “task” is not to be confused with IMPs’ “tasks”, which, as will be discussed in Chapter 3, often lead to information-seeking activity. Similarly, the SDSS’s “domain” should not be confused with the “information
domain” of documents used by IMPs, which, as will be discussed in Chapter 2, influences their information seeking processes (Marchionini, 1995). According to Marchionini (1995), there are 6 main information-seeking factors that shape a user’s information seeking process (ISP), including task and information domain. TDD development allows for the design of a support system that limits its focus to ISPs shaped by similar information-seeking factors.

TDD NLP development as discussed in this dissertation amounts to producing NLP models that are optimized for their role in the designed SDSSs. The most direct way to appraise these models is to conduct extrinsic task-based evaluations. In task-based evaluation, a system is assessed by its ability to perform a specific task. TDD evaluation is an appraisal based on the task and domain the NLP models are optimized for. For instance, TDD evaluation of an SA model designed to detect market reaction can be achieved by measuring the returns of trades that are informed by the SA model. Although TDD evaluation can be automatic, as was the case in the previous example, they can involve task-based human-subject experiments if the assessed results are obtainable directly from end-users. In task-based human subject experiments, human participants often act as potential end users and in different trials, use the results produced by several NLP models under evaluation to perform a task. NLP models are then assessed by how well their users performed the task in the human-subject experiment.

It is also possible to conduct extrinsic evaluations that are not task-based. This can be achieved by relying on human judges’ opinions, without having them perform a real-world task. To perform this form of appraisal systematically, experimenters typically ask human judges to annotate a test set. Systems can then be evaluated by comparing their output with human-produced annotations. For example, it is possible to ask human annotators to label expressed sentiment in a set of news articles without providing training or contextual information and evaluate SA systems by how closely they match the human-produced annotations. As will be discussed, this form of evaluation is used commonly and is referred to in this dissertation with the abbreviation, ENT (Extrinsic but Not Task-based). Non-task-based evaluation of NLP models can also be performed intrinsically against an objective ground truth, independently of how models will be used outside research settings. TDD NLP, on the other hand, aims to support the performance of a specific task, for instance, supporting IMPs’ information-seeking. TDD and TDI NLP are to some extent complementary. TDD NLP research can be informed by findings in TDI NLP research, as well as other areas such as Human-Computer Interaction. By
sheding light on the merits of different modeling techniques in supporting end-users, the results of TDD NLP research can in turn be of interest to the wider NLP community.

TDI SA is currently an important direction of research in NLP. The latest trends include the use of neural networks to achieve maximal intrinsic accuracy (e.g., Yang & Eisenstein, 2017; Nakov et al., 2016; Kumar et al., 2016; Irsoy & Cardie, 2014; Tai et al., 2015; Kim, 2014; Socher et al., 2013). Recent work also attempts to develop models that perform well across datasets in disparate information domains (e.g., Kim, 2014). SA research, for example, has tackled the challenging task of detecting sentiment in short documents, such as tweets (Rosenthal, Farra, & Nakov, 2017), or in sentences and phrases (Kokkinos & Potamianos, 2017). This has required the creation of datasets with annotations for smaller units of text which are significantly more resource-intensive. As will be shown in Chapter 3, TDD SA need not be more resource intensive than TDI SA, although TDD development involving human-subject experimentation can be significantly more resource-intensive and time-consuming. Chapter 6 will discuss one such scenario and will introduce a means to significantly reduce the cost associated with TDD development involving human-subject experimentation.

A distinguishing characteristic of TDI NLP development is its use of non-task-based evaluation. In contrast, TDD NLP development must rely on task-based evaluations. TDI SA is commonly evaluated in an ENT manner by calculating an accuracy over a labelled test set (referred to as TDI accuracy or just accuracy). However, the results in Chapter 5 show that an SA system’s accuracy can notably disagree with its merit in task-based evaluations. ASR transcripts are also evaluated in a non-task-based manner by calculating their edit distance with manual transcripts (referred to as the Word Error Rate or WER). Current research has focused on minimizing WER with substantial success. Yet, as it will be shown in Chapter 6, an ASR system’s WER is not a good indicator of its merit in our task-based evaluations. Chapters 3 to 6 suggest that TDD evaluation and development are more fruitful and reliable than contemporary practice in TDI development. Specifically, the results show that by using TDD development, a) more effective NLP-enabled information systems (SDSSs) can be developed to aid users; and b) determining the usefulness of ASR and SA in the real-world scenarios studied here requires the use of TDD evaluation.
This dissertation starts by providing an overview of ASR and its evaluation (Section 2.1). As will be discussed, with few exceptions, progress in ASR has been measured using WER, which presumes task and domain independence. SA is then discussed, and it is noted that TDI accuracy scores are commonly used to evaluate SA (Section 2.2).

There are some rare studies in the fields of Human-Computer Interaction (HCI), social sciences, and information systems that explore how NLP technologies could be used to help users of spoken documents (Section 2.3). The review of this research begins by introducing Marchionini’s theory of Information Seeking Process (ISP) in electronic environments (Marchionini, 1995). Marchionini’s theory states that the ISP is inherently embedded in a problem-solving process initiated by an information seeker during task performance. The theory also states that the form of ISP can vary significantly depending on the task, the information domain, and other information-seeking factors. Due to this variance, it is prudent to assume that each information task could require its own set of tools. The requisite NLP tools for these systems may thus need to be developed, evaluated, and optimized differently in each domain.

TDD design of SDSSs starts with gaining useful knowledge about IMPs’ ISPs (Chapter 3), and then showing that the knowledge gained can lead to useful designs within the constraints of the studied task and domain (Chapter 4). To study IMPs’ ISPs, a contextual inquiry has been conducted in which IMPs were observed while using spoken records of central bank news conferences and public company earnings calls (Chapter 3). The study reveals that the IMPs’ main purpose is to gain the necessary information (referred to as the Essential Predictive Knowledge or EPK) to predict an institution’s future actions (e.g., whether the Federal Reserve will raise interest rates), the outcomes of said actions (e.g., whether the rate hikes will increase the price of inflation securities), and other market participants’ reactions to the shared content (e.g., how market consensus changes regarding the Federal Reserve’s future actions as a result of the shared content). A taxonomy of information tasks is then inferred. As the taxonomy describes, participants examined both how speakers communicated (i.e., paying attention to expressed sentiment as well as communication tactics) and what they communicated. The EPK includes the studied institution’s past and present actions and outcomes, its executives’ cognitive and affective states, and relevant external factors (e.g., oil shocks, US inflation, etc.). The contextual inquiry has also uncovered a general lack of adaptation of software tools for
supporting spoken document use, as all participants preferred using printed transcripts or tools such as tablets that simulated the experience of interacting with printed transcripts.

To validate the taxonomy uncovered, follow-up participatory design (PD) workshops were conducted which have supported the taxonomy and shown that effective support in the studied domain can be provided by relying on NLP technologies such as SA and Speech Processing (Chapter 4). The findings of the contextual inquiry and PD workshops (i.e., domain knowledge) not only inform a better SDSS design, but also show how NLP technologies can be evaluated extrinsically in a TDD manner.

After showing that TDD development can lead to an effective SDSS design (Chapter 4), one requisite NLP technology, SA, is examined to determine whether currently available techniques can viably play a role in the produced designs. It will be shown that for predicting market reaction—measured by returns of SA-informed trading strategies of NYSE-traded equities—TDI SA systems did not perform significantly better than established trading baselines such as the momentum strategy. Next, it will be shown how TDD development can achieve better results by tuning TDI models to market data. The presented optimization technique not only produces significantly better results than the baselines with annualized returns of up to 70.1%, but also better results than models trained end-to-end on market data (Kazemian, Zhao, & Penn, 2016). The reported task-based evaluations isolate specific feature selection dilemmas where the TDI accuracy errs both in magnitude and sign of the delta in SA usefulness. Collectively, the SA experiments in this dissertation show that TDD development can produce systems that are more effective in predicting market reaction to content than TDI development, and that TDD evaluation is the only unequivocal way of measuring this difference.

Furthermore, this is not just a concern specific to SA accuracy. As mentioned above, ASR has been almost exclusively evaluated intrinsically with WER. In a comprehensive human-subject experiment based on the decision audit task involving spoken records of team meetings, we evaluate 4 ASR transcription conditions (provided by several leading ASR research labs). It will be shown that, similar to SA accuracy (Chapter 5), WER does not correlate with ASR usefulness in the reported task-based evaluation (Favre, et al., 2013). As will be discussed, in some contexts, including the experiments of Chapter 6, TDD evaluation can amount to conducting human-subject experiments that are notably time-consuming. As an alternative, a semi-automatic
TDD evaluation system, Auditor Performance Predictor (APP), is introduced that uses statistical pattern recognition to learn an evaluation (Chapter 6). Our experiments will show that APP is a better indicator of transcript usefulness for performing decision audits than WER (Favre, et al., 2013).

Finally, Chapter 7 concludes the dissertation, offering suggestions for future research. It must also be noted that portions of this dissertation have already been published as Favre et al. (2013) and Kazemian, Zhao & Penn (2016).
Chapter 2
Background and Literature Review

This chapter is composed of three sections. Section 2.1 addresses Automatic Speech Recognition (ASR), elaborating on its evolution in the past decades. Speech retrieval (SR), a subfield of ASR, is then discussed which aims to develop information retrieval capabilities for spoken documents similar to what is already available for text. Section 2.2 provides an overview of sentiment analysis (SA) research. Focus is placed on ASR and SA because observational studies that will be described in Chapters 3 and 4, as well as studies of spoken document use in other domains, have suggested that these technologies may be useful in making spoken content more accessible (Whittaker, Hirschberg, & Nakatani, 1998; Whittaker, Tucker, Swampllai, & Laban, 2008). Section 2.3 describes why theories of information seeking should be considered when designing Spoken Document Support Systems (SDSSs). Details regarding the design and evaluation of such systems are subsequently provided.

### 2.1. Automatic Speech Recognition

Chapter 1 defined an audio or video file containing spoken content as a spoken document, also stating that the primary aim of ASR is to generate word-for-word transcripts of spoken document input. This section explains how ASR achieves this feat by utilizing acoustic and language models. Section 2.1.1 discusses how the accuracy of ASR transcripts is measured. Section 2.1.2 discusses SR, which aims to make speech searchable. Finally, Section 2.1.3 outlines how SR is typically evaluated and examines its role thus far in improving access to spoken content.

ASR takes discrete speech signals as its input and produces a transcript as its output by first breaking the speech signals into overlapping frames. Frames are typically 32 milliseconds long and typically overlap with a step size of 8 milliseconds. Each frame is then converted into a multidimensional continuous-valued feature vector, typically the frame’s perceptual linear predictive (PLP) coefficients (Hermansky, 1990) or Mel-Frequency Cepstral Coefficients (MFCC) (Huang, Acero, & Hon, 2001).
ASR uses an *acoustic model* to associate each feature vector with a probability distribution over all possible sub-word units (e.g., phones, subphones, triphones). It then refines these probabilities using a *language model* in the context of candidate labels from the surrounding frames. The acoustic model consists of a set of statistical models that represent either all the phones (a speech sound or gesture considered to be a single physical event), subphones (pieces of phones), or triphones (a context-sensitive phone that includes the phones that precede and follow the middle phone) in the dataset. Commonly used statistical models for this purpose include Gaussian Mixture models (Huang, Acero, & Hon, 2001) and artificial neural nets (Hinton, et al., 2012). Acoustic models are trained on datasets that may be as large as hundreds or thousands of hours of labeled speech data. The acoustic model produces a probability distribution over all phones or triphones that could have been uttered in the underlying speech. However, in isolation, an acoustic model cannot accurately predict these probabilities because of issues such as variability in speaking styles, variability in recording conditions, noise, and spoken artifacts (e.g., filled pauses, false starts, differences in breathing). To improve recognition accuracy, a language model is used. The language model estimates the probability of observing a word \( w_i \) with the conditional probability \( p(w_i|w_{i-1}, w_{i-2}, ..., w_{i-n+1}) \), i.e., by making the Markov assumption. The model is typically comprised of a set of n-gram (usually bigram or trigram) probabilities trained over large text corpora. Language models can also be trained on phones or triphones. This is required for assembling sequences of phones into words or for transcribing words outside of the ASR’s pronunciation dictionary (described in greater detail later). In this case, words in the training corpus must be converted to sequences of phones before training the language model.

Formally, acoustic models represent conditional probabilities \( p(o_i|l_i) \), describing the acoustic properties of each label \( l_i \). Language models represent the conditional probabilities \( p(l_i|l_{i-1}, l_{i-2}, ..., l_1) \) of observing a label \( l_i \) in the context of previously observed \( n - 1 \) labels. Bayes’ rule is then used to combine the acoustic and language model probabilities into the posterior probability \( p(l_i|o_i) \) of a label \( l_i \) given a speech segment \( o_i \), typically made up of a sequence of feature vectors:

\[
p(l_i|o_i) = \frac{p(o_i|l_i)p(l_i|l_{i-1}, l_{i-2}, ..., l_{i-n+1})}{p(o_i)}
\]
where \( n \) is the sequence length of the n-grams in the language model. The posterior probability represents the ASR’s confidence that \( l_i \) was spoken in \( o_i \). In practice, acoustic and language model probabilities are not equally weighted, and the posterior probability is represented in the log domain:

\[
\log P(r) = FLATw \left[ \frac{1}{LM_W} \log P_{AM}(r) + \log P_{LM}(Word(r)) + \frac{1}{LM_W} \log P_{IP} \right]
\]

where \( r \) is the label, \( \log P_{AM}(r) \) is the acoustic model score of \( r \), \( \log P_{LM}(Word(r)) \) is the language model score, \( LM_W \) is the language model weight, \( \log P_{IP} \) is an insertion penalty, and \( FLATw \) is a flattening weight. \( LM_W, FLATw \), and \( \log P_{IP} \) are typically estimated from training data (Chelba et al., 2007). The posterior probability of a sequence of labels representing a sequence of speech segments is easily calculated by adding individual label log posteriors, i.e., \( \log P(r) \).

Multiple label sequences can attain a high enough posterior probability to be of interest. These sequences are then organized into a lattice (see Figure 2.1). A lattice here refers to a directed acyclic graph where each arc represents a label (typically a phone, triphone, or word) along with a posterior probability, and each node represents a timestamp. Each path in the lattice represents a possible sequence of labels (typically words or phones) that could transcribe the speech data. The most probable path through the lattice is the 1-best transcript (often referred to simply as the transcript). The lattice is referred to as a word lattice if the labels on its arcs are words and a phone lattice if the labels are phones. Using a pronunciation dictionary, phone and word lattices can be converted into one another if necessary. Figure 2.1 shows the high-level workflow of a typical ASR system as described.

### 2.1.1. Measuring ASR transcript accuracy

The traditional method for measuring ASR performance is using word error rate (WER). This is the length-normalized Levenshtein distance, the weighted sum of the number of deletions, insertions, and substitutions needed to correct the ASR transcript. WER can vary from 60% for difficult genres of speech to less than 10% for recorded broadcast news stories. Although research in creating better language models has made a significant impact on ASR (Chelba, et al., 2013), most recent research in ASR has focused on developing better acoustic models with the aim of reducing WER. Examples include a WER reduction of 32% in transcribing
telephone conversations (Seide, Li, & Yu, 2011). ASR research has also aimed to make speech searchable, giving rise to the study of SR, which will be discussed in the following section.

Figure 2.1: Overall workflow of a typical Automatic Speech Recognition (ASR) system (Siegler, 1999).
2.1.2. Using ASR to gain access to spoken content: Speech retrieval

In the mid-nineties, it became clear to the research community that with the availability of cheaper storage and faster networks, spoken documents would become more widely used. Thus, the ASR and information retrieval communities set out to develop techniques that would replicate text search capabilities for spoken documents, assuming that information seekers would use spoken documents similarly to text documents. This research area came to be known as “speech retrieval”.

The assumption that users of spoken document corpora in domains such as broadcast news or recorded lectures used these documents like text documents could not be verified in the nineties because large spoken document corpora in those domains were not generally available to end users. However, voicemail messages were frequently used at that time. Studies that explored how voice messages were used in practice indicated that text information retrieval techniques, such as keyword searching, were not particularly helpful to voicemail users (Whittaker, et al., 2002). This result highlighted the variability of needed support for spoken document use. Hence, SR technology cannot be viewed as the ultimate solution for supporting spoken document use but rather as one family of potentially useful functionalities. SR research has principally focused on solving three problems: spoken document retrieval (SDR), spoken utterance retrieval (SUR), and spoken term detection (STD).

Spoken document retrieval: SDR is the identification of a subset of spoken documents in a corpus that are relevant to a user-provided query. This is similar to the classic text information retrieval problem which has received a fair amount of attention in the past two decades—i.e. SDR is to a corpus of spoken documents what the Google search engine is to text available on the Internet. SDR was extensively studied in Text REtrieval Conferences (TRECs) with the conclusion that transcribing spoken documents and then treating them as text documents in a text information retrieval framework would produce satisfactory results (Garofolo, Auzanne, & Voorhees, 1999). However, it was later pointed out that the documents in the TREC dataset were uncharacteristically easy for ASR systems to transcribe due to unrealistic (i.e., great) recording conditions, the availability of superb datasets for language model training, and the proficiency of the speakers producing the spoken documents (Allan, 2002).
**Spoken utterance retrieval:** SUR is the identification of the sections in a spoken document that are relevant to a user-provided query. Here, *utterance* refers to a sentence-like unit in speech. Utterance boundaries are found when the speech signal’s intensity is lower than a predetermined threshold for a given time span, usually 100 to 200 milliseconds. A spoken document can be viewed as a sequence of utterances. SUR is especially important because it is not possible to browse through speech in its native form with equal facility as browsing through text to identify relevant portions. What constitutes a relevant document (in SDR) or a relevant section of a document (in SUR) has not been precisely defined in the literature. Some experiments use relevance judgments from human-subjects (Garofolo, Auzanne, & Voorhees, 1999), whereas others have defined relevance by the presence of queried terms in a spoken document or utterance (Zhou, Yu, Chelba, & Seide, 2006; Saraclar & Sproat, 2004; Hori, Hetherington, Hazen, & Glass, 2007; Chelba, Silva, & Acero, 2007).

**Spoken term detection (STD):** STD is the identification of all the intervals in a set of spoken documents that contain an occurrence of a user-defined keyword. This problem was initially referred to in the nineties as Word Spotting or Keyword Spotting (Hofstetter & Rose, 1992) but later referred to as STD in the 2006 NIST Evaluation Plan (NIST, 2006) and a growing body of ensuing SR research. Unlike SDR and SUR, which are considered to be more user-centric, STD is aimed at improving retrieval in challenging situations, for instance, in noisy recording conditions. Phrases used to test STD systems are generally chosen not because of their potential utility to end users but rather for the different phonetic properties that they possess.

Although much effort has been devoted to developing SR techniques, with the exception of searching 1-best transcripts, these technologies have not been evaluated in field or user studies. But as will be discussed in Section 2.3, different information-seeking scenarios could require disparate support systems. Field studies that characterize different information seeking scenarios followed by carefully designed user studies can play an important role in determining the usefulness of different SR technologies in various contexts.
2.1.3. Measuring speech retrieval success

There are several ways to evaluate the performance of SDR systems. If SDR is viewed as a binary classification task of distinguishing between relevant and irrelevant documents to a query, then it can be evaluated by the F-measure (Manning & Schütze, 1999):

\[
F\text{-measure} = \frac{2(r \times p)}{r + p} \tag{2.1}
\]

where \( p \) is precision and \( r \) is recall. The former is calculated by dividing the number of correctly labelled relevant documents by the number of documents labeled as relevant. The latter is calculated by dividing the number of correctly labeled relevant documents by the number of documents genuinely relevant to the query in the search corpus.

However, SDR is typically more complicated than a binary classification problem. Much like text information retrieval systems, SDR systems can return a ranked list of documents sorted by their hypothesized relevance to the query. In this scenario, an irrelevant document (false positive) with a high hypothesized relevance can appear before relevant documents in the produced ranked list, affecting the SDR system’s performance more detrimentally than an irrelevant document with a lower hypothesized relevance. Mean average precision (MAP), a measure commonly used by TREC to benchmark SDR and text information retrieval performance accounts for this difference (Garofolo, et al., 1999). It is calculated as follows:

\[
MAP = \frac{1}{Q} \sum_{q \in \Gamma} AP_q \tag{2.2}
\]

where \( Q \) is the number of queries in the test set \( \Gamma \), and \( AP_q \) is the average precision for each query \( q \):

\[
AP_q = \frac{1}{D} \sum_{n=1}^{N} Prec_q(n) \times rel_q(n) \tag{2.3}
\]

In equation (2.3), \( N \) denotes the number of retrieved documents, \( D \) denotes the number of relevant documents to query \( q \), \( Prec_q(n) \) represents the precision of the first \( n \) documents in
the returned list, and \( r_{elq}(n) \) is an indicator function equaling one if document \( n \) is relevant to query \( q \) and zero otherwise. In essence, MAP measures precision at different recall points, from 0–100%. As can be seen in equations (2.2) and (2.3), MAP is sensitive to the position of false positives in the returned list.

As discussed, SUR is closely related to SDR (Zhou, Yu, Chelba, & Seide, 2006; Saraclar & Sproat, 2004). Researchers often utilize the F-measure to evaluate SUR systems because producing ranked lists is usually not required. Another benchmark is utterance spotting performance, similar to the figure of merit (FOM) metric defined by NIST for word-spotting evaluations. This metric measures the detection rate averaged over 0 to 10 false positives per hour of speech for every query (Zhou, Yu, Chelba, & Seide, 2006).

STD is another problem extensively studied in the SR community. To increase retrieval speed in a large spoken-document repository, STD systems first construct an index that represents the content of the spoken spoken-document repository and can later be searched swiftly at retrieval time. STD results are presented as a list of hypothesized hits along with a system-provided confidence score. STD performance is typically measured by FOM, actual term weighted value (ATWV), and upper bound term weighted value (UBTWV). ATWV, defined for the NIST STD evaluations (NIST, 2006)\(^1\), is calculated with the following formula:

\[
ATWV(\theta) = 1 - \text{average}_q \{P_{\text{miss}}(q, \theta) + \beta P_{FA}(q, \theta)\}
\]

where

\[
\beta = \frac{C}{V}(P_q^{-1} - 1)
\]

\[
P_{\text{miss}}(q, \theta) = 1 - \frac{N_{\text{correct}}(q, \theta)}{N_{\text{true}}(q)}
\]

\(^1\) Spoken term detection was introduced in NIST’s 2006 STD evaluation workshop. Systems participating in the evaluation were tested on a large, heterogeneous data set containing recorded meetings, telephone conversations, and broadcast news spoken in Arabic, Chinese, and English. Queries were unknown ahead of time and included 1,000 single and multi-word phrases from each language.
\[ P_{FA}(q, \theta) = 1 - \frac{N_{spurious}(q, \theta)}{N_{NT}(q)} \]

where \( \theta \) is the system-provided confidence threshold, \( N_{correct}(q, \theta) \) is the number of correct detections of \( q \), \( N_{true}(q) \) is the number of occurrences of \( q \) in the underlying speech, \( N_{spurious}(q, \theta) \) is the number of spurious detections of \( q \), \( N_{NT}(q) \) is the number of terms in the document that are not \( q \), \( C \) is the cost value ratio (set to 0.1), and \( P_q \) is the prior probability associated with the query \( q \) (set to \( 10^{-4} \)). \( N_{NT}(q) \) is made proportional to the duration of the spoken data in seconds, \( T_{speech} \):

\[ N_{NT}(q) = T_{speech} - N_{true}(q) \]

The maximum achievable ATWV is 100% while an ATWV of 0 is achieved by systems that output no results. Systems with high values of \( P_{miss} \) and \( P_{FA} \) attain negative ATWV scores. Note that for queries that do not occur frequently in the test set, \( P_{miss} \) and thus ATWV are sensitive to missed occurrences of the query. For instance, missing one occurrence of a query that appears in a corpus twice has an absolute reduction of 0.5 in \( P_{miss} \) for the query and \( \frac{0.5}{N_q} \) in ATWV, where \( N_q \) is the number of queries. On the other hand, ATWV is not overly sensitive to false alarms in large corpora in which the value of \( N_{NT} \) becomes dominated by \( T_{speech} \).

A system’s ATWV score depends on the tunable variable \( \theta \). Removing this dependence, UBTWV (Szoke, 2010) lets \( \theta \) assume different values \( \tilde{\theta}_q \) of each query to minimize \( P_{miss} + \beta P_{FA} \). As such, UBTWV is calculated as:

\[ UBTWV = 1 - \text{average}_q\{P_{miss}(q, \tilde{\theta}_q) + \beta P_{FA}(q, \tilde{\theta}_q)\} \]

The measures discussed so far in this section are intrinsic and thus do not directly assess the evaluated systems’ ability to support information seeking. To make such assessments, future research must appraise SR technologies in human subject experiments. Moreover, it is possible to conduct meta-evaluations that examine whether the described intrinsic measures are good indicators of an evaluated systems’ ability to support information seeking. Establishing such an empirical link enables future evaluators to treat the described intrinsic
measures as effective proxies of usefulness to information seeking. As will be discussed in Chapters 5 and 6, such links cannot be assumed.

Without observational evidence, it is difficult to identify information-seeking scenarios in which SR technologies could play a productive role. Outside of searching 1-best transcripts, observational evidence in Chapters 3 and 4 does not recognize SR as useful functionality. Instead, the evidence strongly points to ASR and SA as potentially useful technologies to support information-seeking. The next section provides a brief overview of SA research and describes the techniques and workflows that will be used in this dissertation’s SA experiments.

2.2. Sentiment Analysis

As discussed, the primary goal of this dissertation is to make spoken documents more accessible to information seekers. Studies have shown that knowing the sentiment expressed in voicemail messages or in meeting records is important to end users (Whittaker, Hirschberg, & Nakatani, 1998; Whittaker, Tucker, Swamplillai, & Laban, 2008). However,
the majority of SA research efforts using spoken (e.g., Neumann & Vu, 2017) or written sources (e.g., Rosenthal, Farra, & Nakov, 2017) do not directly aim to improve access to spoken content and rather take a TDI approach. Section 2.2.1 describes a widely used SA workflow which we will use in Chapter 5. Section 2.2.2 discusses how TDD development of SA is possible and justifies its pursuit.

### 2.2.1. Sentiment analysis workflow

SA is the process by which a document, spoken or written, is automatically classified into one of several possible categories, for instance, positive, negative, or neutral. Although it is common to categorize expressed sentiment into two or three classes, the number of categories can become quite large when differentiating between different emotions and psychological processes, e.g., 44 in Pennebaker, Chung, Ireland, Gonzales, & Booth (2007). To perform SA, each document is often represented using a high-dimensional feature vector with each dimension corresponding to a possible non-stop-word type in the system’s lexicon. Figure 2.2 provides a visual description of this representation. As is the case in Figure 2.2, the number of word types in most documents is smaller than the number of entries in the system’s lexicon. As such, the feature vector is often sparse.

The feature vector in Figure 2.2 uses the raw frequencies of word types in a document. These raw frequencies are referred to as term frequency or $tf$. As many experiments have confirmed, $tf$ is not the most effective strategy to represent a document’s content for SA (Pang, Lee, & Vaithyanathan, 2002; Pang & Lee, 2008; Lin & He, 2009; Mohammad, Kiritchenko, & Zhu, 2013) Instead, *term presence*, defined as $I(tf > 0)$, where $I(\ldots)$ is the indicator function, allows for a more accurate SA classification. In a comprehensive study of the effects of sub-linear scaling of $tf$ on SA accuracy, Paltoglou and Thelwall (2010) demonstrated that a BM25 weighting scheme allows for the best overall SA accuracy. The nonlinearity of BM25 is managed using two parameters $k_1$ and $b$:

$$BM25_i(d) = \frac{(k_1 + 1)tf_i}{tf_i + k_1(1 - b + b \frac{l(d)}{avg_{dl}})}$$

where $avg_{dl}$ is the average document length, $tf_i$ is the number of times word $i$ occurs in document $d$ (referred to as term frequency of $i$), and $l(d)$ is the length of document $d$. Here,
$b$ controls the document length normalization and $k_1$ controls the degree of sub-linearity of BM25 with respect to term frequency. As will be described in Chapter 5, $tf$, term presence, and BM25 weighting schemes are used in our SA experiments.

It is not necessary nor productive to include every word type in a document’s feature vector. Researchers instead commonly exclude the most frequently used word types in a language from feature vectors because it is believed that these words, especially in a language like English, are not content- or sentiment-bearing. As such, the presence of their features can only dampen sentiment. This technique is called *stoplist filtering* and is used for SA as well as for other NLP tasks (Androutsopoulos, Koutsias, Chandrinos, Paliouras, & Spyropoulos, 2000). One may also choose to filter words that do not often appear in a representative corpus, with the rationale that rare words are unlikely to accurately reflect the sentiment of the entire document (Pang, Lee, & Vaithyanathan, 2002). Focus may be placed on lists of words associated with certain emotions (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007; Hirst, Riabinin, Graham, & Boizot-Roche, 2014), although using all the words in the document, with or without stoplist filtering or rare word removal, produces better SA accuracy results (Hirst, Riabinin, Graham, & Boizot-Roche, 2014).

Once a document has been transformed into a feature vector, supervised machine learning techniques can be used for its classification. Much SA research has been dedicated to experimenting with different classification algorithms, such as Naive Bayes (Pang, Lee, & Vaithyanathan, 2002), Support Vector Machines (SVMs) (Pang, Lee, & Vaithyanathan, 2002; Mohammad, Kiritchenko, & Zhu, 2013), and Latent Dirichlet Allocation (Lin & He, 2009). Support vector machines consistently produced some of the best results in SA, including in the SEMEVAL 2013 Competition. Thus, the experiments reported in Chapter 5 use SVMs for sentiment classification.

Given a labelled training dataset containing documents which are represented by feature vectors associated with binary labels, an SVM creates a linear decision hyperplane separating vectors belonging to the two different classes (Cortes & Vapnik, 1995). To ensure this hyperplane properly generalizes to the test data, it is positioned equidistantly from the boundary training samples of each class. These boundary training samples are referred to as support vectors. As seen in Figure 2.3, the described positioning of the decision hyperplane ensures that the margins between the decision hyperplane and training vectors from either
class are maximized. The decision hyperplane can be represented in the Euclidean space with a vector \( \mathbf{w} \) and an offset \( b_0 \):

\[
\mathbf{w} \mathbf{x} + b_0 = 0
\]  

(2.5)

Constraining equation 2.5 such that the support vectors’ distances from the decision hyperplane is 1, the width of the margin to be maximized becomes equal to the projection of a support vector from the first class onto \( \frac{\mathbf{w}}{|\mathbf{w}|} \), summed with the projection of a support vector from the second class, again on \( \frac{\mathbf{w}}{|\mathbf{w}|} \), creating an overall margin of \( \frac{2}{|\mathbf{w}|} \). Maximizing \( \frac{2}{|\mathbf{w}|} \) can be achieved by minimizing \( \frac{1}{2} \mathbf{w}^2 \) with the previous constraints for the support vectors:

\[
\mathbf{w} \mathbf{x} + b_0 = 1.
\]

Differentiating the Lagrangian of this statement with respect to \( \mathbf{w} \) and setting it to 0 yields:

\[
\mathbf{w}_0 = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i
\]  

(2.6)

where \( \mathbf{x}_i \) is a support vector, \( \alpha_i \) is a Lagrange multiplier, \( y_i \in \{-1, 1\} \) is a class label, and \( \mathbf{w}_0 \) is the optimal decision hyperplane. Equation 2.6 can be used to further simplify the Lagrangian to:

\[
L(\lambda) = \lambda^T \mathbf{1} - \frac{1}{2} \lambda^T \mathbf{D} \lambda
\]

where \( \lambda = (\alpha_1, \alpha_2, ..., \alpha_l) \), \( \mathbf{1} \) is an \( l \)-dimensional unit vector, and \( \mathbf{D} \) is a symmetric \( l \)-by-\( l \) matrix with elements \( d_{ij} = \mathbf{x}_i \mathbf{x}_j y_i y_j \). \( L(\lambda) \) can be solved using quadratic programming.

If the training data are not linearly separable, it is possible to apply the kernel trick. That is, a kernel function can be used to project vectors represented by \( \mathbf{x}_i \) into a higher dimensional kernel space in which the data are linearly separable. A decision hyperplane can then be found in the kernel space using the method described above. If applying the kernel trick is not possible or desirable, it is possible to train the decision hyperplane using a soft margin constraint in which parameter \( \xi_i \) is introduced as a slack variable to represent misclassified vectors in the training set. In this case, the optimal decision hyperplane must simultaneously maximize the margin while minimizing the number of misclassified vectors:
\[ \frac{1}{2} \mathbf{w}^2 + CF \left( \sum_i \xi_i \right) \] (2.7)

where \( \xi_i \) is zero if \( \mathbf{x}_i \) is correctly classified and larger otherwise, \( C \) is a constant, an \( F(u) \) is a monotonic convex function. Setting \( F(u) = u^k \) and subjecting equation 2.7 to constraints \( \xi_i \geq 0 \) and \( \mathbf{w} \mathbf{x}_i + b_0 \geq 1 + \xi_i \) yields the following Lagrangian:

\[
W(\lambda) = \lambda^T \mathbf{1} - \frac{1}{2} \left[ \lambda^T \mathbf{D} \lambda + \frac{\delta^2}{C} \right]
\]

subject to constraints:

\[
\lambda^T \mathbf{Y} = 0, \delta \geq 0,
\]

\[
0 \leq \lambda \leq \delta \mathbf{1}
\]

with

\[
\mathbf{w}_0 = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i
\]

Figure 2.3: Optimal decision hyperplane (dotted line) maximizes the margins between training vectors from the two classes (Cortes and Vapnik, 1995)
Once again, the SVM’s parameters, namely $w_0$ and $b_0$, can be calculated using quadratic programming. Unlike other pattern recognition paradigms, including neural nets, whose parameters can only be locally optimized, SVMs’ parameters are globally optimized, making them attractive classifiers. Moreover, SVMs do not impose a limit on the number of dimensions a feature vector can have and so they can work natively with very high dimensional feature vectors such as $tf$, term presence, and BM25.

2.2.2. Using SA to support information seeking in the financial domain

The end goal of this dissertation is to support information seekers. As will be discussed in Chapters 3 and 4, information seekers in the financial domain need to predict the market reaction to content shared by an institution. These information seekers believe that market reaction is guided by the authorial sentiment expressed in the content, and moreover, that systems capable of extracting authorial sentiment can be useful in supporting their ISPs.

SA and more generally text classification research in the financial domain has often relied on ENT evaluation (Devitt & Ahmad, 2007; Ahmad, Cheng, & Almas, 2006; Drury & Almeida, 2011; O’Hare, et al., 2009; Koppel & Shtrimberg, 2004). But such extrinsic evaluation can be undependable if the labelled datasets used for evaluation are produced by annotators who are unfamiliar with the domain. In specialized domains such as financial reportage, this is due to the complexity and context-dependence of the discourse in which the same phrase, e.g., “this portfolio has an upside risk in high inflation environments”, can express both negative and positive sentiment depending on potentially unexpressed contextual factors. It is thus prudent to conduct extrinsic evaluation that is task-based or rely on annotation produced by trained human subjects. If the conducted research follows the TDD paradigm, as is the case here, task-based evaluations become central.

An example of a real-world task for TDD evaluation of SA is equity trading. Studies confirming the relationship between media and market performance date back to at least Niederhoffer (1971) who looked at headlines from the New York Times and determined that large market changes proved more likely following world events than on random days. Established theories of financial economics have also examined investor sentiment, indicating that negative investor sentiment leads to downward pressure on stock prices (De Long, Shleifer, Summers, & Waldmann, 1990). Note that investor sentiment here refers to
the general state of investors’ belief about a particular security, which differs from expressed or authorial sentiment that will be investigated in Chapter 5. Engle and Ng (1993) looked at the effects of news on volatility, showing that bad news introduces more volatility than good news.

More recently, Chan (2003) claims that prices are slow to reflect bad news while stocks that are covered widely by news media exhibit momentum. This study examines the effects of news on stock movements but does not actually perform precise quantitative analysis of the news documents’ content. Antweiler and Frank (2004) showed that there is a significant negative correlation between the number of messages on financial discussion boards about a stock and its returns, but that this trend is economically insignificant given plausible transaction costs. Using the Naive Bayes algorithm, they then classified the messages as containing buy, sell, or hold signals based on their authorial sentiments. Antweiler and Frank (2004) then showed that disagreement amongst posted messages in their dataset of Yahoo and Raging Bull messages was associated with increased trading volume. However, positive sentiment in messages did not have a statistically significant effect on stock movements. They conclude that this method is not directly applicable to the task of equity trading. In contrast, Chapter 5 shows that an SVM-based TDD SA system can indeed be directly used to trade stocks and generate annualized returns of up to 70.1%.

Tetlock (2007) examined media pessimism and concluded that high media pessimism predicts downward prices. In his experiments, Tetlock measured media pessimism from the content of the Wall Street Journal’s “Abreast of the Market.” He first converted text documents into 77-dimensional vectors where each dimension represented a word category from General Inquirer’s Harvard IV-4 dictionary. He then performed principal component analysis (PCA) to find the largest contributor to variation in the documents. The variation turned out to be tightly coupled with negative-sentiment word categories. In a vector autoregression study, this factor was shown to be a significant predictor of downward pressure on prices. Tetlock also developed a simple trading strategy that bought shares in the Dow Jones Index when negative sentiment words in the trading day’s document were in the bottom third of the last year’s negative word distributions and sold shares in the Dow Jones Index if negative sentiment words were in the top third of the last year’s distributions, achieving modest annualized returns of 7.3% (Tetlock, 2007).
Das and Chen (2007) analyzed discussion board posts to detect expressed sentiment. Using this analysis, they built a *sentiment index* that computed the time-varying sentiment of the 24 stocks of the Morgan Stanley High-Tech Index (MSH) and tracked how well their sentiment index followed the aggregate price of the MSH itself. Their sentiment analyzer was based on a voting algorithm that consolidated decisions made by five simple analytical classifiers. Their baseline, the Rainbow algorithm (McCallum, 1996), also came within 1 percentage point of their reported accuracy. This is one of the few studies that has evaluated SA itself (as opposed to a sentiment-based trading strategy) against market returns (versus gold-standard sentiment annotations). Das and Chen (2007) focused exclusively on discussion board messages and their evaluation was limited to stocks on the MSH. The study in Chapter 5 focuses on *Reuters* newswire and evaluates a wide range of NYSE-listed stocks and market capitalization levels.

Butler and Keselj (2009) attempt to determine sentiment from corporate annual reports using both character n-gram profiles and readability scores. They also developed a sentiment-based trading strategy with high returns but did not report how the strategy works nor how they computed the returns, making the results difficult to compare. Basing a trading strategy upon annual reports also calls into question the frequency with which such a trading strategy could be exercised.

Zhang and Skiena (2010) analyzed both financial blog posts and financial news to detect expressed sentiment, forming a market-neutral trading strategy whereby companies were ranked each day according to their reported sentiment. They calculated sentiment using the Lydia SA System (Bautin, Vijayarenu, & Skiena, 2008) which essentially calculates sentiment polarity as the normalized difference of the number of words that co-occur with known positive entities and the number of words that co-occur with known negative entities. Positive (negative) score of each entity is determined by the ratio of positive (negative) adjectives it co-occurs with. The strategy then goes long and short on an equally valued basket of positive- and negative-sentiment stocks, respectively. They conducted their trading evaluation between 2005 and 2009 and reported an annual return of ~30% when using news data, and an annual return of up to ~80% when using Twitter and blog data. Furthermore, they traded based on sentiment ranking rather than pure SA scoring. That is, instead of
trading based on the raw sentiment score of the document, they first ranked the documents and then traded based on this relative ranking.

Zhang and Skiena (2010) compare their strategy to two strategies, termed Worst-sentiment Strategy and Random-selection Strategy. The Worst-sentiment Strategy trades in an opposite manner to their strategy’s suggestions, going short on positive sentiment stocks and going long on negative sentiment stocks. The Random-selection Strategy randomly picks stocks to go long and short on. As trading strategies, these baselines set a very low standard. The evaluation in Chapter 5 compares our SA-based strategy to standard trading benchmarks, such as momentum trading and holding the S&P 500, as well as to oracle trading strategies over the same trading days (these baselines are defined in Chapter 5). Moreover, Zhang and Skiena (2010) do not optimize their SA model to the task. As it will be shown in Chapter 5, this is a missed opportunity.

As will be discussed in Chapters 3 and 4, analyzing spoken documents’ sentiment to predict market reaction to spoken content is one of several tasks performed by information seekers in the studied domain. Observational studies in Chapters 3 and 4, upon which this dissertation is grounded, are based on Marchionini’s theory of information seeking, which we turn to next.

### 2.3. Information Seeking Process

Sections 2.1 and 2.2 described ASR and SA in some detail and demonstrated that progress in both fields has focused on TDI development. TDD development starts with determining how and why spoken documents are used in the first place. This can be accomplished by characterizing users’ ISPs in order to derive a more specific role for potentially useful NLP technologies (e.g., ASR and SA). This section proceeds by reviewing ISP theory and subsequently states its consequences for designing SDSSs. Research with the objective of improving access to spoken content is then reviewed and analyzed from the information seeking theory perspective.

Marchionini (1995) defined what still remains the most prominent information-seeking paradigm for electronic search systems, describing the ISP as being composed of the following sub-processes:

1. Recognize and accept an information problem.
2. Define and understand the problem.

3. Choose a search system.

4. Formulate a query.

5. Execute the search.

6. Examine the results.

7. Extract information.

8. Reflect, iterate, and stop.

This process is dependent on highly interactive *information-seeking factors* described below:

- **The information seeker**: The *information seeker* is the person who initiates the ISP, defines the task in a domain, uses the search system, assesses the outcomes, and determines when the process is complete.

- **The task**: The *task* is the problem statement, typically articulated as a question. The task evolves as the search progresses due to the interactive nature of information seeking. The task does not necessarily need to be external, such as finishing assigned work. It can be internal to a seeker, e.g. altering a seeker’s understanding of a particular phenomenon.

- **The search system**: The *search system* contains information and provides tools and/or rules for users to access the information. It is possible to think of the search system as having a backend that contains information (a database) and an interface that communicates with the user. Examples of search systems include: people, books, libraries, and Internet search engines.

- **The domain**: The *domain* is a body of knowledge, such as chemistry, contemporary dance, or computer science. Domains can vary in complexity, clarity, size, age, and level of organization.

- **The setting**: The *setting* is the situation in which information seeking takes place. The setting describes the physical, conceptual, and social context of the ISP.
• **The outcomes:** The *outcomes* refer to artifacts that contain information (such as text or audio documents) and can change the state of the information seeker’s knowledge. The ISP or an iteration of it can also be considered an outcome as it changes the user’s knowledge. For instance, it can improve the information seeker’s understanding of a search system’s operations.

Marchionini (1995) understands the information seeker’s actions and motivations through the models presented by Dervin (1977), Belkin et al. (1980; 1982), and Kuhlthau (1988). Dervin’s model relates information seeking to the seeker’s need to make sense of the world. In her model, information seeking occurs across three stages (Dervin, 1977). First, the information seeker establishes the context for the information need, called a situation. The seeker then realizes that there is a gap in his or her understanding of the situation. The information seeker then addresses this gap and uses the newly gained knowledge to move to the next situation. Belkin and his colleagues introduce the anomalous states of knowledge (ASK) model in which the information seeker is concerned with solving a problem that is typically not clearly understood. In fact, the information needed to solve the problem is also not understood. The information seeker, aided by a search system, goes through an iterative process of trying to clarify the information need until the search system responds to the information need. Kuhlthau’s model (Kuhlthau, 1988) describes how students search for information for writing assignments. This model explains the affective states of the information seekers and describes the ISP with seven sub-processes:

1. task initiation,
2. topic selection,
3. prefocus exploration,
4. focus formulation,
5. information collection,
6. search closure, and
7. writing the assignment.
The ISP described above has consequences for the design and development process of search systems, including systems that improve access to spoken documents, i.e., SDSSs. First, it is important to note that in all the models discussed, information seeking is phrased as a problem-solving process dependent on a task and a search system. As such, the first step in analyzing and supporting information seeking is to learn about the task that initiates these processes. Moreover, these models are user-centred, placing emphasis on the importance of understanding the cognitive abilities, emotional states, attitudes, and motivations of a user who interacts with an SDSS. Such factors emphasize that supporting the ISP must be preceded by studies that help designers learn about their target user group. Other information-seeking factors, such as information domain, setting, and information-seeking outcomes must also be studied before the design and development of SDSSs. It is thus reasonable to conclude that better SDSSs could be produced if design is preceded by studies that investigate information-seeking factors.

These studies can broadly be divided into two categories: practice-based and functionality-elicitation-based (Lalanne & Popescu-Belis, 2012). In practice-based studies, researchers learn about the current information practices of their target users through surveys, questionnaires, or, more effectively, through observational studies. These studies aim to reveal challenges that are faced by end users and identify gaps in the functionality of the currently-existing tools. Practice-based studies of spoken document use are reviewed in Section 2.3.1. Functionality elicitation is made possible through a variety of techniques such as surveys, semi-structured interviews, and participatory design workshops (Schuler & Namioka, 1993). In a participatory design workshop, a subset of target users is asked to take part in designing the end product. Although this method can lead to more specific designs, it requires that users be guided in the design process, for instance, by introducing them to available technology that can be used in the designs.

Unfortunately, functionality-elicitation-based studies of SDSS design are rare. As reviewed by Lalanne and Popescu-Belis (2012), a number of survey and interview studies have been conducted in which users were explicitly asked about the type of information they would seek when interacting with a hypothetically “intelligent” system (e.g., Cremers et al., 2007), or about the formal queries that they would make of the system (e.g., Banerjee et al., 2005). Similar to practice-based studies, these studies established that users of meeting-record
browsers sought arguments for decisions, tasks to complete, summaries, meeting agendas, and the names of participants. Additionally, these studies revealed that users also aimed to understand the main topics of discussion and get a summary of the interactions between participants in the forms of points of disagreement and questions.

Noticeably lacking from these studies is the use of participatory design, which has been used effectively to help end users influence the design process by articulating what functionality the support system needs to implement, (e.g., Massimi, Baecker, & Wu, 2007), and how the support system should interact with end users and their existing tools to provide the required support (Vines, et al., 2012). The study reported in Chapter 4 uses this methodology in addition to the contextual inquiry from Chapter 3 to describe end users’ information practices (the what), and then provides specific examples of how NLP technology can be incorporated into the design of effective SDSSs.

Once new TDD SDSSs are designed, these systems can be evaluated in a task-based manner. As discussed, although most evaluations of technology related to spoken documents are TDI, there are some notable exceptions that are reviewed in Section 2.3.2. All evaluations reported in this dissertation (Chapters 5 and 6), except baselines, are task-based and domain-dependent.

The models reviewed in this section describe the ISP as inherently interactive. In these models, information needs are met through a series of communication acts between a user and the search system. Spoken documents come with their own interactivity challenges. These challenges, as well as the field’s solution to them, are discussed in Sections 2.3.3 and 2.3.4. Finally, Section 2.3.5 describes and motivates the experiments presented in the remainder of this dissertation.

2.3.1. Examining information-seeking factors affecting the use of spoken documents

Spoken documents are often records of some form of communication. Examples include voicemail, recorded lectures, and recorded meetings. This section reviews observational studies aimed at understanding users’ information practices when interacting with these spoken records.
Whittaker, Hirschberg, & Nakatani (1998) conducted a field study to identify tasks that are important to voicemail users. Using interviews and surveys of 148 high-volume users, the study found two main problems in accessing voicemail: scanning, that is, inter- and intra-message navigation to find relevant information (also referred to as relevance judgments), and fact-finding which is the discovery of specific facts in messages. Their data shows that, at least subjectively, voice messages contain a significant amount of information: half of their participants reported message durations of 30 to 60 seconds, while the other half reported durations of 1 to 2 minutes, containing as much information as a “whole memo” or a “huge email.” To accomplish their tasks, 72% of participants almost always took notes, either fully transcribing messages or at least noting down key information. Participants also preferred not to rely only on their notes and often kept the audio records for reference. They were notably concerned about losing intonational information from which they could extract the speaker’s expressed sentiment. Expressed sentiment in turn was used by participants to assess whether messages needed immediate attention. These findings guided the development of an SDSS called SCANMail that aimed to assist users with fact-finding, gisting, and relevance judgment tasks. SCANMail was shown to better support voicemail users than the then-current voicemail technology in follow-up task-based evaluations (Whittaker, et al., 2002).

Jaimes et al. (2004) studied how and why users review meeting records. They did so by conducting a large-scale Internet survey and interviews with 15 meeting participants. The Internet survey showed that video and photographic records of meetings were rarely used (91% of respondents never used video and 84% of respondents never used photographs), while distributed documents, memos to self, and minutes were most often relied upon. Although respondents did not often use video records of meetings, they believed such records could be useful for verifying what had been communicated, understanding missed content, re-examining content from a different perspective, retaining a complete record of the interactions for unforeseen purposes, and recalling content not captured in other records. Moreover, through an interview study, Jaimes et al. (2004) discovered that participants often remembered the meeting rooms, table layouts, names of other participants, the locations in which participants sat and major topics discussed in meetings better than, for instance, the meetings’ dates and times. Based on these findings, the authors designed a cue-based meeting retrieval system that used room layout, room selection, meeting participants, and discussion results to more effectively browse and retrieve video records of meetings.
Cremers et al. (2007) conducted an interview study (preceded by a questionnaire) with nine participants to gain insight into how team members communicated and accessed information when working on design projects. Their observations revealed that the use of audio-video records was limited at best when preparing for a meeting, during a meeting, and after a meeting, highlighting the inadequacy of existing support systems. While available state-of-the-art systems such as Ferret (Wellner, Flynn, & Guillemot, 2005) were deemed overwhelming by participants, the possibility of reviewing missed meetings provided by such meeting browsers was deemed useful. This study also determined that the automatic extraction of: (1) tasks to be completed, (2) ideas, (3) decisions, (4) arguments, and (5) summaries (minutes) are required features of an effective support system.

In a later field study, Whittaker et al. (2008) identified gisting (i.e. identifying the goals of meetings and main contributions, key points, decisions, actions, and overall atmosphere or sentiment) and fact-finding (i.e. extracting specific comments, opinions, and contributions) as the main tasks accomplished by participants when using written meeting records in corporate settings. After analyzing observations from seven meetings (12 hours in total duration) held by two teams in two different service firms (one in mail delivery and the other in software services), as well as 25 hours of interviews with the meeting participants over the course of 3 months, the authors noted that participants used (a) public notes or meeting minutes, and (b) personal notes to document events and discussions taking place during meetings. According to the study, meeting minutes lacked detail, as well as sensitive and peripheral information. Reading these minutes did not capture the experience of being in the meeting. Personal notes had disadvantages too, lacking accuracy and comprehensibility, often with an esoteric nature that made them difficult to use by those absent from the meeting. Whittaker et al. (2008) then examined whether a state-of-the-art SDSS, namely the Ferret Meeting Browser, which implements most features present in modern state-of-the-art meeting browsers, provides sufficient support for gisting and fact-finding. They concluded that, although Ferret did not provide the necessary support, Speech Excision technology (Nenkova & Passonneau, 2004) was effective in assisting users with the performance of gisting and fact-finding.

By exposing relevant information about information-seeking factors (i.e., task- and domain knowledge), the reviewed observational studies all led to the design of more effective SDSSs. Moreover, these studies highlight the important role of NLP technologies (e.g.,
speech excision) in SDSS development. Our observations in Chapters 3 and 4 confirm the importance of NLP technologies, including speech recognition, speech segmentation and alignment, SA, machine comprehension, and topic detection and tracking, in making spoken content more accessible in the financial domain.

Moreover, these studies—also supported by Bertini and Lalanne (2007)—confirm the widespread use and many limitations of traditional meeting records such as minutes and personal notes. They also point to the limited usefulness and lackluster adaption of raw audio-visual records while outlining how, with appropriate support, these records can become useful. Similar to these studies, the work reported in Chapter 3 also shows that users in the financial sector are not effectively supported in utilizing spoken records of critical events such as central-bank press conferences and public-company earnings calls. The work presented in Chapter 4 outlines how more effective support can be provided by a TDD SDSS design.

### 2.3.2. Task-based evaluations of SDSSs

Whittaker et al. (1999) conducted a task-based evaluation of SCAN, an SDSS designed to help users access a 47-hour Broadcast News (BN) spoken document corpus. It is worth examining all the interactivity features of this system (Figure 2.4) in some detail as they have become stable constituents of the SDSS designer’s toolbox—See Larson & Jones (2012) for examples. The interface has a search bar. Documents deemed relevant to the searched query are presented in a results pane as a list with each item represented by the spoken document name (e.g., National Public Radio’s *All Things Considered*), the date, the number of news reports, the relevance score, the document duration, and the number of query hits in the document. Clicking on each document highlights the document in the list and populates the overview and ASR transcript panes with the selected document. The overview pane contains a histogram. Each histogram column corresponds to one paratone (audio paragraph) in the document. The height of each histogram bar is proportional to the weighted frequency of the corresponding query-word in the paratone (a similar calculation to term frequency times inverse document frequency, TF-IDF, but also normalized by paratone length) and its width is proportional to the paratone’s relative duration. Bars for different query terms within a paratone are colour-coded and vertically stacked. Clicking on each bar begins playing the spoken document from the beginning of the corresponding paratone. The ASR transcript
pane, designed to support fact-finding, represents each paratone as a paragraph enclosed in quotation marks. Query terms are highlighted and colour-coded. Finally, random access is provided through a progress bar below the transcript pane (not shown in the figure) along with stop and pause buttons. The player cursor can be positioned anywhere in the spoken document by clicking on the appropriate part of the progress bar.

To evaluate SCAN, Whittaker et al. (1999) designed a randomized within-subject study with 12 participants. The SCAN interface was compared to a baseline UI which was missing the overview and ASR transcript panes. The results indicated that using the SCAN interface significantly improved solution quality, time-to-solution, and perceived task difficulty. Mean solution quality improved from 66.7% to 78.3%, mean time-to-solution was reduced to 414.7 seconds from 500.7 seconds, and mean perceived task difficulty moved from 2.77 to 3.5 in a 1-5 scale where one denoted hard. SCAN was particularly successful in supporting fact-finding and relevance-judgment tasks. For instance, solution quality for fact-finding improved from ~74% to ~97%. However, SCAN failed to assist users with gisting. According to the authors, inaccurate ASR transcripts were not trusted by users in accomplishing the gisting task. Participants felt that the transcripts did not get the “slant” of the story, though they were still good enough for fact-finding tasks. The authors recommended automatic summaries as a feature to help with gisting. They believed that using imperfect summaries would still be better than going through entire spoken documents for this task.

It can be helpful to establish a usefulness threshold on the accuracy of ASR. Munteanu et al. (2006) showed that transcripts with high WER (close to 45%) were still effective for simple fact-finding tasks in the lecture domain, when complemented with accessible PowerPoint slides. Munteanu et al. (2006) also showed that transcripts with a WER of 25% or lower are statistically indistinguishable from perfect transcripts for their task. However, two transcripts with the same WER of 25% can contain drastically different content. Additionally, given that WER is TDI, it may or may not be a good predictor of transcript usefulness for performing real-world tasks such as assisting users’ ISPs. Chapter 6 examines whether WER is a good indicator of transcript quality in experiments similar to Whittaker et al.’s (1999). As will be discussed, our results demonstrate that WER is not a good indicator of transcripts’ usefulness in assisting users’ task performance. We also introduce a more effective TDD alternative which is a better predictor of transcript quality than WER.
In another study, Whittaker et al. (2002) designed a system called SCANMail to support voicemail users in performing scanning, fact-finding, tracking the status of current messages, and managing voicemail archives. This was a follow-up study to (Whittaker, Hirschberg, & Nakatani, 1998) discussed in Section 2.3.1. As shown in Figure 2.5, the SCANMail UI contains a search bar and a scrollable list of voice messages in the mailbox, represented by their header information, which includes the caller name, phone number, subject, and voice message size. A thumbnail compactly represents the clicked document’s header and ASR transcript. Double-clicking any voice message populates the transcript pane with its contents. Information-extraction techniques are used to label names and phone numbers in the transcript. The transcript is also synchronized with the underlying audio, allowing users to navigate through the audio by clicking on different parts of the transcript. Tape-recorder-style play-back and pause buttons, as well as a progress bar, can also be used for navigation. The playback speed can be adjusted using a speed slider widget. Finally, users can archive voice messages into folders accessible from the interface or send them as emails with the audio file attached and the ASR transcript inserted in the body of the email. Results of each query are presented in a separate window, similar to Figure 2.5, but with a query box in place of the archive management tools. The results window is a scrollable list of thumbnails, each representing a relevant voice message. The query terms present in each voice message are highlighted using a colouring scheme in both the thumbnails and the transcript pane. Whittaker et al. (2002) tested SCANMail in a laboratory user study and later in a field study and demonstrated that it outperformed Audix, a state-of-the-art voicemail system, in supporting scanning and fact finding.

2.3.3. Interacting with spoken documents

One of the challenges unique to interacting with spoken documents is that, in their native form, spoken documents can only be listened to. This form does not enable a user to visually inspect the content or employ pre-existing navigation tools available for browsing text documents. It is thus common to present spoken documents visually.
Visual representations of spoken content can help users better accomplish relevance judgments, fact-finding, and gisting tasks (Section 2.3.2). As seen in Section 2.3.2, one way of representing spoken content is to use manual or ASR transcripts. Manual transcription of spoken documents can be laborious, time-consuming, expensive, and intrusive; thus automatic transcription of spoken documents is often considered. On their own, automatic transcripts can be relatively unreadable compared to text documents. As shown in Figure 2.6, utterances that are the building blocks of a spoken document are not necessarily cohesive units of meaning. Moreover, single units of thought can be broken across more than one utterance. Furthermore, unlike text documents, topic changes are not as easily detectable. In text documents, topic transitions typically coincide with paragraph transitions, although not

![Figure 2.4: The SCAN user interface (Whittaker et al., 1999)](image-url)
all paragraph transitions are necessarily topic transitions. Most automatically generated transcripts are also unpunctuated and contain incorrectly abbreviated expressions, making them more challenging to read. Finally, if a spoken document represents a dialogue, the tangled nature of undiarized utterances from speakers included in the conversation can further complicate the transcript, making it even more difficult to understand.

To make automatic transcripts more readable, one can segment the spoken document into a sequence of topically coherent sections. A large body of research in topic detection and tracking (TDT) has focused on spoken-document segmentation with annual evaluation conferences having run from 1998 to 2004 (Allan, et al., 1998; Allan, et al., 2005). Although Werff (2010) has shown that topic segmentation can improve SDR in broadcast news corpora, as measured by MAP, the merit of topic segmentation for transcript readability remains largely unexplored. While SDSSs have been built that segment ASR transcripts (Glass, et al., 2007), no task-based user studies have been conducted to examine the effects of topic segmentation on ASR transcript readability or the added utility of segmented transcripts in supporting users’ task performance. Glass et al. (2007) used a graph-partitioning algorithm that represents candidate segments with distributions of word hypotheses and chooses segment boundaries that maximize similarity within segments while minimizing similarity across segments. Characterizing the created segments as well as examining the qualitative and quantitative similarities and differences between segmentations produced by graph
partitioning, both the unsupervised methods outlined in Werff (2010), and the more established statistical methods from the Topic Detection and Tracking (TDT) community (e.g., Hearst, 1997, Reynar, 1999), are interesting directions of future research in this domain. Such research can enable SDSS designers to make more informed decisions when choosing between segmentation techniques.

Punctuating ASR transcripts can also help information seekers. With high enough accuracies, automatic punctuation of transcripts can only make them more readable and thus beneficial. F-measures as high as 93.19% can be achieved when punctuating ASR transcripts of broadcast news with WERs of 20% using a conditional random field (CRF) trained on lexical features (Lu & Ng, 2010). Unfortunately, automatic punctuation of other information-rich domains, such as academic lectures, has not been explored despite the likelihood that these harder-to-read transcripts could become far more readable when punctuated.

Information related to transcript accuracy is also important. Without this information, the reader is left unassisted in identifying potential errors in the transcript and must rely solely on his or her intuition. To provide automatic assistance, it is possible to rely on ASR confidence

![Figure 2.6: A sample automatic transcript of a meeting between four participants taking part in designing a remote control. The transcript was produced by AMIDA ASR engine with a WER of 38.9% (Renals et al., 2007).](Image)
scores. One visualization technique is to render recognized words in greyscale with their brightness set according to the inverse of their recognition confidence (Chelba, Hazen, & Saraclar, 2008). Unfortunately, this representation still leaves out information. For instance, it does not communicate alternative transcriptions that can be mined from the recognition lattice. Furthermore, depending on the visual composition of greyed and solid portions, readers can get the impression that some parts of the transcript have been highlighted to denote salience. This can confuse readers and make a transcript ultimately less comprehensible (Silvers & Kreiner, 1997). Another technique to distinguish low-confidence regions of ASR transcripts is to represent their spoken content as a phone sequence instead of a word sequence (Basson, et al., 2002; Coursand-Moreau, Crepy, & Destombes, 2000). But this representation has not been evaluated empirically and so its usefulness to information seekers is currently unknown.

Recognition errors can significantly affect the way information seekers use ASR transcripts. Presenting the information seeker with a fairly accurate but still imperfect ASR transcript may convince them to rely solely on the transcript for examining the content of a spoken document, leading them to overlook the occurrence of misrecognized words (Carmichael, et al., 2008; Whittaker, et al., 2002). This behavior negatively impacts the information seekers’ task performance if recall is important or if the words of interest are out of vocabulary (OOV). Content representation that prevents this behavior can generate a tangible improvement in spoken content accessibility. On the other hand, high-error-rate regions of ASR transcripts are difficult to comprehend for most users who are accustomed to written text devoid of recognition errors. Buried in recognition errors, however, there may be accurately transcribed words that can help. Filtering low-confidence words in hard-to-recognize portions a transcript can prove beneficial although the resulting representation is no longer a faithful word-for-word record of the underlying spoken content.

There are alternatives to visualizing spoken content entirely and serially using transcripts. Tables of contents (TOCs) and overview bars (as used by Whittaker et al. 1999) are two such alternatives. Dufour et al. (2011) have shown that TOCs can effectively assist users with gisting. Whittaker and Hirschberg (2003) have shown that overview bars can assist with relevance judgment and fact-finding tasks more effectively than transcripts with a mean
WER of 28%. Both TOCs and overview bars can be designed to draw attention to relevant portions of a spoken document. These representations are also more robust to recognition errors and spoken document quality. As will be discussed in Chapter 4, non-transcript representations of spoken content were used frequently in our participatory design workshops, supporting the suggestion that a word-for-word transcript may not be the best representation of spoken content in all scenarios.

2.3.4. Representing a collection of spoken documents

Typically, SDSSs present search results as a list of surrogates, each representing a document or a document fragment (Larson & Jones, 2012). Figure 2.7 shows how surrogate lists can be rendered. The list in Figure 2.7 is produced by Buchenwald SDSS in response to the query “bezetting” (Ordelman, Heeren, Huijbregts, Jong, & Hiemstra, 2009). Each surrogate contains a key frame from the spoken document, a portion of the ASR transcript relevant to the query (with query occurrences highlighted), and other metadata related to the document. The surrogates can be clicked to navigate to the underlying spoken document. Content representation in the surrogate must be biased towards the query. It is recommended that the relevance of the document to the query be made explicit, otherwise users may skip over the document (Tombros & Crestani, 1999). When query words appear in ASR transcripts, the relevance of a document or document fragment can be communicated by highlighting the occurrence of query terms in the surrogate (Alberti, et al., 2009; Ordelman, Heeren, Huijbregts, Jong, & Hiemstra, 2009; Whittaker, et al., 2002). However, query words may not appear in the ASR transcript if lattice-based retrieval methods, dimensionality reduction, or other information retrieval techniques, such as query expansion are employed (Manning, Raghavan, & Schütze, 2008). Determining methods of displaying a document’s relevance to a query in the absence of explicit occurrences of query terms is an important and unaddressed issue.

As discussed, it is possible to have surrogates in a results list that point to both documents and document fragments. In this case, multiple items in the list (surrogates representing the documents and their relevant fragments) may point to the same document. Such effects can be confusing to an end user. To avoid this problem, it is possible to display only one surrogate per spoken document by aggregating the document’s relevant fragments under one surrogate. The disadvantages of this method are that each surrogate takes up more screen
space, and the ranked order may be perturbed. The relevance score of documents and
document fragments can also confuse users. The Buchenwald search system shows only one
surrogate per document. Each surrogate shows the number of fragments that contain the
query but does not provide a way for users to directly examine these fragments in the results
page. Users must navigate to the document page to view a list of relevant fragments.

If the application of the SDSS is known, it is possible to present the results within a useful
context, making it easier for users to deduce the relevance of the returned documents. For
instance, a news retrieval system can present its results on a timeline that is also labeled with
major world events. Given that each document will likely reference events that have taken
place before or after the document’s publication, it is possible to anchor each document to its
publication date (with a larger anchor), as well as to dates of the events it references (with
smaller anchors). This is the approach taken by Alonso et al. (2010) who used crowdsourcing
to identify candidates as significant world events along with temporal expressions
representing centuries, decades, years, months, weekdays, dates, and relative expressions of
time (e.g., “last year”, “tomorrow”). Next, by finding consensus among annotators, authors
identified one or two significant events for each temporal expression and displayed them on
the timeline. Note that the significant events displayed on the timeline are static and do not
change with user queries. It may be interesting to conduct task-based evaluations to examine
whether varying the significant events displayed on the timelines with user queries would
improve the system’s ability to support information seekers.

News documents can also be presented on a map, as is the case in NewsStand (Teitler, et al.,
2008). The aim of such an approach is to allow users to make two types of queries: “Where
did story X happen?” and “What is happening in location Y?” To accommodate a large
number of news articles on the map, the authors employed an online clustering method to
group stories. Authors then presented the clusters either with markers indicating their news
type (e.g., sports, politics, business, etc.) or with their most salient keywords. By panning and
zooming, users are able to immediately understand the geographical sources of a news piece
or, alternatively, all news documents originating at a place of interest. Unfortunately, the
usefulness of this interface for supporting information seeking was not evaluated.
One property common to all SDSSs examined so far is that the presentation format of the search results is independent of the query. For instance, the result presented by these systems for the query, “which president was involved in the Lewinsky scandal?”, or, “what was the scandal involving Lewinsky and the president?”, will be the same because the words “everything”, “about”, “what”, “is”, and “the” are stopword-filtered. However, it is conceivable that users may prefer to see only the most relevant document about the Lewinsky scandal in response to “what is the Lewinsky scandal” and a more comprehensive list covering many subtopics in response to “everything about the Lewinsky scandal.”

Additionally, optimal query-result presentation can vary by task and domain. In examining meetings, for instance, Whittaker et al. (2008) found that the goal of information seekers is often to track the progress of a work item over several meetings—a process that is not directly supported by current SDSSs—while, with voicemail, information seekers are mostly
concerned with extracting phone numbers and actionable items. Customizing a search system’s response to query content, users’ task, and information domain has the potential to make SDSSs more useful.

2.3.5. Grounding experiments in ISP theory

Marchionini’s ISP theory (Marchionini, 1995) is the basis of the work presented in this dissertation. As has been discussed, our goal is to help users in the financial domain make use of spoken documents. The study begins by providing a positive example. As discussed in this chapter, it is prudent to begin such a process with a contextual inquiry to gain a proper understanding of information-seeking factors, including tasks performed by the target user group (Chapter 3). This knowledge, along with data collected from participatory design workshops, is employed to produce a novel prototype (Chapter 4). Collectively, the results of the contextual inquiry and the participatory design workshops identify speech processing and SA as two NLP technologies needed in SDSSs capable of supporting users in the financial domain.

SA is subsequently examined to determine whether currently available TDI techniques can viably play a role in the designs produced. As reviewed in Section 2.2, TDI development is central to current SA research. It will be shown that for predicting market reaction—measured by returns of SA-informed trading strategies of NYSE-traded equities—TDI SA systems did not perform significantly better than established trading baselines while TDD development achieves better results by tuning TDI models to market data. The optimization technique presented produces significantly better results than both the baselines and models trained end-to-end on market data (Kazemian, Zhao, & Penn, 2016). The task-based evaluations reported isolate specific feature selection dilemmas where TDI accuracy errs both in the magnitude and sign of the delta of SA usefulness.

Given the importance of ASR in building SDSSs and the prominence of TDI evaluation is the speech community as described in Section 2.1, WER is subsequently examined in order to determine if it is a good proxy of transcript quality in real-world scenarios (Chapter 6). In a comprehensive human-subject experiment based on the decision audit task involving spoken records of team meetings, we evaluate 4 ASR transcription conditions (provided by
several leading ASR research labs). It will be shown that, similar to SA accuracy, WER also
does not correlate with usefulness in the reported task-based evaluation (Favre, et al., 2013).

In some contexts, including the experiments in Chapter 6, TDD evaluation can amount to
conducting human-subject experiments that are significantly more time-consuming than TDI
evaluation. As an alternative, a semi-automatic TDD evaluation system, Auditor
Performance Predictor (APP), is introduced that uses machine learning to learn an evaluation
(Chapter 6). Our experiments will show that APP is a better indicator of transcript usefulness
for performing decision audits than WER (Favre, et al., 2013).
Chapter 3
Investigating Spoken Document Use by Investment Management Professionals

This chapter reports on the TDD development of an SDSS in the financial domain. TDD development is defined here as a two-step process: the systematic characterization of ISPs to be supported, i.e., gaining domain knowledge, followed by allowing the gained knowledge to lead the design and development process. In this chapter, the intended users are investment management professionals who make use of spoken documents in their professional activities. To characterize their ISPs, a contextual inquiry (Beyer & Holtzblatt, 1997) was conducted which led to the derivation of a taxonomy of information-seeking tasks (Section 3.4). The reported findings here made it possible to conduct a series of participatory design (PD) workshops in which the described taxonomy was validated (Chapter 4). PD workshops made it possible to explore how the information tasks uncovered here could effectively be supported with an SDSS prototype (Section 4.5).

Previous field studies have investigated the workflow and information practices of investment management professionals, producing taxonomies of the information transfer process from different parties (Ramnath, Rock, & Shane, 2008) or details of accounting practices (Bouwman, Frishkoff, & Frishkoff, 1995). Nevertheless, these do not capture investment management professionals’ information-seeking needs when using spoken records from earnings calls and news conferences. Moreover, these studies do not describe how investment management professionals interact with information systems to satisfy their information needs, nor do they explain how technologies such as NLP can be employed to better support their information practices. The contextual inquiry in this chapter precisely answers these questions.

As reviewed in Chapter 2, HCI and information-science researchers have conducted both practice-centric observational studies and functionality elicitation studies to better understand users’ information practices when using spoken records of meetings or voice messages. As discussed, these studies have led to the design of more effective SDSS systems. However, as demonstrated by numerous studies (e.g., Toms, Freund, Kopak, & Bartlett, 2003, Vakkari, 2003), due to the variability of ISPs, different types of support are required by information
seekers in disparate information-seeking scenarios. Thus, to build more effective SDSSs for investment management professionals, the findings of the practice-centric or functionality elicitation studies from other domains cannot be relied upon exclusively, motivating us to conduct the reported in-domain contextual inquiry.

3.1. Background

There are many stakeholders and agents that interact within the space of the financial market. Of these, investment management professionals play a prominent role. On a macro scale, investment management professionals can be responsible for the long-term strategies of institutions such as mutual funds, pension funds, sovereign wealth funds, etc. At the core of their activities lies information-seeking: being well informed on everything from understanding market trends as captured by external reports to developing their own predictive models based on thorough statistical analysis of large and varied sources of data.

Within the technological sphere, there has been some NLP research related to the financial industry but only to the extent of securities trading based on sources such as company quarterly reports, financial blog posts, and social media text (Bollen, Mao, & Zeng, 2011; Butler & Keselj, 2009; Zhang & Skiena, 2010), and even then without much reflection on the ISP of professionals engaged in this activity. Nevertheless, natural language documents such as these are highly valued by experienced analysts as these can yield nuanced insights not available in aggregated, numerical data. As one of our participants bluntly explained:

“The thing about having a job in the market is at all times you’re trying to not lose money and hopefully gain money. At any point when relevant information comes out, you need to know. For example, what Yellen\(^2\) said, everyone needs to know, if there is a loser who doesn’t know, he is going to lose money at the expense of his ignorance” (P1).

Our critical survey of major financial-analysis software (e.g., Bloomberg Terminal, FactSet) reveals, however, that while this software is ubiquitous, its use of tools that could amplify

\(^2\) Previous Chairperson of the Federal Reserve Bank
understanding or enable discovery within natural-language sources is extremely conservative. This is particularly noticeable against the backdrop of a general trend in the financial sector towards automation of information processes, and an abstract awareness that ever-expanding available datasets can facilitate more nuanced decision making (Flood, Jagadish, & Raschid, 2016).

Central banks such as the US Federal Reserve (FED) or the European Central Bank (ECB) play a prominent role in deciding the monetary policy of a jurisdiction (Bernanke & Kuttner, 2005). The leaders of central banks hold several press conferences a year to inform the public about their activities and provide guidance as to how they may act going forward. Similarly, publicly traded companies play a significant role in capital markets by providing investment and risk-mitigation opportunities to financial organizations. Public companies are required to hold regular earnings calls to update the public on their activities. To investment management professionals, such events are critical for risk-mitigation efforts and the transcripts of these calls or conferences are a valuable information resource.

Investment management professionals (referred to here also as financial analysts) make investment decisions on behalf of an organization (buy-side analysts) or provide investment advice to large financial institutions (sell-side analysts). The scale and complexity of decision making for these types of analysts sets them apart from retail analysts who provide investment advice to individuals and small businesses.

This chapter examines how buy-side and sell-side analysts make use of records from central banking news conferences and earnings calls (referred to here simply as “natural language documents” or “spoken records”) in their professional activities. The findings, summarized in the taxonomy presented in Section 3.4, motivate and inform the PD study that will be presented in Chapter 4.

3.2. Methodology and data

To understand how and why spoken documents are used by financial analysts, a solid understanding of their information-seeking factors is necessary (Marchionini, 1995). This study thus formulated research questions to examine these factors:

Q1. What social and physical context is the ISP situated in? (setting)
Q2. What tasks or real-world problems bring about information problems? (task)

Q3. What background knowledge and personal information infrastructure do the information seekers possess? (information seekers)

Q4. What knowledge domain is being explored during the ISP? (domain)

Q5. What search systems are already in use by the information seekers? (search systems)

Q6. What are these information problems about? (outcome)

To answer these research questions, we conducted a contextual inquiry to observe how financial analysts working on Wall Street utilized the described spoken records. Eleven experienced analysts (four women and seven men) who actively use spoken documents responded to our participation call distributed through the author’s professional network and word of mouth. All participants have more than 5 years of experience in their profession and were working on Wall Street (New York, USA) with hedge funds, asset management firms, central banks, and large multinational investment banks at the time of the study. The study was conducted at places the participants would typically interact with spoken documents: their firms’ offices, home offices, living rooms, and firms’ libraries. Participants were instructed to choose spoken records they would be interested in reading and that they would have eventually read as part of their professional activities. The chosen documents included transcripts of earnings calls of publicly traded companies and news conferences given by leaders of central banks.

The author observed the participants as they read the chosen documents and later conducted a semi-structured interview with participants designed to gain insights into the research questions. The interviews were recorded and transcribed. The study’s data consist of these transcripts as well as the observation notes.

Users were observed as they interacted with colleagues, the surrounding environment, tools, and reading material. Analysis of observational notes made by the researcher revealed that of the 11 participants, 7 read paper transcripts and indicated that they had always preferred to read printed transcripts, while 3 participants used a laptop and indicated no preference for
reading transcripts on an electronic device or on paper. One participant read the transcript on a tablet and remarked that for him, reading on a tablet (with a stylus), was the “same thing” as reading on paper with the aid of a pen. It must be noted here that, in practice, to access spoken content, our participants went directly to the transcripts, discarding the audio files altogether along with any assistive tools that could have been used to extract information from them, even when electronic devices were used. The electronic devices were used (similar to analog devices) to highlight text and make marginal notes. Moreover, when participants were asked in interviews about existing tools that could aid their information seeking, they only noted tools that helped them interact with transcripts, such as note-taking and browser-search functionalities. Because transcripts are the de facto representations of spoken content in this domain, the phrases “spoken documents” and “transcripts” are used interchangeably for the remainder of this chapter.

Analysis of observation notes also revealed that all participants were generally highly focused when reading transcripts. Of the seven who read printed transcripts, six used either a pen or a highlighter to underline parts of the transcript, to write notes, or both. Two of the three participants who used a laptop, and the participant using a tablet, were simultaneously examining related documents and reading. All but two participants skimmed through and skipped portions of the transcript that they deemed to be unimportant. These practices reflect the time pressure IMPs typically work under, a condition that will be discussed in more detail in the following section.

After observing participants’ interactions with the spoken documents, the author then conducted semi-structured interviews which were recorded and transcribed. A thematic analysis (Braun & Clarke, 2006) was then conducted to extract relevant information about our research questions from the produced dataset. In a thematic analysis, a theme captures an important aspect of the dataset related to the research questions. The study did not begin the analysis with pre-existing theories or assumptions concerning information-seeking factors. Instead, the aim was to begin from a blank slate and extract information about information-seeking practices of financial analysts. Thus, the analysis was inductive and, as will be discussed later, codes were produced from the dataset, similar to the coding approach in grounded theory (Glaser & Strauss, 1967). Moreover, the analysis was performed at the semantic level. The current study did not attempt to study the latent level by identifying
ideologies, assumptions, or conceptualizations that would affect participants’ responses because this form of analysis requires reliance on some form of constructionist theoretical framework. To the best of our knowledge, such a theoretical framework has not been developed for the population group that our participants belonged to. The thematic analysis presented here can thus be categorized as a semantic and essentialist qualitative analysis, which is the most common form of thematic analysis (Braun & Clarke, 2006).

We adapted Braun and Clarke’s (2006) six-step process for the presented analysis:

1. The researcher familiarizes him/herself with the dataset by reading all transcripts and listening to the audio recordings of interviews while taking notes on potential initial codes for the data.

2. Codes are produced from the notes taken in the previous stage and the entire dataset is coded.

3. The relations between all codes are examined. Groups of interrelated codes are linked to form themes.

4. Themes are reviewed to ensure codes exhibit internal homogeneity and external heterogeneity (Patton, 1990). The themes must not only explain associated coded data extracts but also reflect all of the meaning present in the dataset. At this point, the analysis produces a “thematic map” of the dataset.

5. Themes are defined and refined to get at their essence by exploring how each theme captures an aspect of the data related to the research questions.

6. Relations between all themes are considered for the data analysis to produce an overall “story” connected to the research questions, as presented in the following section.

The thematic analysis presented in Section 3.3 is supplemented with information obtained during observation, to provide the richest possible descriptions of the information-seeking factors that shape the financial analysts’ ISPs. The following section describes the main themes identified in the dataset for each research question; the notation T_i labels observations
supporting these themes. These labels map into the taxonomy that is described in Section 3.4 (Figure 3.3).

3.3. Analysis

3.3.1. Q1: Setting

Our participants (i.e. readers) all work for organizations that are market participants, i.e., entities that buy and/or sell assets in the investment markets, such as, investment banks, hedge funds, pension funds, and central banks. It is important to stress that when describing their professional duties, participants noted that they made decisions exclusively in a group, typically called a desk. Eight of the eleven participants explicitly mentioned that they typically update their team about what they have learned after reading transcripts. Our participants also noted that they often hold desk meetings to discuss investment ideas and action plans. Unfortunately, similar to other industries, financial analysts also rely on traditional meeting records such as personal notes, meeting minutes, and shared documents to recall information from meetings (Whittaker, Tucker, Swampillai, & Laban, 2008). These findings depict financial analysts’ ISPs and decision-making processes as highly collaborative, denoting the importance of supporting financial analysts’ collaboration and interactions. In Chapter 6, we examine the role of NLP in making the spoken content of analysts’ recorded interactions during meetings more accessible for future information-seeking activity.

The ISPs of participants take place in a fast-paced environment. Participants cited two main reasons for this. First, in investment markets, every fraction of a second can make a difference in the amount of money a transaction will bring in (or not). For instance, P1 discussed an important decision that had to be made in his team about investing in a security following a statement by the previous Chairperson of the U.S. Federal Reserve Bank (“the Fed”), Janet Yellen, concerning US inflation. The sooner a product was to be purchased, the cheaper it would be and the less P1’s firm would be exposed to currency risk. Second, at any point in time, there are numerous investment opportunities available to an analyst and his/her team. The faster a team can evaluate these opportunities, the more alternatives they can consider before arriving at a final decision. As mentioned by P4, the very fact that these analysts need to consider many alternatives makes it difficult or impossible to listen to earnings calls when they are live, an important reason for using transcripts in the first place.
The fast-paced nature of participants’ ISPs was also reflected in their reading of transcripts. Most skimmed and skipped sizeable portions. As one participant mentioned, a tool that would condense a transcript to its essential points would be of great value. Across interviews, participants often talked about the portions of transcripts that they found important, describing this information as: (a) consequential to the investment markets and (b) helpful to their learning. In other words, important information in the transcripts contributed to an information-seeking outcome related to investment markets. More detailed characteristics of such information are discussed in Section 3.3.6.

3.3.2. Q2: Tasks

Our participants made it unequivocally clear that their meta-goal is to make as much money for their institution as possible. This goal is entangled with users’ understanding of investment markets and the economy, an understanding which needs to be constantly updated and expanded upon. To achieve this, participants must develop insights about the future actions of the organizations they study (e.g., whether the Fed would raise rates) and be able to anticipate the outcomes of said actions (e.g., whether a company’s stock value would appreciate over the next year). Moreover, participants need to develop a solid understanding of the market’s expectations of these actions and outcomes. A reader has a far better chance of success if she/he has a more accurate grasp of an organization’s future actions and outcomes than other market participants.

To construct such insight, readers must stay up-to-date with the latest available information. This is why they make use of spoken records from earnings calls or central bank speeches as soon as they are made publicly available—it allows them to hear the latest messages from an entity as well as the market reaction to the rehearsed content. This task is formulated by
going through the content of these spoken records and extracting key takeaways that can be referenced or shared with colleagues:

“I put down 5–6 bullet points if there are recommendations on the investment that I have, I also put that down and that usually starts a discussion in our group” (P4).

As can be seen, the meta-goal of making money for the participants’ institutions initiates a hierarchy of tasks (Figure 3.1) that eventually require the use of spoken documents.

3.3.3. Q3: Information seeker

Responses to questions targeted at revealing the readers’ information infrastructure\(^3\) showed that readers do not and cannot begin without a proper background. Readers must have a good understanding of investment markets as well as the economy and should be familiar with issues that can have a tangible effect on the investments they are examining. Moreover, they need to have an adequate understanding of available information about the investment from a number of sources, such as SEC filings, investor presentations, previous earnings calls, previous research, and related news articles. Finally, they need to know about the “dialogue” between prominent financial analysts and market participants concerning the potential content of a spoken document before its release. For instance, as P5 noted, to understand the significance of statements in the ECB’s latest news conference,

“[you have to] know what the dialogue has been in the past several meetings [of the European Central Bank] to kind of know what to extract from these transcripts” (P5).

As will become evident, the Q&A section of a transcript provides an excellent snapshot of this ongoing “dialogue” \((T_{1w})\).

---

\(^3\) Marchionini (1995) refers to an information seeker’s understanding of the search system, knowledge about the information seeking domain, and cognitive abilities as his/her information infrastructure.
All responses describe the readers’ ISPs as accumulative. As P4 said, “the transcript is an incremental arrow of knowledge about the investment”. The thematic analysis revealed that the process is also iterative with two steps. In the first step, before reading the document, readers form expectations about the content of the document. In the second step, they begin reading, either challenging or confirming their expectations and initiating a form of learning that helps readers update expectations about the following document they will read. These expectations are also related to investment decisions. For instance, the ratio of a public company’s loss associated with a financial scandal over P4’s expectation of the loss would influence whether P4 would advocate for purchasing bonds from said company. Expressions of surprise are not uncommon when readers’ expectations are not met, typically prompting the readers to do more research before making an investment decision.

3.3.4. Q4: Domain

Participants are interested in examining all publicly available documents related to their investments. These can be divided into two broad categories: primary and secondary sources of information. Primary sources include information that comes directly from a company or institution (e.g., earnings call, regulatory filing, speech by the head of a central bank). Secondary sources include information released by other entities that have studied primary sources and subsequently make their analysis public (e.g., an equity analyst’s take on an earnings call, a news article on a central bank decision, or a sell-side economist’s understanding of a central banker’s statements during a press conference). While aware that such secondary sources are easier to access, participants noted several shortcomings including, but are not limited to, biases of certain analysts or journalists, inaccurate assessments of what was said, or omissions of important details. This demonstrates why primary material, including spoken sources, is so indispensable.

Users were adamant in stating that spoken records are only one of many sources that they regularly examine: “In isolation, it would be bullshit to say that I could tell you I drew some conclusion based on this [earnings call transcript] really, it’s not just this, but one of the things in here is that you can clearly tell what the strategy is” (P10). At the same time, it is noted that transcripts of earnings calls or news conferences are nonetheless important sources of information: “There is a lot of information that’s in the transcript which isn’t anywhere else,
a lot of information” (P10). Some participants identified types of information that they believed could only be found in transcripts or was best presented in transcripts ($T_{1b}$). These included qualitative and quantitative guidance about a company’s future performance, information about a company’s management style, speakers’ attitudes towards certain issues, and market perceptions evident from Q&A content. The presence of Q&A in the studied transcripts makes them a unique primary source. They not only have content and messages that institutions would like to communicate, but also include questions and answers which contain other market participants’ reactions to some of what has been communicated ($T_{2a}$, $T_{3c}$). Thus, these spoken documents allow users to hear not only what an entity has said, but also to encapsulate the markets’ reaction to a given statement. As will be discussed, both the “message” and the markets’ reaction are important to financial analysts’ task performance.

3.3.5. Q5: Search system

Participants named the following tools and information systems that they use for their professional activities:

- To get information: the Bloomberg terminal, Reuters Eikon, FactSet, Google Finance, Google searches, web searches.
- For reading support: Adobe Acrobat Reader, Microsoft Word, Microsoft Excel, Tablet, pen, paper, and highlighter.

The Bloomberg Terminal, Reuters Eikon, and FactSet seem to enjoy wide popularity in the industry. According to a recent market survey of commercial software products used in the investment management industry, the Bloomberg terminal and Reuters Eikon alone take up 62.3% of market share. The top 10 software solutions, which includes FactSet, hold 85.2% of market share (Deschamps & Strempel, 2015).

3.3.6. Q6: Information-seeking outcomes

The majority of discussions in the study’s dataset revolved around content that contributed to participants’ information-seeking outcomes. The analysis presented here refers to the organizations that hold the transcribed event as the organization (unless otherwise qualified). The representatives of the organization in the transcribed event who, in all cases, were the
leaders of the organizations are sometimes referred to as *speakers*. Finally, when unqualified, *outcomes* refer to the potential results of the organizations’ actions.

Table 3.1 summarizes the information our participants tried to extract when reading transcripts. We refer to them collectively as the essential predictive knowledge (EPK). To assess future actions and outcomes of an institution, as well as market reactions to them, readers looked for information related to the organization, the speakers, and external factors ($T_{lb}$).

Table 3.1: Sought-after knowledge for predicting future actions and outcomes

<table>
<thead>
<tr>
<th>External Factors</th>
<th>Speakers</th>
<th>Studied institutions</th>
<th>Past</th>
<th>Present</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market phenomena</td>
<td>Market reaction</td>
<td>Actions and outcomes</td>
<td>• Market phenomena</td>
<td>Analysts’ questions</td>
<td>• Market reaction</td>
</tr>
<tr>
<td>Previous remarks</td>
<td></td>
<td>Provided guidance</td>
<td>• Previous remarks</td>
<td>Speakers’ questions</td>
<td>• Speakers’ influence in institutions’ actions and outcomes</td>
</tr>
<tr>
<td>Professional history</td>
<td>Speake...</td>
<td>Actions and outcomes</td>
<td>• Professional history</td>
<td>Speakers’ cognitive and affective states</td>
<td>• Actions and outcomes</td>
</tr>
<tr>
<td>Previous remarks</td>
<td></td>
<td>Provided guidance</td>
<td></td>
<td></td>
<td>• Actions and outcomes</td>
</tr>
</tbody>
</table>

### 3.3.6.1. Actions and Outcomes

A significant portion of the discussions with all participants revolved around the organizations’ actions (e.g., “what [the] FED does [about the federal funds rate]”, P2) and their tangible outcomes in the capital markets (e.g., “the value of dollar would go up or down”, P4). Table 3.2 provides more examples. The goal of these discussions was to gain insight about *future* actions and outcomes. This in turn will allow readers to more effectively assess the profitability of an investment. For instance, after predicting the future actions of the studied organization (i.e., generation of free cash in 2016, Table 3.2 row 4), P6 is able to make the assessment that investing in the studied organization can be a profitable endeavor.

When predicting the future actions and outcomes of an organization, readers pay close attention to provided guidance, an organization’s official prediction about its future actions (e.g., plans for future acquisitions) and outcomes (e.g., the company’s future quarterly
profits). Changing a single word in a guiding statement would merit an unprompted explanation:

“Where she says ‘today’s modification of our (the Fed) guidance should not be interpreted as though we’ve decided on a date [for the interest rate hike] so just because we removed the word patience from our guidance doesn’t mean we’re going to be impatient. ’ That’s important information” (P1).

The importance of such guiding statements can also be inferred from the fact that ambiguity in statements is met with clarifying questions. For instance, when the Fed chairperson had originally used the term ‘patience’, and later ‘couple’, analysts at the news conference were not satisfied with the ambiguity of these terms and asked for clarifications:

“Because she had said patient means we’re not going to raise rates in the next couple of meetings and couple means two, she was explicitly asked what you mean by couple. This is how much they pay attention to what she says” (P9).

All participants reading FED transcripts in our study knew about this incident and brought it to the researcher’s attention to highlight the importance of guiding statements. Participants also discussed the importance of comparing the institutions’ actual actions and outcomes with guidance they had provided in previous conferences. This was certainly the case with P6 as he repeatedly stressed the importance of analyzing the divergence between guidance and actual actions and outcomes (T3b).

Although most discussions were about current and future actions and outcomes, past actions and outcomes were also considered in interviews. Often, past actions were discussed either to justify present outcomes: “Basically he’s trying to justify the rate cut, because he got a lot of negative attention for that” (P5); or to contextualize the discussion about future actions: “So they did not pay a dividend this year, an analyst brings it up and asks whether they’ll pay it next year” (P6). All participants compared current and anticipated outcomes with past outcomes at some point during the interviews (T3a). Some explicitly anchored their opinion about the future profitability of a company on a comparison of the company’s past and current outcomes. For instance:
“Company’s leverage was 39% in 2013 and closed 2014 at 48%. This is very important. Usually when they are doing annual calls they compare with previous years; those trends are very important... that’s not a good thing. These all reflect a company that had a bad year” (P4).

This assessment was part of P4’s justification for an investment decision about a company holding an earnings call. The importance placed on comparing current outcomes to past outcomes can also be deduced from the fact that virtually all quantifiable past outcomes of a company or an economy can be visualized graphically in a time axis-centric feature-rich environment in most software systems available to financial analysts. Figure 3.2 displays a graphic from the Bloomberg Terminal showing price, volume, and momentum indicators of the EUR-USD exchange rate in the past year.

3.3.6.2. Speakers

Due to the significant influence they exert over the actions of their organizations, part of the discussion about organizations’ actions revolved around the speakers themselves. The readers wanted to become more familiar with the speakers’ characters as to have a better understanding of how they will lead their organizations in the future. Discussions about the speakers revolved around their background, cognitive states, and affective states.

Participants talked about speakers’ professional backgrounds to better understand how they think. A speaker’s professional background, such as previously-held positions, was important when trying to predict the reactions of interested market participants\(^4\) to their communication or when assessing a speaker’s ability to influence their audience (P6). Participants also discussed speakers’ reputations concerning hawkish or dovish remarks:\(^5\) (T\(_{2b}\))

“If a guy is a hawk and is saying the economy is weak and we need expansionary economic policy, that sends a bigger message” (P8).

\(^4\) Other financial analysts, investors, and market observers.

\(^5\) A hawkish remark states confidence in the status of the economy and the stability of markets and often advocates for raising the federal funds rate and against expansionary economic policy. A dovish remark is the opposite.
As this statement demonstrates, the salience and meaning of statements are assessed in relation to speakers’ previous remarks.

Discussions about speakers’ cognitive states revolved around how they thought about important issues. Many issues of interest to our participants are inherently complex, making it possible to approach them from a variety of perspectives. Often, speakers make use of measures for a complex issue that are contingent upon a particular perspective. For instance, to understand a social media company’s user-engagement strategy, analysts on the earnings call wanted to know the measures the company uses to evaluate their strategy:

“He doesn’t have one metric for engagement, um, DAU (Daily Average Users) is one measure of engagement, um, DAU to monthly average users’ ratio ... [but] timeline [usage] is one that they do not look at” (P11).

Knowing what measures are used by speakers informs readers about how speakers think about an issue of interest, and in turn how they might act in the future. Sometimes, discussions about speakers’ views elucidate a particular goal and their expectations about it. For instance, P9 made the following assessment:

“I [want to] ... see how Draghi and his team are thinking about the recovery, ... is it lagging behind expectations? Because if it is lagging behind expectations, there is a chance that they would continue [the stimulus program] past 2016”.

Table 3.2: Examples of discussions of actions and outcomes

| “So they cut rates by 25 basis points and they thought it is going to be transformative for the economy in the sense that it really loosened financial conditions”.
| “They just closed an acquisition what is that going to have the business do in the next year or two”.
| “The fact that they are raising debt to buy back stock is always a good sign, that you’re giving cash back to investor”.
| “They are going to generate free cash for 2016 which makes me think this might be a catalyst for the stock to trade up”.
| “They moved from a regular software in a box to software as a service model which Wall street doesn’t like”.
| “Basically, he’s trying to justify the rate cut because he got a lot of negative attention for that”.
| “They have some highlights for exploration and production so the proven reserves are 16.4 billion barrels”.

Here, we see that a participant has inferred the speaker’s goal of bringing about economic recovery. Identifying a speaker’s expectations about an issue of interest (i.e. the goal of achieving economic recovery) helps the participant determine the next steps (actions). Such expectations are held about market phenomena, such as fluctuations in oil prices, and macroeconomic phenomena, such as low inflation. In all cases, speakers’ expectations, or more generally, views, provided participants with a glimpse into the cognitive processes that assessed an issue and planned future actions.

Finally, participants paid attention to how speakers expressed their feelings about issues of consequence. The following were judged to be important clues: enthusiasm, hesitation, comfort, and happiness. For instance:

“It makes me think that he’s comfortable with the level of financial stability issues ... so yeah, maybe he’ll cut again, he’s open to that” (P3).

Assessments of speakers’ affective states also allowed participants to gain insight about speakers’ future actions as leaders of their organizations. Currently, limited support is
provided by existing systems for performing this assessment. The designs produced in Chapter 4 address these limitations.

3.3.6.3. External Factors

Factors external to the speakers’ direct control, but consequential to an organization’s actions and outcomes, were discussed at length in the interviews. Such external factors can be phenomena related to specific organizations and legal entities. For instance, an active shareholder can influence a company to adopt cost-cutting initiatives and share buy-back programs (P7). External factors can also include market phenomena, such as, for instance, a widely held belief by market participants at large, also referred to as market consensus (e.g., “the CEO is perceived as overpaid” P7), or changes in the prices of market-traded securities (e.g., a decrease in the price of crude oil). The effect of cheaper oil prices on the actions of central banks and companies in the petrochemical industry were the subject of many discussions. The ability of a company to deal with external factors was also considered. For instance, P4 talked about the ability of a petrochemical company to turn over a profit given the low price of gasoline.

The organization’s activities can also be affected by the actions of external organizations. For instance, competition from other companies can affect a firm’s revenues: “Electronic medical records [companies] are in some ways eroding [their company’s sales]” (P7). Here, P7 notes that players in the medical records industry have segmented the company’s market, reducing its sales revenue. Market phenomena such as working capital cycle can also affect outcomes:

“They are losing money... December through January they are losing money and then in the next two quarter they make money in q2 and q3. So there’s working capital cycle for this business” (P6).

These external effects can have a large impact on the readers’ decision-making processes. For instance, before reading an earnings call transcript of a media company, P6 was

---

6 Working capital cycle describes cyclical and predictable factors such as seasonality that affect a company’s annual costs and revenues. To form a good understanding of a company’s long term profitability, these effects must be separately considered from non-cyclical factors, such as commodity prices and low inflation.
considering an investment in the company’s issued debt. But he decided to postpone the investment after reading about the company’s working capital cycle.

Market reaction to transcript content is yet another critically important external factor (T$_{3C}$). For this, participants pay close attention to the Q&A section. In fact, the Q&A section of transcripts contains the most valuable information according to participants. A significant portion of the discussions revolved around content from the Q&A: “Q&A gets you some of the most interesting information, but it’s always random because it depends on the questions” (P10). As P10 suggests, ultimately the quality of content is determined by the usefulness of questions asked during the event by analysts who are present. These analysts are also typically authors of research reports that exert significant influence in shaping the market reaction. Questions asked by these analysts strongly hint at what they will write about in their research reports:

“If the go-to analyst expressing skepticism about the guidance coming down... it doesn’t seem it was FX (Foreign Exchange) related; so why? It’s not so much about the answer but a question being asked. If you do not have a slam dunk answer for it then in tomorrow’s research report, you might have Morgan Stanley expressing skepticism about this. Yeah, so I usually learn more from Q&A, not necessarily long-term thesis, but how the market is going to react in short term” (P7).

As P6 stresses, there is a difference between the market reaction (largely shaped by research reports) and his thesis about future actions and outcomes of the examined company. Although participants (P1–2, P4–7, P10–11) pay attention to research reports and believe that they play a strong role in shaping short-term price movements in reaction to transcript content, they do not form theses based solely on them because “analysts can get it wrong” (P4). This highlights the importance of examining primary sources. According to P10, a successful information-seeking session includes analyzing not just new information but also assessing how analysts and the market at large are viewing the information and, if possible, extracting important data from the transcripts that are missed by the market:

“you... [note] interesting stuff from Q&A, and then also reading research... if you read three or four research reports, you very quickly see what people are
focusing on and... you might be, ah, I'm surprised they are not paying
attention to this, this seem like a bigger point” (P10).

These findings depict financial analysts’ ISPs as differential: at any point in time, participants try to maintain an understanding of what others think about an investment as well as the differences in opinion they hold. It is this differential process that places emphasis on gauging market reaction to transcript content, discussed extensively by participants during the interviews. The differential nature of participants’ ISPs also explains one of P10’s stated goals, namely to have a better “insight” into investment or economic issues than other analysts.

3.4. Taxonomy of financial analysts’ information practices

The current study’s analysis shows that none of what is communicated by speakers is viewed by participants (or by other sophisticated observers) as ground truth. Instead, content is interpreted through comparisons to previous communications from the examined organizations and speakers, as well as in the context of the organizations’ activities and market perceptions. Rather than taking what is communicated at face value, participants “read between the lines” in an attempt to pick up “hints” or “signals”:

“so Company A is taking over and doesn’t use them, you could read between the lines and say Company B is going away, if I was covering Company B that would inch me in that direction” (P6).

This widely held belief among participants was directly stated by P1 during the interviews: “the important stuff nobody says, you have to pick up hints.”

The speakers representing the institution understand this complexity. Their aim is to send carefully drafted messages to their audience; messages that may or may not be supported by all of the institutions’ actions and outcomes or related external factors (T2a). To effectively and believably send these messages, speakers play a communication game with their audience (i.e., the market), including the participants of the current study, who aim to satisfy their information needs.
To respond to information needs, participants performed several sub-tasks while using the transcripts which are summarized in the taxonomy depicted in Figure 3.3. First, participants *interpreted facts about the EPK (the what).* This starts with forming a solid understanding of the company’s past actions and outcomes, and included becoming familiar with the “dialogue” about the company. Next, readers take notes on disclosed information and reference related material not shared in the transcript. For instance, P6 directly took several pieces of information from a transcript for his analysis of a company’s attractiveness as an investment, namely the amount the company spent on acquisitions, and, the company’s plan for generating free cash in the year to come. Sometimes, to infer meaningful information, participants referenced related material that was not shared in the transcript. For instance, P1 had to interpret the consequences of Chairperson Yellen’s comments on FX markets against the backdrop of price fluctuations in inflation-linked securities. Or, P7 had to consider the rumors that TECH1 company was losing its number-one institutional client to be able to assess the true implications of the company’s decreased revenues. There are many reasons why a transcript may not contain all the relevant information needed for readers to satisfy their information needs. In addition to practical reasons, such as the predetermined duration of the transcribed events, the sought-after information may not be known to speakers at the time of the event. But, as participants noted, there may also be strategic reasons for not sharing information. For instance, as P4 put it, it could be “embracing” or as P10 described, not in sync with the message speakers are trying to send to the market.

In addition, users *assessed the communication acts in the transcript (the how).* Special attention was paid to the expressed sentiment (P1, P3–5, P7–10) which was described as “bullish” (or “bearish”) (P7, P3, P5), “un-bashing” (or “reserved”) (P4), “positive” (or “less positive”) (P4), “gung-ho” (or “defensive”) (P10), and “hawkish” (or “dovish”) (P1, P8-9). According to these participants, the expressed sentiment was not only a good clue about an organization’s future actions, but also influenced the short-term market reaction to communicated content. Participants also discussed a phenomenon they called *sentiment inflation* which describes the tendency of speakers to “find the silver lining in the clouds” (P4) and “paint the story better than it is” (P4) when discussing their company’s affairs. This appeared to be a widely-held belief by participants as they paid greater attention to negative statements (including negative statements made by analysts in the Q&A session) than
positive statements. This behavior is also in agreement with Engle and Ng’s (1993) observation that negative sentiment in the news has a larger impact on price movements in investment markets than positive sentiment.

Furthermore, participants made note of any communications tactics used by speakers (P1, P4–5, P7, P10–11). For instance, speakers may side-step questions or provide “evasive” responses (P4), for reasons discussed earlier. Other techniques include repeating certain key phrases or concerns to denote their importance to motivating future actions. For instance, according to P1, by repeating her observation about low market measures of inflation, Chairperson Yellen was trying to stress that a rise in market measures of inflation may be a precondition to the FED’s future rate hikes. Other communication strategies include providing more detailed information about a salient subject in the Q&A, for instance, by sharing detailed insights into streaming video market trends. According to P10, this technique was used to signal the company’s competence and readiness to monetize emerging opportunities in the new media market. Inferring useful information from communication tactics along with detecting expressed sentiment, in a similar manner to what has been described, was an important information-seeking subtask which is currently not well-supported by existing tool. Chapter 4 will describe ways in which assessing communication can be supported more effectively.

Finally, participants performed comparative analysis of gained knowledge to form expectations about future actions and outcomes of studied organizations. This analysis
included comparing mined information with past guidance, past actions and outcomes, and market consensus.

3.5. Summary

The presented thematic analysis discusses factors shaping financial analysts’ ISPs when using the described spoken records. The analysis characterizes these ISPs as accumulative and differential with the objective of gaining the EPK. A better grasp of the EPK can enable a financial analyst to predict future actions and outcomes of the studied organization more accurately than other market players can. To develop this edge, financial analysts perform the sub-tasks broken down in the presented taxonomy (Figure 3.3). To support financial analysts’ information practices when using spoken records of earnings calls and central bank news conferences, software tools must incorporate functionality that directly assists with the described sub-tasks. To validate this claim and investigate ways of effectively supporting these sub-tasks, we conducted a series of participatory design workshops, detailed in Chapter 4.
Chapter 4
Participatory Design of SDSSs for Investment Management Professionals

Chapter 3 analyzed factors shaping participants’ ISPs when using spoken records of earnings calls and central bank news conferences, resulting in the characterization of participants’ information needs (EPK in Table 3.1) and the development of an information-task taxonomy (Figure 3.3). The study then claimed that, to support financial analysts’ information practices when using spoken records of earnings calls and central bank news conferences, assistive tools must incorporate functionality that supports the sub-tasks described in the formulated taxonomy. This chapter reports on PD workshops we conducted to examine this claim by investigating the relationship between the designs produced by workshop participants and the proposed taxonomy.

The findings here also describe how different technologies can effectively be used to help information seekers in the financial domain. As will be discussed, an effective SDSS needs to incorporate a host of NLP technologies. These technologies include machine comprehension (Rajpurkar, Zhang, Lopyrev, & Liang, 2016), speech recognition and processing (Budnik, Besacier, Khodabakhsh, & Demiroglu, 2016; Goldman, 2011; Hinton, et al., 2012), and Topic Detection and Tracking (Allan, 2012). The results discussed in this chapter not only inspired the SDSS design depicted in Section 4.5, but also outline how TDD evaluation of the mentioned NLP technologies can be conducted.

4.1. Participants

The competitive nature of the financial markets and the considerable constraints on financial analysts’ time has made this user group largely unavailable to participation in user studies. We were fortunate in recruiting four financial analysts to participate in our PD workshops (three men, one woman; identified as D1–D4). Of the four analysts, D1 and D2 have also participated in the contextual inquiry discussed in Chapter 3. All participants had more than 6 years of experience in the industry. D3 and D4 are sell-side analysts for investment banks and equity firms, while the others are buy-side analysts for large global asset management
firms and hedge funds. Please see Table 4.1 for more information about the participants’ professional backgrounds.

Table 4.1: Participants' professional history in the participatory design (PD) study

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Buy-side/sell-side</th>
<th>Employers</th>
<th>Years of Experience</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Buy-side</td>
<td>Global asset management firm, one of the largest 10 by the size of assets under management.</td>
<td>5+</td>
<td>Portfolio Manager</td>
</tr>
<tr>
<td>D2</td>
<td>Buy-side</td>
<td>Emerging Markets Hedge Fund in NYC, managing more than $5 billion in assets.</td>
<td>5+</td>
<td>Execution Trader</td>
</tr>
<tr>
<td>D3</td>
<td>Sell-side</td>
<td>One of the Big 5 Canadian banks, Acquisitions and Divestiture Advisory Division, as well as a boutique institutional investment firm on Bay Street.</td>
<td>5+</td>
<td>Senior Associate</td>
</tr>
<tr>
<td>D4</td>
<td>Sell-side</td>
<td>One of the Big 5 Canadian banks, Institutional Investing Division. (Different firm than D3’s)</td>
<td>10+</td>
<td>Senior Associate</td>
</tr>
</tbody>
</table>

4.2. Methodology

Four workshops were attended by a facilitator, a participant, and a visual artist. The visual artist’s role was to assist participants with sketching their proposed ideas and help facilitate the visual conversation, mitigating participants’ potential lack of sketching or drawing expertise. In each design workshop, the participant began with a warm-up by reading a transcript as they normally would in their work routine. He/she were subsequently asked questions about how and why transcripts were read. The participant was then introduced to a scenario similar to what was described in Section 3.3.2 which involved examining the content of a transcript relevant to an investment decision that a hypothetical employer was considering. The participant then wrote five takeaways from the transcript, potentially including an investment recommendation, for the purpose of sharing it with her/his team members.
After presenting the scenario, the participant was provided with drawing tools, large sheets of paper, and a new transcript of their choosing and asked to design a prototype (relying on the help of the visual artist for drawing support, if needed) that would better support them in performing the task in the usage scenario than currently existing tools. The participants were told that the “sky’s the limit” for technologies they can incorporate into their designs: visualization, navigation, and AI including NLP. They were also asked to think about features they would like to see in the software they regularly use (e.g., the Bloomberg Terminal or FactSet).

Each workshop, lasting 1.5 to 2.5 hours, was video recorded and produced one design in the form of a sketch. The components of the designed systems (UI features labelled as $F_i$ in our analysis) were identified by examining the produced sketches alongside the sessions’ video recordings. Affinity diagramming (Kawakita, 1982) was then used to categorize the elements into themes present in all the designs.

4.3. Analysis

4.3.1. Content Themes Presented in Design Components

Functionally speaking, the components could be categorized into three groups, with many components providing functionality from multiple categories. As will be discussed, each of the functional categories in fact assisted the performance of one sub-task depicted in Chapter 3’s task taxonomy (Figure 3.3), supporting our hypothesis. The functionality categories included:

1. Elements showing useful shared and unshared information about EPK (“the what”)

2. Elements assessing communication acts (“the how”)

3. Tools for comparative analysis

4.3.2. Showing Useful Shared and Unshared Information about EPK

Some components in this category presented important qualitative data such as management outlook, past and current guidance, and the organization’s performed strategic and corporate actions, in a bullet list to make them easier to access (D1, D3).
Other features augmented disclosed information with additional data to enhance their interpretability (D1, D3). For instance, “cash flow overview” (Figure 4.1) visualized key aggregate performance figures, such as the company’s overall cash flow (D3) along with a breakdown of contributions from different business units. The visualization shows a graph of the historic and forecasted values of key aggregate figures (e.g., the company’s revenue) and their components (e.g., Google Websites, Google Network, Google Play, and Google Cloud), allowing the user to rapidly uncover causes of change:

“this quarter EBITDA\(^7\) went down, why was it? was it because your revenue went down... was management taking out some money um what was it...” (D3)

This feature also facilitated the rapid comparison of quantifiable outcomes across companies, which are often calculated differently within or between industries. Although much of the information presented in D3’s tools exists in products such as Bloomberg Terminal, it could not all be accessed simultaneously, forcing investment management professionals to collect the needed pieces of information into a spreadsheet for analysis.

Another component in this category provided additional detail about the company’s productions facilities, enabling the user to better interpret the consequences of production stoppages on the company’s future profitability (D1). The component visualized different production facilities on a zoomable map, annotating each facility with its production capacity as well as production costs per unit, and highlighting the facilities that were affected by a production stoppage (F1a). D1’s design allows users to rapidly assess the extent to which the company’s profits would be affected. Although information about production stoppages also exists in products such as Bloomberg Terminal, it is typically dispersed amongst multiple text documents. Extraction techniques are required to populate such visualizations, which are now becoming possible given machine comprehension’s success under similar scenarios (Wang, Yuan, & Trischler, 2017; Wang, Yang, Wei, Chang, & Zhou, 2017) (F1b).

\(^7\) Earnings before interest, taxes, depreciation, and amortization (Welch, 2009).
Yet another designed component, named “Sensitivity Analysis” (Figure 4.2), augmented the organization’s guidance about future outcomes (D1). Each predicted outcome (e.g., revenue in Figure 4.2), is based on assumptions made by the company (e.g. oil prices, exchange rates), which may not be reasonable from the reader’s perspective. To alleviate this, the “Sensitivity Analysis” tool extrapolates the provided guidance to a range of alternative values, enabling readers to inspect the sensitivity of guidance to the company’s key assumptions, as seen in Figure 4.2 (F2a).

Sensitivity analysis is currently performed manually by junior analysts on Wall Street (D1). To automate this, one needs to build a model of the company’s outcomes (e.g., revenues) as a function of one or more assumed variables (e.g. oil prices, exchange rates). Such models are currently built using spreadsheets. As participants in both studies have indicated, the information needed to build these models can be found verbatim in earnings call transcripts as well as the company’s filings. This is also true of the map widget discussed above. What is missing in current tools is the effective visualization designed by D, which requires the use of machine comprehension techniques (F2b) to be fully automated. All four of the participants stressed time pressure as the motivation for automation, but not accuracy of the resulting computations, nor recall rates of important information from source material.

4.3.3. Elements Assessing Communication

Visualizing sentiment in transcripts was envisioned as one of the tools for assessing communication (D1, D5). D1 designed widgets to visualize the sentiment score of the transcripts along with the distribution of sentiment scores from the company’s previous communications using a box chart (F3a). The widgets also facilitated the comparison of sentiment information across companies in the same industry (Figure 4.3). Moreover, the widget allowed users to track the evolution of sentiment over time using a popup line chart (F3a), allowing them to account for “sentiment inflation” in companies exhibiting the common habit of finding the “silver lining in the cloud” (D1). The investment management professional can use the widget to determine whether the studied company beats its competitors or peers in positive sentiment. Interestingly, D1’s sentiment visualization
Figure 4.1: Sketch of the "cashflow overview". By showing components that make up key aggregate figures such as revenue, and further showing how these components change over time, this visual can help users explain changes in key aggregate figures. For instance, the sizeable increase in Google revenue in the past quarter can be explained by the increase in Google Website revenue in the same period.
Figure 4.2: Components comparing the document’s sentiment with the company’s previous communication (using box chart and line graph) and with competitors’ communications.

Figure 4.3: Sensitivity analysis extrapolating the value of a company’s important outcomes (e.g., revenue) under different assumptions about key performance factors.
included two copies of the described widget, one representing sentiment in the prepared remarks and another for the Q&A content.

The central feature of D4’s prototype compared expressed sentiment across time on a hawkish-dovish scale ($F_{3a}$). Depending on the value of sentiment, the system would also recommend a set of trades to the end user. D4’s prototype contained four tables: a table containing hawkish terms in the transcript along with their frequencies, a table containing dovish terms along with their frequencies, as well as two similar tables representing frequencies in only the most recent transcript ($F_{3a}$). The component showing the change in sentiment on a scale would be used most often, with the term tables being used only when an explanation was needed about how the system arrived at the computed sentiment. This would work naturally for current sentiment analysis algorithms, which assign sentiment based on frequencies of sentiment-bearing words, e.g., Pennebaker, Chung, Ireland, Gonzales, & Booth (2007). Since investment management professionals often access sentiment to appraise short-term investment opportunities, the successful sentiment analysis technology used in automatic trading algorithms (e.g., Bollen, Mao, & Zeng, 2011; Zhang & Skiena, 2010; Kazemian, Zhao, & Penn, 2016) is an excellent source of the accurate sentiment scores needed to populate these widgets ($F_{3b}$).

The next set of tools described how communication tactics were used. These included a natural language text interface designed to help users locate queried phrases in the transcript (or other related documents) and to provide information about how often they were repeated (D2). These also included a side-panel tool that would show important topics discussed in the company’s past publications that were left out of the current transcript (D1). Topic Detection and Tracking technology (Allan, 2012), which has been successfully used in similar domains to segment written or transcribed text into cohesive, single-topic units, can be used to populate this side-panel tool.

Participants also designed tools to support mining information from the Q&A sections. D3’s design allowed for more rapid access to Q&A content by initially hiding the answers in order to quickly scan all the questions at once before choosing which answer(s) to view. D1’s designs aimed to characterize how speakers responded to questions, by measuring: the amount of time taken by speakers before formulating a response (“did the candidate... dilly-
daily a lot or was he very forthcoming... [with] answers”), the average length of answers relative to with the company’s peers (“what was the average length ... usually they give longer answers when they don't have an answer”), and the percentage of responses that resulted in the disclosure of specific facts or quantifiable information (“what percentage of the time was he BS-ing and what percentage of the time was he giving a clear direct answer”). Companies that have direct and quantifiable responses are viewed by the market as more certain investment opportunities (D1). The goal of these widgets, using similar visualizations to D1’s sentiment widgets (F4a), is to convert qualitatively expressed metadata about a speaker’s communication tactics into a quantitative score depicting the investment attractiveness of the company.

Although D3’s design does not require the use of NLP, D1’s three widgets do. For these widgets, using established tools such as speaker segmentation (Budnik, Besacier, Khodabakhsh, & Demiroglu, 2016) and speech alignment (Goldman, 2011), each transcript portion can be aligned with its underlying audio signal, and to also calculate average duration of responses. Disfluency detection (Liu, et al., 2006) can help find the time taken by disfluencies before a coherent response is produced. Thus, current NLP technology can be used to populate D1’s first two widgets (F4b). However, to the best of our knowledge, state-of-the-art NLP tools such as answer selection (Rao, He, & Lin, 2016) or fact extraction (Pasca M., Lin, Bigham, Lifchits, & Jain, 2006) have not yet been evaluated in scenarios similar to the third widget.

An important observation can be made about the tools designed so far. With the aid of visualization and NLP techniques, these tools extract information from examined transcripts, augment it with information from other sources, and present it to users visually. All users noted the time savings accrued in comparison to manually reading and producing similar visualizations with current software. No user noted that such tools may also help them because they are more accurate or methodical in their detection of implicit information such as sentiment or communication tactics into the analysis. Participants, when questioned, saw no particular advantage to cognitively offloading to a computer the interpretative or analytic
activity that followed upon the information gathering sub-tasks because: 1) they themselves were highly effective at doing it, whereas 2) a computer might make mistakes.

Although all users enlisted the aid of NLP to populate their visualizations, in one case, the use of NLP even here was doubted. D2 reluctantly considered automatic highlighting to mark a transcript’s salient parts, but indicated that this amounts to the system thinking on her behalf. There are areas of NLP such as summarization and information extraction that could indeed be used to highlight text, but this falls within the purview of interpretation, whereas parsing complex syntactic constructions in free-flowing text to identify objective quantities was considered more reliable. D2 remarked that she was only willing to use automated highlighting when extreme time pressure prevented her from reading the entire transcript.

Observations such as this suggest that investment management professionals do not embrace NLP when it removes their own decision-making agency.

This, together with the prior important observation, highlights a key theme running through all the features $F_i$ mentioned: our participants view such designs and the possible underlying NLP technology simply as time-saving tools, and not tools that may enhance discovery or interpretation. This suggests the need to preserve decision making agency when using software that provides assistance during information-seeking tasks – software that must be transparent in the use of the NLP tools.

4.3.4. Tools for Comparative Analysis

Most tools designed by participants compared actions and outcomes to those of the past, to those of the institutions’ peers, and to published projections. Although the described comparisons are not supported for natural language data in current analysis software, comparing curated, quantitative data to historic values or projections is well-supported in current products such as Bloomberg Terminal or FactSet.

Similarly, many components in the sketched prototypes included easy-to-access links to related research reports that complement users’ analysis of market perception and anticipate market reaction to transcript content. Again, links to research reports are available in software such as Bloomberg Terminal and FactSet, but are not integrated with tools for the qualitative analysis of transcripts.
4.3.5. The Involvement Spectrum

Our participants sketched complete prototypes, with multiple widgets, not minor modifications to current product interfaces. These sketches are placed along an involvement spectrum. At one end lie assistive tools for helping readers access the contents of transcripts. At the other, we find tools that process transcripts into an artifact with easier-to-digest augmented content. In one case, the produced artifact was extremely concise: a sentiment score along with a set of investment recommendations. Here, the participant (D4) completely outsourced the reading process but retained the ability to examine how the system produced the artifact. The next point on this spectrum illustrated a three-page report with much of the quantitative information in the transcripts structured within the context of past performance, current market expectations and peer performance (D3). This artifact also included qualitative data mined from the transcript (e.g., management outlook) but did not include any link to the underlying transcript, potentially revealing D3’s willingness to completely trust the system. D1’s designed system was a hybrid, providing both assistive tools for reading (if user had time), and for accessing a brief report if he did not. The report contained similar features to the D3’s artifact and additionally included sections that assess communication acts in spoken records. Finally, D2 did not produce an artifact, but instead provided assistive tools to help users produce their own artifact. In no case was the artifact able to replace or even suggest interpretative judgements as to buying and selling or, more abstractly, the quality of an investment.

At the extreme points of this spectrum, our participants’ designs delimit two types of support for utilizing spoken records: reading assistants and report producers. The participants who are buy-side analysts preferred the reading assistants when they had time and report producers when they did not, while the sell-side analysts were mostly interested in report producers.

4.4. Participatory Design Summary

The PD workshops revealed that the three categories of sub-tasks in Chapter 3’s taxonomy, namely, interpretation of EPK-related facts, assessment of communication acts, and comparative analysis with past actions, outcomes, guidance, and market expectations were all supported by the components designed in the discussed PD workshops. Moreover, it emerged
that current NLP technology including machine comprehension, sentiment analysis, speech alignment, and speaker segmentation already go a long way towards the technically supporting the (effective) functioning of our participants’ envisioned tools.

The presented analysis also highlights the importance of visualization for investment management professionals’ ISPs, and the need to ground design in observational studies that provide more effective tools than the visualizations embodied in current analysis software. The two broad types of support tools that emerge on our spectrum (reading assistants and report producers) are technically feasible. What dictates the choice between them is not entirely clear because of our sample size — possibly buy-side vs. sell-side orientation, possibly familiarity with AI-type technologies, possibly time allocation priorities.

4.5. A More Effective SDSS Design for Transcripts of Earnings Calls and News Conferences

Of the information practices performed by investment management professionals, several are not well supported by existing analysis software. Our studies suggest that investment management professionals need more and more detailed visualizations than what exists in current software. They also suggest that NLP technology is best received as an extraction mechanism to populate such visual aids, in such a way that preserves a sense of agency over decision making.

Figure 4.4 synthesizes our findings, as an extensive taxonomy capturing the high-level information practices, the software functionality that would serve the typical cases envisioned by analysts, the available NLP tools to support this, and the common UI elements in which these tools can be encapsulated (as drawn or described by the PD workshop participants). This figure also reveals that existing software does not sufficiently employ currently available NLP tools in a way that investment management professionals perceive as valuable.
Finally, this investigation has shown the importance of conducting user studies to assess the usefulness of technology for supporting ISPs. Blindly following a “deep-learning” crusade in pursuit of true system intelligence is unlikely to result in widespread adoption by investment management professionals, and yet there is a considerable gap between that vision and the current black-box appropriation of speech recognition and sentiment analysis. To some extent, this is a Catch-22. Current software does not incorporate advanced NLP, so investment management professionals are not aware of its potential, and thus reluctant to give up agency over even the minutest of their decision-making tasks. To adequately support investment management professionals, these tools need to be embedded in designs and

![Image: Figure 4.4: The proposed taxonomy, capturing high-level information practices and related software functionality, along with available natural language processing (NLP) technology and common UI elements that can implement the functionality to support information practices.](image-url)
visualizations that maintain agency and provide not only time value but superior recall and aggregation capabilities. These requirements have guided the design of an SDSS prototype that is illustrated in Figure 4.5.

The proposed prototype is a reading assistant primarily designed to support buy-side analysts’ ISPs. Preference was given to buy-side analysts because they are ultimately responsible for making large investment decisions and thus their information seeking has the additional complexity of adjudicating their beliefs about the future profitability of investments with those of sell-side analysts with whom they collaborate. Moreover, buy-side analysts are typically required to be knowledgeable about a significantly larger set of investments. Thus, the rapid examination of primary sources of information, such as spoken records of important events, is likely more urgently important to buy-side analysts’ overall information-seeking activities. The proposed design effectively combines the widgets designed by the buy-side analysts participating in our PD workshops.

The proposed prototype is designed with financial analysts’ typical workspace in mind. This typically consists of a desktop machine connected to six or eight screens running a number of multi-window applications in parallel. These applications include information systems that provide up-to-the-millisecond information about prices of market-traded securities, market news, or other spontaneous information, such as the Bloomberg Terminal and FactSet, analysis applications such as Microsoft Excel, and writing and communication applications such as Microsoft Outlook and Microsoft Word. It is also common to have a number of information-rich documents open, such as research reports, earnings call transcripts, and SEC filings. Similar to these applications, as can be seen in Figure 4.5, the prototype has multiple windows and multiple panes. The main window presents the transcript along with two panes. The (top right) static pane provides related information to help users interpret the transcript content. The pane presents selected news articles describing important issues facing the
Economic growth was relatively weak late last year and early this year. Some of the factors weighing on growth were expected. For example, exports have been soft, reflecting subdued foreign demand and the earlier appreciation of the dollar. Also, activity in the energy sector has obviously been hard hit by the steep drop in oil prices since mid-2014. But the slowdown in other parts of the economy was not expected. In particular, business investment outside of energy was particularly weak during the winter and appears to have remained so into the spring. In addition, growth in household spending slowed noticeably early in the year despite solid increases in household income as well as relatively high levels of consumer sentiment and wealth. Fortunately, the first-quarter slowdown in household spending was not as large as in household income as well as relatively high levels of consumer sentiment and wealth. Fortunately, the first-quarter slowdown in household spending appears to have been temporary; indicators for the second quarter have so far pointed to a sizable rebound. This recovery is a key factor supporting the Committee's expectation that overall economic activity will expand at a moderate pace over the next few years.

Despite lackluster economic growth, the job market continued to improve early in the year. During the first quarter, job gains averaged nearly 200,000 per month, just a bit slower than last year's pace. And the unemployment rate held near 5 percent even though notably more people were actively looking for work. However, more recently the pace of improvement in the labor market appears to have slowed markedly. Job g-

Figure 4.5: The proposed SDSS design. All windows and components can be closed and added back again from the drop-down menu.

cOMPANY as well as the most recent research reports on the company from top analysts (designed by D1). The second pane, referred to as the dynamic pane (designed by D2, lower right), contains two components. The first is a text box that allows users to input natural
language queries such as “show me all the times Janet Yellen mentioned the dollar in the past two years”. The system responds by opening a moveable, resizable, and minimize-able window in the second component of the dynamic pane. Figure 4.6(a) shows how the response to the example query is rendered. As can be seen, the resulting list contains sentences from different transcripts which can be opened in new tabs in the transcript pane (Figure 4.6(b)).

While this query helps a user assess the speaker’s language use by focusing on repetitions of
Sure. Yeah. Great question. It's Jason. So, yeah, there was definitely an increase in payables and I think that will start to unwind a little bit in Q4. A lot of that is natural. I feel like the production, I believe, it increased 37% quarter-over-quarter, so there's naturally going to be more parts coming into the factory. So I think some of that is just in the course of business.

And the other thing that I think is worth pointing out on the cash flow statement is receivables. We had a lot of deliveries right at the end of the quarter, so we weren't able to collect all of our receivables. We ended up with a fairly large receivable balance on cars that were delivered in that last 10 days or so.

Elon Reeve Musk - Tesla Motors, Inc.

Yeah, definitely also – yeah, with emphasizing that, I mean it's a first approximation you expect payables to increase by 37% if you – production reserve (7:46). And then you have to net out against receivables. And when you do that, I think it's not really a – it's not a material situation.

Figure 4.7: Interacting with tagged production figures in the transcript pane.

(a) Clicking on tagged production figures opens the production popup.

(b) Time series can be selected from the options in the popup.

(c) Production facilities map can be selected from the options in the popup.

a particular phrase, the presented design also requires the system to handle other queries that may come up when users perform sub-tasks depicted in the taxonomy in Figure 4.4. These
queries include, for instance, “show me the price of 10-year inflation break-evens from 2 months before this transcript to 2 months after”, which would provide the market information necessary to interpret the meaning of statements made by Chairperson Yellen, and for instance, “show me a curve of US inflation since the start of the QE\(^8\) program”, which would help a user explore potential effects that the ECB’s QE program could have on inflation through comparative analysis.

The other two windows in the interface (Figure 4.5, bottom) are dedicated to supporting users’ assessment of speakers’ language and communication use. The first window renders the following:

a) A bulleted list of important topics discussed in the institutions’ past publications (in the past 2 to 3 years) that were not addressed in the current transcript, showing the reader important topics that the speaker might be skipping (discussed in Section 4.3.2).

b) A bulleted list of new topics in the presented transcript that are absent from the institution’s previous publications (again designed by D1).

c) Two widgets used to visualize expressed sentiment in the prepared remarks. The two widgets, visualized sentiment in both the audio signal and the transcript, respectively, using the interface techniques illustrated in Figure 4.3.

The second window describes how language was used, specifically in the Q&A section, using the following two sets of widgets:

a) The first set visualized sentiment in the same manner as prepared remarks in the previous window.

b) The second set described how questions were answered by the time taken to produce responses, the duration of responses, and the ratio of questions that resulted in a

---

\(^8\) Quantitative easing refers to non-traditional ways of stabilizing the financial systems, for instance, through large asset purchase programs, in the hundreds of billions of dollars range (the Federal Reserve Bank of St. Louis, 2011).
factual and direct answer (see Section 4.3.2 for more detail). Similar to expressed sentiment, these parameters are visualized using the interface techniques illustrated in Figure 4.3.

As discussed in Section 4.3.2, it is important to note that the described widgets in the bottom two windows of the prototype, can be populated by the data-feeders employing current state-of-the-art NLP technology.

The remaining components in the prototype are enclosed in popups that become visible if a user clicks on a pre-determined set of phrases in each transcript. These include quantifiable EPK, such as oil production, total sales, total revenues, or the names of market-traded goods and securities. Clicking on a company's production numbers in the transcript text displays a popup that can show one of two graphics:

a) A time series bar chart showing how the figure has evolved over time, including consensus forecasts of its future value\(^9\). Such bar charts are ubiquitously available on a variety of platforms, such as the Bloomberg Terminal, FactSet, or StatistaCharts (see Figure 4.8), although we have taken the further step of directly integrating them into the spoken content presentation.

b) The map visualization discussed in Section 4.3.1 which helps users more effectively interpret changes to the production plans of the organizations under study (e.g., production stoppages), for instance, by taking geopolitical factors into consideration.

Figure 4.7 shows how each of these graphics can be selected for display in the popup’s tab view.

Clicking on other quantifiable outcomes (i.e., non-production figures) or security names (e.g., crude oil) brings up a popup that shows a time-series bar chart similar to Figure 4.8. The data needed to populate these charts can be pulled from a number of commercial data

\(^9\) Typically derived by averaging different analysts’ forecasts.
sources\textsuperscript{10} by providing a ticker symbol\textsuperscript{11}. Note, however, that identifying the correct ticker symbol may require processing contextual factors. For instance, if the word “dollar” is mentioned in a FED news conference transcript, it most likely refers to the trade-weighted US Dollar Index\textsuperscript{12}, while if it appears in an earnings-call transcript of a German automotive company, it refers to the USD/EURO exchange rate. Moreover, clicking on forecasted values brings up a popup aimed at illustrating the institution’s credibility in forecasting. In a tabbed view (see Figure 4.7), this popup allows users to either inspect the sensitivity analysis matrix associated with the forecast (see Section 4.3.1), or a graph that compares the institution’s past forecasts with actual figures (Figure 4.9).

Finally, the names of the questioners in the Q&A section can be clicked to bring up yet another type of popup (Figure 4.10). Given that the questioners are often sell-side analysts who publish investment recommendations about the company, the popup shows a graph of the company’s stock value, annotated with the analyst’s investment recommendations about that stock on the time axis. The aim of this graph is to visualize the potential influence of the analyst in moving the company’s stock price. The popup also displays the name of the institution (typically an investment bank) that the analyst belongs to, the average stock price movement after the analyst’s positive and negative recommendations, and finally, the analyst’s standing in the Institutional Investor\textsuperscript{13}. As can be noted, the proposed prototype makes heavy use of popups to augment transcript content. This choice strictly reflects participant designers’ strong preference for this form of information display.

So far, the discussed interactive features assist users with interpreting the transcript content by providing relevant contextual information. Much of the supplementary information was

\textsuperscript{10} Bloomberg, FactSet, Reuters, and Capital IQ all offer such data sources that can be queries on the web programmatically.

\textsuperscript{11} A unique abbreviation used to identify a traded security in a public market. Examples include “GOOG” for Alphabet Inc. and “CAD/USD” for the CAD to USD exchange rate.

\textsuperscript{12} The trade-weighted US dollar index measures the value of the US dollar against a collection of reputable world currencies.

\textsuperscript{13} Available at \url{http://www.institutionalinvestor.com/}, this website ranks sell-side analysts based on their reputation and influence in the capital markets.
presented in popups, a choice that reflected our participants’ preference. On the other hand, the final proposed interactive feature supports users in assessing language use. This feature is a natural language search box that can be accessed from the context menu once a portion of the transcript text is highlighted. This context menu search box, mimics the functionality of the search box in the Dynamic Pane (see Figures 4.5 and 4.6) with the exception that the user query can reference the highlighted text.

As can be seen, the proposed design supports the performance of all information sub-tasks depicted in the taxonomy of Figure 4.4. The proposed design integrates all of the components of P1’s and P2’s (the two buy-side analysts) prototypes in the PD workshops. This integration seemed appropriate as the participants in the PD workshops consistently preferred feature-rich multi-component systems that could be customized, allowing the user to simultaneously view as many, or, as few of the available components as desired. This preference is also reflected in the software they routinely interact with. Bloomberg Terminal, Reuters’ Eikon, and FactSet are all feature-rich, multi-component, and highly customizable. Given that all the components in the proposed design can be easily closed and added back from the menu bar, it is argued that the proposed SDSS design is as effective to the participants as the prototypes they designed. Given that the participants deemed their prototypes more effective than currently available tools, we then argue that our design is also more effective than the currently available tools.

4.6. Building on Observational Evidence

The contextual inquiry of Chapter 3 and the PD workshops of the current chapter have led to the following discoveries:

- Sentiment analysis is a useful technology for supporting financial analysts’ ISPs that involve spoken records of earnings calls and news conferences. More specifically:
  - Expressed sentiment was used to gauge short-term market reaction to the content. This is in contrast to how it was used in the voicemail domain, i.e., to detect a voice message’s urgency, and the meeting domain, i.e., to sense a meeting’s overall atmosphere.
As can be inferred from Figure 4.3, sentiment analysis is a regression problem, allowing for the quantitative treatment of an otherwise intuitive process.

- Speech processing plays a nuanced role in supporting financial analysts’ ISPs:
  
  o To make the content of earnings call and news conference records more accessible, speech technologies such as speech alignment, disfluency detection, and speech segmentation have a greater impact than speech recognition.
  
  o The collaborative nature of financial analysts’ ISPs and decision-making processes is worth highlighting. These processes include attending many internal investment meetings. Similar to other industries, information seekers currently rely on traditional recording means, such as personal notes, meeting minutes, and shared documents. Speech recognition could have a transformative impact if it could help information seekers make use of multimedia records of investment meetings to review discussed investment decisions, i.e., perform decision audits.

Based on these findings, we focus on sentiment analysis (Chapter 5) and speech recognition (Chapter 6) to examine how they can be developed to make the spoken records of earnings calls, news conferences, and investment meetings more accessible to information seekers. Specifically:

  a) Sentiment analysis is viewed as a regression problem with the objective of detecting market reaction to content. SA systems are embedded in simple trading algorithms and evaluated and optimized using generated returns.

  b) Speech recognition experiments are focused on the meeting domain, and evaluate systems by their ability to support the decision audit task.
Figure 4.8: A bar chart showing historic, present, and forecasted values of one of Microsoft’s production figures: annualized commercial cloud revenue run rate.

Figure 4.9: Line chart comparing a company’s actual vs projected total revenue.
Figure 4.10: Analyst profile. This includes analyst’s employer, rating on analyst rating websites, average percent change in stock value after publishing positive or negative reports, and a time chart showing the stock’s movement annotated with times the analyst has published positive or negative reports about the company.
Chapter 5
Utilizing Task and Domain Knowledge in Sentiment Analysis Development and Evaluation

Financial analysts review spoken records of important events, such as central bank news conferences and earnings call transcripts in part to predict market reactions to the studied spoken content. As discussed in Chapter 4, the financial analysts participating in the PD workshops suggested that the SDSSs include components that detect the expressed sentiment of the spoken content. This inclusion was motivated by the widely held belief amongst investment management professionals, that the sentiment expressed in institutions’ communications guides short-term market reaction. Thus, in this chapter, we investigate models that can detect expressed sentiment and evaluate these models, in a TDD manner, by their ability to generate returns when integrated into a simple trading algorithm. This trading algorithm buys or sells stocks in accordance with the evaluated SA models’ output. A distinguishing characteristic of the task-based evaluations described in this chapter is the use of competitive trading baselines. It is not uncommon that published evaluations of SA-based trading algorithms (reviewed in Section 2.2.2) will not include a comparison with baselines (e.g., Tetlock, 2007), or include only a comparison with weak baselines (e.g., Zhang & Skiena, 2010).

The belief that expressed sentiment is a good indicator of market reaction and that it is a useful signal to include in trading algorithms is widely spread amongst market professionals.14 It is not uncommon to use SA in conjunction with other strategies (e.g., Conomos, Platt, & Hamilton, 2010), or as the primary determinant of trading algorithms (e.g., Tetlock, 2007). It must also be noted that not all the innovation in this area has been published. Investment management professionals, primarily interested in generating revenue for their employers, are not always motivated to publish what they know of SA-based trading. The work described here, builds on what has been published and what we have learned through our own interactions with numerous investment management professionals with extensive experience in the industry. This chapter examines whether current “off-the-

---

14 Indeed confidential reports made available to this author describe innovative ways of incorporating sentiment analysis into trading algorithms generating annualized returns of approximately 10%.
“shelf” SA systems produced by TDI development can effectively be used to create a trading algorithm. As will be shown, although this system is competitive in terms of intrinsic accuracy (as defined in Chapter 2), its generated returns do not differ statistically from common trading baselines (Section 5.3.1).

Next, we will examine whether TDD optimization can improve the results (Section 5.3.2). This work is most similar to Zhang and Skiena’s (2010) who experimented with news sentiment to inform simple market neutral trading algorithms, and produced an impressive maximum yearly return of around 30%, even more when using sentiment from blogs and Twitter data. They did so, however, without an appropriate baseline, making it difficult to appreciate the significance of this number. Using a standard, and in fact somewhat dated sentiment analyzer, we are regularly able to garner annualized returns over twice that percentage, and in a manner that highlights two of the better design decisions that Zhang and Skiena (2010) made, viz., (1) their decision to trade based on numerical SVM scores rather than upon discrete positive or negative sentiment classes, and (2) their decision to go long (resp. short) in the \( n \) best- (worst-) ranking securities rather than to treat all positive (negative) securities equally.

On the one hand, we also use a test documents’ raw SVM score, rather than its relative rank within a basket of other securities as Zhang and Skiena (2010) did, and we experimentally tune a threshold for that score that determines whether to go long, neutral or short. We sampled our stocks for both training and evaluation in two runs, one without *survivor bias*, the tendency for long positions in stocks that are publicly traded as of the date of the experiment to pay better using historical trading date than long positions in random stocks sampled on the trading days themselves. Most of the evaluations of sentiment-based trading either unwittingly adopt this bias, or do not need to address it because their returns are computed over very brief historical periods. We also provide appropriate trading baselines as well as Sharpe ratios (Sharpe, 1966) to attempt to quantify the relative risk inherent to our experimental strategies. As tacitly assumed by most of the work on this subject, our trading strategy is not portfolio-limited, and our returns are calculated on a percentage basis with theoretical, commission-free trades.

To evaluate models, the larger SA community relies on a variety of measures, including classification accuracy or hypothesis testing scores such as F-measures, SARs, kappas or Krippendorff alphas that rely on TDI labeled datasets produced by annotators. The human-
subject annotators recruited are typically untrained. Given the complexities of annotating authorial sentiment in the financial domain, as described in Section 2.2.2, the annotations produced can thus be noisy. In Section 5.3.3, we examine this conjecture. Evidence to the contrary could justify the use of measures relying on TDI annotations produced by untrained human-subjects as a good measure of expressed sentiment in this domain. Such a result would in turn motivate similar experiments in other domains to examine the versatility of untrained annotators in detecting sentiment in disparate contexts. To test the veracity of annotations produced by untrained human subjects, we examine whether a computed accuracy that relies on said annotations is a good proxy for task-based evaluation in our domain (Section 5.3.3). We exhibit one particular modification to our experimental financial sentiment analyzer that, when evaluated against a labeled test set sampled from the same pool of human-subject annotations as the analyzer’s training data, returns poorer performance than when evaluated against actual market returns. The results in Section 5.3.3 show that measures using said annotations may at some points not be accurate even up to a determination of the delta’s sign in task-based performance. These results suggest that at least, in this domain, reliable evaluations of SA models must either be task-based or rely on annotations that are carefully elicited from trained annotators.

On the other hand, the results reported here should not be construed as an indictment of sentiment analysis as a technology or its potential application or even of using noisy, labeled training datasets. In fact, one of our baselines alternatively attempts to train the same classifier directly on market returns without utilizing our noisy, labeled sentiment training dataset, and the experimental approach handily beats that, too. For TDD development, our findings suggest that it is important to train on human-annotated sentiments, but then it is equally important to tune, and eventually evaluate, on an empirically grounded task-specific measure, such as market returns. This chapter thus presents, to our knowledge, the first real proof that sentiment is worth analyzing in the financial domain, as earlier attempts were either not sufficiently grounded in a task, or evaluated against poor baselines.

A likely machine-learning explanation for this experimental result is that whenever two unbiased estimators are pitted against each other, they often result in an improved combined performance because each acts as a regularizer against the other. If true, this merely attests to the relative independence of task-based and human-annotated knowledge sources. A more HCI-oriented view, however, would argue that direct human-subject annotations are highly problematic unless the annotations have been elicited in a manner that is ecologically valid.
When human-subjects are paid to annotate quarterly reports or business news, they are paid regardless of the quality of their annotations, the quality of their training, or even their degree of comprehension of what they are supposed to be doing. When human-subjects post film reviews on websites, they are participating in a cultural activity in which the quality of the films under consideration is only one factor. These sources of annotation have not been properly controlled in previous experiments on SA.

Regardless of the explanation, this is a lesson that applies to many more areas of NLP than just sentiment analysis, and to far more recent instances of sentiment analysis than the one that we based our experiments on here. Indeed, we chose sentiment analysis because this is an area that can set a higher standard; it has the right size for an NLP component to be embedded in real applications and to be evaluated properly. This is noteworthy because it is challenging to explain why recent publications in sentiment analysis research would so dramatically increase the value that they assign to sentence-level sentiment scoring algorithms based on syntactically compositional derivations of “good-for/ bad-for” annotation (Anand & Reschke, 2010; Deng, Choi, & Wiebe, 2013), when statistical parsing itself has spent the last twenty-five years staggering through a linguistically induced delirium as it attempts to document any of its putative advances without recourse to clear empirical evidence that PTB-style syntactic derivations are a reliable approximation of semantic content or structure.

We submit, in light of our experience with the present study, that the most crucial obstacle facing the state of the art in SA in the financial domain is not a granularity problem, nor a pattern recognition problem, but an evaluation problem. For more accurate assessments, researchers in this domain can rely on annotated data produced by trained annotators and conduct task-based evaluation viz. securities trading. The latter choice is recognized as a good measure of expressed sentiment by domain experts and in fact, needs no manual annotation, given the prevalence of low-cost historical trading data.

5.1. News Data

Our dataset combines two collections of Reuters news documents. The first was obtained for a roughly evenly weighted collection of 22 small-, mid-, and large-cap companies, randomly sampled from the list of all companies traded on the NYSE as of 10th March 1997. The second was obtained for a collection of 20 companies randomly sampled from those companies that were publicly traded in 10th March 1997, and still listed on 10th March 2013.
For both collections of companies, we collected every chronologically third Reuters news document about them from the period 10\textsuperscript{th} March 1997 to March 2013. The news articles prior to 10\textsuperscript{th} March 2005 were used as training data, and the news articles on or after 10\textsuperscript{th} March 2005 were reserved as testing data. We split the dataset at a fixed date rather than randomly in order not to incorporate future news into the classifier through lexical choice.

In total, there were 1256 financial news documents. Each was labelled by two human annotators as being negative, positive, or neutral in sentiment. The annotators were instructed to gauge the author’s belief about the company, rather than to make a personal assessment of the company’s prospects. Only the 991 documents that were labelled twice as negative or positive were used for training and evaluation.

5.2. Task- and Domain-Independent SA

For each document, all punctuation and the most common 429 stop words were first filtered out. The sentiment analyzer is a Support Vector Machine with a linear kernel function implemented using SVM\textsuperscript{light} (Joachims, 1999). The experiment examined different feature sets, including, raw term frequencies, binary term-presence, and term frequencies weighted by the BM25 scheme—a scheme which had the most resilience in a study on information retrieval weighting schemes for SA by Paltoglou and Thelwall (2010). A 10-fold cross-validation was performed on the training data with the folds constructed such that each contained an approximately equal number of negative and positive examples. This ensured against accidental bias in a fold.

Pang et al. (2002) use word presence features with no stop list, instead excluding all words with frequencies of 3 or less. Pang et al. (2002) normalize their word presence feature vectors, rather than term weighting with an IR-based scheme like BM25, which also involves a normalization step. Pang et al. (2002) also use an SVM with a linear kernel on their features, but they train and compute sentiment values on film reviews rather than financial texts, and their human judges also classified the training films on a scale from 1 to 5, whereas ours used a scale that can be viewed as being from -1 to 1, with specific qualitative interpretations assigned to each number. Antweiler and Frank (2004) use SVMs with a polynomial kernel (of unstated degree) to train on word frequencies relative to a three-valued classification, but they only count frequencies for the 1000 words with the highest mutual
information scores relative to the classification labels. Butler and Keselj (2009) also use an SVM trained upon a very different set of features, and with a polynomial kernel of degree 3.

Table 5.1: Average 10-fold cross-validation accuracy of the sentiment classifier using different feature sets and term frequency weighting schemes. The same folds were used across different feature sets.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>term presence</td>
<td>80.164%</td>
</tr>
<tr>
<td>bm25 freq</td>
<td>81.143%</td>
</tr>
<tr>
<td>bm25 freq with sw</td>
<td>79.827%</td>
</tr>
<tr>
<td>freq</td>
<td>79.276%</td>
</tr>
<tr>
<td>freq with sw</td>
<td>75.564%</td>
</tr>
</tbody>
</table>

As a sanity check, we measured our sentiment analyzer’s accuracy on film reviews by training and evaluating on Pang and Lee’s (2004) film review dataset, which contains 1000 positively and 1000 negatively labelled reviews. Pang and Lee conveniently labelled the folds that they used when they ran their experiments. Using these same folds, we obtain an average accuracy of 86.85%, which is comparable to Pang and Lee’s 86.4% score for subjectivity extraction. The purpose of this comparison is simply to demonstrate that our implementation is a faithful rendering of Pang and Lee’s (2004) algorithm.

Table 5.1 shows the performance of SVM with BM25 weighting on our Reuters evaluation set versus several baselines. All baselines are identical except for the term weighting schemes used, and whether stop words were removed. As can be observed, SVM-BM25 has the highest sentiment classification accuracy: 80.164% on average over the 10 folds. This compares favorably with previous reports of 70.3% average accuracy over 10 folds on financial news documents (Koppel & Shtrsimberg, 2004). We will nevertheless adhere to normalized term presence for now, in order to stay close to Pang and Lee’s (2004) implementation.

5.3. Task-based Evaluation

Overall, our trading strategy is simple: go long when the classifier reports positive sentiment in a news article about a company, and short when the classifier reports negative sentiment.

We will embed the aforementioned sentiment analyzer into three different trading algorithms. In Section 5.3.1, we use the discrete polarity returned by the classifier to decide whether go long/abstain/short a stock. In Section 5.3.2 we instead use the distance of the current document from the classifier’s decision boundary reported by the SVM. These distances do
have meaningful interpretations apart from their internal use in assigning class labels. Platt (1999) showed that they can be converted into posterior probabilities, for example, by fitting a sigmoid function onto them, but we will simply use the raw distances. In Section 5.3.2, we also impose a safety zone onto the interpretation of these raw distance scores.

In the experiments of this section, we will evaluate an entire trading strategy, which includes the sentiment analyzer and the particulars of the trading algorithm itself. The purpose of these experiments is to refine the trading strategy itself and so the sentiment analyzer will be held constant. In Section 5.3.3, we will hold the trading strategy constant, and instead vary the document representation features in the underlying sentiment analyzer.

In all three experiments, we compare the per-position returns of the following four standard strategies, where the number of days for which a position is held remains constant:

1. The momentum strategy computes the price of the stock $h$ days ago, where $h$ is the holding period. Then, it goes long for $h$ days if the previous price is lower than the current price. It goes short otherwise.

2. The S&P strategy simply goes long on the S&P 500 for the holding period. This strategy completely ignores the stock in question and the news about it.

3. The oracle S&P strategy computes the value of the S&P 500 index $h$ days into the future. If the future value is greater than the current day’s value, then it goes long on the S&P 500 index. Otherwise, it goes short.

4. The oracle strategy computes the value of the stock $h$ days into the future. If the future value is greater than the current day’s value, then it goes long on the stock. Otherwise, it goes short.

The oracle and oracle S&P strategies are included as toplines to determine how close the experimental strategies come to ones with perfect knowledge of the future. “Market-trained” is the same as “experimental” at test time, but trains the sentiment analyzer on the market return of the stock in question for $h$ days following a training article’s publication, rather than the article’s annotation.
5.3.1. Experiment One: Evaluating TDI SA

Given a news document for a publicly traded company, the trading agent first computes the sentiment class of the document. If the sentiment is positive, the agent goes long on the stock on the date the news is released; if negative, it goes short.

Table 5.2: Returns and Sharpe ratios for the experimental, momentum, S&P, oracle S&P, and oracle trading strategies over 30, 5, 3, and 1 day(s) holding periods.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Period</th>
<th>Return</th>
<th>S. Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>30 days</td>
<td>-0.037%</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>0.763%</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>0.742%</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.716%</td>
<td>0.108</td>
</tr>
<tr>
<td>Momentum</td>
<td>30 days</td>
<td>1.176%</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>0.366%</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>0.713%</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.017%</td>
<td>-0.002</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>30 days</td>
<td>0.318%</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>-0.038%</td>
<td>-0.016</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>-0.035%</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.046%</td>
<td>0.036</td>
</tr>
<tr>
<td>Oracle S&amp;P</td>
<td>30 days</td>
<td>3.765%</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>1.617%</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>1.390%</td>
<td>0.949</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.860%</td>
<td>0.909</td>
</tr>
<tr>
<td>Oracle</td>
<td>30 days</td>
<td>11.680%</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>5.143%</td>
<td>0.809</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>4.524%</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>3.542%</td>
<td>0.630</td>
</tr>
<tr>
<td>Market-trained</td>
<td>30 days</td>
<td>0.286%</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>5 days</td>
<td>0.447%</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>0.358%</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>0.533%</td>
<td>0.080</td>
</tr>
</tbody>
</table>

All trades are made based on the adjusted closing price on this date. We evaluate the performance of this strategy using four different holding periods: 30, 5, 3, and 1 day(s).

The returns and Sharpe ratios are presented in Table 5.2 for the four different holding periods and the five different trading strategies. The Sharpe ratio is a return-to-risk ratio, with a high value indicating good return for relatively low risk. The Sharpe ratio is calculated as:

\[
S = \frac{E[R_a - R_b]}{\sqrt{var(R_a - R_b)}}
\]

where \( R_a \) is the return of a single asset and \( R_b \) is the return of a 10-year U.S. treasury note.
The returns from this experimental trading system are fairly low, although they do beat the baselines. A one-way ANOVA test among the experimental, momentum and S&P strategies using the percent returns from the individual trades yields p values of 0.06493, 0.08162, 0.1792, and 0.4164, respectively, thus failing to reject the null hypothesis that the returns are not significantly higher. Furthermore, the means and medians of all three trading strategies are approximately the same and centred around 0. The standard deviations of the experimental strategy and the momentum strategy are nearly identical, differing only in the thousandths digit. The standard deviations for the S&P strategy differ from the other two strategies due to the fact that the strategy buys and sells the entire S&P 500 index and not the individual stocks described in the news articles. There is, in fact, no convincing evidence that discrete sentiment class leads to an improved trading strategy from this or any other study with which we are familiar, based on their published details. One may note, however, that the returns from the experimental strategy have slightly higher Sharpe ratios than either of the baselines.

One may also note that using a sentiment analyzer mostly beats training directly on market data. This vindicates using sentiment annotation as an information source.

Figure 5.1 shows the market capitalizations of each individual trade’s companies plotted against their percent return with a 1 day holding period. The correlation between the two variables is not significant. Returns for the other holding periods are similarly dispersed.

Figure 5.1: Market capitalization (in millions USD) of each trade with one-day holding period.
The importance of having good baselines is demonstrated by the fact that when we annualize our returns for the 3-day holding period, we get 70.086%. This number appears very high, but the annualized return from the momentum strategy is 70.066%, which is not significantly lower.

Figure 5.2 shows the percent change in share value plotted against the raw SVM score for the different holding periods. We can see a weak correlation between the two. For the 30 days, 5 days, 3 days, and 1 day holding periods, the correlations are 0.017, 0.16, 0.16, and 0.16, respectively. The line of best fit is shown. This prompts our next experiment.

5.3.2. Experiment Two: Using Domain and Task Knowledge

5.3.2.1. Moving the threshold

Before, we labelled documents as positive (negative) when the score was above (below) 0, because 0 was the decision boundary. But 0 might not be the best threshold, $\theta$, for high returns. To determine $\theta$, we divided the evaluation dataset, i.e. the dataset with news articles dated on or after March 10, 2005, into two folds having an equal number of documents with
positive and negative sentiment. We used the first fold to determine $\theta$ and traded using the data from the second fold and $\theta$. For every news article, if the SVM score for that article is above (below) $\theta$, then we go long (short) on the appropriate stock on the day the article was released. A separate theta was determined for each holding period. We varied $\theta$ from −1 to 1 in increments of 0.1.

Using this method, we were able to obtain significantly higher returns. In order of 30, 5, 3, and 1 day holding periods, the returns were 0.057%, 1.107%, 1.238%, and 0.745% ($p < 0.001$ in every case). This is a large improvement over the previous returns, as they are average per-position figures.

5.3.2.1. Using safety zones

For every news item classified, SVM outputs a score. For a binary SVM with a linear kernel function $f$, given some feature vector $\mathbf{x}$, $f(\mathbf{x})$ can be understood as the signed distance of $\mathbf{x}$ from the decision boundary (Boser, Guyon, & Vapnik, 1992). It is then possibly justified to interpret raw SVM scores as degrees to which an article is positive or negative.
As in the previous section, we separate the evaluation set into the same two folds, only now we use two thresholds, $\theta \geq \zeta$. We will go long when the SVM score is above $\theta$, abstain when the SVM score is between $\theta$ and $\zeta$, and go short when the SVM score is below $\zeta$. This is a strict generalization of the above experiment, in which $\zeta = \theta$.

For convenience, we will assume in this section that $\zeta = -\theta$, leaving us again with one parameter to estimate. We again vary $\theta$ from 0 to 1 in increments of 0.1. Figure 3 shows the returns as a function of $\theta$ for each holding period on the development dataset. If we increased the upper bound on $\theta$ to be greater than 1, then there would be too few trading examples (less than 10) to reliably calculate the Sharpe ratio. Using this method with $\theta = 1$, we were able to obtain even higher returns: 3.843%, 1.851%, 1.691, and 2.251% for the 30, 5, 3, and 1 day holding periods, versus 0.057%, 1.107%, 1.238%, and 0.745% in the second task-based experiment.

5.3.3. Experiment Three: Feature Selection

In our final experiment, let us now hold the trading strategy fixed (at the third one, with safety zones) and turn to the underlying sentiment analyzer. With a good trading strategy in place, it is clearly possible to vary some aspect of the sentiment analyzer in order to determine its best setting in this context. We will measure both market return and classifier accuracy to determine whether they agree. Is the latter a suitable proxy for the former? Indeed, we may hope that classifier accuracy will be more portable to other possible tasks, but then it must at least correlate well with task-based performance.

In addition to evaluating those feature sets attempted in Sections 5.3.1 and 5.3.2, we now hypothesize that the passive voice may be useful to emphasize in our representations, as the existential passive can be used to evade responsibility. So we add to the BM25 weighted vector the counts of word tokens ending in “n” or “d” as well as the total count of every conjugated form of the copular verb: “be”, “is”, “am”, “are”, “were”, “was”, and “been”. These three features are superficial indicators of the passive voice. Clearly, we could have used a part-of-speech tagger to detect the passive voice more reliably, but we are more interested here in how well our task-based evaluation will correspond to a more customary classifier-accuracy evaluation, rather than finding the world’s best indicators of the passive voice.
Table 5.3: Sentiment classification accuracy and trade returns (in percentage) of different feature sets and term frequency weighting schemes. The accuracy is the average 10-fold cross-validation accuracy. The same folds were used for different representations. Non-annualized returns are presented in columns 3–6.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Accuracy</th>
<th>30 days</th>
<th>5 days</th>
<th>3 days</th>
<th>1 day</th>
</tr>
</thead>
<tbody>
<tr>
<td>term presence</td>
<td>80.164%</td>
<td>3.843%</td>
<td>1.851%</td>
<td>1.691%</td>
<td>2.251%</td>
</tr>
<tr>
<td>bm25_freq</td>
<td>81.143%</td>
<td>1.110%</td>
<td>1.770%</td>
<td>1.781%</td>
<td>0.814%</td>
</tr>
<tr>
<td>bm25_freq_d_n_copular</td>
<td>62.094%</td>
<td>3.458%</td>
<td>2.834%</td>
<td>2.813%</td>
<td>2.586%</td>
</tr>
<tr>
<td>bm25_freq_with_sw</td>
<td>79.827%</td>
<td>0.390%</td>
<td>1.685%</td>
<td>1.581%</td>
<td>1.250%</td>
</tr>
<tr>
<td>freq</td>
<td>79.276%</td>
<td>1.596%</td>
<td>1.221%</td>
<td>1.344%</td>
<td>1.330%</td>
</tr>
<tr>
<td>freq_with_sw</td>
<td>75.564%</td>
<td>1.752%</td>
<td>0.638%</td>
<td>1.056%</td>
<td>2.205%</td>
</tr>
</tbody>
</table>

Table 5.3 presents returns obtained from these 6 feature sets. The feature set with BM25-weighted term frequencies plus the number of copulars and tokens ending in “n”, “d” (bm25_freq_dnc) yields higher returns than any other representation attempted on the 5, 3, and 1 day holding periods, and the second-highest on the 30 days holding period. But it has the worst classification accuracy by far: a full 18 percentage points below term presence. This is a very compelling illustration of how misleading an intrinsic evaluation can be.

5.4. Conclusion

The current chapter examined the application of SA in predicting market reaction to content. First, a binary sentiment classifier was trained with high accuracy when tested on movie data and financial news data from Reuters. Three task-based experiments evaluated the use of SA in indicating market reaction to content. This was accomplished by embedding SA in trading strategies and measuring produced returns. Although high annual returns cannot be achieved simply by training with sentiment labels, the returns can be substantially improved by optimizing the models to the task using the SVM’s decision score (Section 5.3.2) as well as by including features designed to detect evasive language use (Section 5.3.3). Results here support the hypothesis that TDD development can produce more useful SA systems.

It has also been observed that classification accuracy alone is not always an accurate predictor of task performance. The same changes to the SA system that improved returns degraded classification accuracy (to as low as 62.1%) while changes that degraded returns modestly improved classification accuracy. This calls into question the benefits of using TDI sentiment-classification accuracy, particularly when the relative cost of a task-based evaluation may be comparably low.

Given the prominence of TDI measures in NLP research, it is important to examine whether the limitations of TDI measures observed here also affect evaluations of other NLP technologies needed for building SDSSs. Due to ASR’s important role in the design and
development of SDSSs, its evaluation is closely examined in the next chapter. As noted in Chapter 2, this technology has been evaluated almost exclusively with WER, another TDI measure. Unlike SA accuracy, however, WER possesses an objective ground truth, removing one potential complication from its ability to indicate the usefulness of an evaluated system. The next chapter explores whether this evaluation is an appropriate proxy for ASR usefulness as a component of an SDSS.
Chapter 6
Evaluating Automatic Speech Recognition

Motivated by the conclusion in Chapter 5, that TDI accuracy is not an unequivocal evaluator of SA, this chapter examines whether the TDI evaluation of ASR (using TDI WER) can adequately characterize its performance. As will be discussed, the results show that, similar to TDI accuracy, WER is also not a good proxy for task-based evaluation in our domain. A semi-automatic TDD measure is then introduced, which is a better indicator of task performance than WER.

Sixty years ago, it was not uncommon to regard speech transcription as a worthwhile task in its own right. In part because ASR works so much better now than then, and in part because it is increasingly easier to find, play, and store spoken documents, transcripts are no longer created purely for their own sake. Often, ASR is widely-used with other NLP techniques, such as speech summarization or speech-to-speech translation, to assist information seekers in accomplishing a task. With the exception of accessibility for the hearing impaired, transcription as a speech processing task has been commoditized.

Meanwhile, means of evaluating the quality of a speech recognition system have remained largely unchanged; WER is still the standard. It is clear why this is important for evaluating the quality of a transcript, but its applicability to downstream uses of ASR has rightfully come under some scrutiny, e.g., when ASR is used with speech summarization to help users make use of meeting transcripts (Murray, et al., 2009) or when it is used within spoken language understanding systems (Wang, Acera, & Chelba, 2003). In other words, what is the harm of transcribing some words incorrectly if a user does not see such errors (e.g., when filtered out by a speech summarizer), and the performance of the task remains unchanged? In certain domains at least, relatively high-WER transcripts have been shown to be perfectly usable (Munteanu, Baecker, Penn, Toms, & James, 2006).

The purpose of the current chapter is not to postulate a better alternative to WER for evaluating transcript quality; no better alternative likely exists if the task at hand is taken to be speech transcription for its own sake. Nor is the purpose to suggest that the ASR community’s attempts to improve transcripts’ WER have been in vain. This is clearly not the
case since as noted in Chapter 1, these efforts have led to the creation of successful speech-as-input interfaces now commercially available on most smartphones and operating systems. Rather, the current study asserts that there are many more useful and relevant tasks to be evaluating speech recognition within. One such task, and a chief concern of the current study, is examining ASR’s usefulness as a component technology in an SDSS to make spoken content more accessible. Due to the nature of ISP, it is anticipated that TDD evaluation is more indicative of ASR usefulness as a component of SDSSs than TDI WER. Furthermore, it is asserted that TDD evaluation should adhere to the principle of ecological validity, which requires the participating human-subjects to be placed in situations that closely approximate real-world scenarios. In the current study’s setting, an ecologically valid task is one that realistically describes how ASR output and the embedding SDSS can be used in real-world contexts. An ecologically valid evaluation uses that task to predict the performance of subjects using ASR output, i.e. how useful a given output is to users.

Our contribution is as follows: availing ourselves of the decision audit task as an ecologically valid one, we define the workflow for a metric that involves human-subject experimentation (Section 6.2). Unfortunately, as will be evident, designing and conducting task-based user studies such as the one described later in Section 6.2 can be time-consuming. This limitation makes human-subject experiments less-than-ideal for evaluating progress in the fast-moving field of ASR, in which evaluations take place far more often than human-subject experiments can be designed and conducted. Next, we study how well WER can predict task success, and propose an alternative TDD metric that attempts to simulate the task-based study described in Section 6.2 for a new transcript with the use of statistical models. This metric (Section 6.3) is trained on data collected from the described human-subject experiment and a number of features of the input transcript and audio (Section 6.4). Experimental results show that it outperforms WER at predicting human performance (Section 6.5). Related work is discussed in Section 6.6. As was the case in Chapter 5, experiments here also support the thesis that TDD evaluation is a better indicator of NLP technology usefulness as components of SDSSs.

### 6.1. Decision Auditing and Financial Analysts’ Spoken Document Use

As discussed in Chapter 3, the information-seeking practices of financial analysts and their decision-making processes are highly collaborative. For instance, financial firms hold investment presentation meetings to discuss their portfolios. In these meetings, analysts and economists recommend investments to portfolio managers. As described by our participants
in Chapter 3’s study, questions and discussions in these meetings, which revolve around investment decisions, are critically important for financial analysts to better understand their firm’s investment strategies and priorities (P1, P4, P7). Similar to other industries (Whittaker, Tucker, Swampillai, & Laban, 2008), financial analysts often use traditional records, such as shared documents, minutes, and private notes to review information from meetings that they may or may not have attended. Given the limitations of these traditional records (Whittaker, Tucker, Swampillai, & Laban, 2008), a well-designed SDSS could make a significant contribution to financial analysts’ information-seeking success. Such an SDSS needs to support decision auditing: identifying decisions as well as arguments for and against these decisions in meetings. The importance of decision auditing to information seekers in the investment management industry has been established in Chapter 3. This chapter examines the effect of transcripts’ errors on their ability to support decision auditing.

The study focuses on decision auditing, instead of the information seeking subtasks performed using spoken records of earnings call and central bank news conferences, because the development of ASR systems is at a more critical stage in the meeting domain. Automatic transcription of documents with properties similar to earnings calls and news conferences, e.g., broadcast news, has a low enough error rate to be correctable with minimal human intervention. State-of-the-art ASR systems in the meeting domain, however, produce far more errors than can be feasibly revised by human intervention. Moreover, meetings are far more frequent, making human transcription prohibitively costly, and in the case of confidential meetings, simply ill-suited. Improving ASR in this domain can have a transformative effect on the usefulness of SDSSs. This chapter investigates whether WER can adequately characterize an ASR system’s quality in this regard.

Unfortunately, despite an extensive search, we were unable to find a publicly available corpus of meetings in the investment management industry. Moreover, the author was prohibited from attending, let alone recording, such meetings during the contextual inquiry reported in Chapter 3 due to financial analysts’ fiduciary and legal obligations to preserve the confidentiality of the information exchanged in the meetings. The dataset most similar to this domain, the AMI meetings corpus (McCowan, et al., 2005b), was thus chosen. Focus was given to the dataset’s scenario meetings in which four participants discuss the design of a remote control over four meetings by presenting candidate features and debating their pros and cons.
Even though AMI is the most similar dataset we could find to a repository of recorded investment presentation meetings, there are still notable differences between the two. For instance, although both remote controls and investment portfolios are created to eventually contribute to a holding company’s profitability, the immediate goal of the AMI meetings is to produce the most usable and useful remote controls on the smallest budget, while the goal of investment presentation meetings is to design a portfolio that directly maximizes a firm’s return on investments. These differences clearly influence the topics discussed in the meetings. Major topics discussed in AMI meetings include potential user expectations, material used to make the remotes, aesthetics of the design, functionality, usefulness, and usability. Topics discussed in investment meetings can range widely from economic, political, and market issues, to industry-specific phenomena such as seasonality or specific take-over bids.

Moreover, contextual factors affect these two types of meetings. For instance, the fast-paced and aggressive nature of the investment management industry can influence how meeting participants communicate, potentially creating disparity between the content of meetings in the financial domain and the AMI corpus where volunteers were recruited to create the dataset.

However, there are also notable similarities between these two types of meetings. Firstly, both types of meetings are task-based, where participants aim to produce products featuring different components and attributes. In investment presentation meetings, components include investment instruments and attributes include the portfolio size as well as the associated risks. For remote controls, components include buttons and navigational tools, and attributes include the remote’s colour and shape. Participants of these two types of meeting also assume similar roles and responsibilities. In investment presentation meetings, a portfolio manager often leads the meetings and incorporates input from the participants into an eventual portfolio design. In the AMI meetings, a project manager plays a similar role to a portfolio manager, incorporating input from marketing representatives and designers into a remote control design. Finally, the designs produced in both types of meeting are subject to constraints. In the investment presentation meetings, these constraints revolve around the desired portfolio neutrality to a set of risks, while in remote control design meetings, the constraints are primarily budgetary.
From one perspective, using the AMI corpus instead of a repository of investment meeting records makes an industry-invariant assumption about why and how meeting records are used. This assumption is partly justified for two reasons. First, the records of investment meetings, as described earlier, are used in part to keep track of investment decisions, as well as arguments for or against them. Field studies of corporate meetings in disparate industries, namely, information technology and transportation, noted that meeting records are also used for this purpose (Whittaker et al., 2008). Second, the same studies did not observe variance in the way meeting records were used across industries. The difference was mostly related to a participant’s role in the meetings; i.e., whether or not they played a managerial role. Given these observations, and following previous practice in the area, where field studies of meeting records were followed by task-based evaluations using the AMI corpus (Whittaker et al., 2008), we chose this corpus in the absence of a more appropriate one.

### 6.2. Method

One common form of ecologically valid evaluation in the HCI community is a task-based within-subject experiment in which a large number of human participants are recruited to use the systems under evaluation (here, SDSSs) to perform a task that does not significantly deviate from their daily activities. In this scheme, evaluated systems are scored by how well their users perform the task in the experimental trials. This chapter reports on such a study in

![JFerret interface](image)

**Figure 6.1:** JFerret interface as used by participants in the human-subject experiment. Clicking on each dialogue act seeks the videos to the point in the recording where the dialogue act was spoken. Video record can also be navigated using the play, pause, and stop buttons. Words spoken at the current moment are also highlighted.
which transcripts from several ASR conditions (four automatic and one manual transcription as a topline) are evaluated within a decision audit task.

6.2.1. Task

Our decision audit task is similar to (Murray, et al., 2009). Each study participant\(^{15}\) (the meeting auditor) plays the role of a recently hired executive product manager in a company that manufactures remote controls. The company had asked three independent teams to design the remote control. The auditor needs to catch up with the decisions made by each design team concerning the remote control in meetings that were held and recorded before s/he was hired. The auditor then needs to browse through the recorded meetings (minutes from the meetings are not available). This description of a real-world scenario helped explain the decision audit task to our participants in meaningful and familiar terms. Hence our participants had a more uniform understanding of what they had to do.

6.2.2. Data

In each AMI scenario meeting, four participants discuss the design of a remote control over four meetings: a kickoff meeting, design requirements meeting, conceptual design meeting, and detailed design meeting. Auditors, playing the role of executive product managers with hectic schedules, are only given 25 minutes to browse the final meeting held by each team which lasted 32–48 minutes. All meetings used in the experiment were conducted under the same protocol, had identical goals, and were selected by an experimental design specialist to be as similar to each other according to criteria such as length, number of decisions, flow of dialogue, etc.

6.2.3. Tools and experimental design

To browse meetings, the auditors used the JFerret (Wellner, Flynn, & Guillemot, 2005), a meeting browser that presents audio/video feed from all meeting participants as well as textual meta-data, with a simplified user interface (Figure 6.1). The interface shows both the recorded video as well as an extractive summary of the meeting, presented as a clickable list of transcribed utterances. The auditor is able to navigate through a meeting by clicking on

\(^{15}\) 118 participants were recruited for this study, including pilots. Results presented are based on 98 participants.
utterances in the extractive summary. The interface also included standard play, pause, and stop buttons for video navigation.

To complete the experiment, the auditors had to operate a desktop computer, fill out a form presented to them on the computer, learn how to operate the JFerret interface during the training portion of the experiment, and become familiar with extracting specific information related to a question that they see on the screen. All participants were quite familiar with these sub-tasks. Perhaps the most challenging of them was extracting information from the meeting data. This is quite similar to taking notes during a class or in meetings with a particular goal in mind, such as passing a course. All participants were either students who were already taking notes in class in preparation for assignments and tests, or staff who worked in professional office environments. The fact that participants were familiar with these sub-tasks supports the ecological validity of our experimental conditions.

Each extractive summary contained transcribed utterances from one of five transcription conditions: a reference transcript, as well as four automatic transcripts with WERs averaged over all meetings of 26.8 (ASR1), 28.2 (ASR2), 49.2 (ASR3), and 38.9 (ASR4). The first two were chosen with nearly identical WERs despite having different acoustic and language models. If WER is an accurate predictor of human-subject performance, then auditor scores between these two systems should be similar. In a Latin square experimental design, each of the 98 recruited participants audited three meetings, one with a summary that used a reference transcript and two with summaries that used different automatic transcripts. These conditions thus focused on the ability of experimental ASR systems to generate transcripts from which usable summaries could be obtained.

Auditors were instructed to use an adjacent desktop computer to note down decisions related to functionality, physical properties, components, and design. Also, when possible, they had to record arguments made in the meeting favoring or opposing each decision. To evaluate auditor performance, two judges independently extracted the decisions and arguments by listening to the meetings and viewing the reference transcripts with no time limits. Afterwards, the judges adjudicated their lists to form a final decision rubric. An independent team of markers (who were not judges, experimenters, nor auditors) used these rubrics to mark each report created by auditors. Each audit report was marked independently by two markers who then adjudicated their marking reports to produce a final score. Auditors were assessed according to the number of design decisions they could find, $\alpha_1$, or partially find,
\( \alpha_2 \), the number of positive or negative arguments pertaining to the decisions that were found, \( \beta \), and the number of false alarm decisions, \( \gamma \), that were in fact not made to produce an \( H-score \) for each audit:

\[
H-score = 2\alpha_1 + \alpha_2 + \beta - 2\gamma
\]

The following section measures a transcript’s usefulness by the average \( H\text{-Score} \) of the auditors who used it and examines whether a transcript’s average \( H\text{-score} \) correlates with its WER.

6.3. **Automatic Evaluation Metric**

Running a human-subject experiment such as this one is time-consuming and expensive. Our objective is therefore to find automated means of anticipating the results of running a new human-subject experiment, given a new ASR system.

The *de facto* automatic metric used in the speech recognition community is word error rate (WER), but our repeated measures ANOVA tests failed to demonstrate any statistically significant effect by WER on audit scores. Figure 6.2 shows that WER does not numerically correlate with H-score (\( \rho = 0.017 \)), a result in line with those of similar studies (Wang, Acera, & Chelba, 2003; McCowan, et al., 2005a). The lack of statistical correlation suggests that, at least for this representative sample, WER is in fact not a viable predictor of human performance. There are almost certainly other sources of variability than the quality of the transcript, such as subjects’ abilities to find decisions and their mining strategies (listening vs. reading). Reference transcripts do not always result in better \( H-score \) than ASR systems because transcript quality is not the only reason for task success, even though it has a major impact. A potential explanation for the lack of correlation is that WER does not capture features of the transcript that condition these other sources.

WER has been widely-used and successful at evaluating transcript quality for the sake of measuring how close it is to a gold standard reference. However, in light of the results shown here (Figure 6.2) and in past experiments (Wang, Acera, & Chelba, 2003; McCowan, et al., 2005a), it is likely not an optimal indicator of transcript usefulness to human-performed tasks. In its place, we utilize Auditor Performance Prediction (APP), which simulates the evaluation for a new ASR transcript. In particular, for each decision-bearing dialogue act, the
idea is to extract a number of features and train a binary classifier to predict whether auditors in our user study, would have found the related decision or not.

This form of TDD simulation is justified from the perspective of ISP theory. As discussed in Chapter 2, the ISP is shaped by the information-seeking factors of setting, domain, task, information seeker, search system, and outcomes. The goal of this evaluation, as is generally true of other task-based user studies of information systems, is to examine the effects of one of these factors (a search system) on another (outcomes). This can be achieved only by keeping other information-seeking factors constant, as is the case in our APP simulations.

In section 6.4, we describe the TDD features used to represent each dialogue act in the evaluated spoken documents. In Section 6.5, we examine APP’s ability to predict transcripts’ usefulness against WER as a baseline. We also examine APP’s ability to predict the usefulness of in-sample spoken documents (Section 6.5.1), which simulates running a new human-subject experiment for the spoken documents currently known to APP. In Section 6.5.2, we experiment with APP’s ability to determine ASR usefulness for out-of-sample spoken documents. This requires APP to also predict whether a previously unseen dialogue act contains a decision, i.e., decision detection.
6.4. APP Features

To simplify the simulation of human-subject performance, meetings are treated as a discrete-time sequence of dialogue acts, each of which is represented by feature values calculated to capture the difficulty of the task (task-specific features), corruption of the transcript (WER variants and language modelling features), and the importance of the dialogue act (decision features, and summarizer features).

Let $w$ denote the sequence of words in the transcript of a dialogue act, $w_t$ the subsequence of topic words (Lin & Hovy, 2000), $w_s$ the subsequence of stopwords, and $w_p$ the subsequence of non-stopwords. Let $T$ be the constituent parse tree generated by the Berkeley parser (Petrov, Barrett, Thibaux, & Klein, 2006) over the transcript of the dialogue act, and $T^n = \{ t \in \text{subtree}(T), |t| \leq n \}$ the set of all subtrees of $T$ up to size $n$. Finally, let $G$ be the graph formed from $T$ and $w$ by linking each word to its immediate neighbors in $w$ and to its part of speech in $T$. In the following, $P(A) \approx \prod_{a \in A} P(a)$ denotes the maximum likelihood estimate of the distribution of the assumed-independent elements of $A$. The extracted features are then:
• Task features: The identifier of the user, of the meeting, and of the decision; whether the dialogue act is shown in the extractive summary.

• Word error rate variants: The dialogue-act-level WERs computed separately on $w$, $w_t$, $w_p$, and $w_s$ relative to the reference transcript; the value of the tree kernel between $T_{ref}$ and $T_{asr}$ (Moschitti, 2006), the Jensen-Shannon divergence (JSD) between $P(T^n_{asr})$ and $P(T^n_{ref})$, and the graph edit distance (Justice & Hero, 2006) between $G_{asr}$ and $G_{ref}$.

• Decision features: The number of phrases in $T_{asr}$; number of pronouns in $w$; ratio of the number of 3rd-person pronouns to the number of verbs (indicating a probable use of 1st or 2nd person pronouns) in $w$; depth of $T_{asr}$; token to type ratio; term frequencies of a list of decision-making verbs and modals (“think”, “believe”, “should,” etc.).

• Language modeling features: The ratio of novel substructures compared to a corpus\textsuperscript{16}, \[|T^n_{asr}/T^n_{corpus}|;\] the JSD between $P(T^n_{asr})$ and $P(T^n_{corpus})$.

• Summarizer features: The score and rank of the dialogue act according to the following summarization baselines: Maximal Marginal Relevancy (Carbonell & Goldstein, 1998), KL-divergence, character length, number of words, duration, sum of inverse document frequency, cosine similarity to the centroid of the meeting, topic words, and an acoustic summarization SVM trained on energy, word- and character-normalized duration and pitch.

6.5. Experiments and Discussion

6.5.1. Predicting Auditor Performance

Given the set of features extracted at the dialogue act level, the task then is to train a classifier to predict whether each decision dialogue act was found by the auditor or not. An Adaboost classifier was chosen that iteratively searches for the best combination of 1,000 decision stumps (one-level decision trees; Favre et al., 2007). This classifier has proved useful in a range of tasks and has the advantage of being unaffected by irrelevant features (contrary to SVMs, for instance). As interest lies in predicting human behavior given a new ASR transcript, leave-one-out cross validation was performed in which, for each transcript

\textsuperscript{16} Automated parses from Switchboard and reference parses from the Ontonotes corpora.
condition, a model was trained on the remaining conditions and evaluated on the left-out transcript. Figure 6.3 contains precision-recall curves for the models averaged over each ASR system and overall. The figure compares the predictive ability of APP when a) only WER is used to represent dialogue acts, and b) all features described in Section 6.4 are used. As it can be observed, including non-WER features significantly improves APP’s ability to predict performance.

Figure 6.4 shows the impact on the automatic metric quality of removing different subsets of features from training. The most important subset is the set of task features, capturing that a particular decision is difficult to detect from a dialogue act, or that a particular auditor is not good at finding decisions. The summarizer features, which capture whether a dialogue act is important for the meeting in general, seem to help in predicting auditor success. Interestingly, decision features did not improve performance. This may be because ASR errors disrupt the Berkeley parser, which makes these features less reliable. In any case, many features are affected by the quality of the transcript, including all parsing-related features. On the other hand, features like the length summary baseline operate more independent of the transcript and can only serve as normalizers for the other features. For the sake of comparison, the figure also shows APP performance when dialogue acts are purely represented by their WER.

In these experiments, the conditions where task features were included in the feature set are TDD evaluations (“w/o decision”, “All features”, “w/o WER variants”, “w/o LM”, “w/o summarizers”, in Figure 6.4). The conditions where task features were not used are TDI evaluations (“w/o task, decisions”, “w/o task”, “w/o task, WER variants”, “w/o task, LM”, “w/o task, summarizers”, “WER only”, in Figure 6.4). These results support the claim that TDD evaluation is a more effective indication of transcript usefulness than TDI evaluation.

Note that the results reported thus far assume the prediction model’s training dataset contains all the spoken documents that will be evaluated. Figure 6.5 shows the effect of excluding the evaluated transcripts from the training data for the 3 spoken records that were used in all the experiments, emulating a situation where a transcription of a previously unseen spoken document is evaluated. In this situation, the model must infer the effects of transcription errors on decision audit performance based on the auditor’s interactions with dialogue acts from other spoken documents, whereas previously, the model has the opportunity to learn about the effects of transcription errors on an auditor’s performance for every evaluated
Figure 6.4: Effects of holding out the test meeting from training set, and transcribing the training set using each of the transcription conditions.

Figure 6.5: Feature ablation experiment. F-score when each subset of features is removed from training.
dialogue act. In this scenario, APP’s ability to predict transcript usefulness substantially suffers, producing F-scores of 0.22, 0.08, and 0.24, in comparison to 0.40, 0.58, and 0.34 from the previous condition on the three meetings in Figure 6.4. One potential explanation for this performance drop is that the participants’ decision auditing processes vary more across differing dialogue acts than across different transcriptions of the same dialogue act, making it harder to infer auditor performance for unseen dialogue acts. Excluding the auditor identifier from the feature set further degrades APP, from an average F-score of 0.21 to 0.14, signifying that predicting a specific auditor’s performance based on their previous interactions with other documents is more robust than the same prediction for a generic auditor. However, the average F-Score in this scenario is still decidedly greater than the WER-only condition even when all dialogue acts are seen during training: 0.14 vs. 0.01, respectively.

6.5.2. Decision-Detection

Note that in the experiments thus far, the objective has been to predict whether auditors in our user study would have been able to detect the decisions, given that the dialogue act is decision-bearing. In a realistic scenario, predicting auditor performance in an out-of-sample spoken document would also require a means of first detecting dialogue acts which contain decisions, i.e., decision-detection. This task would require the exclusion of the task features
as a) we are no longer modeling how users performed in the human-subject experiment, and b) the task features’ domains do not overlap in the test and training datasets. Figure 6.6 shows decision-detection performance with no task features when training and testing on reference transcripts. As can be seen, the F-scores vary greatly for the different meeting records and the average F-score of 0.29 denotes that the non-task features are not very good at distinguishing decision-bearing dialogue acts from the rest, at least in the absence of task features.

We then examined the effects of transcription errors on decision detection by first using noisy transcripts during testing, and later during training. Decision-detection performance when testing on noisy transcripts did not degrade in our experiments. However, training on imperfect transcripts degraded performance. Figure 6.7 shows the effect of training transcript WER on decision-detection performance. As can be observed, although decision-detection varies across the different transcription WERs, having a higher WER does not necessarily translate into worse decision-detection performance: an increase in WER only led to an F-score decrease 6 out of 16 times. These results again support the claim that TDI measures such as WER are not unequivocal indications of real-world task performance.

6.6. Related Work

Several previous studies have aimed to quantify the effects of transcript WER on downstream task performance, such as spoken information retrieval, spoken language understanding, and spoken machine translation. Sanders and Le (2002) studied the effects of speech recognition accuracy on dialog systems, and found a high correlation between WER and task completion.
—even for high WERs. Even here, WER was shown to have little effect on user satisfaction when it is less than 35%), however.

There have also been previous attempts at modifying WER to take into account different types of errors in ASR output, e.g., using an error rate that weighs content words or information bearing words more heavily (Garofolo, Auzanne, & Voorhees, 1999). Morris et al. (2004) proposed to use match error rate (MER) and word information loss (WIL) to evaluate recognition performance and represent the proportion of word information communicated. For high error rates, they found that these are more appropriate. Similarly, McCowan et al. (2005a) suggested posing recognition evaluation as an information retrieval problem and thus using a more application-oriented evaluation. Mishra et al. (2011) developed a metric called Human Perceived Accuracy, and showed in a voice mail recognition task that it correlates more highly with human judgments of ASR accuracy than WER does. Their method of using a regression task for predicting recognition performance has some similarity to our study in this paper, although our focus is on the metric of ASR in the context of an ecologically valid task. This is important—without ecological validity, the human judgments and scores themselves are meaningless. Human judgements of ASR accuracy are not ecologically valid.

Current research on MT evaluation has many parallels to the issues facing ASR evaluation. In MT, automated metrics are routinely used in spite of their now well-documented limitations (Callison-Burch, Osborne, & Koehn, 2006) because they, too, provide a rapid, cost-effective means for developers to tune their systems’ performance. To independently validate proposed metrics, NIST Metrics MATR and WMT organizers have conducted shared meta-evaluation tasks alongside the standard MT evaluation tasks (Callison-Burch, et al., 2010). Initially these tested for correlation with human-subject judgments of translation adequacy and fluency—another ecologically invalid pair of tasks that cannot address real-world scenarios where MT provides support to downstream tasks that humans actually perform. Most recently, they have introduced a “quality estimation” (QE) task, which evaluates MT quality in terms of its impact on human post-editors, just as in our approach anchors its evaluation of ASR in a downstream decision-audit task. The majority of the QE system developers also make use of parsers, part-of-speech taggers, named entity recognizers, etc. to derive linguistic features of the source and target language texts, and then train M5P regression trees or SVM regression models on different combinations of these features to estimate the level of human effort required to post-edit MT output.
6.7. Conclusion

TDI WER has a legitimate place in the evaluation of speech recognition systems, but predicting human-subject performance, at least in our usage scenario, may not be one of them. Complementary TDD evaluation measures to WER are necessary in order to determine the effects of a change to an ASR system in an ecologically valid context. Human-subject judgements are expensive to collect, but automatically learning to predict those judgements, as in the dialogue act classifier here, brings such complementary measures closer to the grasp of experimenters who wish to regression-test their ASR improvements on real applications.

A significant remaining problem is predicting user judgements for out-of-sample documents, and the portability of judgements collected on one task to another, which have not been addressed here.

We showed that the TDI WER is not an effective measure of ASR’s usefulness to decision auditors. On the contrary, the TDD APP measure introduced here is a better alternative. These results, as well as results from Chapter 5 that showed a similar shortcoming of TDI accuracy, strongly support the thesis that TDD evaluation is more effective than TDI evaluation in guiding NLP technology development to make spoken content more accessible to investment management professionals.
Chapter 7
Conclusions and Future Work

The work presented in this dissertation was aimed at improving access to the spoken content of the fast-growing corpora of multimedia documents. NLP and ASR technologies have enabled users to interact more naturally with their devices through voice user interfaces (Pearl, 2016). This work examined how these technologies could be used to improve access to spoken content.

According to Marchionini’s theory of Information Seeking (Marchionini, 1995), focusing on a usage domain and, further, focusing on supporting specific tasks within that domain are critically important for providing effective support to information seekers. This is due to the fact that a user’s ISP, as they interact with information systems, can vary significantly, and is shaped by several factors, including task and information domain. This dissertation focused on usage scenarios in the financial domain because it was known to the author that financial analysts already use spoken documents in their professional activities. This enabled the author to conduct a field study to characterize factors shaping financial analysts’ ISPs in order to deduce how such information practices can more effectively be supported.

This dissertation showed that incorporating knowledge about information-seeking factors into the design and development of spoken document search systems (SDSSs) (i.e., TDD development) led to the conception of an effective prototype. It was also shown that SA and speech processing were requisite technologies for developing effective SDSSs. Next, it was demonstrated how the TDD development of SA could be achieved, specifically, by optimizing TDI SA models using TDD feedback. Moreover, it was demonstrated that for both SA and ASR, TDD evaluation is more appropriate than TDI evaluation. Finally, it was shown that in cases where TDD evaluation required performance of time-consuming user studies, these studies could effectively be simulated by using statistical models trained on data collected from one such user study.

This study began by exploring how SDSSs can be designed in a TDD manner to assist users in the financial domain. To gain the necessary domain knowledge for TDD development, a contextual inquiry was conducted in which financial analysts were observed as they interacted with spoken records of important events, e.g., central bank news conferences or public company
earnings calls (Chapter 3). The contextual inquiry uncovered that financial analysts’ motivation for using spoken documents is to gain the Essential Predictive Knowledge (EPK) needed to vaticinate an institution’s future actions, the outcomes of those actions, and other market participants’ reactions to the shared content. A taxonomy of information tasks performed by the participants was inferred. These information tasks included interpreting the transcript content in the context of relevant information, assessing communication, and performing comparative analysis. The described taxonomy was then validated in follow up PD workshops (Chapter 4). The prototypes produced in the PD workshops highlighted the importance of NLP technologies such as SA and speech processing in supporting financial analysts’ ISPs.

Next, we focused on SA to find whether TDI SA systems could be effective component technologies for supporting our users (Chapter 5). This was accomplished by conducting a task-based evaluation in which sentiment analyzers’ outputs were used to inform trading strategies. It was shown that a TDI SA system with similar accuracy to the state-of-the-art in SA research did not outperform established baselines in this domain. The returns were substantially improved when the SA system was tuned to market data, producing annualized returns of up to 70.1%, highlighting the utility of domain and task knowledge, and supporting the claim that TDD development produces more effective SA systems for task performance.

A distinguishing characteristic of TDI SA development is its reliance on non-task-based evaluation. The task-based evaluation of SA showed that it was not possible to infer a SA system’s usefulness based on its TDI accuracy: the system with the lowest accuracy (62.1%) produced the highest annualized returns (70.1%), while the system with the highest accuracy, produced one of the worst returns in the experiments. These results suggest that optimizing with TDI evaluation can be counter-productive, at least in our domain. Task-based evaluation of SA, on the other hand, is a more unequivocal appraisal of the systems’ usefulness in the financial domain.

Given the widespread use of the TDI evaluation in ASR research via WER, and noting ASR’s critical role in making spoken content accessible, this dissertation assessed WER’s ability to predict ASR systems’ usefulness in our domain. In a comprehensive task-based user study based on decision auditing, it was shown that similar to SA accuracy, WER also pointed in the opposite direction of transcript utility in several experimental conditions, once again highlighting the
importance of using TDD evaluation. However, TDD evaluation can be time-consuming to conduct when involving human-subjects (e.g., the reported human-subject experiment in Chapter 6). As an alternative, an automatic domain- and task-dependent evaluation system, Auditor Performance Predictor (APP) was introduced. It simulated a user study with aid of statistical models trained on data collected from one such study. Our experiments indicated that APP was a significantly better indicator of transcript usefulness in supporting decision auditing than WER.

The work presented in this dissertation promotes TDD development of SDSSs and necessary NLP components. The systems presented in Chapters 5 and 6 may be further improved by utilizing distributed semantic representations at the word level. Such distributed representations (e.g., Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014) are derived and used in a TDI manner, although they can become TDD if they were trained or somehow tuned on a task-specific dataset. Using TDD distributed word representations, it is possible to utilize sequence models to represent spoken document content in a TDD manner which could lead to better task performance. For instance, Recurrent Neural Networks (Elman, 1990) variants such as Gated Recurrent Units (Luong, Pham, & Manning, 2015) augmented with attention-based techniques (Wang, Liu, & Zhao, 2016) could perform well. This assumption is based on their enhanced ability to incorporate context present in sequential data (e.g., transcript text) into the classification process. Gated Recurrent Units are more appropriate for TDD development than other recurrent neural network variants such as LSTMs (Hochreiter & Schmidhuber, 1997), due to their ability to work with smaller datasets. In cases where the size of labelled TDD datasets is a concern, it is also possible to augment these datasets with larger unlabeled datasets using semi supervised document representation techniques such as paragraph vectors (Le & Mikolov, 2014), to more effectively utilize the TDD labeled data. Finally, using TDD distributed word representations, it is also possible to derive TDD distance measures which could be utilized in APP.

APP can be used as TDD feedback for ASR optimization, producing models that are more apt to supporting the decision audit task. This optimization can be accomplished by employing discriminative training, which has been successfully used to incorporate feedback from evaluation metrics, such as WER, into the training process (Povey & Woodland, 2002). This would, however, require APP to predict the usefulness of a relatively large training corpus, not just a test set from a user study. As experiments in Chapter 6 have shown, despite APP’s
diminished ability to predict an out-of-sample transcript's usefulness, it is still more robust than WER. Thus, discriminative training of ASR models with APP as an objective function is anticipated to produce more effective models for supporting decision auditors. In general, creating TDD datasets produced by ecologically valid data elicitation methods, and optimizing with TDD evaluation is deemed a productive endeavor for future NLP research geared to support users, be it in information seeking using spoken documents, or, other tasks involving natural language content.
Bibliography


