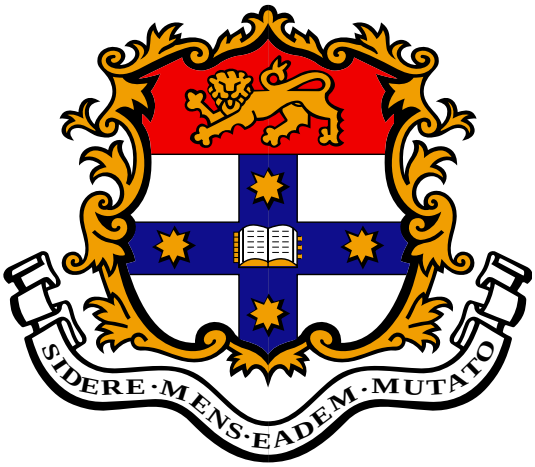
Tracking Information Flow in Financial Text

WILL RADFORD
SID: 200029009



Supervisor: James Curran and Ben Hachey

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Science (Honours)

School of Information Technologies
The University of Sydney
Australia

14 October 2011

Abstract

Information is fundamental to Financial markets, and understanding how it flows between news sources sources is a central problem. Readers need to digest information rapidly from high volume news feeds which often contain duplicate and irrelevant stories. This thesis aims to track information flow from primary to secondary sources in Australian capital markets. Specifically, official announcements from the Australian Securities Exchange (ASX) and stories from the Reuters NewsScope Archive (RNA).

Our task is to identify information flow – the presence of information from one document in another. We present a scheme that codifies information flow and encodes the journalistic contribution of the story. This is whether a story: is the first to report an announcement, adds background details or adds novel analysis. In conjunction with the scheme, we present an annotation tool and show that the scheme can be applied to Finance text with high inter-annotator agreement.

We propose a novel approach to the information flow problem: as a text categorisation task over the intersection of a pair of texts. We use set-based lexical features and similarity scores to model similarity at fine and coarse grains. Word sequences found in both texts are represented to capture longer passages of repeated text. Numeric text is explicitly handled, with similarity modelling and distinctiveness of repeated numbers. News data is fundamentally time-based and we model the publication time of the texts. This broad range of features is combined using Maximum Entropy modelling for classification.

The final contribution of this thesis are that our system identifies information flow at 89.5% F-score and the three types of journalistic contribution at 73.4% to 85.7% F-score. It also performs well in direct evaluation against human performance. We have demonstrated that it is feasible to automatically track information flow in Finance text, which will have a substantial impact on automated financial surveillance and trading systems.

Acknowledgements

I would like to thank my supervisors James Curran and Ben Hachey for their invaluable support and advice over the year and often late into the night.

The Capital Markets CRC deserves credit for accommodating and supporting this research, particularly Maria Milosavljevic. Dominick Ng and Silvio Tannert were immensely helpful during the annotation project, preparing data and wrangling annotators.

Finally, I would like to thank my family and Kylie for their faultless support, without which, this would not have been possible.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

Financial news is an important resource for capital market participants and plays a central role in how they interact with the market. Participants that are alert and responsive to incoming information are at a financial advantage (Zaheer and Zaheer, 1997). It is thus important to understand how information flows in Financial text. Information flow is defined as when a fact from one document is found in another. The focus of this thesis is to investigate information flow in Finance text, specifically from primary to secondary news sources.

Our primary and secondary sources are the Australian Securities Exchange (ASX)[1] official announcements and the Reuters NewsScope Archive (RNA)[2]. Companies listed with the ASX must continuously disclose any information "a reasonable person would expect to have a material effect on the price or value of the entity's securities" (ASX, 2008). The official announcements are thus a primary, canonical source of Finance text news for the ASX and often the first place information is released. Secondary news sources such as RNA report information from many sources, including the ASX. They typically have a wide audience and, more importantly, identify and summarise key information from primary sources. Furthermore, they may add background material to contextualise the events for the reader, or feature novel commentary and analysis. The speed at which Reuters reports events and the value their stories add is vital to the reader. If information flow can be automatically identified, and duplicate or irrelevant stories removed, readers will be able to absorb and react to news more quickly and effectively.

The goal of this thesis is to identify information flow in Finance text. We frame the task as a pairwise text classification task, essentially over the intersection of the texts. Chapter 2 outlines existing research that informs our approach and covers a wide range of Natural Language Processing (NLP) areas. Chapter 3 discusses information sources in general and analyses ASX announcements and RNA stories along

[1] http://asx.com.au

[2] http://thomsonreuters.com/products_services/financial/financial_products/event_driven_trading/newsscope_archive

with their associated issues. In Chapter 4, we present the first main contribution: a scheme and tool for annotating information flow in Finance text. Both were developed in close consultation with our annotators to ensure maximum reliability and efficiency. The reliability of the task is demonstrated using agreement statistics and a held-out dataset is created so that our approach can be compared against human performance. Finally, the annotated datasets used for training classification models are outlined.

Chapter 5 shows how we model information flow. We use techniques drawn from different NLP research areas and combine them using Maximum Entropy modelling. The features presented comprise another contribution and use textual and temporal similarity to represent information flow. Our experimental methodology and results are presented in Chapter 6 and explain how models of information flow are compared and evaluated. As well as measuring experimental performance, we classify the held-out evaluation dataset to directly compare against human performance.

1.1 Contributions

This thesis contributes a new approach to the information flow problem in a Finance domain. The problem of identifying information flow is treated as a text classification task over the *intersection* of two documents. We also draw on techniques from other NLP research areas to provide a generalisable model of textual similarity.

We present a scheme and tool for annotating information flow. These are the result of substantial consultation with Finance annotators and were designed to allow efficient annotation and navigation of complex collections of stories. Good kappa scores show that the scheme can be applied reliably by an annotation team to provide training and evaluation data for automated approaches.

We successfully combine approaches from different NLP research areas to address the information flow problem. Textual and temporal features are combined using Maximum Entropy modelling to accurately represent information flow and journalistic contribution.

We are satisfied that our model performs well under experimental conditions. Furthermore, this performance makes it immediately useful within other systems. Tannert (2009) uses a prototype of our model in research into the Financial implications of information flow. Automatic identification of information flow allows statistical analysis of a much larger dataset than if the task had been performed manually.

A paper based on this research has been accepted to the Australasian Language Technology Workshop ALTA 2009, demonstrating that it makes a contribution to the wider field of NLP.

CHAPTER 2

Literature Review

The information flow problem overlaps with many other research areas. The core proposition is that textual similarity relates to semantic similarity; documents with the same text contain the same information. Related areas differ in the type of information described, their approaches and evaluation methodologies. We propose that sentences are ultimately the most appropriate textual unit for information flow. Words are fine-grained and prone to ambiguity whereas long documents can cover many items of information. Information typically consists of references to specific events and facts; sentences are a proxy for these. Although there is not always a one-to-one mapping between sentence and information, this is true for the majority of cases. Documents, however, are the traditional unit of news consumption and the format in which experimental data is often made available. Furthermore, many evaluation techniques use collections of documents. Many similarity algorithms involve matching, which can be computationally intensive and so any reduction in the number of comparable items is advantageous. For example, if we are trying to identify similar sentences, then each sentence should be compared against the others, yielding far more comparisons than document-wise matching would. While we investigate sentence-oriented approaches, pragmatic considerations such as performance and evaluation mean that our study focusses on document-level textual similarity.

This chapter details some of the existing approaches which can be used to model information flow. Information Retrieval (IR) provides robust document-level analysis and is used as a baseline approach. The Text REtrieval Conference (TREC) has been the context for research that aims to capture events and information as they are published in a news source, specifically the Topic Detection and Tracking (TDT) task and the Novelty track. Textual similarity is central to Plagiarism Detection, the problem of identifying copied (and perhaps slightly edited) text. Sentence alignment re-phrases the task as a preliminary stage of a Machine Translation training process whereby texts in different languages must be aligned, often through non-linguistic means, before linguistic analysis can begin. Perhaps the most directly relevant area of research to information flow is that of text reuse, which examines reuse of

specific fragments of text. Apart from textual similarity, we briefly explore research that addresses the temporal aspect of information flow, particularly through large systems.

2.1 Information Retrieval

Information Retrieval provides many techniques useful for the text similarity problem. Perhaps the most simplistic of these is *bag-of-words*, which treats a document as an unordered collection of its words that acts as "a quantitative digest" (Manning et al., 2008). Weighting functions map words onto real values and can range from simple methods, noting whether a particular word appears in a document, to more sophisticated metrics that account for variations in document length and word distribution across a collection of documents (i.e., a corpus).

Word distribution presents further issues to statistical NLP techniques. Zipf's law is a reasonably accurate model for word distribution. It states that the frequency of a word in a corpus decreases rapidly with its frequency rank such that the frequency of the *ith* most common word is proportional to *1/i*. The most important consequence is that words that serve a grammatical function will be overwhelmingly more common than information-bearing words.

$$\text{tfidf}_{t,d} = \text{tf}_{t,d} \cdot \log \frac{N}{\text{df}_t} \tag{2.1}$$

Equation 2.1 shows the TFIDF weighting function (Spärck Jones, 1973), which states that the weight of a term (i.e., word) in a document is the product of the term frequency (TF) in that document and the inverse document frequency (IDF). The IDF component is the negative log of the proportion of documents the term appears in (*N* is the number of documents in the corpus). Its role is to penalise commonly occurring terms (e.g., "the") and consequently rewards terms that are characteristic of the document they appear in. Bags-of-words with TFIDF weightings are extremely pervasive and often used as a baseline approach.

Vector space models provide a metric for measuring the similarity between two documents. The bags-of-words calculated above are treated as vectors, where words map to a (possibly weighted) count. To calculate similarity between two documents, a vector is created for each document using the union of the words in both. Words that appear in only one document will take a zero value in the other vector.

$$\text{similarity} = \cos\theta = \frac{A \cdot B}{||A||||B||} \tag{2.2}$$

Equation 2.2 shows how cosine similarity (Salton et al., 1975) is calculated from the two vectors. It states that similarity between two vectors can be modelled by the dot product divided by the product of euclidean lengths (which normalises for different document lengths). Thus a pair of identical documents would have high scores, reflecting the cosine of the zero angle between their vectors. When applied to bags-of-words, the cosine function will range from 0 to 1 since word counts cannot be negative.

As well as scaling the word frequency, it is also common to subject the words themselves to shallow linguistic processing such as stemming and stop-word removal. Stemming attempts to account for the fact that different words sharing a common stem refer to the same concept and so ideally a stemming process might convert all of the following *buying, buyer, buyers* to *buy*. Though this technique does not account well for unrelated words with shared prefixes, for instance *general, generation*, it requires less intensive processing than deeper linguistic techniques. Stop-word removal is a pre-processing step whereby grammatical function words are simply removed. As mentioned above, these tend to be far more frequent than information-bearing content words. This can remove words that are critical for deeper linguistic analysis such as discourse markers, but is nonetheless an effective IR technique.

We frame the information flow task as supervised binary classification. Suppose we have a task to predict the class of instances: *A* or *not A*, where we know the actual class for each instance. To evaluate performance, precision, recall and F-score are calculated from the count of true positives, false positives and false negatives. True positives are instances that are correctly classified, *A* or *not A*. False positives are *not A* instances which the model classifies as *A*, where false negatives are *A* instances which the model misses, classifying them *not A*. Equation 2.3 shows how precision is calculated and is the proportion of instances classified as *A* that were actually *A*. Equation 2.4 defines recall, this is a measure of how many *A* instances we managed to retrieve and classify as *A*. In a classification task, it is trivial to achieve total recall by classifying all instances as *A*, but this will come at a cost to precision, since some instances will be classified incorrectly (assuming they are not all *A* to begin with). F-score is designed to capture both precision and recall and is the weighted harmonic mean of the two, shown in equation 2.5. It should also be noted that while precision, recall and F-score are most often used for evaluation, different task constraints may preclude their use or may require other techniques. This thesis makes extensive use of these metrics to measure classification performance.

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{2.3}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{2.4}$$

$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.5}$$

2.2 Plagiarism Detection

Plagiarism detection contributes a slightly different interpretation of similarity, concentrating first and foremost upon text, the trivial example being a document that is an exact copy of another. Approaches that rely on exact string matches can be easily circumvented by inserting or deleting single words. A plagiarist's strategy relies on maintaining the semantic similarity through minimal effort expended on textual changes and its detection has led to more sophisticated methods defining less exact similarity. The problem is distinctive also from the point of view that it can require almost *n-by-n* document comparison since each document can be similar to another and though fingerprint-based techniques use indices to reduce the computational complexity to *n-by-f* document to fingerprint comparisons, the major computational challenges have resulted in techniques that place a stong emphasis on efficiency.

The COPS system (Brin et al., 1995), takes the fingerprint approach whereby documents are separated into meaningful units, typically sentences, and sequences of these units known as chunks are hashed for later comparison against new documents. It is important to note that the chunk selection strategy does not necessarily use logical unit sequences (i.e., contiguous sentences) and Brin et al. experiment with variations in chunk selection for fingerprint building and sampling for retrieval. The ability to constrain the comparison problem to between documents and fingerprints is crucial to scaling such a system to large collections of documents as are the techniques used to combine chunk data into fingerprints to represent document similarity.

Closely related to this is the detection of co-derivative documents, a set of documents that are derived from a common source document which may also be in the set. One example of this data was examined in Red Hat Linux documentation, where "long blocks of text are largely preserved, but possibly with intermittent modifications, and some original text is added"(Hoad and Zobel, 2003). IR-based ranking

approaches are compared to fingerprinting over the task of returning a set of documents co-derivative to 53 manually chosen query documents from more than 80,000 documents. The fingerprinting approaches differ in unit granularity and chunk selection strategy. The ranking approaches included standard IR techniques such as cosine similarity, but also added *identity* measures oriented towards the comparison of similarly-sized documents, as opposed to the traditional query-document comparison, following the intuition that "similar documents should contain similar numbers of occurrences of words". The system uses five identity measures that vary the extent to which they reward documents with similar word-occurrence patterns. An absolute score representing the similarity to the query document allows further evaluation beyond precision and recall. The study defines the notion of *Highest False Match* (HFM), the highest score of an incorrect document and *separation*, the difference between the lowest correct result and HFM, measured after the retrieval of *n* documents to together characterise a system's effectiveness. The interaction between HFM and separation is such that the ratio of HFM to separation is a useful overall indicator, since a high HFM is acceptable as long as separation is also high. The results show that while cosine-based ranking approaches were unsuccessful, an identity measure has the highest precision, recall (measured after 20 documents) and HFM/separation ratio with 97%, 97% and 2.05. The best fingerprinting approach manages 92%, 92% and 1.14 and only after considerable effort tuning unit granularity.

Plagiarism Detection requires the analysis of similarly-sized texts with an emphasis on textual similarity and efficiency. It is somewhat at odds with other textual approaches in that topic, or in fact any notion of the *type* of information, is immaterial.

2.3 Text Reuse

Text Reuse is concerned with identifying where short segments of text are reused in different documents and is closely related to information flow. Metzler et al. (2005) define a *similarity spectrum* from exact text matches and broad relatedness. Motivated by the concentration of research at either end of this spectrum, the aim is to track text and facts through corpora, expecting them to be "in general, cohesive structural units" best found in sentences. As a framework to explore different approaches, they define a similarity function that takes a query (Q) and candidate sentence (R), returning a similarity score. Word overlap as defined in Equation 2.6 is used as a baseline, defining the size of the intersection of words in the query and candidate sentences divided by the size of the set of query words. Equation 2.7 defines an overlap scaled by the cumulative IDF of each word in the intersection to up-rank 'interesting'

TABLE 2.1: Sentence similarity spectrum.

Rank	Description
5	Exact match
4	Minor revision
3	Major revision
2	Specific topic
1	General topic
0	Unrelated

words. A reformulation of TFIDF from the TREC Novelty track (discussed later in this chapter) is used to weight the similarity by word distribution and is defined in Equation 2.8. This sums, for each word in the intersection, the weighted product of the word's TF in Q, TF in R and IDF component. The equations for the more complex measures are not stated here, but these include a relative frequency measure from co-derivative research and two statistical language models that treat reuse as a translation process.

$$S(Q,R) = \frac{|Q \cap R|}{|Q|} \tag{2.6}$$

$$S(Q,R) = \frac{|Q \cap R|}{|Q|}(\sum_{w \in Q \cap R} log\frac{N}{df_w}) \tag{2.7}$$

$$S(Q,R) = \sum_{w \in Q \cap R} log(texttf_{w,Q} + 1)log(\mathrm{tf}_{w,R} + 1)log(\frac{N+1}{\mathrm{df}_w + 0.5}) \tag{2.8}$$

Using 50 single-sentence topic queries from the TREC question answering track, they evaluate each system over a collected TREC newswire corpus of 848,481 documents including Associated Press, Wall Street Journal and Financial Times stories. The documents have stop-words removed and the top 25 ranked results per query were assessed by two of the authors as to their position on a six category similarity scale (see table 2.1).

The principal metric for evaluation is mean average precision (MAP), the mean of the precision measured after each of the 25 documents were retrieved. The highest MAP increases from approximately 57% to 97% as the similarity spectrum is climbed. The relative ordering of the measures at different similarity ranks does not suggest that any particular approach was clearly better than the other, especially given the baseline word-overlap technique's good performance.

TABLE 2.2: Document similarity spectrum.

Rank	Description
3	Identical
2	Sufficient overlap that there must have been common source material
1	Some overlap, but nothing to suggest that there was common source material
0	Unrelated

To apply sentence similarity in a bottom-up fashion to documents, they evaluate several approaches to combining scores for sentences in each document, finally choosing a metric that combines only the best sentence-wise similarity scores between the two candidate documents. These are evaluated against baselines of a language-model derived query likelihood function for identifying topical similarity and DECO, a system for identifying co-derivative documents (Bernstein and Zobel, 2004). The corpus was drawn from 40 TREC newswire documents known to share facts and evaluated in the same manner as above. Rather than using the fine-grained similarity spectrum (see table 2.1), a less detailed document similarity spectrum is adopted (see table 2.2).

The bottom-up method does not improve on the baselines, though the gap is less at the second rank of the document similarity spectrum. This bodes well for future work improving both the sentence similarity functions and the algorithm to combine them. (Metzler et al., 2005)'s main contribution is to provide a framework within which to describe different levels of similarity at sentence and document levels. Although the results at a document level are somewhat disappointing, it is, at least validation that it is promising area of research.

The problem of identifying text reuse can also be approached as a text classification task. The MEasuring TExt Reuse (METER) corpus concentrates on newswire text reused in newspaper stories (Clough et al., 2002). The motivation behind this is to identify newspaper stories that reuse newspaper copy, essentially a form of legitimate plagiarism. A number of explanations for copy-use are proposed, including short deadlines, prescriptive writing practices, size limits, readability and house styles. Indeed, the same copy can be presented differently depending on whether the story is to appear in a tabloid or broadsheet newspaper. Automated text reuse detection might allow newswire companies to measure the extent to which their stories are used by other parties (not only newspapers). Court and entertainment stories are taken from the UK Press Association newswires and nine British newspapers and classified as wholly, partially or not derived from the newswires. 77% of their 944 newspaper stories were thought to have used part of the 772 copy stories in some way.

To represent text reuse, three techniques from different areas are used to score similarity between the copy and story: n-gram overlap, greedy-string tiling and sentence alignment. N-gram overlap is simply the size of the intersection of both documents' tokens divided by the size of the story's token set. The choice of denominator is significant for this particular problem, since this would allow a perfect overlap of 1 if all n-grams in the story appeared in the copy. They also explore a variant that only considers hapax legomena (words that appear only once) from the copy that overlap. Drawn from plagiarism detection, greedy string tiling efficiently finds larger "tiles" of common subsequences and is robust to small changes to conceal plagiarism (Wise, 1996). The ratio of total length of tiles against the size of the story forms a similarity score. Finally, three weighted sentence alignment metrics are summed to produce a third similarity score.

These similarity scores are used as features for experiments using Naïve Bayes classifiers. Initial experiments simply used the score from one of the metrics as the single feature in a three class problem (wholly/partially/not derived) and each approach exceeded baseline performance with 77.4% and 65.4% F-score for wholly and partially derived. Interestingly, the hapax n-gram variant performing as well or better than more complicated tiling or alignment approaches. By cascading classifiers and attempting the 'easiest' class first, their system performs at 70.3% F-score over all classes, significantly better than the best single-feature equivalent. Possible reasons suggested for the relative utility of the simple overlap approaches are that newspaper stories are heavily edited and, as such, approaches that treat words in isolation are more suited to the problem.

In sum, the METER project is interesting since it explores textual similarity through text classification and employs a variety of approaches to journalistic data with an aim to analyse news production and consumption.

2.4 Other Textual Methods

The above methods are those most related to information flow, the fundamental nature of textual similarity means that many areas potentially contribute. Although most work in textual similarity places strong emphasis on the words themselves, methods that do not consider them can be, paradoxically, quite effective.

Sentence alignment is typically performed over parallel corpora - collections of documents and their translations - and may be used as the first pass of an automated Machine Translation training process

before further analysis. One challenge, of course, posed by parallel corpora in different languages is their textual dissimilarity leading to a number of approaches based on sentence length. Brown et al. (1991) used the sentence lengths (measured in words) to align English and French sentences from Canadian Parliamentary Hansards and despite use of document-structure anchors to pre-align sets of sentences, reported good alignment results. The corpora were treated as sequences of sentence lengths punctuated by paragraph markers (¶) and their alignment modelled as beads. The eight categories of beads specified the ways in which sentences could align: *e, f, ef, eef, eff, ¶e, ¶f, ¶e, ¶f* where *eff* refers to an English sentence aligning with two French sentences. Probabilities of generating sentences of a given length from each corpus were calculated and it was assumed that the log of the ratio of English lengths to French lengths was normally distributed. Brown et al. developed a hidden markov model (HMM), a statistical model that assumes an unseen (i.e., hidden) state machine generating the observed output, in their case, alignments. They used an expectation maximisation algorithm to estimate HMM parameters and from them found that the sentence length correlations might be sufficiently strong to align solely on their basis with a low error rate.

For evaluation, a contrived corpus is generated then automatically aligned, the error rate being the fraction of automatically aligned *ef* beads that didn't correspond to the generated alignments. Over thousands of simulations, the base error rate is 0.9% and omitting anchors or paragraph raises this to 2% and 2.3% respectively. Omitting both leads to an error rate of 3.2% which supports the EM parameter correlations. Application to the Hansard corpora and analysis of a sample of 1000 sentence pairs reveals 6 mis-alignments, consistent with the 0.9% error rate above. In related work, Gale and Church (1991) uses the number of characters in a sentences, addressing a claim that character-based metrics are less varied than word-based equivalents. Gale and Church (1991)'s work in Cantonese and English has demonstrates that this approach is not limited to pairs of languages belonging to Indo-European language families such as English and French. The consideration of content at only structural levels has the advantage that it is quick and efficient to compute and language independent. Though it seems unlikely that the correspondence between primary and secondary sources is as strong as direct translation, especially given journalistic input, the success of these relatively simple approaches is compelling.

The scope of this thesis does not extend to interpreting the meaning of flowing information, merely its existence. Once the latter problem is tackled, semantic analysis of the information would be a natural next step. Dagan et al. (2006) define the problem of Recognising Textual Entailment as "recognizing, given two text fragments, whether the meaning of one text can be inferred from (be entailed by) the

other.". Since this essentially refers to the mapping between text and meaning, it is central to many NLP areas and is, as such, the focus of much current research activity. Though the end-result of this task addresses semantic inference, a pre-requisite is that the fragments refer to the same unit of information and it is clear that better understanding of information flow might eventually benefit textual entailment research.

2.5 Topic Detection and Tracking

The research areas presented above consider their corpora as a static resource. Finance news, in common with all news text, has a strong temporal dimension that is relevant to information flow. The Topic Detection and Tracking (TDT) task was part of the TREC programme between 1997 and 2000 (Allan et al., 1998a). Its overall goal was to organise news information, but event tracking was an initial focus. Events are defined as "something that happens at a particular time and place" (Allan et al., 1998b), though the authors note that this definition can present issues in the case of long-running events that occur in many places. Event *identity* – the notion of similarity between events – is less problematic. Hence an event might be "The Eruption of Mt. Pinatubo on June 15th, 1991" and the task requires its detection and tracking through a body of news. The news formats include transcribed speech from broadcast news and newswire story text.

There are three distinct subtasks of the TDT pilot: tracking of known events, detection of unknown events and segmentation of a news source into stories. Each task is designed to provide a context to approach the natural components of a TDT application. The task corpus consists of nearly 16,000 stories evenly drawn from CNN broadcast news transcripts and Reuters newswire from July 1994 up to and including June 1995. 25 events were identified from the corpus and documents were tagged with respect to their coverage of each event: *YES*, *NO* or *BRIEF* where less than 10% of the story covered the event.

The tracking subtask is more similar to a traditional IR task, given a set of documents describing an event, the system has to identify further references in the corpus. The 'seed' set of event documents is not available in the detection task and is more similar to an unsupervised clustering task. Different methods were used to approach the retrospective and on-line versions of the task, the latter allowing access to each document as it arrives. Segmentation tackles the issue of segmenting streams of data into distinct stories. This is particularly useful for the CNN transcript data, since a transcript of a news

television program may discuss different stories and needs to be segmented before further processing. This suggests how stories might be segmented to identify information flow at a sub-document level.

Evaluation techniques vary substantially across the different tasks – the discussion here focusses on retrospective event detection. The output of a system is a description of how the documents clustered together. Evaluation of these systems requires matching these clusters against hand-annotated clusters. Allowing documents to belong to more than one cluster might have increased performance in some cases, but this is limited to one to simplify evaluation. Once the systems' choices are identified, miss rate, false alarm rate, precision, recall and F-score were calculated. Decision Error Trade-off curves (Martin et al., 1997) are used to show the trade-off between misses and false alarms on the y and x axes respectively. Since both percentages increase from the origin, better performing systems plot closer to the axes and the curves have the advantage over other methods of clearly describing the trade-off between false positives and false negatives.

The pilot phase of the TDT task has produced encouraging results, with segmentation shown to be tractable, and clustering shown to perform well within detection subtasks. Allan et al. (1998b) implement an online tracking system which examines each story on arrival. A query representation is built and compared against previous queries in the system. Queries with a match below a threshold are assumed to refer to new topics and the query added to memory before handling the next document in the stream. Their approach has a strong focus on the temporal nature of news stories. For instance, thresholds including a time penalty factor are developed for each query accounting for the fact that stories released in quick succession are likely to refer to the same events. They also incrementally update the IDF component of a TFIDF measure to model word use at a point in time. Using this method, words that have not been seen before are likely to discuss new events, while those that have not been used recently may indicate new stories in an existing event.

Later TDT tasks add *first story detection*, an extension of detection that returns the first rather than all matching documents and *link detection*. Given a pair of documents, a link detection system will predict whether they are linked or not. This task is analogous to our framing of the information flow problem as a pairwise classification task. In our case, we are detecting information flow links between ASX announcements and RNA stories.

Lavrenko et al. (2002) use relevance models for the link detection task to perform query expansion, increasing the scope for overlap between story text. This assumes that the co-occurrence of a word and

a query can approximate the probability of observing that word in the documents relevant to that query. In the case of the link detection problem, a model is calculated for each document and their similarity is assessed using information gain. The notion of query clarity, the extent to which the model diverges from a model of general English is also used and taken into account in the final similarity score. Lavrenko et al. (2002)'s relevance model system outperforms a vector-space model using cosine similarity and TFIDF.

Other approaches explore linguistically motivated features in judging similarity between short texts of similar size. Hatzivassiloglou et al. (1999)'s rationale is that shorter units of text required deeper exploration of semantic features such as word co-occurrence, noun phrase heads (i.e., the *subject* of a phrase), synonyms and verb semantic classes. Combinations of these features are used with a rule induction classifier over the Reuters part of the TDT pilot corpus to assess whether pairs of paragraphs were similar. Two reviewers manually performed this task over 10,345 pairs of paragraphs and attained favourable kappa agreement scores (Cohen, 1960). Random cases were also manually checked by a further two reviewers. There are known issues in applying kappa to agreement on *rare* events, in this case approximately 3% of all events, and their kappa of 0.59 may be more fairly represented using later modifications to kappa by Di Eugenio and Glass (2004). When measured against a baseline system using TFIDF features, linguistically-motivated features substantially increase recall from 32.6% to 60.5% at equivalent precision.

Moreover, TDT has been used as a component technology for other NLP applications such as Newsblaster (McKeown et al., 2002). After identifying documents that match a topic, the system summarises their content. Pre-classification into broad categories (e.g.*US, Financial, International, Entertainment, Science and Technology, Sports*) is achieved using cosine similarity. A secondary stage of clustering uses TFIDF and some linguistically motivated features such as proper nouns and noun phase heads. The authors report comparable performance with other participants in TDT-2. This demonstrates a potential use-case where an information flow detection engine might drive a Finance-specific news application. Overall, the TDT task has added elicited solutions that merge textual and temporal similarity, something that is central to our approach to tracking information flow.

2.6 Novelty Detection

Related to TDT is the TREC Novelty track that ran from 2002 to 2004 (Soboroff, 2004) which models "an application where a user is skimming a set of documents, and the system highlights new, on-topic information". The input data is a corpus of documents, each of them associated with a topic description and realised as a sequence of sentences. Systems first complete the task of identifying sentences relevant to the topic, similar to the TDT task, except addressing smaller, sentence-level units of text. This marks a shift from document emphasis of the TDT task and requires systems to handle smaller textual units. The second task requires identifying those sentences that present novel information about the topic. This extension is vital for tracking long-running events where there is a high degree of repetition and the most important sentences are those that include new information. Although this thesis considers information flow at a document level, this finer-grained approach should be the eventual goal of an information flow system.

The final year of the novelty task made the substantial change of adding irrelevant documents to the corpus, which was drawn from the AQUAINT collection of three news source: New York Times, Xinhua and Associated Press, with the period between June 1998 and September 2000 covered by all sources. This additional redundancy, combined with the irrelevant documents led to a lower proportion of novel information makes the task more difficult. Assessors created 50 topics: 25 events and 25 opinions on contentious issues and then retrieved 25 relevant documents per topic. The dataset also includes some irrelevant documents that are close content matches to the relevant documents. The four subtasks vary in the amount of relevant and novel sentences available for training and are designed to allow participants to address relevance and novelty detection separately. The sentences were manually judged by two assessors with respect to relevance and novelty, and resulted in 19.2% of all sentences judged as relevant and 42% of those judged as novel. The second assessor only viewed sentences judged relevant by the first assessor. The system is evaluated against the second assessor's judgements using precision, recall and F-score. In this case, the assessor's judgements play the role of real classes.

Most approaches model relevance as similarity to topic and novelty as dissimilarity to prior relevant sentences. Substantial use is made of IR techniques such as TFIDF, though some participants explore linguistic features, for example, Litkowski (2003) use discourse-related nouns and verbs to increase recall when identifying relevant sentences , while Eichmann et al. (2003) use named entities (e.g., people, places and organisations) and synonyms. Overall, most systems operate with low precision and high recall, essentially identifying too many sentences as novel. In searching for an explanation of this,

Soboroff (2004) suggests that inclusion of irrelevant documents could not be the only explanation and perhaps over-tuning to the previous dataset is more important. F-measures range from 36% to 40% for relevant sentence identification and between 18% and 21% for novel sentence identification. Precision is uniformly low and recall is higher and more varied. Event topics are more successfully identified than opinions. The latter seem to require all relevant seed sentences for good performance. Additional training data does not seem to improve relevant sentence retrieval. Although Novelty detection has been shown to be a difficult problem, its suggests that deeper linguistic analysis is essential to identify information flow at a finer-grained level than between documents.

2.7 Other Temporal Methods

Apart from TDT and Novelty Detection, the above areas almost exclusively address textual similarity. Other areas place more emphasis on the timing and structure of information flow. The internet is now a major news medium and is the subject of research into modelling flow and diffusion of information through it. Much of this research focuses on hyperlink analysis and deeper examination of the temporal nature of information sources. These "information streams" can be effectively modelled as bursty time-series (Kleinberg, 2003) using a hidden markov model over hidden states that specify an emission rate. The states are zero-indexed at the base state and each subsequent state has an increased index and exponentially higher rate. Intensity of emission j is modelled by periods at which the state index is higher than j and burstiness characterised by the skipping of states. Once a collection of emitted documents has been collected and analysed, a inverted index can be build that maps words to bursts involving the documents they appeared in. Enumerating bursts ranked by their burstiness tends to reveal emerging technical terms and language change. As well as this, "landmark" documents, those which mark the start of bursts of particular words, might be viewed as analogous to first story detection within the TDT area. Other analyses have considered information flow as an epidemic (Gruhl et al., 2004) to further model the interactions between more than a pair of information sources. Amongst other features such as hyperlinks, commonly occurring sequences of words and proper nouns, they used a modified TFIDF (TCIDF) to take into account average *cumulative* word use in previous days.

2.8 Summary

In conclusion, tracking information flow is an important problem, especially in the context of capital markets. Different research areas have addressed issues of textual and temporal similarity and inform our approach. IR offers robust techniques such as bags of words and cosine similarity, as well as solid metrics for evaluation. Plagiarism Detection examines ways of modelling longer sequences of text, while Text Reuse explores combining a broad range of similarity metrics. TDT and Novelty Detection address the temporal nature of information flow at a document and sentence levels. The following chapters will outline how we draw on this research to track information flow.

CHAPTER 3

Data

Financial information flows to market participants in many different forms. These range from stock prices and other numeric data to less structured textual sources which can be more challenging to interpret. Participants depend on the ability to process high volumes of incoming information and make decisions quickly to maintain an advantage over their competition. This chapter describes the data over which we track information flow. Firstly, we discuss the primary and secondary sources that we use: Australian Securities Exchange (ASX) announcements and Reuters NewsScope Archive (RNA) stories. We then compare them in terms of volume, content and publication timing. Finally, we outline how we process the trade data that supports our annotation process.

3.1 News sources

Finance text plays an important role in capital markets and is central to tracking information flow. This thesis is concerned with the relationship between primary and secondary sources of information. By definition in the continuous disclosure regulations, the ASX announcements are considered the primary, canonical source for ASX information. News agencies are important secondary sources since their stories draw information from many sources, including the primary source. Moreover, they summarise and add further content, making them more valuable (although less timely). The relationship between the ASX and RNA sources is a microcosm of the wider news ecosystem. The entire system can be modelled as a graph of sources that receive and emit information, possibly altering content or adding information of their own. It is significant that the ASX is the canonical source, since it allows us to study the ASX and one other secondary source (i.e., RNA) in isolation. Indeed, this is not possible in markets that do not require centralised disclosure. There are, of course, edge cases to consider. The proposition that all price sensitive information is released to the market via the ASX official announcements is optimistic at best. Insider trading occurs when information is leaked to some market participants (or already known

Year	ASX	RNA Stories
2003	66,233	1,901,722
2004	80,570	1,954,259
2005	90,484	2,053,525
2006	102,235	2,298,462

TABLE 3.1: Document count per for complete ASX and RNA datasets from 2003 to 2006.

to insiders) before others and they take undue advantage. This is signalled when a stock price rapidly increases or decreases *before* the release of an announcement. There is also potential for more complex patterns of information flow. For example, an RNA story might cause the ASX to request clarification from a company concerning a published announcement. The request and response themselves will take the form of announcements. Despite these minor issues, simplifying the information flow relationship is advantageous and allows deeper study of the content and timing.

To develop automated approaches, we first collect documents from primary and secondary sources covering the same time span and stock tickers. In addition to the documents' textual content, metadata that specifies the title, related stock tickers and publication time are collected. The textual content from each source poses different extraction and preprocessing challenges that are detailed in the next sections. Sirca[1] provides finance-related data to the academic and private sector for research under a subscription model. Sirca supplies the primary and secondary source data used in this study: Australian Securities Exchange (ASX) official announcements and Reuters NewsScope Archive (RNA).

Table 3.1 shows the increasing volume of ASX and RNA documents over time. This trend, even when we study just two sources, shows the vast scope of the information flow problem. These RNA story counts include every story that Reuters releases globally and the counts are correspondingly higher than the ASX. Reuters is a large newswire company and so we make make the assumption that they report newsworthy events promptly. Newsworthiness is a loose concept and we do not intend to define it further here, suffice to say that we expect Reuters to report on the most important events of the day for the most important companies as and when they occur. Consequently, only some of the ASX announcements will be reported, but those that are reported sooner are more likely to be important. However, despite this implicit filtering there is still a high volume of stories for readers to navigate. An extra benefit of identifying information flow is that redundant stories within the *same* source can be filtered for the reader.

[1] http://www.sirca.org.au

Source	Count	Text (%)
ASX	10,404	83.9
RNA	8,277	99.6

TABLE 3.2: Document type distribution and text coverage in the *sample* corpus.

Year	Faxed	Scanned
2003	12.1	3.1
2004	9.5	3.1
2005	7.9	1.9
2006	6.6	1.6

TABLE 3.3: Percentage of faxed and scanned ASX announcements by year in the complete datasets from 2003 to 2006 (a superset of the announcements included in the *sample* corpus.

To study information flow, we created a *sample* corpus of ASX announcements and RNA stories from 18 months between 2005 and mid-2006. The ASX200 is an index published by Standard and Poors[2] that shows the top 200 companies by value at a point in time. To capture the most important companies over our sample timespan, we obtained the ASX200 list on the last day of each year from 2002 to 2008. 403 tickers appear at least once in these lists and are highly traded and closely followed companies. Table 3.2 shows the number of documents from each source in our initial dataset. It also shows the proportion of document where we were able to extract text. The following sections explain specific features of the ASX and RNA documents and discuss extraction issues.

3.1.1 ASX Announcements

The broad scope of the ASX's continuous disclosure rules means that almost any type of document can be part of an announcement. While these are all in PDF format after 2002, the dataset includes short letters, corporations law forms, long annual reports and presentation slides. We used the PDFBox[3] Java libraries to extract text from 83.9% of the PDF documents in our experiment dataset. The remaining 16.1% are made up of cases where errors occur in the extraction process, the extracted text is empty or the documents were marked as faxed or scanned.

Faxed and scanned announcements pose major problems to text extraction, since they typically have poor image quality and would produce noisy text. SIRCA marks these documents using a filename suffix so

[2] http://www.standardandpoors.com
[3] http://incubator.apache.org/pdfbox

it is possible to isolate them from the other announcements. Initial experiments using Optical Character Recognition did not produce accurate results. These documents were considered to have missing text, but not removed from the dataset to maintain as much realism as possible. Table 3.3 shows that the proportion of faxed and scanned announcements is decreasing, which is encouraging for any work in the future. In addition, the ASX text can contain artefacts of the PDF extraction process that cause problems for any subsequent text processing. For example, the extracted text does not contain newline characters, and sometimes has inconsistent whitespace placement. The following snippet from PDFBox shows that simple tokenisation will fail to capture information about the company name.

O I L S E A R C H L I M I T E D (Incorporated in Papua New Guinea)

This is problematic for two reasons, firstly that the company name 'Oil Search Limited' is made up of isolated capital letters. Secondly, the fact that capital letters are used is evidence that the text is significant. Our processing ignored all extraction issues issues since fixing them would be non-trivial and we were able to extract sufficient text using simple methods.

Figure 3.1 shows the second page of an ASX announcement from Rinker, a materials company. The entire document starts with a cover letter page addressed to the ASX followed by three pages of trading update textual content, then 11 pages of supporting tables and financial figures. This illustrates some of the difficulties involved in processing announcements. Often the useful information can be found in just part of the document, not necessarily the first page. Specific terminology is also used, for example EBIT, EBITDA and ROFE and the lack of a consistent author or editing guidelines means that a broad range of language and formatting features are found.

Sirca includes plain-text metadata with each announcement, specifying the publishing timestamp, title and related tickers of each announcement. The ASX labels announcements according to their price sensitivity with respect to different tickers. For example, an announcement that discusses two companies may contain information that sensitive to only one of them. 23% of ASX announcements from 2003 – 2006 are marked as price sensitive. 92.2% of announcements are associated with one ticker, which means that when we consider an ASX announcement at a document level, it is unlikely to contain confounding information about different companies. Table 3.4 shows how the ASX's Comnews categorisation scheme

RINKER™

News Release

Rinker third quarter net earnings up 33% in US$, 25% in A$

Rinker Group Limited (Rinker) today announced net profit after tax (PAT) for the three months ended 31 December 2004 of US$95 million[1], up 33% (A$124 million, up 25%) on the December quarter 2003.

Earnings per share (EPS)[2] rose 35% to 10.1 US cents (13.1 Australian cents, up 26%). Earnings per ADR were US$1.01. Earnings per share prior to the amortisation of goodwill[2] were 11.4 US cents (up 30%) and 14.9 Australian cents (up 21%).

Other key measures:

- Earnings before interest and tax, depreciation and amortisation (EBITDA)[3] rose 26% to US$227 million (A$297 million; up 19%)
- Earnings before interest and tax (EBIT)[3] rose 34% to US$163 million (A$213 million; up 26%)
- Trading revenue was up 18% to US$1,095 million (A$1,426 million; up 11%)
- Return on funds employed (ROFE)[4] was 21.2% in US$ for the year to December, up from 16.3% for the prior year (22.6% in A$, up from 18.6%). All business segments sharply increased ROFE.
- Return on equity[5], pre goodwill amortisation, was 17.1% in US$ for the year to December, up 2 pp from 15.1% for the prior year (18.2%, up from 17.1% in A$)

Rinker Materials Corporation in the US, which produces around 80% of group earnings and revenue, delivered another consistently strong performance, with US$ EBITDA up 31%, and revenue up 19%. Readymix EBITDA rose 2% in local A$ currency, with revenue up 10%. Normalised EBITDA – excluding a A$3.5 million land sale in the December quarter 2003 – grew 8%.

For the nine months to end December, Rinker PAT was up 36% to US$302 million (A$418 million, up 26%). EPS rose 36% to 32.1 US cents (44.3 Australian cents, up 26%). EPS prior to goodwill amortisation rose 30% to 36.0 US cents (49.8 Australian cents, up 21%).

- EBITDA rose 26% to US$697 million (A$961 million; up 17%),
- EBIT was up 34% to US$509 million (A$703 million; up 24%)
- Trading revenue rose 16% to US$3,228 million (A$4,447 million, up 8%), and
- Operating cash flow was steady at US$476 million, with free cash flow[6] down slightly due mainly to the timing of tax paid and volume driven increases in working capital.

For the nine months, Rinker Materials' EBITDA was up 24% and EBIT up 35% in US$. In A$, Readymix EBITDA rose 23% and EBIT 21%.

"All of the group businesses continued to grow and improve their performance again during the quarter," said Rinker CEO David Clarke. "The strong improvement in ROFE supports our commitment to working our assets harder and using our capital as efficiently as possible."

The group's balance sheet continued to strengthen. Net debt[7] at end December was US$416 million (A$535 million), down 31% from US$601 million (A$796 million) at the year-end in March. EBIT interest cover[8] for the 12 months to end December was 20.7 times in US$. Gearing or leverage (net debt/net debt plus equity)[9] was 14% while net debt/equity[9] was 17%.

FIGURE 3.1: The second page of an ASX announcement from the company Rinker titled: "Trading Update for the Quarter ended 31 December 2004"

Comnews main category	Count
Takeover announcement	5,784
Security holder details	65,242
Periodic reports	113,489
Quarterly activities report	7,373
Quarterly cash flow report	5,827
Issued capital	67,998
Asset acquisition & disposal	8,859
Notice of meeting	23,605
ASX announcement	9,823
Dividend announcement	17,886
Progress report	38,832
Company administration	18,116
Notice of call (contributing shares)	21
Other	16,898
Chairman's address	5,244
Letter to shareholders	4,830
ASX query	1,410
Structured products	11,156
Commitments test entity quarterly reports	4,265

TABLE 3.4: Count of the main Comnews category labels applied to ASX announcements from 2003 to 2006. ASX announcements can be tagged with multiple labels and we aggregate sub-category counts. Categories named 'Reserved for Future Use' are omitted.

[4] is applied to the announcements. The scheme is a two-level hierarchy and each announcement is labelled with up to 10 subcategories depending on its broad topic. The table shows the aggregated counts of the labels applied to ASX announcements in each main category.

The counts vary substantially over the categories and demonstrate the range of document types found in the ASX announcements. Reports of different types make up the majority and, depending on the company, these can cover many different topics and events. There are substantial numbers of security-related categories such as 'Security holder details', 'Issued capital' and 'Dividend announcement'. In summary, the ASX announcements are a corpus that is diverse in format and content.

[4] See the Comnews manual for more details: `https://www.asxonline.com/marketinfo/Doco/comnews_manual.pdf`.

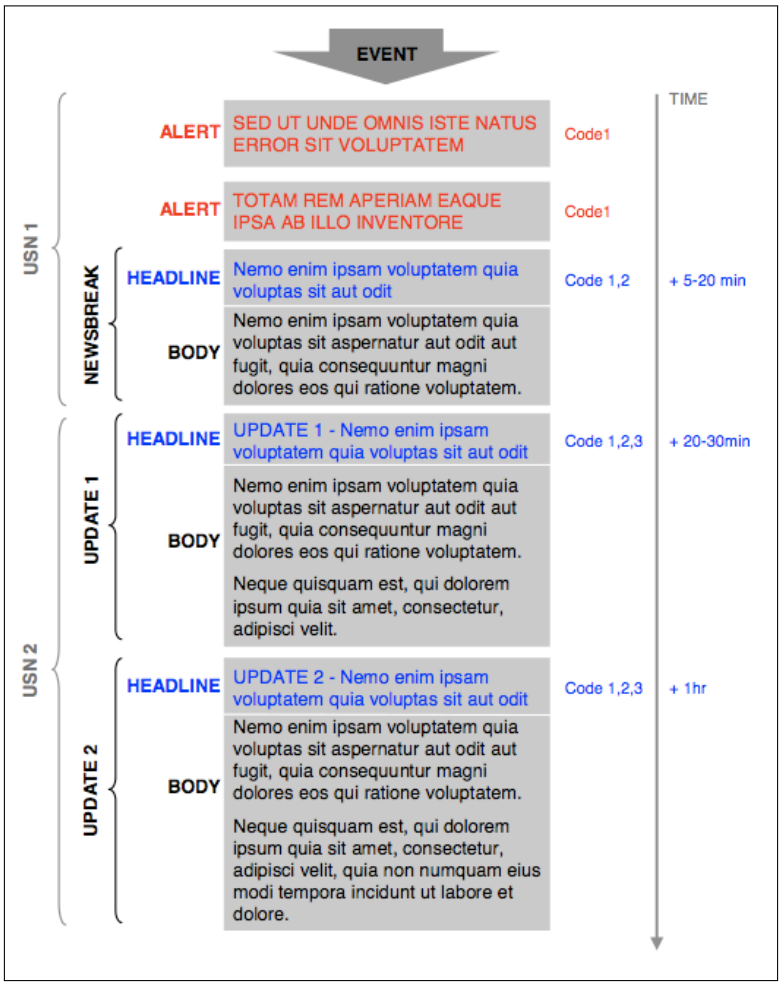


FIGURE 3.2: The evolution of an RNA story.

3.1.2 RNA Stories

The RNA data collects together stories published on the global Reuters news feed. Reuters operates in over 100 countries[5] and publishes stories in many different languages. Each story comprises a sequence of distinct *messages* based on the Reuters workflow. Figure 3.2, taken from the RNA documentation, shows an example of the evolution of a story.

The first time a story is reported is typically through an *alert*, a short upper-case sentence containing the key information. This is followed five to twenty minutes after by a *newsbreak*, which expands on the details of the story with a headline and two to four paragraphs of body text. The original alert and newsbreaks remain, but as events change, *updates* are released and include further information. Each

[5] `http://thomsonreuters.com/about/`

of these contains a *story key* and this is used to compile these messages into distinct stories. Update messages can add or replace text as required and while the protocol is well defined, there are occasional errors in the data and thus RNA stories without text. We use the text and title as of the final message and the timestamp of the first message. In addition to the text and title, each story is tagged with lists of relevant tickers, languages, topics and geographical areas. It is worth noting that RNA stories are often tagged with multiple tickers. Only 55.7% of RNA stories are tagged with one ticker contrasted with 92.7% of ASX announcements. Consequently, RNA story tickers provide less specific information than for ASX announcements. To select RNA stories specific to the ASX market, we use only those marked with ASX tickers and the English language tag.

Figure 3.3 shows a collated story with information from the announcement in Figure 3.1. The key information is clearly reported: "Rinker Group Ltd. <RIN.AX> reported a 33 percent climb in third-quarter net profit on Thursday" with mention of the strong U.S. market as a reason. It also includes background content, placing it in context with other companies in the sector, James Hardie and Boral. A quotation from an investment analyst is reported that explains the market's expectation. Artefacts of the journalistic workflow are clearly visible, marked with parentheses: the byline (author name), exchange rate, and message that indicates the content of the latest change: `(Adds details, fund manager comment, updates shares).`

Table 3.5 shows the top 30 categories and percentage of RNA stories they label from a sample of 8,277 stories. Note that the 'English Language Code' and 'Reuters Securities News Pool' are highly ranked since they are part of the selection criteria. The labels quickly reduce in frequency as the rank increases, but they specify financial news categories and some of Australia's main trading partners[6]. Where ASX announcements are predominantly associated with one ticker, this is not the case with RNA stories. Table 3.6 shows that RNA stories are more likely to have more tickers than ASX announcements (92% of which have only one ticker). This means that stories are more likely to discuss different companies and contain more items of information. This may be due to explicitly 'noisy' story types such as 'Hot Stocks', which appears in 10.32% of the sample above, and discusses a range of different interesting stocks.

[6]China is conspicuously absent from this 2004–2005 top 30 and appears in only 8.6% of stories.

(Adds details, fund manager comment, updates shares)
MELBOURNE, Jan 27 (Reuters) - Australian heavy building materials company Rinker Group Ltd. <RIN.AX> reported a 33 percent climb in third-quarter net profit on Thursday, built on strong U.S. construction markets, pushing its stock to a record high.

Rinker raised its outlook for operating earnings growth in its U.S. business to 35 percent for the year to March 2005, up from an earlier forecast of 30 percent, while sticking to its forecast for 20 percent growth in its Australian business.

"The result itself is pretty much in line with expectations, but they've obviously upgraded their guidance in the U.S. I think that's the main thing that people are focusing on," said BT Financial Group investment analyst Misha Collins.

Rinker Chief Executive David Clarke said it was too early to forecast profit growth in 2005/06, but said its order books showed solid demand. He expected price rises on Rinker's products to help offset higher fuel, power and raw materials costs.

"Our strong cash generation should enable us to finance both acquisitions and organic development, and simultaneously some form of capital management, as appropriate," Clarke said in a statement.

Rinker shares rose as much as 3.8 percent to an all-time high of A$11.16 after the result was released. Its shares have risen 70 percent since the start of 2004, heavily outperforming Australian building materials makers James Hardie <JHX.AX> and Boral Ltd. <BLD.AX>.

Collins said Rinker shares were also bolstered by comments from a rival, Florida Rock Industries Inc. <FRK.N>, which said it sees very strong price rises coming through for concrete and aggregates in Florida, a key market for Rinker.

Rinker, which makes about 80 percent of its earnings in the United States, reported a net profit of $95 million for the three months to Dec. 31, including a $15 million pre-tax loss on the sale of its small U.S. pre-stress concrete manufacturing arm.

Analysts' third-quarter forecasts ranged between $90 million and $99 million.

Revenue from its U.S. materials business rose 19 percent, with earnings before interest, tax, depreciation and amortisation up 31 percent, based mainly on sharp price increases on products like concrete and cement and cost cuts.

Growth in its Australian Readymix business slowed in the third-quarter following a land sale a year earlier, with revenue up 10 percent and earnings before interest, tax, depreciation and amortisation up 2 percent in Australian dollar terms.
($1=A$1.29)
((Reporting by Sonali Paul, editing by Richard Pullin; sonali.paul@reuters.com; Reuters Messaging: sonali.paul.reuters.com@reuters.net; +61 3 9286 1419))

FIGURE 3.3: An RNA story entitled "UPDATE 1-Australia's Rinker Q3 profit up 33 percent" containing the same information as the announcement in Figure 3.1.

Category	Percentage
English Language Code	100.00
Reuters Securities News Pool	94.30
Securities International News Service	89.85
Australia	88.01
Australian Domestic Financial News Service (Reuters)	83.98
Asia	79.70
Domestic News Pool	75.66
Reuters Markets News Pool	75.15
Reuters Corporate News Pool	55.91
North American Securities Domestic News Service (Reuters)	30.51
Euromarket Domestic News Service (Reuters)	27.58
New Zealand	25.07
New Zealand Domestic News Service (Reuters)	24.82
Mergers and Acquisitions (including changes of ownership)	24.55
United Kingdom	20.31
Corporate Results Forecasts	20.13
UK Focus Domestic News Service (Reuters)	19.31
Reuters Treasury News Pool	19.19
Financial and Business Services	18.76
Stock Markets	18.73
Treasury International News Service	17.31
Western Europe	17.30
Europe	17.05
Debt International News Service	16.89
Market Focus International News Service	16.73
United States of America	15.67
Major breaking news	15.49
Commodities International News Service	15.28
New Zealand stocks	14.51
Australia stocks	14.51

TABLE 3.5: The top 30 RNA story categories from a sample of 8,277 stories.

Number of tickers	RNA
1	67.05
2	15.41
3	7.13
4	4.01
5	2.96
6	1.06

TABLE 3.6: The distribution of tickers from a sample of 8,277 RNA stories.

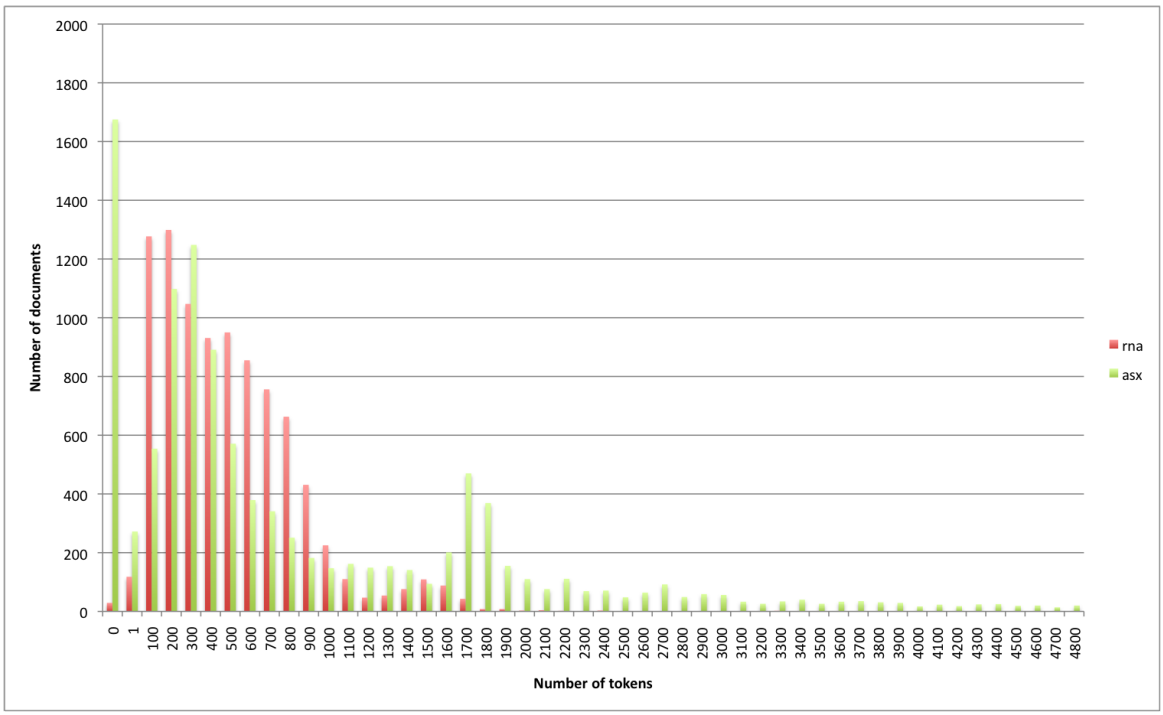


FIGURE 3.4: Histogram showing the document size distribution for ASX and RNA. Note that the tail continues until 9,800 tokens, but has been omitted for clarity.

3.2 Comparison

Though they handle similar information the ASX and RNA data sources differ in the quantity and frequency at which they release text. This impacts on how we choose to model information flow.

Figure 3.4 shows the document size distributions for the ASX and RNA in our experiment dataset. The first points to note are that the empty ASX announcements are the most frequent document size and the next most common for both ASX and RNA is between 200 and 300 tokens. The RNA stories then form a reasonably tight distribution with the majority under 1000 tokens. This is most likely due to a combination of editorial guidelines, space and time constraints. The ASX announcement sizes are far more varied with a longer tail, indeed, the longest are over twice the maximum graph value at 9,800 tokens. This reflects the longer deadlines available to announcement authors, large variety of announcement types. Effective RNA stories, on the other hand, *must* be a concise summary of the announcements they report on. If stories do not report the facts more efficiently, readers could simply read the announcement itself. Further research would confirm whether certain announcements types are longer than others, but inspection shows that the longer announcements tend to be reports with extensive technical detail.

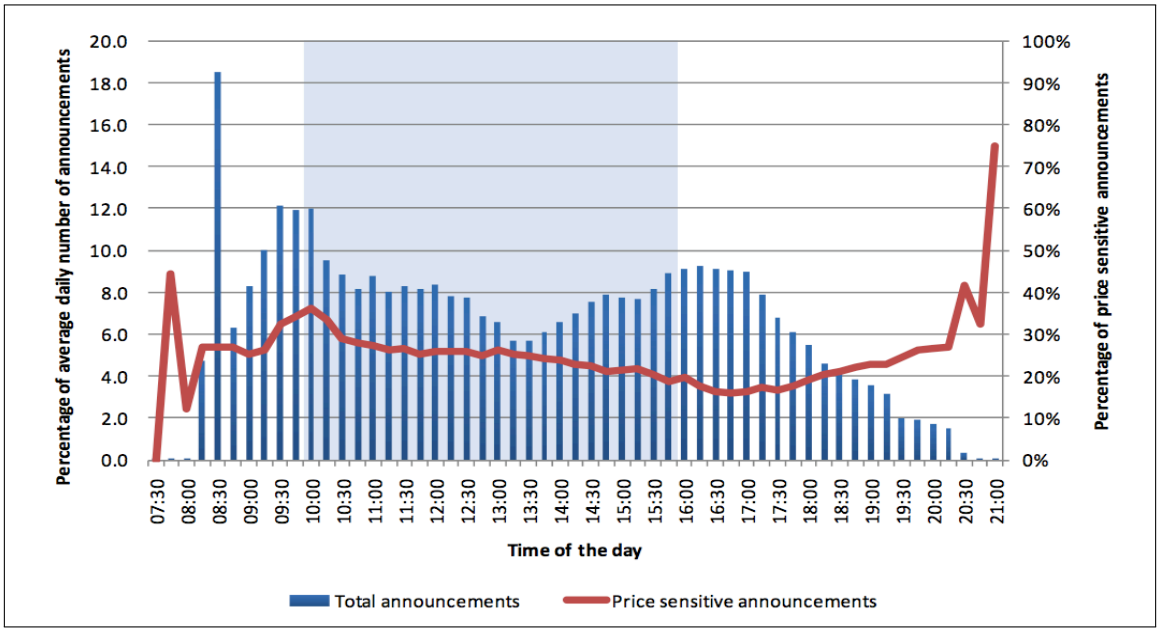


FIGURE 3.5: Distribution of ASX announcements over the day 2004–2006, trading hours (10am–4pm) are shaded. Reproduced from (Tannert, 2009).

The ASX limits trading to between 10am and 4pm[7] and accordingly, we do not display data outside of those times. This has wider implications on the timing of announcement and story release. Tannert (2009) explores the distribution of announcements and stories in the Australian market. This large-scale longitudinal analysis is made possible by automated information linking research from this thesis. Most financial analysis is limited to tedious manual study and our automated approach allows drawing of stronger conclusions based on larger sample sizes.

Figure 3.5 shows the daily distribution of ASX announcements with trading hours shaded. The bars show the average percentage of announcements released over the day and the line shows the proportion of those that are classified price sensitive for any ticker. The distribution shows two peaks either side of trading hours (9:30am, 4pm) with fewer announcements released *during* trading. Moreover, barring outliers where few announcements are released, price sensitive announcements are most common immediately before trading. These outliers occur at the beginning and end of the day when there are few announcements, but the few sensitive announcements comprise a higher proportion of those released *at that time*.

We expect the distribution of RNA stories to be less tethered to the ASX trading hours. This is because the ASX is only one of the sources that RNA follows and these other sources may release news at different

[7] http://www.asx.com.au/resources/education/basics/trading_hours_asx.htm

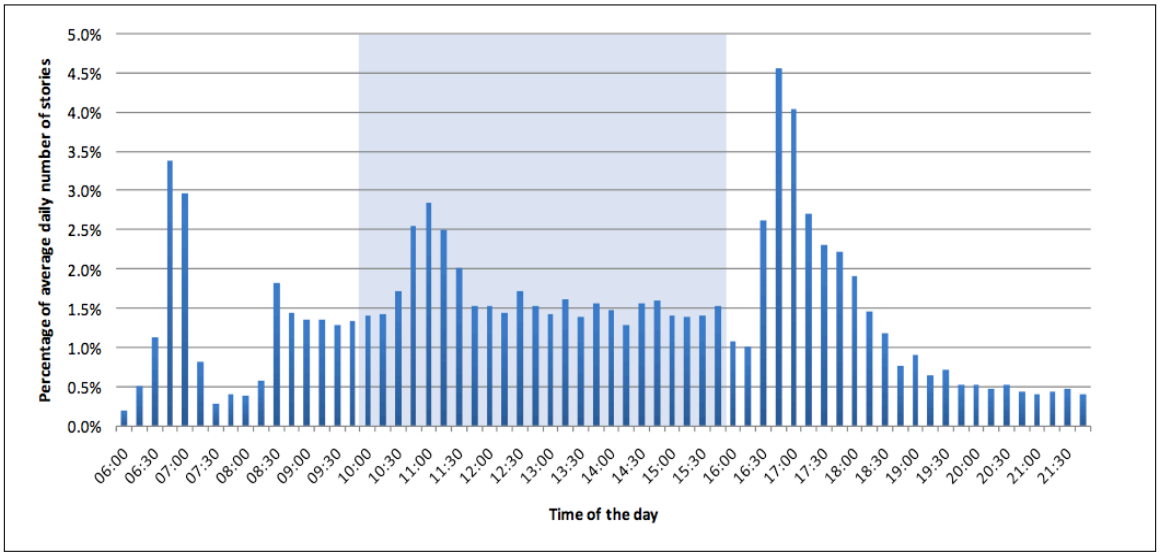


FIGURE 3.6: Distribution of RNA stories over the day from 2006, trading hours (10am–4pm) are shaded (Tannert, 2009).

times. Figure 3.6 shows that the distribution of RNA stories is spread over a wider time period with peaks at different places. It may be tempting to assume that peaks at 11am and 5pm correlate with the ASX announcement peaks, but this remains unclear without further analysis.

3.3 Trade data

We obtained ASX trade summary data from SIRCA to cover the same period as the ASX announcements and RNA stories[8]. The data consists of records that have a timestamp, a stock ticker, flags representing if it is a trade or a quote, prices and volumes. We process the records and store summary information about the price of a stock ticker at a particular point in time.

The ASX operates an "order-book" trading system, where traders offer to buy or sell volumes of stock at particular prices (i.e., quotes). These bids and asks would traditionally have been recorded in a ledger (hence book) and the best bid (highest) and ask (lowest) noted. Current systems automate this process and issue trades when the best bid converges with the best ask, trading as much volume as possible.

We use the best bids and asks to provide an indication of the market's value of a ticker. This is not used in any automated processing, but is used to give annotators more context about the announcements and stories they read as part of the annotation described in the next chapter.

[8]We use the TAQTIC data, but this service has been replaced with Reuters DataScope Ticker History (RDTH)

3.4 Summary

We have described the sources of financial information used in the current work. We discuss information sources and motivations for studying our two sources. The data obtained from Sirca are diverse in their content and presentation, raising a number of processing challenges. The 10,404 primary ASX announcements and 8,277 secondary RNA story sources constitute a solid platform upon which to identify and model information flow.

CHAPTER 4

Annotating Information Flow

Manual annotation of information flow is central to our study and experiments. Firstly, using an iterative scheme and tool development process deepens understanding about how to identify information flow. If the task cannot be reliably performed by human annotators, automation is of limited value. Secondly, modelling of the information flow problem also benefits from more exposure to the data and annotator feedback. Finally, annotation provides data for training and evaluating automated approaches. During development of machine learning models, it is critical to test how well the features model the information flow problem. This can be achieved by comparing human annotator decisions in labelled training data to those of the system. We conducted a pilot annotation project to develop the annotation scheme and tool, followed by a longer project where we collected our experimental data. In both cases, we selected Finance students as annotators since they are familiar with the domain subject matter and presentation.

This chapter outlines the first contributions of this thesis: the information flow annotation scheme and annotation tool. We explain the scheme and shows some examples of how it is applied to ASX-RNA pairs. We then describe the tool that our team of Finance students used to annotate the ASX-RNA pairs. Next, we detail the process itself and discuss the data we use to model and evaluate automated information flow tracking.

4.1 Annotation Scheme

The information flow annotation scheme describes two phenomena: whether an RNA story contains information from an ASX announcement (LINK) and the journalistic contribution that the RNA story makes. LINK applies if the RNA story in an ASX-RNA pair *repeats* information from the ASX announcement. Information, in this case, is defined as company details that are legally required to be disclosed *first* through an announcement to the ASX (i.e., continuous disclosure).

Link	Text
FIRST	...Record BHP profit of $2.45 million...
BACK	...BHP has been moving into NSW...
ANLY	...The profit exceeds expectation, said...

TABLE 4.1: Examples of RNA story journalistic contribution given the ASX announcement information: *BHP posted record annual profits of $2.45 million.* .

Journalistic contribution refers chiefly to the RNA story and can be indicated by any of FIRST, BACK or ANLY if (and only if) LINK also applies. FIRST indicates that an RNA story focusses on an ASX announcement and that the story is the first story to report on it. This means that it describes the story's position relative to the other stories that also report the information. BACK refers to background information regarded as common knowledge and publicly available before the release of the announcement. ANLY indicates new information added by the news source, such as analysis, editorial content and new quotes from industry commentators. Table 4.1 shows examples of the RNA story labels. Note that the distinction between BACK and ANLY is somewhat subjective and is the main cause for annotator confusion. ASX-RNA pairs that are unrelated have no labels, but if information flow occurs, then the LINK label applies and any of FIRST, BACK, ANLY (perhaps all) may also apply to the RNA story.

The scheme also defines an label unrelated to information flow: DIGEST. The Reuters dataset includes stories that contain snippets of news that relate to multiple events and companies, often a daily market report of reviews of 'Hot stocks'. These RNA stories were annotated with the DIGEST tag and are important for two reasons that might merit special treatment. Firstly, they tend to report information from many sources about many tickers which might be ambiguous. For example, an RNA story might legitimately contribute ANLY, but not concerning the ASX announcement. Secondly, the summary nature of these stories may be important and a ticker's inclusion (or even position) in the report might indicate importance or newsworthiness. The example below shows an excerpt from a digest story containing coverage of multiple stocks:

> - - David Jones <DJS.AX> chief executive, Mark McInnes, has signed a new contract that will see him lead the department store group until mid-2008. David Jones yesterday announced details of the new contract, which will see Mr McInnes' base salary rise from A$ 856,096 to A$ 1.45 million and includes performance bonuses of up to an additional A$ 2.2 million. Page 16.

> - - Crane Group <CRG.AX> yesterday forecast flat earnings growth in 2005-06 due to high raw materials costs and a slowing domestic building cycle. The plumbing supplies maker and distributor said it intended to focus on improving margins and efficiency during the current financial year. 'The market is going to do us no favours, so we have to be careful on costs and margins and tight on working capital,' said Crane managing director, Greg Sedgwick. Page 16 .

It is interesting to note how the tickers are identified explicitly in the text (i.e., "David Jones <DJS.AX>") and the use of formal cues to demarcate the sections of the story. There is no automated extraction or partitioning of the text since we only consider whole document processing at this point. We do not report agreement scores for DIGEST label annotation since it is not used in our classification tasks, although the DIGEST labelled stories are still included in the datasets.

4.2 Annotation tool

We designed an interface to allow annotation of the ASX-RNA pair within the information flow scheme. The tool's main design requirement was to allow efficient annotation of the information flow between ASX-RNA pairs. This efficiency relies on the ability of annotators to quickly navigate complex lists of time-aligned documents and minimise the most time-consuming part of the process; reading the documents.

To develop the tool as quickly as possible, we implemented it using Python's Django[1] web-framework with a PostgreSQL[2] relational database backend. This allowed us to quickly implement and release changes suggested by the annotators during the initial phases of the project. A web-based tool was more suitable for long-term annotation since it doesn't require that the annotators work only in a physical office during business hours. This was particularly useful for our annotators who were balancing study and other work constraints.

The main unit of annotation work is a *screen* displaying the ASX announcements and RNA stories released for a ticker over a fortnight. It allows annotators to view the pairs in context and annotate information flow links. Figure 4.1 shows the three main sections: a time-aligned navigation panel spanning

[1] `http://www.djangoproject.com/`
[2] `http://www.postgresql.org/`

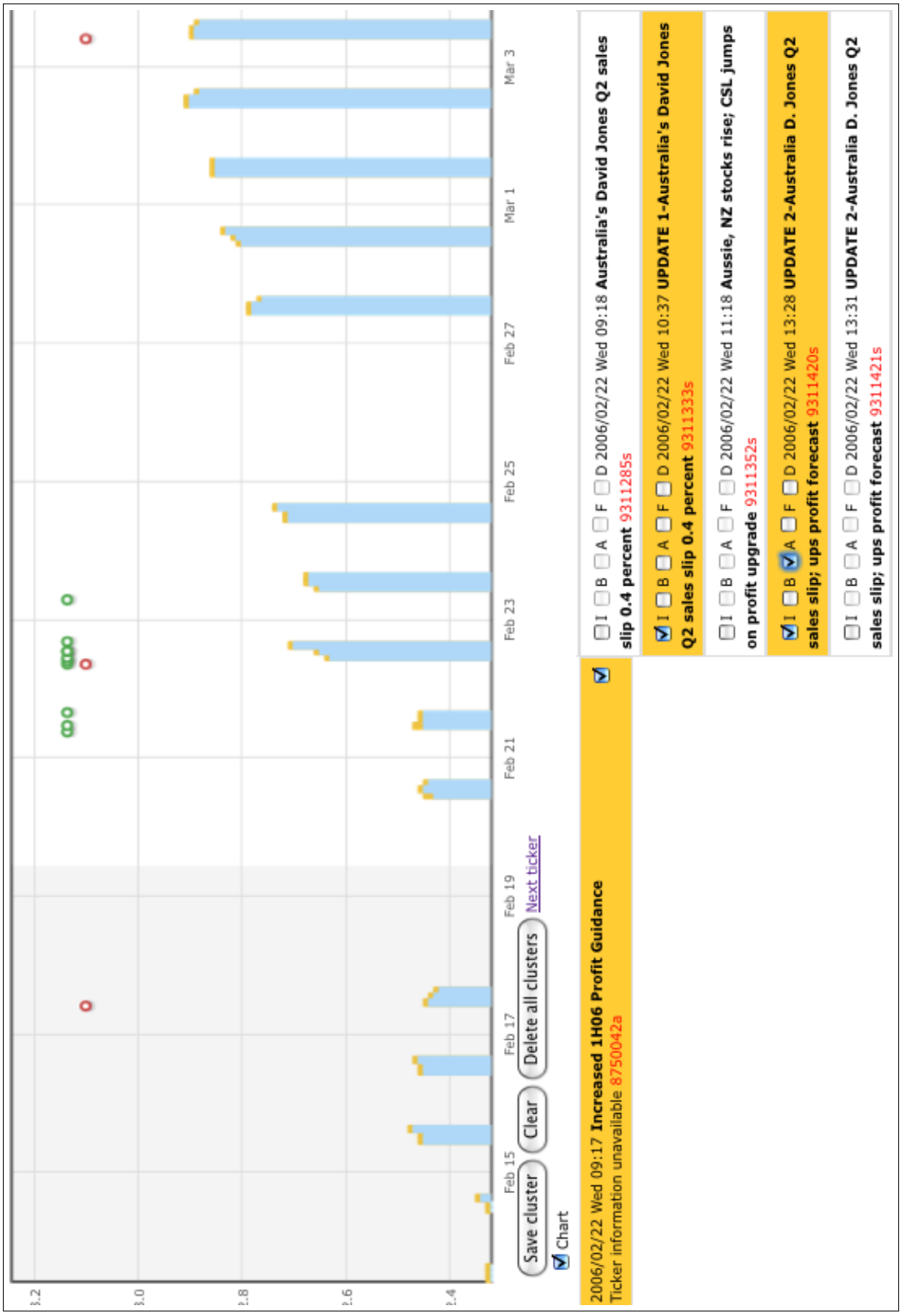


FIGURE 4.1: A *screen* from the annotation interface showing links between an announcement and two stories.

Label	Checkbox
LINK	I
FIRST	F
BACK	B
ANLY	A
DIGEST	D

TABLE 4.2: Legend mapping information flow labels to tool checkboxes.

the top, then two vertical document lists below. The top panel shows the fortnight of interest and a context week either side (though the right context week is not shown in this figure due to space constraints). The vertical bars show the stock price over that period of time, with the coloured tip showing difference between the best bid and ask price. The rows of dots indicate an ASX announcement or RNA story; in this example, the ASX announcements on the bottom row are followed by a burst of RNA stories on the top row. This presentation quickly gives annotators context about the placement of the documents within the fortnight and the market reaction to them while periods when the market is closed (e.g., weekends) are implicitly represented by the lack of trading data. The two lists of documents on the bottom of the panel are ordered by timestamp and show the ASX announcements on the left and RNA stories on the right. The titles and timestamp are always visible and annotators can click to reveal the RNA story text or open the ASX PDF file. Checkboxes corresponding to the information flow labels are on the left side of the story list. Table 4.2 shows how the information flow labels map to the checkboxes. When the cursor is placed over a dot on the timeline or a document in the list, the corresponding item is highlighted in the other panel, which helps annotators to navigate complex screens.

Although all evaluation here applies to ASX-RNA pairs, the list presentation format allows annotators to cluster related documents. We define a cluster as an announcement and one or more stories related to that same announcement. The highlighted cluster in Figure 4.1 consists of the ASX announcement on the left and second and fourth RNA stories on the right hand side. The RNA `I` checkboxes show cluster membership; the ASX-RNA pairs that include these stories (and the announcement) will be labelled with LINK. The `A` checkbox shows the ANLY label for the second story. The RNA story label checkboxes are disabled until `I` is active, ensuring that only LINK-labelled stories can be annotated with the other labels. The main benefit of this strategy was efficiency since the top-down view allows annotators to easily isolate clusters without re-reading documents and see which clusters they had previously created. The 'save cluster' button lets the annotator save the documents marked with the checkboxes as a cluster. This marks each document with a coloured box, so that all clusters that have been annotated on a screen

are visible. If the annotator has made a mistake, they can clear all checkbox selection using the 'clear' button, or remove all annotated clusters using the 'delete all clusters' button.

While unusual, it is also possible to add multiple ASX announcements to the same cluster, for example when a meeting announcement is followed by a set of presentation slides. However, the main constraint is that clusters be as minimal as possible and any announcements containing new information should form new clusters. For example, a company takeover might span several months of offer and counter-offer but the annotators should pick out the individual stages of the overarching process as individual clusters. Moreover, the minimality constraint encourages conceptual clarity and mirrors the way information is released piecemeal while still allowing later aggregation of clusters if required.

Once an annotator has finished with a screen, they can click on the 'next ticker' link to continue, or simply close the browser to finish. Perhaps the most popular request during the development phase was a feature that allowed annotators to review clusters and fix errors. Unfortunately, the short span of the project did not allow for this, though this issue was solved to some extent by the 'delete all clusters' button. Other suggestions included the ability to filter announcements or stories using regular expression matches in the title or text, or automatic scrolling of the document lists.

4.3 Annotation

The entire process was split into two distinct phases: pilot and final. The pilot phase was more exploratory and used to refine the scheme and tool. Once complete, we moved onto the final annotation phase of the process which was used to generate the data we use to train and evaluate our information flow models. Throughout the process, we measured how well the group of annotators were able to interpret and apply the annotation scheme. High inter-annotator *agreement* scores are an important indicator that the task is sufficiently well-defined. Additionally, it is difficult to evaluate the performance of an automated approach to a task that humans cannot reliably perform.

4.3.1 Pilot task

In the pilot, 120 screens were annotated by 5 Finance PhD students and the results manually checked. The emphasis was on cluster membership (i.e., correct LINK labelling) and provided a qualitative measure of the annotators' understanding of the task. It is important to note that all subsequent agreement is measured on a pairwise basis. For example, the pairwise metrics consider that two annotators agree

if they give the same answer (whether labelled or not) for a ASX-RNA pair, whereas cluster membership would require that clusters consist of the same ASX announcements and RNA stories.

Feedback from the annotators was very informative and mainly concerned cluster size and what stories should be included. The annotators tended to create larger clusters that addressed a topic rather than discrete events. For example, a company takeover might yield several announcements as details of different prices and conditions from the negotiation are released to the exchange. These announcements, and the stories that report on them, are all part of the same overarching process, but clustering them all together would create less specific clusters that span long periods of time. This problem is addressed by the cluster minimality principle in the final version of the scheme and the notion that there should be one item of information (perhaps found in more than one announcement) per cluster. Furthermore, annotators had difficulty identifying information in the announcements and disambiguating between closely related clusters. This is a subjective issue that is difficult to resolve and so this was left to the annotators' judgement. The annotators noted that stories might make cursory mention of an announcement despite focussing on different news. The decision was taken to cluster these stories that contribute *minor information* as LINK but not FIRST.

Early versions of the scheme placed emphasis on the temporal ordering of the ASX announcement and RNA story which had a negative effect on annotation. Insider trading is the focus of much Finance research, and relates to the situation where price-sensitive information is leaked before it is made public. In our dataset, this would be manifested as a RNA story that was released before the ASX announcement that is the official source of its information. Aware of these Finance-specific distinctions, annotators would spend time trying to ascertain whether story content was contained in a *later* announcement or simply a rumour. And while this is meaningful, the extra time taken slowed annotation pace.

The final two issues relate to wider problems designing information flow experiments. Though we only consider two data sources, the primary ASX announcement and the secondary RNA stories, market participants are exposed to far more information at different times. Inside company sources might preempt announcements by revealing key information via email or online forum. Different news sources, online forums and weblogs might report and comment on stories. Examining only information flow from the ASX and RNA source is reasonable because they are seen to be accurate sources with wide audiences, however there are cases that are problematic. As mentioned in the previous chapter, the ASX will occasionally request more information from a company, which is then required to file a response through the usual announcement channels. In the cases where the ASX is the source of the announcement (not

the company), care must be taken to check that their announcement contains the original information before clustering. Also, an announcement in response to action taken by an external party (e.g., a ratings agency report on the company) might trigger a story containing information external to the announcement. The scheme dictated that these rare cases not be clustered, though a case could be made that this is still constitutes information flow.

4.3.2 Final task

There are two objectives for the second phase of annotation: annotation of as much data as possible for training, and validation of inter-annotator agreement. A second team of Finance students was hired and trained to use the scheme and tool using materials developed during the pilot phase. We measure inter-annotator agreement using the Kappa statistic, a chance-corrected pairwise agreement measure (Cohen, 1960). Equation 4.1 shows how kappa between two annotators is calculated with $P(agree)$ the proportion of times they agree and $P(error)$ the probability of chance agreement.

$$\kappa = \frac{P(agree) - P(error)}{1 - P(error)} \tag{4.1}$$

There has been substantial discussion on the suitability of the kappa statistic (Carletta, 1996; Di Eugenio and Glass, 2004), but we consider kappa scores above 0.67 to be reasonable and above 0.8 to be good (Krippendorff, 1980). We chose five annotators with consistently high average kappa scores to continue.

Having established that the group has high agreement, the rest of the task is split into three: 215 screens to be completed individually, 50 shared screens and a final 215 individual screens. Each of the screens were randomly sampled from the 18 month period from January 2005 to June 2006. A shared task midway through the project allowed re-checking of agreement figures and is also used as held-out evaluation data. The central placement of the shared task is somewhat of a compromise between measuring agreement and annotating useful amounts of data. The most reliable strategy would be to have one annotator, but this would take too much time. On the other hand, the potential for inter-annotator disagreement and management effort increases with team size. Due to time constraints, the annotators did not complete all screens and the final count was 1779 individual and 42 shared screens.

Inter-annotator agreement for our five annotators is assessed using Cohen's Kappa over the shared task of 42 screens. The bottom row of Table 4.3 shows that acceptable Kappa scores are achieved for LINK

	LINK	FIRST	BACK	ANLY	*mean*
a	**0.80**	0.73	0.73	0.60	0.72
b	0.77	0.68	0.71	0.61	0.69
c	0.73	0.68	0.66	0.62	0.67
d	0.72	0.74	0.54	0.39	0.60
e	0.74	0.70	0.67	0.56	0.67
mean	0.75	0.71	0.66	0.55	

TABLE 4.3: Mean kappas by annotation and task. Scores are marked: **good**, reasonable and bad.

(a) LINK

b	**0.84**			
c	**0.82**	0.73		
d	0.75	0.75	0.66	
e	**0.80**	0.77	0.70	0.70
	a	b	c	d

(b) FIRST

b	0.67			
c	0.68	0.63		
d	**0.82**	0.74	0.71	
e	0.76	0.67	0.68	0.70
	a	b	c	d

(c) BACK

b	**0.83**			
c	0.74	0.70		
d	0.57	0.57	0.52	
e	0.77	0.74	0.69	0.49
	a	b	c	d

(d) ANLY

b	0.64			
c	0.78	0.68		
d	0.33	0.51	0.37	
e	0.65	0.62	0.63	0.33
	a	b	c	d

TABLE 4.4: Kappa scores for each annotator {a,b,c,d,e} for the shared task. Scores are marked: **good**, reasonable and bad.

and FIRST, borderline reliability for BACK (0.66) and ANLY substantially under the acceptable cut-off of 0.67. Each cell is the mean for an annotator over a label with aggregated means on the bottom row and rightmost column. Reading across the relevant rows and columns shows the ranges of average kappa scores for the different labels. BACK and ANLY have ranges of 0.19 and 0.23 and are an order of magnitude higher than LINK and FIRST (at 0.09 and 0.07), showing some of the agreement patterns within the group. Annotator d has the lowest inter-annotator agreement on average, mostly due to extremely low scores on BACK and ANLY. ANLY was the hardest information flow type to annotate with an average of 0.55 and is consistent with annotator feedback during scheme development. ANLY labels were reportedly the most difficult to disambiguate from BACK since the distinction between existing and new information proved subjective. Even including ANLY, all annotators apart from d performed at reasonable kappa levels over the tasks. Removing d would result in higher average agreement for BACK (0.69) and ANLY (0.60), but also less data for training and evaluation.

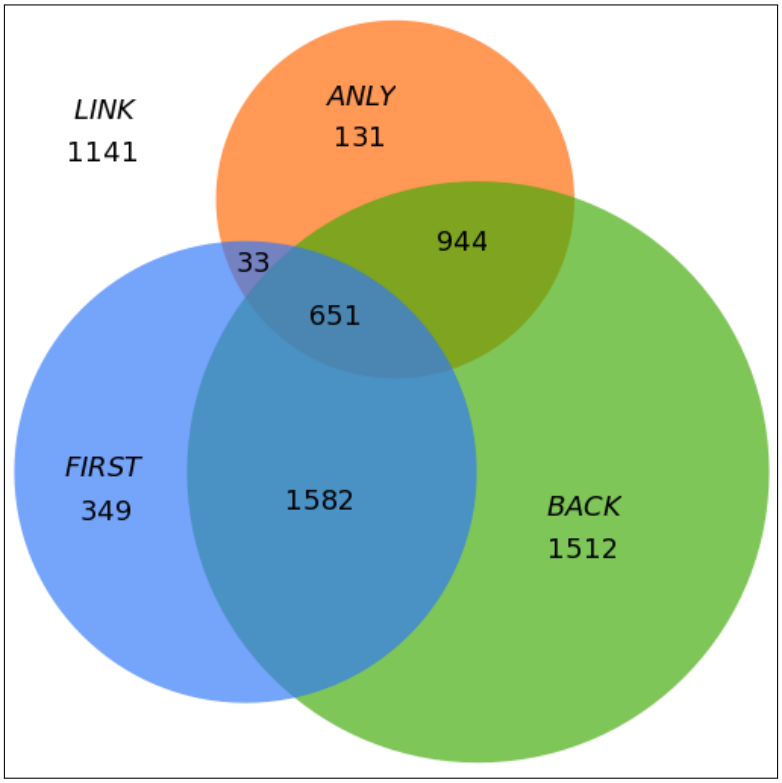


FIGURE 4.2: Venn diagram showing the proportions of all 6,343 ASX-RNA pairs annotated with information flow labels. Note that sizes are indicative only.

Tables 4.4(a-d) show the detailed pairwise kappa scores for each label in the shared task and these results follow the summary above. There are some general indications of pairs of annotators that have consistently high agreement, but there seems not to be any strong relationship. These kappa results show that while annotators could reliably annotate LINK and FIRST, task and annotator level issues mean that BACK and ANLY have lower agreement scores.

We also examine the co-occurrence of FIRST, BACK and ANLY in the annotated data. Figure 4.2 shows the extent to which ASX-RNA pairs are annotated with combinations of FIRST, BACK or ANLY. The linked ASX-RNA pairs comprise about 5% of the 135,537 total. BACK seems to be a common factor in most pairs, with very few pairs containing FIRST and/or ANLY *without* BACK. BACK has the highest proportion of type-singleton pair (i.e., those labelled *only* with BACK), 47% compared to 15% and 8% for ANLY and FIRST. The relatively high level of crossover is not unexpected with a document-level annotation task, since all of FIRST, BACK, ANLY might apply to a story; indeed this is the case with 651 stories (approximately 10%). The low agreement on ANLY (Table 4.3) means it is difficult to draw definitive conclusions without deeper error analysis. The main consequence of this overlap is that if an

(a) LINK

Annotator	Precision	Recall	F-score
a	85.7	97.4	91.2
b	97.1	86.3	91.4
c	85.4	77.4	81.2
d	97.7	72.2	83.0
e	78.8	92.3	85.0
Mean	88.9	85.1	**86.4**

(b) FIRST

Annotator	Precision	Recall	F-score
a	84.9	89.9	87.3
b	74.4	84.1	78.9
c	81.5	76.8	79.1
d	96.9	89.9	93.2
e	65.0	94.2	76.9
Mean	80.5	87.0	**83.1**

(c) BACK

Annotator	Precision	Recall	F-score
a	84.7	94.5	89.3
b	91.2	89.0	90.1
c	82.2	78.7	80.4
d	97.5	48.2	64.5
e	66.1	100.0	79.6
Mean	84.3	82.1	**80.8**

(d) ANLY

Annotator	Precision	Recall	F-score
a	65.3	88.6	75.2
b	81.1	85.7	83.3
c	61.7	94.3	74.6
d	67.4	41.4	51.3
e	77.5	78.6	78.0
Mean	70.6	77.7	**72.5**

TABLE 4.5: Average precision, recall and F-score agreement between *majority* and each annotator. The average F-scores used as upper bounds are marked in bold.

RNA story is annotated with one label, it is more likely to be annotated with another. The results may also indicate that the scheme could be further refined to allow annotators to apply it more consistently, particularly in the case of ANLY.

To evaluate our systems against annotator performance, we rephrase the agreement in more traditional classification task terms and create a *majority* annotator that represents the group. Rather than the kappa statistic, we calculate precision, recall and F-score over agreement for a pair of annotators. One annotator is taken to be the gold standard and the other compared against them and true positives, false positives and false negatives are calculated as explained in Chapter 2. Equations 2.3, 2.4 and 2.5 are used to calculate precision, recall and F-score. To construct the *majority* annotator, we take the majority response for each ASX-RNA pair and each label. For example, if three of the five annotators agree that a ASX-RNA pair should be labelled LINK and BACK, then this becomes the majority annotation. Table 4.5 shows the agreement scores between each of the five annotators and the majority as well as the average. During evaluation, we treat the majority annotation as the gold standard and calculate F-score agreement with our system's responses; the *system/majority* agreement score. Comparing this with the annotator/majority average directly evaluates how the system performs to the upper bound. If, in the best possible case, system/majority is equal to the average annotator/majority, then the system is indistinguishable from the human annotators.

	Training	Evaluation
Lag < week		
Pair count	125,915	9,620
LINK count	6,034	258
LINK attenuation	48	3
Lag < day		
Pair count	30,249	1,621
LINK count	5,596	231
LINK attenuation	486	30

TABLE 4.6: Impact of different lags on the datasets.

4.4 Datasets

The product of the annotation phase of the project are two datasets: training and held-out. These ASX-RNA pairs are the instances used to train and evaluate automated information flow tracking. The training dataset is made up of pairs annotated by one annotator and the evaluation dataset is made up of pairs from the shared task. Where numbers of labels in the evaluation dataset are discussed, this follows the majority annotation rules whereby at least three of the five annotators must agree for a label to be counted.

To transform the annotations into data instances, for each screen visited by the annotators, we extract all ASX announcements from the central fortnight and generate a ASX-RNA pair for each RNA story within a time lag window. The size of the window is important and larger windows mean that more pairs are generated since more stories are 'in range' of an announcement. From the experimental point of view, a smaller window is preferable since this increases the prior probabilities of pairs that *have* information flow. This follows the assumptions made earlier that important announcements are reported promptly and, as such, we expect to find information flow more often in ASX-RNA pairs with a short lag. It is also consistent with the cluster minimality constraint; annotators were encouraged to split clusters that spanned too long a time. A balanced prior probability class distribution provides more evidence for classification and increasing the information flow class prior should improve recall. Furthermore, a shorter lag improves computational performance since fewer pairs need comparison. This does, however, come at a cost since we cannot predict information flow over pairs outside the lag window and this places an upper bound on classification performance. Table 4.6 shows the effect of different lags on our datasets: one week and one day. The three sets of rows show the count of pairs, linked pairs and lost pairs (i.e., attenuation) for each dataset at each lag. At a lag window of one week, 48 pairs are lost from

Label	Training	%	Evaluation	%
Total pairs	30,249	100.0	1,621	100.0
LINK	5,596	18.5	231	14.3
FIRST	2,394	7.9	81	5.0
BACK	4,118	13.6	166	10.2
ANLY	1,472	4.9	72	4.4

TABLE 4.7: Distribution of labelled pairs in the training and evaluation datasets (lag < day).

the training datasets and 3 lost from the evaluation and so 0.8% and 1.1% of linked pairs will never be classified correctly. We used a lag window of a day in our experiments and the attenuation rates are accordingly higher: training set attenuation is 8% and evaluation 11.5%. There is clearly a trade-off between higher prior probabilities and lower attenuation rates. Initial experiments showed that higher priors performed better despite having more pairs out of lag range. In the interests of realism, we use the day lag window throughout all following work, but are at an immediate disadvantage due to attenuation during evaluation.

Table 4.7 shows how the information flow labels are distributed in the datasets. Despite the day lag window, the priors are all still relatively low. We would expect experimental performance to be best on LINK classification, followed by BACK, then ANLY or FIRST. Although the screens on which the pairs are based were randomly selected, there are differences in the label distributions. The main difference is that all labels are slightly rarer in the evaluation dataset. A fundamental assumption when constructing training and evaluation datasets is that they represent the same population. This means that a model based on the training set will allow inference of classes in the evaluation set. The less similar the datasets are, the less successful classification will be. It is not clear, in this case, whether the difference in priors is significant. There are far fewer ASX-RNA pairs in the Evaluation dataset and so a difference in 4% for LINK distribution only requires about 70 more LINK labels, which might easily be accounted for by annotation error. Table 4.8 shows number of documents from each source in both datasets as well as the text coverage (i.e., the number of documents which have extracted text). The ASX announcement text coverage is much lower for the evaluation dataset, which may limit the classification performance if those announcements are involved in linked ASX-RNA pairs. Also, there are more RNA stories than ASX announcements in the evaluation set. These differences mean that performance on the evaluation dataset may be worse than on the experimental dataset.

Source	Training	Text coverage (%)	Evaluation	Text coverage (%)
ASX	6,350	83.7	220	63.6
RNA	5,374	99.7	295	100.0

TABLE 4.8: Number of documents and text coverage in training and evaluation datasets (lag < day).

4.5 Summary

This chapter presented the first main contributions of this thesis: the annotation scheme and tool. We outlined a framework for manually identifying information flow in Finance text and presented the interface we built to do so. Both scheme and tool were developed in an iterative manner, incorporating substantial feedback from the annotators. We described the annotation process itself, emphasising the statistical tests that show that information flow (LINK) can be reliably annotated at a mean kappa of 0.75. The majority annotator statistics were calculated to provide an upper bound of 86.4% F-score on system performance. Finally, we present the resulting experiment and evaluation datasets, comprising a total of 31,870 pairs of ASX announcements and RNA stories.

CHAPTER 5

Modelling Information Flow

We define information flow as when information from one document is found in a second document. We frame this novel task as text classification and consider a broad range of approaches when modelling information flow. The literature review outlines approaches based on textual and temporal similarity and these inform our work. We use our annotated ASX-RNA pairs to train classifiers to predict the information flow, representing the pairs using features extracted from the ASX announcement and RNA story. Our modelling is exploratory and we expect that some features will represent the same underlying phenomena. It is thus important to use a classifier that does not require features to be conditionally independent.

This chapter outlines Maximum Entropy modelling and its use for text classification, showing it to be a suitable choice. We then discuss the ASX-RNA pair pre-processing and explain how we model information flow using temporal and textual features.

5.1 Maximum Entropy

Maximum Entropy modelling has been used in a variety of Natural Language Processing tasks (Ratnaparkhi, 1998). It is used in a classification context to model the conditional probability of an instance class with respect to evidence from a dataset. Supposing our task is to classify a news story into three genre classes: *sports, politics, entertainment* and we know that 60% of stories containing the word `team` are *sports* stories. While we might reasonably assume that the probability of the class *sport* given the word `team` is 0.6, we have no evidence to suggest how much of the remaining probability mass of 0.4 should be distributed to *politics* and *entertainment*. The Maximum Entropy principle states that in the absence of evidence for these classes, this probability mass should be distributed uniformly (i.e., 0.2 per class) to the effect that our model has the maximum possible entropy. Intuitively, it makes sense to only

make assumptions based on evidence from the data that we observe, and otherwise prefer a uniform distribution.

Equation 5.1 gives the maximum entropy model in sum form, returning a conditional probability for a class given context. This is essentially the sum of features ($f_i(\text{class}, \text{context})$) with each weighted by a parameter (λ_i). Equation 5.2 shows the normalisation that allows us to consider 5.1 as a probability. Features are represented as a function over the contextual predicate and class. This is outlined in equation 5.3 and in terms of our example, this feature will be active (i.e., have a value of 1) only in *sport* documents containing `team`.

$$p(\text{class}|\text{context}) = \frac{1}{Z(\text{context})} exp(\sum_i \lambda_i f_i(\text{class}, \text{context})) \tag{5.1}$$

$$Z(\text{context}) = \sum_{\text{class}} exp(\sum_i \lambda_i f_i(\text{class}, \text{context})) \tag{5.2}$$

$$f(sport, data) = \begin{cases} 1 & \text{if } \texttt{team} \text{ is in data} \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

Having defined the conditional probability $p(\text{class}|\text{context})$, the learning task requires that we find the optimum model; that which maximises entropy. Equation 5.4 gives the entropy of the conditional probability and we thus need to find the model (p^*) of all possible models (C) that maximises the entropy (i.e., Equation 5.5). The resulting model consists of feature weights (λ_i) that indicate what contribution those features make to the classification. For example, when classifying ASX-RNA pairs as LINK or not, a high positive feature weight means that the feature is a good indicator of LINK with respect to the model's training data. High negative weights indicate the feature strongly favours not labelling the pair.

$$H(p) \equiv - \sum_{\text{class,context}} \tilde{p}(\text{context}) p(\text{class}|\text{context}) \log(p(\text{class}|\text{context})) \tag{5.4}$$

$$p^* = \underset{p \subset C}{\operatorname{argmax}} H(p) \tag{5.5}$$

Berger et al. (1996) have shown this to be a constrained optimisation problem and various algorithms have been developed to find the global maximum, including Generalised Iterative Scaling (GIS) (Darroch and Ratcliff, 1972) and Limited Memory BFGS(LM-BFGS)(Nocedal and Wright, 1999). These algorithms are the most time-consuming stage of the training process and various alternatives exist to GIS (Malouf, 2002). It is also useful to note that for classification, the normalising factor (Equation 5.2) and the exponential can be omitted from Equation 5.1 if a probability is not required. We use the MegaM (Daumé III, 2004) Maximum Entropy package for ASX-RNA pair classification since it uses an efficient implementation of LM-BFGS. We use the `binomial` parameter and represent each pair as a class and a list of its active (1-valued) features. We experiment with no other parameters, placing instead emphasis on modelling information flow using extracted features.

The strength of the Maximum Entropy approach is that it offers a framework for using diverse "knowledge poor" features (Ratnaparkhi, 1998). While they may be linguistically simple in isolation, these features can be combined to represent complex linguistic phenomena. This is possible since, unlike Naïve Bayes, the features need not be class-conditionally independent and overlapping features can be used without risk of "double counting" evidence (Nigam et al., 1999). For example, we use features inspired by different research areas to model textual similarity. Many of these features will represent the same underlying phenomenon (i.e., the similar text) in slightly different ways. Maximum Entropy modelling allows them to be combined

5.2 Preprocessing

To model information flow, we need to pre-process the ASX-RNA pairs, which involves extracting text, title and timestamp from each document. The previous chapter outlines how the pairs that make up the datasets are chosen. We represent feature values as binary attributes, consistent with most Maximum Entropy problems. Where features have real number values, we split them into bins over a range of values. In most cases, the bins are equally sized and the ranges are empirically determined.

The text and title of the documents are available as `utf-8` encoded strings. Version 2.0b5 of the Natural Language Toolkit (NLTK) is used to tokenise the texts. Default `Tokeniser` modules are used in both cases: `TreebankWord` for word tokenisation and `PunktSentence` to detect sentence boundaries, the latter an implementation of the Punkt algorithm (Bird et al., 2009; Kiss and Strunk, 2006). Artefacts of the extraction process (see Chapter 3) are not removed or normalised. Removal of stopwords is

common in text classification since, for many tasks, words that serve grammatical functions are less useful: we use NLTK's list of 127 English stopwords.

The outcome of the preprocessing steps is that each document can be represented as a sequence of sentences, themselves sequences of tokens. Tokenised versions of each text are cached to avoid recalculating for each ASX-RNA pair it appears in. Tokyo Cabinet[1], a fast, lightweight key-value database, is used to store the tokenised text.

5.3 Textual features

We propose that textual features, and in particular, features based on textual similarity, will be good indicators of information flow. These features fall into two broad categories: fine and coarse-grained. Fine-grained features, such as any bag of words based feature, are specific indicators of the document content. These features will allow us to model relationships between particular word use and the information flow links, but are tightly coupled to the training dataset. For example, we might construct a model from one feature for each word that appears in both ASX-RNA texts. If the word *stock* appears in the ASX announcement and RNA story in a ASX-RNA, then the corresponding feature is active. However, we might use our model to classify unseen ASX-RNA pairs and find the word *tsunami* in both texts. If it does not appear in the training set, which is possible for such rare events, then our model is unable to take this into account, regardless of how indicative it may be. Word frequency follows a Zipfian distribution such that a word's frequency is inversely proportional to its frequency rank. As such, fine-grained features are likely to be rare, but *specific* when they are found. On the other hand, coarse-grained, *generalised* features are likely to be more common, but at the cost of specificity. We explore combinations of fine and coarse-grained features to strike a balance and robustly model information flow.

5.3.1 Pair set bags-of-words

The bag-of-words similarity features follow the intuition that, if two texts are similar, they should contain the same words. We consider unigrams (individual words) and bigrams (pairs of adjacent words) as a fine-grained model of information flow. First, unigrams and bigrams are extracted from the ASX announcement and RNA story. Any tokens that appear in the stopwords list are removed, including

[1] `http://1978th.net/tokyocabinet/`

bigrams where either word is a stopword. We also collapse all tokens to lower case to increase the potential for tokens to match.

We use the set operations: intersection ($\cap$) and set difference ($\setminus$) to create three classes of bag-of-words features that specify: ASX $\cap$ RNA, ASX $\setminus$ RNA and RNA $\setminus$ ASX. In the context of these features, set difference: ASX $\setminus$ RNA is defined as the tokens that are in the announcement and not in the story. Table 5.1 gives some examples of the feature values extracted from the ASX announcement and RNA story introduced in Chapter 2. For each token found in both ASX-RNA documents, a feature value (ASX$\cap$RNA TEXT-1G) is created, as well as (ASX$\cap$RNA TEXT-2G) for bigrams. This is expected to be useful for representing LINK and FIRST. However, this is unlikely to be a good indicator of the RNA story's contribution: information *added* in the story. The unigrams and bigrams found only in the RNA story are represented as (RNA$\setminus$ASX TEXT-1G) and (RNA$\setminus$ASX TEXT-2G). This should represent BACK and ANLY information present (i.e., added) in the RNA story. The inverse case, unigrams and bigrams in the ASX announcement but not in the RNA story is mainly added for completeness (ASX$\setminus$RNA TEXT-1G and ASX$\setminus$RNA TEXT-2G). There may be content in the announcement that indicates newsworthiness, but is not included in the story due to the editing process. However, the imbalance between document sizes (ASX announcements are typically longer and of more variable length) means that some ASX announcements will yield an overwhelming number of these feature values, perhaps reducing the effectiveness of these features.

The same feature values are calculated for the title as well as body text. Titles play an important role in the representation of announcements and stories since they are a summary of the body text content. Although it is not always the case, particularly for some announcement, the title can be seen as a distillation of the document topic. Unigrams and bigrams from the title are used as described above to make the feature types: ASX$\cap$RNA TITLE-1G, ASX$\cap$RNA TITLE-2G, RNA$\setminus$ASX TITLE-1G, RNA$\setminus$ASX TITLE-2G, ASX$\setminus$RNA TITLE-1G, ASX$\setminus$RNA TITLE-2G. Titles are short and similarly sized, and as such, should not be as susceptible to the overwhelming size issue.

5.3.2 Similarity scores

Similarity scores can provide coarse-grained features giving an overall indication of how similar the ASX-RNA pair texts are. We apply some similarity functions based on (Metzler et al., 2005)'s work in Text Reuse (see Chapter 2) to different text from the ASX-RNA pairs: body text, title and numeric tokens. In the following equations, ASX and RNA stand in for their respective components.

Feature	Example	Feature	Example
ASX∩RNA TEXT-1G	`Rinker`	ASX∩RNA TITLE-1G	*empty*
ASX∩RNA TEXT-2G	`Rinker Group`	ASX∩RNA TITLE-2G	*empty*
RNA\ASX TEXT-1G	`Reuters`	RNA\ASX TITLE-1G	`Q3`
RNA\ASX TEXT-2G	`Analyst's third-quarter`	RNA\ASX TITLE-2G	`UPDATE 1-Australia`
ASX\RNA TEXT-1G	`EBIT`	ASX\RNA TITLE-1G	`Update`
ASX\RNA TEXT-2G	`News Release`	ASX\RNA TITLE-2G	`Quarter ended`

TABLE 5.1: Examples of pair set bag of words feature values. Each feature may create far more values for a given ASX-RNA pair. These features are taken from Chapter 3's examples (see Figures 3.1 and 3.3.

Equations 5.6 and 5.7 give the symmetric and asymmetric overlap scores. We include symmetric overlap for completeness. The latter is the number of tokens in the intersection divided by the number of tokens in the RNA story. This accounts for size imbalance in the case where an ASX announcement (e.g., an annual report) is much larger than an RNA story reporting on it. An asymmetric measure does not penalise a smaller story and it is possible to have perfect similarity if every word in a story occurs in the announcement. Term weighting functions are also used in the similarity scores to indicate distinctiveness.

$$S_{symmetric}(\text{ASX}, \text{RNA}) = \frac{|\text{ASX} \cap \text{RNA}|}{|\text{ASX} \cup \text{RNA}|} \tag{5.6}$$

$$S_{asymmetric}(\text{ASX}, \text{RNA}) = \frac{|\text{ASX} \cap \text{RNA}|}{|\text{RNA}|} \tag{5.7}$$

Equation 5.8 shows asymmetric overlap weighted by the summed inverse document frequencies of the intersecting tokens. We calculate TFIDF for body and title text over the announcements and stories in the dataset. Care must be taken not to double count where these documents appear in multiple ASX-RNA pairs. We use Laplace smoothing (i.e., add one) when term frequency is unknown.

$$S_{assymetric-idf}(\text{ASX}, \text{RNA}) = \frac{|\text{ASX} \cap \text{RNA}|}{|\text{RNA}|}\left(\sum_{w \in \text{ASX} \cap \text{RNA}} log\frac{N}{df_w}\right) \tag{5.8}$$

Equation 5.9 states the TFIDF overlap score we use. We also include two cosine similarity scores not used by Metzler et al. (2005), both calculated using Equation 5.10. The vectors for one use unweighted word counts and the other uses TFIDF weighting.

$$S_{tfidf}(\text{ASX}, \text{RNA}) = \sum_{w \in \text{ASX} \cap \text{RNA}} log(texttf_{w,\text{ASX}} + 1) log(\text{tf}_{w,\text{RNA}} + 1) log(\frac{N+1}{\text{df}_w + 0.5}) \quad (5.9)$$

$$S_{cosine} = \frac{\text{ASX} \cdot \text{RNA}}{||\text{ASX}||||\text{RNA}||} \quad (5.10)$$

To create the feature values for the inputs: text, title or numeric tokens, we extract a list of stopped tokens and calculate the score for each metric. All these metrics return real values, between 0 and 1 for standard overlap and cosine (as explained in Chapter 2), and these values are grouped into bins to produce binary features. We experimented with different numbers of bins and use 55 bins for all measures. We concatenate the metric and bin names to create a feature value for each metric. These scores should provide coarse-grained features that model similarity in a general manner and are a good counterpoint to the fine-grained bag-of-words features. It is worth re-emphasising the role of Maximum Entropy modelling in the context of similarity scores. The scores clearly overlap and it is all the more important to use models that do not double count the features' evidence.

5.3.3 Common sequence matches

The bag of words and similarity score features are limited to short sequences of text, maximally two adjacent words for bigrams. The longer a shared sequence of tokens is, the more likely it is to be an indicators of information flow. Techniques from Plagiarism Detection use sequence matching to identify slabs of copied text. We use Python 2.6's difflib library[2] to identify common sub-sequences. The sequences include stopwords, are converted to lower-case and have a minimum length of three since we already calculate bigrams. Stopwords are retained to maximise the length of matching sequences. Figure 5.1 shows the sequences common to the example documents from Chapter 3. The first sequence is taken from a quotation by Rinker's Chief Executive and is both long and is a very strong indicator of information flow. The second sequence is a Finance idiom, and while it might indicate discussion of figures, the salient information is not clear. Both the lengths and counts of these sub-sequences matched to bins, producing features values. For example, the sequences in the example would yield the feature values: `4, 4x1, 23, 23x1`. These model the fact that there are *some* common sequences of length 4 and 23 as well as more detailed information about the exact number of them (i.e., 1). These count features are reasonably coarse-grained, they do not describe anything about the matched content.

[2] `http://docs.python.org/library/difflib.html`

```
% First sequence - length 23.
"our strong cash generation should enable us
 to finance both acquisitions and organic
 development and simultaneously some form of
 capital management as appropriate"
% Second sequence - length 4.
"for the year to"
```

FIGURE 5.1: Common sequences from Chapter 3's example documents (see Figures 3.1 and 3.3).

However, longer matches should be so rare as to be quite indicative of information flow when they do occur.

A feature that counts the number of entire sentences that match is also produced. Although verbatim copying is less likely given the rewriting and editing process that produces RNA stories, this should be considered a special case when it does occur. This is because a sentence can be considered amongst the smallest well-formed unit of information and its direct repetition is significant. The number of matching sentences feature will duplicate information from shared sentence-sequences, but this will not be double-counted by the Maximum Entropy model. Note that neither sequence in figure 5.1 is a complete sentence and the count of sentences in our example is zero.

5.3.4 Precision hashing

Numeric values are central to finance text and play an important role in information flow. Annual reports and earnings updates offer market participants a window into a company and the balance sheet figures are closely studied. However, finance text necessarily contains many numbers and the particular treatment is required to process them. In addition to the similarity metrics over numeric tokens mentioned above, we represent the precision of numeric values that occur in both texts. For example, repetition of the number `3,454` is more interesting than `1000` and we are far less likely to see the former repeated in two documents than the latter. For each token consisting only of digits, commas or full-stops that appears in both ASX-RNA pair texts, we replace characters to represent its precision. Table 5.2 shows some examples of the hash process. If the number is evenly divisible by ten, then its digits (including commas and full-stops) are replaced with `0`, otherwise they are replaced with `#`. The unique list of these hashed numbers each form a coarse-grained feature values. The example ASX-RNA pair would contain feature values for `#, ##, ###, ####, 00`. These are coarse-grained features and very

Original	Hashed
500	000
5	#
5434	####
50	00
12	##
54.3	####

TABLE 5.2: Examples of precision hashing.

generalised, they do not even refer to specific numbers themselves, simply their length and precision. This follows the intuition that if we find a ASX-RNA pair where both texts contain the same long, precise numbers, then this is a strong indicator of information flow.

5.4 Temporal Features

Timing is an important factor in news and we propose that the placement of announcements and stories in the news cycle is significant for judging information flow. Chapter 2 contrasted the daily frequency of ASX announcements (figure 3.5) and RNA stories (figure 3.6). Unsurprisingly, the ASX announcement release frequency is closely related to the trading hours with peaks immediately before and after trading. The RNA story frequency distribution appears less related to ASX market trading hours but is more interesting relative to announcement release time. If the assumption is made that newsworthy stories are reported promptly, then a short lag between ASX and RNA timestamp is a good indicator of information flow. Conversely, a long lag might indicate that the story does not contain information from the announcement or it has already been reported.

5.4.1 Time of Day

We model the time of day for each ASX-RNA pair using two features based on the ASX and RNA timestamps. These are localised to Australian Eastern Standard Time and rounded into 30 minute bins. Table 5.3 shows the timestamps of the Chapter 3 example documents and the feature values produced. The resulting feature values represent the placement of the announcement and stories with the ASX market news cycle. Larger periods of time could also be significant for representing information flow, for example, weekly story release or annual reports released at the end of the financial year. We do not consider these periods of time since as these periods of time increase, it is less probable that sufficient data can

Document	Timestamp	Feature value
ASX	2005-01-26 09:39:31	9:30
RNA	2005-01-27 11:05:11	11:00

TABLE 5.3: Examples of time of day examples.

be collected for each case to make definitive conclusions. Although we have not analysed this deeply, it is likely that the ASX Comnews categories would reliably identify these announcements.

5.4.2 Lag

Time lag between a ASX-RNA pair is modelled by one feature value: the binned difference between the RNA timestamp and the ASX timestamp. ASX announcements typically precede RNA stories and so this value is usually positive. Unlike the binning strategies used above, the bins are of increasing size since we propose that the longer the time between ASX announcement release and RNA story, the less important the exact time lag is. A number of bin sizes were experimented with and the final configuration uses an initial bin size of 5 minutes and the size increases by 5 minutes with each bin. For example, the bins around zero are: `[-15...-5)`, `[-5...0)`, `[0...5)`, `[5...15)` and are left-closed and right-open so that a pair released at the same time will have a feature value of `[0...5)`. This feature should be a good indicator of Reuters' responsiveness, in turn an entirely temporal indicator of information flow.

5.5 Feature frequencies

Table 5.4 details the frequency of each feature type over the combined dataset. The empirical significance of these features is explored further in the next chapter, but this table gives an insight into the levels of granularity of each feature by showing the number of unique types and the total number of feature values. Coarse-grained features such as text similarity have far fewer values than fine-grained features like the bags of words. Large models that use many features impact the time and storage required to process a ASX-RNA pair. Although this is not a primary requirement in our system, these factors are reason enough to prefer smaller models. Moreover, frequencies give some indication of the utility of each feature. Features that are active in few ASX-RNA pairs might not provide widespread enough evidence for the classifier. However, the frequencies in the table do not indicate overly sparse features. The ASX\RNA TEXT-1G and ASX\RNA TEXT-2G features together account for more than all the other

Feature	Unique types	Frequency
ASX∩RNA TEXT-1G	11,245	823,906
ASX∩RNA TEXT-2G	15,185	104,530
ASX∩RNA TITLE-1G	473	2,975
ASX∩RNA TITLE-2G	101	265
TEXT SIMILARITY	110	30,249
TITLE SIMILARITY	118	181,494
NUM SIMILARITY	162	181,494
NUMBER PRECISION	11	68,050
SENTENCES	12	30,249
SEQUENCES	69	34,367
TIME LAG	257	181,494
TIME OF DAY	74	60,498
ASX\RNA TEXT-1G	326,804	15,656,162
ASX\RNA TEXT-2G	1,936,410	25,853,019
ASX\RNA TITLE-1G	3,802	127,352
ASX\RNA TITLE-2G	7,140	82,741
RNA\ASX TEXT-1G	56,134	6,786,972
RNA\ASX TEXT-2G	283,548	6,703,802
RNA\ASX TITLE-1G	3,751	182,545
RNA\ASX TITLE-2G	11,272	129,887

TABLE 5.4: Feature frequencies over 30,249 ASX-RNA pairs in the combined dataset.

features combined. Careful analysis would reveal if the benefit of these specific fine-grained features outweighs the cost in model size. The motivation behind combining different approaches is to create a model with specific *and* generalised features, and this requires a balance between computational and classification performance.

5.6 Summary

This chapter introduced the framework and features used to model information flow. Textual features range from coarse to fine granularities, allowing general and precise representations of similarity, using approaches from IR, Text Reuse and Plagiarism Detection. Set-operation bags of words model the specific words involved in information flow, whereas similarity scores capture the same phenomenon in a general manner. We use sequence matching features to capture longer snippets of shared text, such as quotes. Similarity of numeric text is explicitly modelled, as well as a measure of its precision to capture unusual numbers appearing in both ASX-RNA pair documents. The temporal aspect of news release and the responsiveness of the news agency are also represented by the time of day announcements

and stories are published. Finally, we use the lag between the publishing timestamps as a feature, following the intuition that important information flows quickly. Maximum Entropy modelling allows us to concentrate on feature design since the model will not double count evidence, even when features are not conditional independence.

CHAPTER 6

Results

We frame information flow as a supervised classification task over ASX-RNA pairs, formulating the LINK, FIRST, BACK and ANLY tasks. This allows us to test our classification performance and compare different models for information flow. This chapter presents the results of our experimentation and evaluation. First, we outline our methodology and how we construct tasks to predict information flow labels. We then demonstrate good performance relative to baseline approaches and discuss the best combination of features for each task. Results from our evaluation directly compare our system to human performance. Finally, we analyse some error cases to gain insight into how to improve our model. These results constitute the final contribution of this thesis: that it is feasible to track information flow in Finance text.

6.1 Experimental Methodology

There are two stages to the experimental program: cross-fold validation using the training dataset and direct comparison against a majority annotator using the held-out evaluation dataset. Cross-fold validation experiments are used to test the effectiveness of different feature combinations. The dataset used consists of the ASX-RNA pairs annotated by one annotator. These pairs are split into n parts and n trials are run for each experiment. In each trial, a model is trained using the fraction of the dataset not in the *nth* part. This model is used to re-predict the label of each ASX-RNA pair with one of four results. Assuming the task is LINK classification, a true positive is when the annotation and predicted label are LINK and true negative if they are both not LINK. A false positive occurs when the annotated label is not LINK, but a LINK link is predicted. The final error class is a false negative - when an annotated LINK pair is mis-classified as not LINK. The counts of each ASX-RNA pair result are aggregated over all n trials and precision, recall and F-score calculated as introduced in Chapter 2 (see the Equations 2.3, 2.4 and 2.5). The purpose of the n-fold cross validation is to account for any bias in the dataset distribution

that might cause us to overfit our model to the dataset. It also allows us to use as much training data as possible (90% in each fold) to test features while maintaining enough held-out data for evaluation.

Although the cross-validation experiments will allow comparison of different feature sets, it is also important to evaluate against human performance at information flow tracking. We evaluate our systems against the majority annotator compiled from the annotator decisions. The best performing combination of features in the cross-fold validation experiments is used to train a model for each task using the entire development dataset. These models are then used to predict the label for each of the ASX-RNA pairs in the evaluation dataset and these predictions are scored for agreement with the majority annotator; the system/majority score. Chapter 4 explains the how this and the annotator/majority scores are calculated. These two agreement scores allows us to directly compare the system's performance against the human annotators. It should also be restated that during evaluation, the system is only presented ASX-RNA pairs that fall within a day lag window (see Chapter 4), with any other pairs classified as not linked.

We selected two simple feature combinations to provide a baseline performance in each of the tasks. Firstly, a text baseline using: ASX∩RNA TEXT-1G, ASX∩RNA TEXT-2G, ASX∩RNA TITLE-1G, and ASX∩RNA TITLE-2G. These minimal bag of words features should be robust, and are a common choice for baseline approaches to text classification. The time baseline uses TIME LAG and TIME OF DAY features and is intended to show the impact of exclusively temporal features.

6.2 Cross-validation Results

Table 6.1 summarises the experimental results for both baselines and our best combinations of features for each task. The left hand column indicates the task and prior distribution of the corresponding label in the dataset. This is significant since we expect tasks with lower prior distributions to be more difficult (see Chapter 4 for discussion). Precision, recall and F-score are listed for each combination of task and feature set. The scores for the best feature set are emphasised using a bold font.

We perform substantially better than the baselines for all tasks. Our best result is an F-score of 89.5% for LINK classification and BACK and ANLY both have F-scores above 80%. FIRST has a slightly lower F-score of 73.4%, but also low text baseline scores, which we discuss below. While higher F-scores were achieved, for the most part, in the tasks with higher prior link probabilities, scores in ANLY were surprisingly high given its particularly low prior of 4.87. Precision is higher than recall in all cases, but the F-score increases gained by the best systems are due to recall improvements. The average

Link	Features	P (%)	R (%)	F (%)
LINK (18.5%)	Time	62.3	33.6	43.6
	Text	85.0	73.1	78.6
	Best	**90.9**	**88.1**	**89.5**
FIRST (7.9%)	Time	70.3	22.9	34.5
	Text	66.0	43.9	52.7
	Best	**77.0**	**70.1**	**73.4**
BACK (13.6%)	Time	60.9	17.0	26.5
	Text	83.4	67.1	74.4
	Best	**88.4**	**83.2**	**85.7**
ANLY (4.9%)	Time	0	0	0
	Text	78.9	56.0	65.5
	Best	**86.7**	**75.0**	**80.4**

TABLE 6.1: Precision, recall and F-score for cross-validation experiments. This summarises the results for the text and time baselines, as well as our best combination of features. The prior class distribution for each label in the training set is also supplied.

increase between the text baseline and the best systems is 7% for precision, but 35.5% in recall. This is consistent with the addition of coarse-grained features to the entirely fine-grained baseline. The time baseline performs worse than the text baseline over all tasks, but the difference is least for FIRST. This is unsurprising given FIRST's "first reuse" emphasis. The time baseline fares badly on other tasks to the point that all ASX-RNA pairs are classified as not linked for ANLY, resulting in zero scores. This poor performance confirms that while announcements and stories are essentially time-series data, information flow can only be effectively modelled by a combination of temporal and textual features.

Table 6.2 shows the best performing (by F-score) feature combinations for each task. To test the contribution of each feature, subtractive analysis was performed on the best performing feature set for each link type. Features used are marked with $\cdot$, while features marked with $\star$ or $\star\star$ if their removal results in significantly worse F-score (at $p<0.05$ and $p<0.01$ respectively). An experiment is conducted that uses all but one feature and the results compared to best using approximate randomisation (Chinchor, 1995) to assess whether adding the omitted feature results in a statistically significant improvement.[1]

The first observation to make from the table is that the tasks can be separated into two groups on the set of features that was most successful: LINK/BACK and FIRST/ANLY, though this may also be related to the different prior link probabilities, higher and lower for each group in this case.

[1]We adapt a parsing evaluation script `http://www.cis.upenn.edu/~dbikel/software.html`

Link	REL	FACT	BACK	ANLY
ASX∩RNA TEXT-1G	·	·	·	·
ASX∩RNA TEXT-2G	⋆⋆	·	⋆⋆	·
ASX∩RNA TITLE-1G	·	·	·	·
ASX∩RNA TITLE-2G	·	·	·	·
TEXT SIMILARITY	⋆⋆	⋆⋆	⋆	·
TITLE SIMILARITY	·	⋆⋆	·	·
NUM SIMILARITY		·		·
NUMBER PRECISION		⋆		·
SENTENCES	·	·	·	·
SEQUENCES	·	⋆⋆	·	·
TIME LAG	⋆⋆	⋆⋆	⋆⋆	·
TIME OF DAY	·	·	·	⋆
ASX\RNA TEXT-1G				
ASX\RNA TEXT-2G				
ASX\RNA TITLE-1G	⋆⋆		⋆⋆	
ASX\RNA TITLE-2G	⋆⋆		⋆⋆	
RNA\ASX TEXT-1G		·		·
RNA\ASX TEXT-2G		·		⋆⋆
RNA\ASX TITLE-1G	·	·	⋆⋆	·
RNA\ASX TITLE-2G	⋆⋆	·	⋆	·

TABLE 6.2: Feature combinations for the best performing development experiments. Features significant from subtractive analysis are annotated ⋆ ($p<0.05$) and ⋆⋆ ($p<0.01$).

Features based on the text are broadly useful, both modelling information flow and journalistic contribution. Although ASX∩RNA TEXT-1G and ASX∩RNA TEXT-2G (intersection unigrams and bigrams) appeared in all feature sets, the only significant use was ASX∩RNA TEXT-2G for LINK and BACK. One reason might be that they more effectively model topic-level textual similarity while being less susceptible to single words appearing by chance in both texts. Of the text set difference features, only RNA\ASX TEXT-2G was significant and only then for ANLY, suggesting that the feature effectively represents commentary text. Interestingly, neither ASX\RNA TEXT-1G nor ASX\RNA TEXT-2G was used in any well-performing experiment. One potential explanation is that the wide variety of text sizes is simply too noisy a feature for the model to generalise. These features are likely to be sensitive to different use of terminology. The following terms are taken from the bag of words features from the previous chapter (Table 5.1): `Q3, third-quarter, Quarter`. These all refer to the same period of time, yet only appear as difference features in the table. Correctly processing date expressions is one way which these features could be generalised.

Titles play an important role in announcements and stories, summarising the event that they report on. Rather than intersection, the set difference features proved to be more significant. ASX\RNA TITLE-1G and ASX\RNA TITLE-2G were significant for LINK and BACK tasks, and RNA\ASX TITLE-1G and RNA\ASX TITLE-2G only significant in some cases. This may indicate that cues of ASX announcement newsworthiness may appear in their titles, but are not repeated in the titles of stories that report on them. Conversely, title terms that indicate that a story reports directly on an announcement may not be found in that announcement's title. In addition to this, titles are often constrained by space and the need for concise communication and are less likely to contain unindicative terms.

The similarity features are designed to summarise similarity at a coarse-grained level. TEXT SIMILARITY significantly improved performance for LINK, FIRST and BACK (albeit to a lower significance level). This is representative of the difference between LINK and FIRST, which track *repeated* information, and BACK and ANLY, which indicate *extra* content. TITLE SIMILARITY is significant only in the FIRST task, which raises an interesting question of how well a system that only used metadata was able to perform. This would be advantageous from a performance point of view since there is less textual data to process. Furthermore, this metadata is more easily obtainable than the entire text and tends to require less pre-processing and extraction from arbitrary file formats.

Common sequence and sentence matching was used for all tasks, but proved significant only for FIRST. This may be because the first story to report on an announcement is more likely to quote verbatim since there is less time to edit and synthesise new material and this may play a role in the feature's success. When longer sequences are present, however, they are very strong indicators of FIRST and the high significance bears this out.

Numbers are central to information flow in finance and NUM SIMILARITY and NUMBER PRECISION were present in the FIRST and ANLY experiment (NUMBER PRECISION was mildly significant for FIRST). Though these initial results are encouraging, the importance of numbers to Finance text means that more work is required. An initial improvement might be to correctly normalise numbers in text so that `5.5 million` would be equivalent to `5,500,000`. The initial form, of course, is still important to detect verbatim reuse. Tables of numbers are also common in ASX announcements, for example in lengthy annual reports. Processing tabular data is non-trivial, especially from PDF-extracted text, which is already noisy. It is clear that some form of table processing is important, however, many tables are sufficiently complex for humans to process that the important numbers and figures are repeated in the text. Document zoning may be useful in this case, at the very least to identify and ignore tabular data,

and preferably to model the situation where a particularly important figure exists in both a table *and* the surrounding text. Analysts' forecasting and expectant interest in company balance sheets means that often just one figure summarises an entire annual report. As such, they demand further research.

Finally, news has a strong temporal dimension and we expected the lag feature to be significant for all link types. While it was for LINK, FIRST and BACK, the time-of-day feature was more significant for ANLY. That the analysis and commentary are the only link types sensitive to their placement in the news cycle points, potentially, to less time critical stories released at regular times.

6.3 Evaluation Results

Having established the best feature combination for each of the tasks, we compare the system to the majority annotator decision. Table 6.3 shows the majority annotator agreement scores of the text baseline and best models over the evaluation dataset. We do not compare the time baseline due to poor performance in the cross-fold validation experiments. The upper bound is the annotator/majority average introduced in Chapter 4. Again, our system exceeds the baseline performance in every task, with significantly better recall and F-score for LINK and BACK ($p<0.01$). This is an important result since it shows that the improvements to recall delivered by the features generalise to our held-out evaluation set. The recall and F-scores showed borderline significant improvement at between 0.07 and 0.08, which is also encouraging. This good performance is despite the extra constraints we placed on evaluation in the interests of realism. The reduction of the lag window and inclusion of documents without text (e.g., faxed and scanned announcements, text extraction errors) all make good classification performance more difficult. Another interesting trend in the evaluation task is the extremely high precision in the ANLY task: 95.5% for the best and 89.5% for the text baseline. This indicates that the very specific features in the text baseline (e.g.bags of words) are very strong indicators of commentary or editorial content and when they are present, they closely model this type of information flow.

6.4 Error Analysis

The experiments above measure how successful our features model information flow. This does not give us any insight into how we might improve performance. To gain this, we examine cases where the system misclassified ASX-RNA pairs. Specifically, we address false positives, pairs where the system incorrectly identified information flow and false negatives, where the system failed to detect it. The

Task	Features	P (%)	R (%)	F (%)
LINK	Text	80.0	51.3	62.5
	Best	⋆84.5	⋆⋆70.1	⋆⋆76.6
	Upper	88.9	85.1	86.4
FIRST	Text	58.8	29.0	38.8
	Best	64.6	44.9	53.0
	Upper	80.5	87.0	83.1
BACK	Text	80.0	43.9	56.7
	Best	75.0	⋆⋆62.2	⋆⋆68.0
	Upper	84.3	82.1	80.8
ANLY	Text	89.5	24.3	38.2
	Best	95.5	30.0	45.7
	Upper	70.6	77.7	72.5

TABLE 6.3: System/majority agreement for the best and text baseline models. Upper is the mean annotator/majority agreement. ⋆⋆ indicates that the best system performs significantly better than the text baseline ($p<0.01$).

output of the MegaM classifier specifies the probabilities for the label it assigns. For example, in the LINK classification, any ASX-RNA pair with a probability of 0.5 and more will be assigned LINK. Table 6.4 shows the distribution of correct and incorrect cases. The distributions of the true positives and negatives (tp and tn) skew towards the ends of the probability range; most correct cases are classified with strong (positive or negative) probabilities. Incorrect cases (fp and fn) show a mostly uniform distribution, except for a high number of false negative pairs classified with low (i.e., strong in this case) probability. We select the *most* incorrect pairs relative to the model for further analysis. That is, the false positives with the highest probability and the false negatives with the lowest probability.

A sample of the worst 20 pairs for each error type in the LINK task were reviewed by in a pairwise manner to check whether the annotation was correct and identify any potential issues for the classifier. This is slightly problematic in that it does not faithfully replicate the annotation process, since we do not present any other clustered pairs. Table 6.5 shows the results. DIGEST stories were anticipated to be problematic, since they include content that relates to different events and we report the count of these among the error pairs, as well as the number in which the annotation was wrong and ambiguous given the limited context. 15 of the 20 false positive cases were wrongly annotated; the classifier was correct predicting them as LINK-linked. 3 of these were also labelled as DIGEST, but this does not seem too significant. The high error rate is not unexpected since a link can consist of just one mention of information and mis-annotation can be the result of a lapse in annotator concentration. Mis-annotation is a less common explanation for false negative errors with only 4 of the false negative ASX-RNA pairs

Probability	tp	tn	fp	fn
0.00-0.05	0	22052	0	119
0.05-0.10	0	827	0	71
0.10-0.15	0	406	0	62
0.15-0.20	0	232	0	67
0.20-0.25	0	178	0	65
0.25-0.30	0	132	0	53
0.30-0.35	0	113	0	43
0.35-0.40	0	73	0	62
0.40-0.45	0	71	0	58
0.45-0.50	0	73	0	62
0.50-0.55	79	0	62	0
0.55-0.60	88	0	58	0
0.60-0.65	92	0	48	0
0.65-0.70	106	0	62	0
0.70-0.75	143	0	35	0
0.75-0.80	139	0	36	0
0.80-0.85	195	0	41	0
0.85-0.90	333	0	39	0
0.90-0.95	562	0	38	0
0.95-1.00	2593	0	71	0

TABLE 6.4: Classification probabilities for the best performing LINK model over the training dataset. This includes numbers of correctly classified LINK pairs (true positives and negatives) and incorrect pairs (false positives and negatives) at each probability level.

		Annotation		
Error	Digest	Correct	Incorrect	Ambiguous
False positive	30	20	75	5
False negative	70	75	20	5

TABLE 6.5: Error analysis for the 20 most wrong false positives and negatives for LINK classification, showing the percentages of error cases that were labelled with DIGEST and distribution of incorrect annotation.

affected. The count of digest documents is, however, much higher in the false negative error pairs (3 of 4) and it is likely that legitimate textual similarity is lost in the noise of the irrelevant content. There are two main outcomes of this error analysis. Firstly, there is a very high rate of mis-annotation and that our system's precision could be higher than reported. Secondly, DIGEST stories reduce our system's recall since the content of the story includes multiple threads of information. The similarity between the announcement and *relevant* information is simply overwhelmed by irrelevant content.

```
   Singapore Telecommunications Limited ("SingTel") wishes to
announce that SingTel Asia Pacific Investments Pte Ltd (formerly
known as SingTel Mobile Sales Pte Ltd) ("STAPI"), a
wholly-owned subsidiary of SingTel, has incorporated a
wholly-owned subsidiary in Mauritius known as Viridian Limited.
   Viridian Limited has an initial issued paid-up capital of US
$2 divided into 2 ordinary shares of US $1 each.
The principal activity of Viridian Limited is investment holding.
```

FIGURE 6.1: Excerpt from the ASX announcement titled: "Incorporation of Subsid.& Increase in Capital of Subsid.".

```
    SINGAPORE, June 17 - Singapore Telecommunications Limited
 (SingTel) wishes to announce that it has deposited US$40 million
 with the Privatisation Commission, Government of Pakistan (GoP)
 as earnest money in connection with the privatisation of Pakistan
 Telecommunication Company Limited (PTCL).
    The GoP is in the process of privatising PTCL by selling a 26
 per cent equity stake (Equity Stake) to a strategic investor. The
 Equity Stake comprises 1,326,000,000 class B ordinary shares in
 PTCL, with a transfer of management control.
    As earlier announced by the GoP, SingTel is one of the 9
 pre-qualified parties to participate in the privatisation. The
 bid date is scheduled for 18 June 2005.
```

FIGURE 6.2: Excerpt from the RNA story titled: "TEXT-SingTel deposits $40 mln for Pakistan bid".

There are also subtle error cases, for example the following false positive. Both the announcement and story refer to company-level events for "Singapore Telecommunications" (SingTel). Figures 6.1 and 6.2 show excerpts and SingTel is announcing a new company and a bid for a takeover respectively. The documents are about the same size and the excerpts shown comprise most of the text. What makes this pair distinct is that both documents are formulaic, published with a short lag (20 minutes) and consist of SingTel "wish(ing) to announce that it has" performed an action involving "ordinary shares". The main differences are that the action and target are different. We do not explicitly model financial-specific actions or named entities and this may explain the system's failure in this case.

The error analysis has revealed several interesting trends. Incorrect annotation seems reasonably widespread in certain sections of the dataset. Re-annotation of these worst cases would allow us to model information flow more consistently. DIGEST labelled stories indicate that further efforts to track information flow show concentrate on segmenting documents to remove confounding information. Finally,

subtle errors suggest that deeper linguistic analysis is warranted, especially targetted towards Finance text.

6.5 Summary

This chapter presented the major results of this thesis. Our best model for classifying ASX-RNA pairs as LINK performs with an F-score of 89.5% and the other journalistic contribution tasks between 73.4% and 85.7% F-score. Several of the features that we presented in Chapter 5 improve results with strong statistical significance. Our models also perform well in evaluation against a held-out dataset, performing significantly better than baseline in almost all cases. The contributions made by these results demonstrate that we can feasibly track information flow in Finance text.

CHAPTER 7

Conclusion

This thesis represents a study of information flow in Finance text. We present a review of existing research that motivates our modelling of textual and temporal aspects of information flow. The primary and secondary news sources that we study are then introduced: ASX announcements and RNA stories. The annotation scheme defines broad information flow (LINK) and additional story labels that indicate the first mention of information (FIRST), added background (BACK) and analysis (ANLY) content.

We frame classification tasks over ASX-RNA pairs where the task is to predict whether a type of information flow applies between the pair documents. Different NLP areas motivate information flow *features* which are combined using Maximum Entropy modelling. Bag of words features from IR precisely represent the text used and we define set-theoretic variants tailored to document pairs. Specifically, we create distinct bags of unigrams and bigrams corresponding to the intersection and set differences of the two documents. Similarity metrics used in Text Reuse give a more general indication of textual similarity. Common subsequences and entire sentence matches features are derived from Plagiarism Detection and model longer sections of repeated text. We also model features specific to the ASX and RNA data, namely topic and price sensitivity categories. Numbers are important to Finance text and we calculate similarity of numeric tokens and also represent the precision of numbers that occur in both documents. News data has a temporal dimension and we model the time lag between the ASX announcement and RNA story publication as well as the release time of day.

7.1 Future work

This thesis can be seen as an investigation of information flow and there are promising extensions. The internet provides an immense volume of news sources varying in genre, format and influence, including established news sites such as Reuters, forums and weblogs. Scaling the system to these larger, more

diverse datasets represents significant algorithmic and engineering challenges. The current system assumes a primary and secondary source and while this is pragmatic for a short project, it is more accurate to consider information flow across a network of news sources. We would instead model an *information graph*, whereas the current model is equivalent to studying just two nodes (primary and secondary). The current system is able to leverage the fact that Reuters closely follows the official ASX announcements and that newsworthy announcements are reported soon after their release. An information graph means that this will no longer be possible since information may flow through multiple nodes to reach the destination, requiring study of the wider news ecosystem. This may reveal closely related sub-regions, or *cliques*, that correspond with sources through which news spreads quickly. Furthermore, it may be possible to identify redundant news sources that simply syndicate information rather than add value.

Additional work would certainly require deeper computational linguistic analysis of the text. The current system's naïve performance can be loosely described as conservative and while information flow cases are precisely identified, there is more scope for improvement in its ability to recognise *more* of the information flow cases. Given the importance of numeric expressions in Finance text, it is important that they are correctly processed. The current system, for example, would not recognise the similarity between *2.45 million* and *2,450,000*. The identification task is at a document level, where in reality, documents can contain many different items of information. Sub-document processing and higher-level linguistic representation of the information will allow the system to generalise more effectively and identify the cases which it currently misses. For example, if the system can describe events in a more abstract manner, then the precise wording of the information, the surface textual form, is less important. This will come at a higher processing cost, underlining the importance of developing efficient algorithms.

In addition to these issues, practical applications might require a slightly different approach. The current dataset consists of a snapshot of stories from between 2005 and 2006. While a static dataset is a critical aspect of designing experiments to compare different systems, it is less useful for other modes of operation. TDT's online event detection and Novelty Detection provide real-world scenarios where systems monitor news streams, identifying new events or information; clearly a useful application for an information flow system. The current system partially represents information novelty using the FIRST label. This isn't entirely appropriate since FIRST is a document pairwise attribute that describes a cluster phenomenon. A FIRST story precludes *other* stories in the same cluster being marked FIRST. Ideally, an online version of our system would need to incrementally process announcements and stories as they

are released. Modelling word distribution over time would allow us to model new information (i.e., previously unseen words) and redundant information (i.e., recently mentioned words). When combined with sub-document processing, these methods would support powerful tools for monitoring information flow in finance text. One example of this is a tool that emphasises fresh and filters redundant information to allow investors or surveillance analysts to absorb and react to information more quickly. The existing system would be suitable for identifying information flow in a body of Finance text, creating hyperlinks to allow convenient exploration for audit purposes.

7.2 Results

Our system classifies information flow in ASX-RNA pairs (i.e., the LINK task) at 89.5% F-score. It has lower performance in tasks classifying the story's journalistic contribution. The FIRST task identifies pairs where the story reports information for the first time and the system performs at 73.4% F-score. Background (BACK) and analysis (ANLY) content are classified at 85.7% and 80.4% respectively. The results are all substantially above baseline approaches that use textual and temporal features in isolation. In evaluation experiments using a held-out test set, the system achieves an F-score 9% lower than human performance for FIRST classification and is significantly above baseline performance. Performance at recognising journalistic contribution is lower again, but above baseline in all cases (significantly for BACK).

7.3 Contribution

This thesis has formulated a new approach to the information flow problem in a Finance domain. The problem of identifying information flow is treated as a text classification task over the *intersection* of the two documents. We also draw on techniques from other NLP research areas to provide a generalisable model of textual similarity. These textual features are combined with temporal features to model information flow.

We present a scheme and tool for annotating information flow. These are the result of substantial consultation with Finance annotators and were designed to allow efficient annotation and navigation of complex collections of stories. Good kappa scores show that the scheme can be applied reliably by an annotation team to provide training and evaluation data for automated approaches.

We successfully combine approaches from different NLP research areas to address the information flow problem. Textual and temporal features are combined using Maximum Entropy modelling to accurately represent information flow and journalistic contribution.

We are satisfied that our model performs well under experimental conditions. Furthermore, this performance makes it immediately useful within other systems. Tannert (2009) uses a prototype of our model in research into the Financial implications of information flow. Automatic identification of information flow allows statistical analysis of a much larger dataset than if the task had been performed manually. We have demonstrated that it is feasible to automatically track information flow in Finance text, which will have a substantial impact on automated financial surveillance and trading systems.

Bibliography

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998a. Topic detection and tracking pilot study final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

James Allan, Ron Papka, and Victor Lavrenko. 1998b. On-line new event detection and tracking. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM, New York, NY, USA.

ASX. 2008. Continuous disclosure. *ASX Listing Rules*, Chapter 3.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Y Bernstein and J Zobel. 2004. A scalable system for identifying co-derivative documents. *Lecture Notes in Computer Science*.

Stephen Bird, Edward Loper, and Ewan Klein. 2009. Natural language processing with python.

Sergey Brin, James Davis, and Héctor García-Molina. 1995. Copy detection mechanisms for digital documents. In *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 398–409. ACM, New York, NY, USA.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. *Proceedings of the 29th annual meeting on Association for Computational Linguistics*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Nancy Chinchor. 1995. Statistical significance of muc-6 results. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 39–43. Association for Computational Linguistics, Morristown, NJ, USA.

Paul Clough, Robert Gaizauskas, Scott S. L. Piao, and Yorick Wilks. 2002. Meter: Measuring text reuse. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 152–159. Association for Computational Linguistics, Morristown, NJ, USA.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

I Dagan, O Glickman, and B Magnini. 2006. The pascal recognising textual entailment challenge. *Lecture Notes in Computer Science*.

J Darroch and D Ratcliff. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*.

Hal Daumé III. 2004. Notes on cg and lm-bfgs optimization of logistic regression.

B Di Eugenio and M Glass. 2004. The kappa statistic: A second look. *Computational linguistics*.

D Eichmann, P Srinivasan, M Light, H Wang, and X Qiu. 2003. Experiments in novelty, genes and questions at the university of iowa. *TREC Notebook Proceedings*.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184. Association for Computational Linguistics, Morristown, NJ, USA.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. 2004. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, New York, NY, USA.

Vasileios Hatzivassiloglou, Judith L. Klavans, and Eleazar Eskin. 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212. SIGDAT, College Park, Maryland.

Timothy C. Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *Journal of the American Society for Information Science and Technology*, 54(3):203–215.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Newbury Park, CA, USA.

Victor Lavrenko, James Allan, Edward DeGuzman, Daniel LaFlamme, Veera Pollard, and Stephen Thomas. 2002. Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research*, pages 115–121. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Kenneth C. Litkowski. 2003. Use of metadata for question answering and novelty tasks. In *TREC*, pages 161–176.

R Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. *International Conference On Computational Linguistics*.

Chris Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. *Cambridge University Press New York, NY, USA*.

A Martin, George Doddington, T Kamm, and M Ordowski. 1997. The det curve in assessment of detection task performance. *Fifth European Conference on Speech Communication and Technology*.

K McKeown, R Barzilay, and D Evans. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. *Proceedings of the second international conference on Human Language Technology Research*.

Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524. ACM, New York, NY, USA.

Kamal Nigam, John Lafferty, and Andrew Mccallum. 1999. Using maximum entropy for text classification. In *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67.

Jorge Nocedal and Stephen J. Wright. 1999. *Numerical Optimization*. Springer, New York, USA.

Adwait Ratnaparkhi. 1998. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis, Philadelphia, PA, USA. Supervisor-Marcus, Mitchell P.

Gerard Salton, A Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Ian Soboroff. 2004. Overview of the trec 2004 novelty track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*.

Karen Spärck Jones. 1973. Index term weighting. *Information Storage and Retrieval*, 9(11):619–633.

Silvio Tannert. 2009. *Information Transformation in Financial Markets: Evidence from Australia*. Master's thesis, Technische Universität Darmstadt, Darmstadt, Germany.

Michael J Wise. 1996. Yap3: improved detection of similarities in computer program and other texts. *SIGCSE Bull.*, 28(1):130–134.

Akbar Zaheer and Srilata Zaheer. 1997. Catching the wave: alertness, responsiveness, and market influence in global electronic networks. *Manage. Sci.*, 43(11):1493–1509.