# Corpus and Sentiment Analysis 

## Tai Wai David Cheng

Submitted for the Degree of Doctor of Philosophy
May 2007

## Unis

Department of Computing School of Electronics and Physical Sciences<br>University of Surrey<br>Guildford,<br>Surrey<br>GU2 7XH, United Kingdom


#### Abstract

Information extraction/retrieval has been of interest to researchers since the early 1960's. A series of conferences and competitions have been held by DARPA/NIST since the late 1980's has resulted in the analysis of news reports and government reports in English and other languages, notably Chinese and Arabic. A number of methods have been developed for analysing 'free' natural language texts. Furthermore, a number of systems for understanding messages have been developed, focusing on the area of named entity extraction, templates for dealing with certain kinds of news. The templates were handcrafted, and a lot of ad-hoc knowledge went into the creation of such systems. Seven of these systems have been reviewed. Despite the fact that IE systems built for different tasks often differ from each other, the core elements are shared by nearly every extraction system. Some of these core elements such as parser and part of speech (POS) tagger, are tuned for optimal performance for a specific domain, or text with pre-defined structures. The extensive use of gazetteers and manually crafted grammar rules further limits the portability of the existing IE systems to work language and domain independently. The goal of this thesis is to develop an algorithm that can be used to extract information from free texts, in our case, from financial news text; and from arbitrary domains unambiguously. We believe the use of corpus linguistics and statistical techniques would be more appropriate and efficient for this task, instead of using other approaches that rely on machine learning, POS taggers, parsers, and so on, which are tuned to work for a predefined domain. Based on this belief, a framework using corpus linguistics and statistical techniques, to extract information as unambiguously as possible from arbitrary domains was developed. A contrastive evaluation has been carried out not only in the domain of financial texts and movie reviews, but also with multi-lingual texts (Chinese and English). The results are encouraging. Our preliminarily evaluation, based on the correlation between a time series of positive (negative) sentiment word (phrase) counts with a time series of indices produced by stock exchanges (Financial Times Stock Exchange, Dow Jones Industrial Average, Nasdaq, S\&P 500, Hang Seng Index, Shanghai Index, and Shenzhen Index) showed that when the positive (negative) sentiment series correlates with the stock exchange index, the negative (positive) shows a smaller degree of correlation and in many cases a degree of anti-correlation. Any interpretation of our result requires a careful econometrically well grounded analysis of the financial time series - this is beyond the scope of this work.


## Acknowledgments

I would like to express my sincere thanks to Professor Khurshid Ahmad, for his encouragement and supervision throughout my research. My thanks also go to the EU co-funded project - Generic Information based Decision Assistant (GIDA - IST-200031123), and the FinGrid Project (ESRC Project Number RES-149-25-0028) - for financial backing.

My thanks are further extended to my colleagues, Lee Gillam, Tugba Taskaya, Paulo Oliveria, Saif Ahmad and Hayssam Traboulsi, who have been working in the GIDA project. Thanks also to Dr. Andrew Hippisley who has provided linguistic insights when I began my Ph.D.

Most importantly, I thank my parents, my sisters, and, particular thanks due to my friend Pensiri Manomaisupat, in the preparation of this work. Without their patience, understanding and support, it would not be possible for me to finish this thesis. To them, I will always be grateful, and to them, I dedicate this thesis.

## Contents

List of Figures ..... iii
List of Tables ..... iv
1 Introduction ..... 1
1.1 A Corpus-based Study of Market Movement ..... 5
1.2 Structure of the Report .....  .6
2 Motivation and Literature Review ..... 8
2.1 Information Extraction ..... 8
2.2 Sublanguages and local grammars ..... 23
2.3 Sentiment Analysis ..... 29
2.4 Sublanguages and Idiom Principle ..... 34
2.5 Conclusion ..... 38
3 Method ..... 41
3.1 Introduction ..... 41
3.2 Algorithm ..... 42
3.2.1 Select corpora ..... 44
3.2.2 Generate wordlist and statistical information for each token. ..... 45
3.2.3 Extract candidate terms ..... 47
3.2.4 Extract key collocates for each candidate term ..... 48
3.2.5 Extract local grammar from collocates ..... 52
3.3 Exemplar ..... 52
3.4 Conclusion ..... 57
4 Case Studies and Evaluation ..... 60
4.1 Case Studies ..... 60
4.1.1 Sentiment Analysis for Financial Texts in English ..... 61
4.1.2 Cross-validation: Sentiment Analysis for Financial Texts in Chinese ..... 73
4.1.3 Sentiment Analysis for Film Reviews ..... 85
4.2 Evaluation ..... 92
4.2.1 The case for English financial news ..... 92
4.2.2 The case for Chinese financial news ..... 97
4.2.3 The case for film reviews ..... 98
4.3 System Description ..... 100
4.3.1 The Wordlist Component ..... 101
4.3.2 The KWIC Component ..... 103
4.3.3 The Collocation Component ..... 104
4.4 Conclusion ..... 106
5 Conclusions and Future Work ..... 109
5.1 Conclusion ..... 109
5.2 Main Contribution ..... 110
5.3 Future Work ..... 113
Bibliography ..... 117
Appendix A - Publications ..... 125
Appendix B ..... 127
Appendix C ..... 128
Appendix D ..... 131
Appendix E ..... 133
Appendix F. ..... 134
Appendix G ..... 135

## List of Figures

Figure 2．1．Typical Architecture of Information Extraction Systems． ..... 10
Figure 2．2．LaSIE－II system architecture． ..... 17
Figure 2．3．System architecture of the Information Extraction System developed by the ECRAN project． ..... 18
Figure 2．4．System architecture of the InfoXtract system． ..... 20
Figure 2．5．System architecture of the PV Identification Module． ..... 21
Figure 2．6．A local grammar representation of the adverb speaking． ..... 25
Figure 2．7：A local grammar for the identification of date expressions． ..... 26
Figure 2．8．A typical example of a news story from Reuters Newsfeed，published on 31 August 2006. ..... 33
Figure 3．1．Stage 1：Selecting input corpora． ..... 45
Figure 3．2．Stage 2：Generate wordlist and statistical information for each token in CORPUS ${ }_{c}$ ..... 47
Figure 3．3．Stage 3：Extract candidate terms from CORPUS $_{c}$ ． ..... 48
Figure 3．4．Stage 4：Extract key collocates for each candidate term in CORPUS $_{c}$ ． ..... 51
Figure 3．5：Upward and downward collocates of fell． ..... 56
Figure 3．6．Local grammar of the candidate term fell． ..... 56
Figure 4．1：Left and right neighbours of collocate percent． ..... 69
Figure 4．2：A local grammar to extract positive market sentiment in English． ..... 71
Figure 4．3：A local grammar to extract negative market sentiment in English． ..... 72
Figure 4．4：Left and right neighbours of collocate 百分之． ..... 81
Figure 4．5：A local grammar to extract positive market sentiment in Chinese． ..... 83
Figure 4．6：A local grammar to extract negative market sentiment in Chinese． ..... 83
Figure 4．7：A local grammar to extract market sentiment in Arabic． ..... 85
Figure 4．8：Left and right neighbours of collocate film． ..... 90
Figure 4．9．User interface of the Text Analysis System． ..... 100
Figure 4．10．Filtering and Sorting the wordlist． ..... 102
Figure 4．11．Concordance of the candidate term percent． ..... 103
Figure 4．12．Statistical information for the collocate percent in summarised view． ..... 104
Figure 4．13．Statistical information for the collocate percent in full details． ..... 105
Figure 4．14．Collocates of the Candidate Term percent． ..... 106
Figure 4．15：SATISFI in action：Correlation between news and instrument time series． ..... 107

## List of Tables

Table 1.1: A cross-domain and cross-lingual case studies. ..... 7
Table 2.1: Performance of the FASTUS System in MUC-4 ..... 13
Table 2.2: Performance of the LOUELLA PARSING System in MUC-6 ..... 14
Table 2.3: AutoSlog heuristics (grammar rules) and examples from the terrorism domain. ..... 15
Table 2.4: Comparative results of the AutoSlogs. ..... 16
Table 2.5: Performance of the LaSIE-II system in MUC-7 ..... 18
Table 2.6: Performance of the corpus based information extraction system from the ECRAN project. ..... 19
Table 2.7: Performance of the InfoXtract system. ..... 21
Table 2.8: Performance of the PV identification system. ..... 22
Table 2.9: Definitions of the three major analysis systems used in financial trading. ..... 30
Table 2.10. Objects being analysed in the three major analysis systems used in financial trading. ..... 30
Table 2.11: Health metaphors identified by Knowles (1996) ..... 32
Table 2.12: Examples of the lexical content of a news story and the implied sentiment. ..... 34
Table 2.13: Sample of concordance output for fell. ..... 36
Table 3.1: Diachronic distribution of text in the Reuters Corpus ..... 44
Table 3.2: Statistical information for the top 10 tokens that co-occur with the candidate term shares ..... 49
Table 3.3: Filtering effect by using frequency and frequency $z$-score, or weirdness and weirdness $z$-score. ..... 53
Table 3.4: Filtering effect by using the combination of frequency and frequency $z$-score, and weirdness and weirdness $z$-score. ..... 54
Table 3.5: Collocations for the candidate term fell. ..... 55
Table 4.1: Diachronic distribution of text in Reuters corpus ..... 62
Table 4.2: Distribution of the first 100 most frequent tokens in the BNC according to the cumulative frequency of ten tokens at a time. ..... 63
Table 4.3: Distribution of the first 100 most frequent tokens in the Reuters corpus according to the cumulative frequency of ten tokens at a time. ..... 64
Table 4.4: Distribution of tokens ranked between 101-200 in the Reuters corpus, grouped ten at a time. ..... 65
Table 4．5：Distribution of tokens ranked between 101－200 in the BNC grouped ten at a time． Note that there are some abbreviations in this table，e．g．the title mr and erm（European exchange Rate Mechanism） ..... 65
Table 4．6：Comparison between system generated candidate terms and user selected candidate terms，ranked by frequency and grouped ten at a time ..... 67
Table 4．7：Ten most frequent collocates of percent generated by the Text Analysis System ..... 68
Table 4．8：Tri－gram collocates of percent，with frequency $>=100$ ..... 69
Table 4．9：Distribution of the first 100 most frequent tokens in the $\mathrm{TaBE}+$ LDC corpus according to the cumulative frequency of ten tokens at a time． ..... 74
Table 4．10：Distribution of the first 100 most frequent tokens in the Ming Pao Financial Corpus according to the cumulative frequency of ten tokens at a time． ..... 75
Table 4．11：Distribution of tokens ranked between 101－200 in the Ming Pao corpus according to the cumulative frequency of ten tokens at a time． ..... 76
Table 4．12：Distribution of tokens ranked between $101-200$ in the TaBE＋LDC corpus，grouped ten at a time． ..... 77
Table 4．13：Comparison between system generated candidate terms and user selected candidate terms，ranked by frequency and grouped ten at a time． ..... 79
Table 4．14：Ten most frequent collocates of 百分之 generated by the Text Analysis System． ..... 80
Table 4．15：Tri－gram collocates of 百分之，with frequency $>=20$ ..... 81
Table 4．16：Distribution of the first 100 most frequent tokens in the IMDB training corpus according to the cumulative frequency of ten tokens at a time． ..... 86
Table 4．17：Distribution of tokens ranked between 101－200 in the IMDB training corpus，grouped ten at a time． ..... 87
Table 4．18：Comparison between system generated candidate terms and user selected candidate terms，ranked by frequency and grouped ten at a time． ..... 88
Table 4．19：Ten most frequent collocates of film generated by the Text Analysis System． ..... 89
Table 4．20：Ten most frequent collocates of movie generated by the Text Analysis System ..... 90
Table 4．21：$N$－gram collocates expressing reviewer＇s sentiments，with frequency $>16$ ． ..... 91
Table 4．22：Distribution of news articles between 01 August and 31 August in the Reuters Newsfeed． ..... 93
Table 4．23：Correlation between financial instruments and all the news articles from Reuters Newsfeed in August 2006 using RAW data ..... 94
Table 4．24：Correlations between financial instruments and all the news articles from Reuters Newsfeed in August 2006 using Return and Volatility data ..... 95
Table 4．25：Distribution of the U．S．and U．K．news articles in Reuters Newsfeed between 01 August and 31 August． ..... 95
Table 4.26: Correlations between financial instruments and country specific news articles from Reuters Newsfeed in August 2006 ..... 96
Table 4.27: Correlations between financial instruments and country specific news articles from Reuters Newsfeed in August 2006 using Return and Volatility data ..... 96
Table 4.28: Distribution of news articles published by Xinhuanet between 02 July and 27 July 2007 ..... 97
Table 4.29: Correlations between financial instruments and news articles from Xinhuanet within July 2007. ..... 98
Table 4.30: Correlations between financial instruments and news articles from Xinhuanet in July 2007 using Return and Volatility data. ..... 98
Table 4.31: Evaluation result for detecting sentiments in film reviews ..... 99
Table 4.32: Evaluation result for detecting sentiments in film reviews, after removing neutral reviews ..... 99

## Chapter 1

## 1 Introduction

The behaviour of financial markets is governed by the gains and losses of the traders and investors in the markets. An additional, qualitative, largely intuitive, and often disputed factor is the mood of the investors or traders. The mood of a given investor or the overall investing public, can be either bullish or bearish ${ }^{1}$. The metaphors, bullish and bearish, so-called animal metaphors, refer to the aggressive or recessive (shy) mood of the investors and perhaps of the traders.

The disputed, qualitative factor is termed 'sentiment' or 'market sentiment' in the investors' lexicon. The term sentiment is related to the more quantitative measures expressed in equally colourful terms including market volatility index. This index is based solely on the increment and decrement in prices of financial instruments. (A financial instrument is defined as any tool that can be used in order to implement economic policy, hence having monetary value or recording a monetary transaction). The market volatility index is "used as an indicator of investor sentiment, with high values implying pessimism and low values implying optimism".'

Sentiment analysis has started at Surrey since 1997. In two EU-IST projects (Analyst Control Environment 1997-1999 and Generic Information based Decision Assistant 2001-2003) the scope of sentiment analysis was extended to include the analysis of full news texts (Ahmad, Vrusias and Ledford, 2001; Gillam, Ahmad, Ahmad, Casey, Cheng, Taskaya, Oliviera and Manomaisupat, 2002; Ahmad, Cheng, Taskaya, Ahmad, Gillam, Manomaisupat, Traboulsi and Hippisley, 2003). Supported by ESPRIT and the EPSRC, Surrey pioneered (1985-96) a corpus-based approach to terminology extraction using techniques like frequency analysis, concordance and collocation analyses (Ahmad and

[^0]Rogers, 2001). Subsequently, these techniques have been refined and applied to ontology learning (Gillam, Tariq and Ahmad, 2005) where conceptual hierarchies can be automatically extracted from arbitrary texts. These approaches have been developed and tested in the software of the System Quirk framework for language analysis, developed in close collaboration with applied linguists and translators (Kugler, Ahmad and Thurmair, 1995).

The key objective of the EU-sponsored GIDA project ${ }^{3}$ is to investigate how market sentiment is reported in the public-domain information sources; for example, financial news wire services and monetary institutions (including the US Federal Reserve, Bank of Japan, Bank of England), and how market sentiment might influence the trading in financial instruments. One of the specific objectives of the GIDA project is to identify trends in the news reports related to (a specific) market sector and to correlate these trends with aggregated numerical indices, like the Financial Times Stock Exchange (FTSE) index. These aggregates are computed at discrete intervals, essentially being the geometric mean of the share-values of selected enterprises within a market. A time series of individual aggregates is produced.

The goal of this thesis is to develop an algorithm that can be used to automatically extract information from free texts, in our case, from financial news texts; and from arbitrary domains unambiguously. We believe the use of corpus linguistics and statistical techniques will be more appropriate and efficient for this task, rather than using other approaches that rely on machine learning, part of speech (POS) taggers, parsers, and so on, which have been tuned to work for a pre-defined domain. When switching to an arbitrary domain, such approaches may either fail to work, or require an inevitable retraining from scratch to adapt to the new domain. Based on these, a framework using corpus linguistics and statistical techniques to extract information unambiguously from arbitrary domains was developed. The framework includes:

- a tool to handle text documents in different formats including $X M L, H T M L, P D F$, $D O C, R T F, X L S, P P T$, and $T X T$;
- a tool to identify and extract domain specific terms automatically, which was done manually by the domain experts in the past;

[^1]- a tool to generate concordance for the automatically identified terms;
- a tool to generate collocates for the domain specific terms;
- a tool to visualise the results akin to Excel spreadsheet style;
- a local grammar constructed from the collocation patterns which can extract domain specific information unambiguously;

A Text Analysis System was implemented in Java programming language. The Text Analysis System automatically identifies and extracts domain specific terms and their associated collocates from arbitrary domain using corpus linguistics and statistical techniques such as weirdness, collocation, $z$-score, and so on. Based on the collocation patterns generated, a local grammar can be constructed to extract domain specific information unambiguously.

The hypothesis, which was used to develop the algorithm, can be expressed as follows: the expression of opinions about the behaviour of an entity, say, an enterprise or a collection of enterprises, can be diffuse and ambiguous. Hence, methods have been developed to visualise this behaviour in as simple terms as is possible: the work in financial time series analysis, which according to Kendall (1976) dates back as far as the $19^{\text {th }}$ century, exemplifies this. One now sees quite complex "charts" ${ }^{4}$ that incorporate a wealth of information graphically displayed, giving, for example, opening, closing, high and low values of an instrument on it is traded, many times, throughout the day. These charts are typically embedded within a report in a financial newspaper or have hyperlinks to a linguistic description of the behaviour of the instrument. Commentaries on these charts refer to events and objects that may or may not have stabilised, or which may have precipitated change in the value of the instrument. One can argue that external references influence investor or trader sentiment, and that the references to these influences are couched in words that may express optimism or pessimism.

The expression of optimism or pessimism relies on a choice of words whose meaning is generally understood. This is not to say that the words used in the expression of market sentiment have been standardised in the way the terminology of science and technology have been standardised and where the standardisation is enforced by the editors of

[^2]scientific/technological publications. Rather, there is a general consensus on how to express optimism or pessimism about an instrument. Enterprises nowadays provide guidance about their profitability over a financial year; the (now only marginally-used) term used to be prediction; the term for market crisis used to be market crash.

Financial reports, especially summaries of stock market behaviour on a daily basis, are reported as stocks that rose the most and stocks that fell the most - risers and losers as they are commonly called by some. Market movement is described in terms of patterns, trends and cycles, and the patterns form heads of compound terms like uptrend, downtrend, boom-and-bust cycles and peaks and troughs (of cycles). The consensual words then make their way into (more) written texts and are noted by lexicographers, first of the specialist variety - the terminologists - and subsequently those who record and teach language, then, more generally - the dictionary makers.

The words rise and fall have many senses and both cross grammatical categories. Much the same can be said of their inflections. However, financial report writers constrain the meaning by encoding the words within specific grammatical patterns. The verb rose has many senses as found in the Collins Bank of English:

| access was willed, the barrier <br> On the far side pale grey cliffs | rose <br> rose | : the horror of the scene thus |
| :--- | :--- | :--- |
| , their tops level with the ground on |  |  |

but financial reporting has appropriated the word by co-locating its use with a cardinal number:
stood at 22.7 per cent. The shares rose 14 p to 254 p
The Commerce Department says imports
1988 and 1994 it operating expenses
now are good and the summer's
up $£ 1.5$ million as the index
rose $41 / 2$ percent
rose 50 per cent
rose 7
rose to 21,500 points
. [p] [h] Spring Ram to an eight-month to \$ 2 billion, with a last-minute rush. [h] . But in September

Similarly the use of the word "fell" has been appropriated:
> industry for raw materials and fuel any certainty. Second-section prices government says sales of new homes Michael Clark [/b] [p] Abbey National

> Average price of new properties
in October, depressing on three million shares last month, helping to as a large line of shares in the two years

We have carried out an analysis of one year collection of Reuters news wires in order to study the behaviour of a number of verbs, adverbs and prepositions that are used to indicate how a financial instrument has changed its value. This change can be then used to infer the so-called market sentiment.

### 1.1 A Corpus-based Study of Market Movement

What does the word "corpus" mean? In principle, any collection of more than one text can be called a corpus, corpus being Latin for "body" (Oxford English Dictionary); hence a corpus is any body of text. In this case, when we speak of the "Reuters corpus", we are referring a collection of texts published by Reuters. On the other hand, in the Dictionary of Linguistics and Phonetics, corpus is defined as a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language. Sinclair (1991) refers to a corpus as a collection of naturally occurring language texts, chosen to characterize a state or variety of a language. Considering the three definitions, corpus is a text collection designed for the purpose of linguistic analysis, a way to represent elements of a language, or some part/characteristics of a language; it is not merely a collection of any texts. Therefore, what a corpus is meant to represent depends on the design and structure of the corpus.

Corpora of texts are increasingly being used to investigate a range of literary and linguistic phenomena: from authorship attribution to genre analysis; from lexicographic evidence to language change; from the study of dialects to syntactic and semantic analysis; from optical character recognition studies to language development and second language acquisition. Corpora of texts can also be used to investigate trends in science and technology, particularly through the analysis of texts produced in the different parts of the world and then relating the tokens in the text (including terminology usage, authorship and institutional aspects) to major developments in scientific methods, product developments and innovation management. The advent of the World-Wide Web (WWW) comprising texts, images and sounds from many countries and in many languages, has added and will continue to add to the amount of available texts. This is
reflected in the availability of the network 'crawlers', programs designed to access texts across the WWW based on keyword search. It is, for example, possible to collect literally hundreds of texts in the languages of numerically-determined minorities in a range of disciplines, genres and so on. Thus, by collecting a large corpora of texts used in a specialist domain, and then analysing that corpora of texts to create a lexicon, and through collocation and colligation, we can find patterns of usage of critical lexical items which are specific to the domain. In the financial domain, these patterns may provide us some insights about the changes in value of financial instrument(s).

### 1.2 Structure of the Report

In Chapter 2, a literature review covering topics in information extraction (IE) and performance of the recent IE systems will be discussed. These IE systems include FASTUS from SRI International; AutoSlog and AutoSlog-TS by Ellen Riloff; LOUELLA PARSING System by LockHeed Martin; LaSIE-II (Large Scale Information Extraction) developed at the University of Sheffield; a corpus based multilingual information extraction system used in the ECRAN project; the InfoXtract system that supports information discovery; and the Phrasal Verbs identification system for extracting multiword terms. The common factors amongst all the systems are to do with the reliance on POS tagging, parsing, and the extensive use of gazetteers. Such dependence limits the portability of the existing IE systems to work on arbitrary domains. Alternative approaches to information extraction using corpus linguistics and sublanguages will also be discussed.

In Chapter 3, a corpus based approach for identifying domain specific terms will be discussed. Our approach briefly has been to collect a large corpora of texts used in a specialist domain, then analyse that corpora of texts based on frequency values, z -scores, weirdness and Smadja's statistics for collocations, from which patterns associated with domain specific terms, can be automatically extracted. For instance, in the finance domain, some patterns used by financial journalists and reporters to report the changes in value of financial instrument(s) will be described. Through the use of local grammar, information can be extracted unambiguously. We will describe an algorithm for
automatically extracting domain specific terms and their collocation patterns from an arbitrary domain using corpus linguistic and various statistical techniques.

In Chapter 4, we will demonstrate the Text Analysis System that implements the algorithm described earlier in Chapter 3, and illustrate how our Text Analysis System can aid linguists to verify their hypotheses, to deal with multi-lingual texts, in arbitrary domains, through three case studies: sentiment analysis in both English and Chinese financial news texts, and sentiment analysis in movie reviews, as illustrated in Table 1.1 below:

Table 1.1: A cross-domain and cross-lingual case studies.

|  | Finance | Film Reviews |
| :--- | :---: | :---: |
| English | $\checkmark$ | $\checkmark$ |
| Chinese | $\checkmark$ | X |

In addition to the case studies, effectiveness of the algorithm (in our Text Analysis System) will be evaluated. For the financial sentiment analysis, extracted sentiments from the financial news will be aggregated as a time series, and then correlated to financial instruments. While for the film reviews taken from the Internet Movie Database (IMDB), the results will be compared against unsupervised learning approach employed by Turney (2002) and the machine learning approach employed by Pang, Lee and Vaithyanathan (2002) who have pioneered research in this particular domain.

Finally, Chapter 5 concludes the thesis by summing up the work done so far. Limitations of the algorithm, suggestions for improving the system, and possibilities for future research will also be discussed in this chapter.

## Chapter 2

## 2 Motivation and Literature Review

### 2.1 Information Extraction

Information extraction and retrieval has been of interest to the researchers since the early 1960's. However, due to the unprecedented advancement in technology, network infrastructure and disk storage since the late 1990's, it has become a norm for organisations to publish information through the Internet. This is particularly common for news vendors like Reuters, Bloomberg, CNN, BBC, and so on. The availability of online news wire services and the online newspapers has led to a profusion of studies of language based on collections of news reports. These studies on the whole relate to the extraction of information about specific topics: terrorism, drug trafficking, sports are amongst the popular topics. The automatic analysis of the news reports for gathering information about specific topics has been the basis of substantial amounts of work undertaken by the US Defense Advanced Research Projects Agency (DARPA) and the US National Institute of Standards and Technology (NIST).

A series of conferences and competitions have been held by DARPA/NIST since the late 1980's has resulted in the analysis of news reports and government reports in English and other languages, notably Chinese and Arabic. DARPA/NIST have organised the well attended Text REtrieval Conferences (TREC). A number of methods have been developed for analysing 'free' natural language texts. The Message Understanding Conference (MUC), also sponsored by DARPA lead to the development of a number of systems for understanding messages, and here the focus was on named entity extraction, templates for dealing with certain kinds of news. The templates were hand-crafted, and a lot of ad-hoc knowledge went into the creation of such systems.

IE systems have been developed for texts ranging from structured text with tabular information to free text such as news stories. The key component of such systems is a set of extraction rules for which relevant information can be identified and extracted. This has been widely described by Soderland (1999). There are two main approaches to the design of IE systems - the Knowledge Engineering Approach (KEA) and the Automatic Training Approach (ATA) (Appelt and Israel, 1999). In the KEA, grammars expressing rules for the system are constructed by hand using knowledge of the application domain. The skill and expertise of the knowledge engineer regarding the application domain is essential to the level of performance of the system, as hand crafted systems often performs the best. However, such development processes can be very lengthy as manually investigating domain-relevant texts is required, and it is difficult to accommodate changes in system specifications. Moreover, the required expertise can be problematic to obtain in some circumstances.

For the ATA, instead of requiring system experts when customising the system for a new domain, someone with sufficient knowledge that is capable of annotating a set of training documents of the domain is sufficient. The system is able to analyse novel texts once a training corpus has been annotated and a training algorithm is run. Since this approach relies on training data, sometimes it may be problematic and expensive to obtain a well annotated corpus.

Despite the fact that IE systems built for different tasks often differ from each other, the core elements are shared by nearly every extraction system, regardless of whether they follow the Knowledge Engineering or Automatic Training approach. Figure 2.1 below shows the typical architecture of IE systems, as described by Hobbs (1993), and later by Appelt and Isreal (1999:11):


Figure 2．1．Typical Architecture of Information Extraction Systems．

The four primary modules that every information extraction system shares，are Tokenisation，Morphological and Lexical Processing，Syntactic Analysis and Domain Analysis（specific to application）．Extra modules are likely to be required in addition to the four primary modules，depending on the requirements of a particular application．For example，in Chinese and Arabic texts，there is no word boundary in the sentence．The lack of an explicit token delimiter makes the otherwise simple problem of tokenisation and segmentation（compared with English）extremely difficult，requiring the development of a word segmentation module．The example below will illustrate this．

Let us look at the sentence
這名記者會說國語。 This journalist（記者）can（會）speak（說）Mandarin（國語）．
may be tokenised and segmented to：
／這／名／記者會／說／國語。The press conference（記者會）speaks（說）Mandarin（國語）．
clearly，this is an incorrect segmentation，as it is not possible for the＂press conference＂ itself to speak Mandarin．The correct segmentation should be：
／這／名／記者／會／說／國語。This journalist（記者）can（會）speak（說）Mandarin（國語）．
as it is not a surprise that a journalist (himself/herself) can speak Mandarin.
Approaches to Chinese word segmentation have been both symbolic (rule-based), for example, Yeh and Lee (1991), and statistical, for example, Chen and Liu (1992); Yao and Lua (1998); Peng and Schuurmans (2001).

Morphology is the study of word structure. The term word is ambiguous in common usage. Take the example of computer and computers, there is one sense that these two are the same "word" (same object, one being singular and one being plural), and the other sense in which they are different words (they cannot be used in the same sentences without modifying other words; the determiners this and these, the verbs is and are, as in This computer is powerful and These computers are powerful). The first sense of "word" - computer and computers are "the same word" is called the lexeme, and the second sense is termed the word form. Words are generally accepted as being the smallest units of syntax; it is clear that in most (if not all) languages, words can be related to other words by rules. For example, English speakers recognise that the words fish, fishes, and fisherman are closely related.

On the other hand, within the field of information extraction, lexical analysis is the processing of an input sequence of characters to produce, as output, a sequence of symbols called tokens. Furthermore, parsing is the process of analysing a sequence of tokens in order to determine its grammatical structure with respect to a given grammar.

Performance of IE systems can be evaluated by precision and recall metrics. Recall measures the information that has been correctly extracted, and precision measures the extracted information that is correct. In other words, recall refers to the amount of information that was correctly extracted, while precision refers to the reliability of that information extracted.

Precision is defined as:

$$
\begin{equation*}
\text { precision }=\frac{\# \text { correct answers }}{\# \text { answers produced }} \tag{1}
\end{equation*}
$$

and recall is defined as:

$$
\begin{equation*}
\text { recall }=\frac{\# \text { correct answers }}{\# \text { total possible corrects }} \tag{2}
\end{equation*}
$$

When comparing the relative performance of different IE systems, $F$-measure which combines both precision $(P)$ and recall $(R)$, is used:

$$
\begin{equation*}
F=\frac{\left(\beta^{2}+1\right) P R}{\beta^{2} P+R} \tag{3}
\end{equation*}
$$

The parameter $\beta$ determines how much to favour recall over precision. Usually, $\beta=1$, which means weighing precision and recall equally. For various aspects of the Information Extraction tasks, interannotator agreement has usually been in the 60-80\% range. This gives some idea of how difficult a task is. Another bit of evidence, of course, is how well the competing systems have done at MUC (Message Understanding Conference).

We will briefly review some of the major information extraction systems that have been used to analyse news stories amongst other genre of texts. Many of the systems have been used in the TREC and MUC series of competitions and they have a large number of commonalities. The major lesson we learn from these systems is that there is an important component which requires syntactic analysis in conjunction with part of speech tagging. Many systems use gazetteers, or pre-built thesauri of proper nouns, or some grammatical manipulation to find out named entities, and then there is considerable amount of parsing involved using traditional grammars, generally referred to as deep parsing. In addition, there are claims to detect relationships between objects within the text, or objects and events in the texts. Furthermore, some systems are claimed to be able to resolve coreferences and aliases.

There are seven systems that will be reviewed. The first one is the FASTUS system by SRI International (see Hobbs, Applet, Bear, Israel and Tyson, 1992) for extracting information from natural language text. We will then look at the LOUELLA PARSING System developed by General Electric Corporate Research and Development, subgroup of the Lockheed Martin Management and Data Division (see Childs, Guthrie and Sider, 1995). Next will be Ellen Riloff's (1996) AutoSlog System that can automatically generate extraction patterns from untagged text. Fourth, we will look at the LaSIE-II (Large Scale Information Extraction) system developed at the University of Sheffield using GATE (General Architecture for Text Engineering), which was used in the MUC
competition (see Humphreys, Gaizauskas, Azzam, Huyck, Mitchell, Cunningham, and Wilks, 1998). Fifth, we will look at Thierry Poibeau's (2000) corpus based information extraction system. After this, we will look at the InfoXtract developed by Srihari, Li, Niu , and Cornell (2003), which they claim to be a customisable information extraction engine. Finally, there is the Phrasal Verbs identification system due to Li, Zhang, Niu, Jiang, and Srihari (2003) for which they make similar claims. The common factors amongst all the systems is to do with parsing, or POS tagging, which is basically based on earlier work by Brill and other colleagues, which takes a statistical approach to POS tagging. We will begin with the description of FASTUS.

FASTUS, or Finite State Automaton Text Understanding System, is a system developed by the SRI International for extracting information from free natural language texts, primarily in English, but potentially for other languages as well. Analysis performed by FASTUS is comprised of four steps: Triggering, Recognising phrases, Recognising pattems and Merging incidents. But beforehand, preprocessing the input text into a standardized format is required. The Triggering step deals with identifying relevant sentences using trigger words manually defined. After that, finite state automaton are used for recognizing noun groups, verb groups, conjunctions, relative pronouns, and so on in these sentences. In the third stage, a set of hand-crafted patterns are used to extract relevant information. Finally, incidents like killing, kidnapping, bombing that were identified previously are merged together. Given that only a short development time was invested ( 6 months), performance of FASTUS is quite promising, as can be found in Table 2.1 below:

Table 2.1: Performance of the FASTUS System in MUC-4.

| Task | Precision | Recall | F-Measure |
| :--- | ---: | ---: | ---: |
| TST3 | $55 \%$ | $44 \%$ | $48.9 \%$ |
| TST4 | $52 \%$ | $44 \%$ | $47.7 \%$ |

Next, we will look at the LOUELLA PARSING System developed by Lockheed Martin Management and Data Division (Childs et al., 1993).

The LOUELLA PARSING System participated in three of the four MUC-6 evaluations: Named Entities (NE), Template Elements (TE) and Scenario Templates (ST). The system performs a sequential analysis:

- Text Tokenisation and Segmentation: performs the character string to word translation; and interprets the SGML markers and common punctuations;
- Lexical Look-up: identifies named entities and tags every word in the text;
- Text Reduction: use finite state machine to reduce elements like noun phrases to more complex IN_AND_OUT objects;
- Reference Resolution: link recognised named entities to their appositives or prenominal phrases;
- Information Extraction: extraction information of interest and send such information to the text organizer where relationship between object and events will be established;
- Postprocessing: last chance to apply any heuristics that may seem helpful to produce better results;
- Template Generating: the use of an object-oriented mapping scripts for generating templates.

The LOUELLA developers also appear to have developed a system based on what they regarded as the simplification of incoming news stories to syntactic analysis and other useful devices. The authors have also analysed sequence of temporal events empirically. It turns out that the named entity techniques that have been used in the LOUELLA PARSING System and projects of that kind had had their influence on subsequent projects in this field. The authors claim to have the following precision and recall statistics:

Table 2.2: Performance of the LOUELLA PARSING System in MUC-6.

| Task | System | Precision | Recall | F-Measure |
| :--- | :---: | ---: | ---: | ---: |
| Named Entities | A | $91 \%$ | $91 \%$ | $90.84 \%$ |
|  | B | $93 \%$ | $93 \%$ | $93.09 \%$ |
| Template Elements | A | $76 \%$ | $77 \%$ | $76.29 \%$ |
|  | B | $79 \%$ | $77 \%$ | $78.24 \%$ |
| Scenario Templates | A | $43 \%$ | $64 \%$ | $51.63 \%$ |
|  | B | $45 \%$ | $64 \%$ | $53.20 \%$ |

A: Official performance
B: Unofficial performance - after bug fix

Third, we will look at Ellen Riloff's (1996) AutoSlog System. AutoSlog was developed aiming at reducing the bottleneck in creating dictionary of domain specific extraction patterns. For example, Riloff estimated that it will take about 1,500 person-hours to build the UMass/MUC3 dictionary by hand. However, it only took 5 persons-hours to build a comparable dictionary using AutoSlog given an appropriate training corpus. The training corpus required by AutoSlog is a collection of texts where noun phrases are specially tagged, and the noun phrases must be labeled with their conceptual role and event type. For example, in the following sentence (Riloff, 1999:445-446):

The governor was kidnapped by terrorist commandos.
"the governor" should be labeled as a victim in a kidnapping event, and the "terrorist commandos" as a perpetrator in a kidnapping event. A parser is used for the recognition of clause boundaries, subject, verb, direct object, and prepositional phrases of each clause. Based on this, a set of manually crafted heuristics grammar rules for a specific domain is created. For example, Table 2.3 below shows some of the grammar rules, together with examples from the terrorism domain:

Table 2.3: AutoSlog heuristics (grammar rules) and examples from the terrorism domain.

| Linguistic Pattern | Example |
| :---: | :---: |
| <subject> active-verb | <perpetrator> bombed |
| <subject> passive-verb | <victim> was murdered |
| < subject> verb infinitive | <perpetrator> attempted to kill |
| <subject> auxiliary noun | <victim> was victim |
| active-verb <direct-object> | bombed <target> |
| infinitive <direct-object> | to kill <victim> |
| verb infinitive <direct-object> | threatened to attack <target> |
| gerund <direct-object> | killing <victim> |
| noun auxiliary <direct-object> | fatality was <victim> |
| noun preposition <noun-phrase> | bomb against <target> |
| active-verb preposition <noun-phrase> | killed with <instrument> |
| passive-verb preposition <noun-phrase> | was aimed at <target> |

As information extracted using the above grammar rules has a pre-defined structure (case frames) already, it can be used directly. For instance, Riloff reported that dictionary of
case frames constructed by AutoSlog achieved $98 \%$ of the performance of the handcrafted dictionaries used in MUC-4 evaluation. Despite the benefits offered by AutoSlog, Riloff found that generating an annotated training corpus is a tedious task, in terms of time and difficulty. For example, if the noun phrase to be annotated is part of a conjunction, should the user annotate all conjunctions, or just one? To avoid such problems, AutoSlog-TS was developed with the addition of incorporating statistical feedback. In this case, only a pre-classified text corpus (no annotation is required) comprising two sets of texts are required: one set that is relevant to the domain; and one set that is not. Again, a parser is used for the identification of all the noun phrases in the texts. Grammar rules from Table 2.3 are applied for each noun phrase to generate extraction patterns. Table 2.4 shows performance of AutoSlog and AutoSlog-TS, it can be seen that AutoSlog-TS has a better overall precision and F-measure values, at the expense of recall.

Table 2.4: Comparative results of the AutoSlogs.

|  | AutoSlog |  |  | AutoSlog-TS |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
| Slot | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| Perpetrator | $27 \%$ | $62 \%$ | $38 \%$ | $30 \%$ | $53 \%$ | $38 \%$ |
| Victim | $33 \%$ | $63 \%$ | $43 \%$ | $39 \%$ | $62 \%$ | $48 \%$ |
| Target | $33 \%$ | $67 \%$ | $44 \%$ | $39 \%$ | $58 \%$ | $47 \%$ |
| Total | $31 \%$ | $64 \%$ | $42 \%$ | $36 \%$ | $58 \%$ | $44 \%$ |

Fourth, the LaSIE-II system was developed by Professor Wilks' group at Sheffield. Wilks approach to natural language processing (NLP) is idiosyncratic in that he has focused on semantic preference which is closely related to the notion of collocation and colligation developed by John Sinclair (1991) and the others since then. However, Wilks' work is based deeply on the computational linguistics tradition following existing grammar and grammar rules. The LaSIE-II and GATE has been developed in parallel; GATE provides the general purpose lexical, syntactic, and semantic processors and resources. The LaSIE-II system which participated in MUC-7 has the following architecture (Humphreys et al., 1998:2):


Figure 2.2. LaSIE-II system architecture.

The LaSIE-II system assumes the existence of a number of templates: Template Elements (TE) for representing attributes or values of key entities inserted in a given event - people, places, organisations, objects and concepts; Template Relations (TR) for encoding n-ary relations between the entities; Scenario Template (ST) for detecting and representing relations between entities in a particular type of event. As shown in Figure 2.2, there is an emphasis on parsing, but then there is something called Discourse Interpretation based on a particular discourse model. The discourse model relies on the existence of predicate argument representation of objects and events within the system. The buchart parsing is done on two passes: one is made on the named entities, and the second pass is made on a whole range of tokens which may belong to different grammatical categories except for named entities. The buchart parser generates a predicate argument representation which is regarded as a quasi-logical form (much in the tradition of prolog representation of grammars). The logical form is then to link onto a predefined semantic network of concepts and hierarchy in the system, which is then used to generate results fulfilling the various templates. There are a number of advantages of LaSIE, and the system (QALaSIE for question answering) still remains popular between competitors as the LaSIE team still participating in the TREC 2004 and 2005 (see Gaizauskas, Greenwood, Hepple, Roberts and Saggion, 2004; Gaizauskas, Greenwood, Hepple, Harkema, Saggion, and Sanka, 2005 for details).

On a theoretical level, again the whole notion of generating predicate argument representation and then linking up to a semantic network requires a great deal of manual
handcrafting, certainly of the linkages. There are ways of finding these conceptual hierarchies which we will show later on when we come to discuss our method which is based on frequency count of collocation patterns. In the MUC-7, the LaSIE-II performed pretty well: $6^{\text {th }}$ on the named entity task ( 12 in total); $1^{\text {st }}$ on the coreference task ( 5 in total); and consistently the $3^{\text {rd }}$ on template elements ( 9 in total), template relations ( 5 in total) and scenario templates (6 in total). Scores for each of the tasks can be found in Table 2.5 below (Humphreys et al., 1998:12):

Table 2.5: Performance of the LaSIE-II system in MUC-7.

| Task | System | Precision | Recall | F-Measure |
| :--- | :--- | ---: | ---: | ---: |
| Named Entities | A | $89 \%$ | $83 \%$ | $85.83 \%$ |
|  | B | $94 \%$ | $87 \%$ | $90.41 \%$ |
| Coreference | B | $69 \%$ | $56 \%$ | $61.80 \%$ |
| Template Elements | B | $80 \%$ | $75 \%$ | $77.17 \%$ |
| Template Relations | B | $82 \%$ | $41 \%$ | $54.70 \%$ |
| Scenario Template | B | $42 \%$ | $47 \%$ | $44.04 \%$ |

A: Official performance - initial system
B: Unofficial performance - after modifications being made

Fifth, the corpus based information extraction system from the ECRAN project is closely related to the work at Sheffield. Indeed, the ECRAN project also uses the GATE system developed at Sheffield. The structure is very similar to LaSIE, except for the fact that the system deals with different languages - English, French and Italian.


Figure 2.3. System architecture of the Information Extraction System developed by the ECRAN project.
Figure 2.3 (Poibeau, 2000:255) shows the same kind of grammatical resources being used, which includes a morphology analyzer, a part of speech tagger, supported by a dictionary which includes a gazetteer and named entities. And this is used to fill
templates and generate results. The system was presented at the $16^{\text {th }}$ International Joint Conferences on Artificial Intelligence (IJCAI-99). Table 2.6 (Poibeau, 2000:258) below gives the performance of the system when dealing with terrorism events from the French newspaper Le Monde:

Table 2.6: Performance of the corpus based information extraction system from the ECRAN project.

|  | Precision | Recall | F-Measure |
| :--- | ---: | ---: | ---: |
| Structure of the text (1) | 1 | $89 \%$ | $94 \%$ |
| Linguistic Analysis (2) | 1 | $63 \%$ | $78 \%$ |

Column (1) reflects the scores concerning the quality of the information acquisition process from the structure of the text; column (2) the quality of the information acquisition process by linguistic analysis.

Very high levels of recall and precision are cited for this. For instance, the system achieved $89 \%$ accuracy when extracting the exact value for weapons and the number of injured/death victims during a terrorist event.

The InfoXtract engine is there to extract facts from a repository of texts, so called the Document Pool. The architecture of the InfoXtract is shown in Figure 2.4 (Srihari et al., 2003:52). Here we can see that in terms of orthography, a Case Restoration module is required to handle case insensitive input. In terms of dictionary on lexicon approach, we see there are two dictionaries, one general purpose lexicon, followed by a thesaurus for named entities detection. Then there a number of syntactic tools which help to carry out shallow and deep parsing, and a syntactic and semantic system which is claimed to detect relationships between objects and events.


Figure 2.4. System architecture of the InfoXtract system.

There are some innovative features here which are called Time and Location Normalisation that normalise general events. In addition, there is a profile and event linking and merging feature. These processes appear to generate, particularly, name and correlated entities which are critical for information extraction together with information related to the grammatical categories of these entities. Furthermore, it appears that the system can differentiate between a general event and predefined events. A general event could be a celebration such as weddings, birthdays or marriages; predefined events may include the announcement of a stock index, and so on.

This is a knowledge rich approach to information extraction in that it relies on an extensive repository of grammar rules including mark-up lexica for common and proper nouns. Using the MUC-7 datasets, the performance of InfoXtract system was measured; and Table 2.7 (Srihari et al., 2003:54) below provides the results:

Table 2.7: Performance of the InfoXtract system.

|  | Precision | Recall | F-Measure |
| :--- | ---: | ---: | ---: |
| SVO | $90 \%$ | $82 \%$ | $85.41 \%$ |
| CE | $96 \%$ | $83 \%$ | $88.90 \%$ |

The high performance ( $90 \%$ accuracy) part-of-speech tagging and named entity tagging forms the basis for capturing relationships between entities in the Correlated Entity (CE) modules. For example, relationships such as affiliation between a person and his employer, position of a person in an organisation, and so on, are handled correctly $96 \%$ of the time.

Finally, we will discuss the Phrasal Verbs (PV) identification system by Li et al. This system again is similar to the others in that there is a general lexicon to provide lexical resources, a syntactic analyser to provide part of speech tagging and there is a reliance on extracting named entities followed by shallow parsing, as illustrated in Figure 2.5 (Li et al., 2003:516) below:


Figure 2.5. System architecture of the PV Identification Module.
The innovative aspect of Li et al's work is the PV expert lexicon. The PV Expert Lexicon is closest to our work in that it relies on local grammar formalism for extracting
multi-word expressions which might include idioms, collocation and compounds, together with other phrasal verbs and complex semantic entities. Their formalism is similar to the INTEX tool developed by Silberztein (2000) and IDAREX by Breidt, Segond and Valetto (1994).

The intention here is to identify these multi-word expressions, and to find a set of patterns of grammar which are obeyed by these multi-word expressions such that none of the multi-word expressions thus identified would fall out of these patterns specified by the local grammar. This is a data driven approach where one analyses the data, extracts the rules and then predicates the existence of the local grammar. The results of their analysis for three representative PVs - look for, turn...on and blow...up (including their inflected forms) in the TREC corpus, show a precision and recall well over the $90 \%$ margin, as can be seen from Table 2.8 (Li et al., 2003:519) below:

Table 2.8: Performance of the PV identification system.

|  | Precision | Recall | F-Measure |
| :--- | ---: | ---: | ---: |
| 'look for' | $99.6 \%$ | $93.7 \%$ | $96.6 \%$ |
| 'turn...on' | $93.4 \%$ | $100.0 \%$ | $97.5 \%$ |
| 'blow...up' | $100.0 \%$ | $95.2 \%$ | $97.5 \%$ |

Among these seven systems, there is considerable reliance on traditional information extraction system on syntactic parsing, which is balanced by a lexical approach to language processing whereby a large dictionary of commonly used words, together with a gazetteer and named entities descriptions, or manually crafted grammar rules, are used to extract information from text. There is also a move to generate a knowledge-base derived from the data by asserting the results of lexical syntactic analysis in terms of representation schema whether it is a quasi-logical form as in LaSIE-II or a local grammar formalism for phrasal word expression as in Li et al. We call all these approaches knowledge rich approach as they require extensive knowledge of the rules of the grammar of a given language, and they also need an extensive and well prepared lexicon. We will try to describe how our work is at variance with modern themes in information extraction in that it relies totally on our data. One can see why our approach is more useful in that one finds very few instances in information extraction, or fewer instances in information extraction where the language being processed is
orthographically different - English and Chinese, for instance. There are newer developments in processing of Arabic and Chinese languages. However, the constraint here again is the availability of large parsers and lexical databases which may or may not exist for other languages. In addition, the use of gazetteers may introduce other problems. Take the example of the named entity task in MUC-7, the LaSIE-II system failed to identify Arlington as a city because Arlington was in the organisation gazetteer. We can see that even if named entities are present in the gazetteer, it is still possible to misclassify them.

Furthermore, English simplifies questions related to orthography in writing systems because most of the segmentation, or what it was referred in the four systems we discussed about tokenisation, mainly follows the assumption that white spaces exist between words. When we move on to orthographically complex languages like Chinese and Arabic, then we have to give serious thought to segmentation because white spaces are no longer our guide (see the discussion for segmenting Chinese texts earlier at the beginning of this section). Thus, to finish this section: with a knowledge rich approach there is plentiful availability of grammatical resources, lexical resources and knowledge bases, and even lexical semantic resources like gazetteers, and it is possible to obtain fairly good results on carefully selected training and testing samples. However, to deal with arbitrary texts in arbitrary languages, one has to start at least with a bottom up approach as we replicated in our research for this thesis. Our approach, briefly, sets about collecting a large corpora of texts used in a specialist domain, and then analyses that corpora of texts to create a lexicon, and through collocation and colligation constructs a local grammar to find patterns of usage of critical lexical items which are specific to the domain.

### 2.2 Sublanguages and local grammars

Grishman (2002) has argued that information extraction researchers can benefit from work in the "studies of sublanguage and of sublanguage information structures". He stresses the need for better quality linguistic analysers in order to go beyond the frequency-based metrics for building information extraction systems.

The term sublanguage was introduced by Harris as long ago as 1968. Harris defines sublanguage as a subset of natural language that differs from other subsets of the same language syntactically and/or lexically:

Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of $i$.
(Harris, 1968:152)

Later, Hirschman and Sager refined Harris's definition:
[A sublanguage is] the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles on a technical specialty or science subfield), in which the authors of the documents share a common vocabulary and common habits of word usage.
(Hirschman and Sager, 1982:28)

Compared with the language as a whole, the circumscribed subject area is often referred as specialist language/domain.

Lehrbeger suggested that sublanguages have the following properties (as cited by Pearson, 1998:31):

1. limited subject matter;
2. lexical, semantic, and syntactic restrictions;
3. deviant rules of grammar;
4. high frequency of certain constructions;
5. text structure; and
6. use of special symbols.

Halliday and Martin (1993), and Hoey (1991) also pointed out that in specialist domains, the author uses a restricted set of vocabularies to keep the focus of the reader to specific objects and events. This restricted vocabulary set is repeated and tends to dominate the text as such. The frequency of domain specific terms has been referred to as weirdness in Ahmad (1995); Ahmad and Rogers (2001). Ahmad (1995) claims that the domain specific terms can be identified by contrasting the frequency of words used in the
specialist corpus with the frequency of the same tokens in a representative corpus of general language.

Zellig Harris and Maurice Gross have suggested that in certain types of texts (sublanguages), one may find local grammars in operation. Harris (1991) defines local grammar as a way to describe the syntactic behaviour of certain subsets of sentences of the language that is closed under some or all of the operation in the language. Harris illustrated this point by citing examples of recursive noun-phrases used in biochemical literature to either refer to complex biochemical compounds or complex biochemical processes.

Taking the example of adverbs as described in Gross (1997:332), one can see that many adverbs derived from adjectives are accepted in the left context of speaking and of no other forms. However, the same adverbs are not used in the context of saying, calling, talking, despite the fact that these verbs are morphologically and semantically similar to speaking. For instance, the adverb:
(broadly + generally + roughly) speaking
is considered as a set of (semi-)frozen sequences, and can be compactly represented by the following finite state automaton (FSA):


Figure 2.6. A local grammar representation of the adverb speaking.

A collection of local grammars can be combined and represented by a more complex finite automaton by taking the union of the simpler local grammar automata (Gross, 1993, 1997). Gross (1993:35) focussed on how we specify time and date and showed cardinal numbers used to denote time and calendrical expressions (day / month / year, century) embedded in their own local grammar. Figure 2.7 overleaf shows such a local grammar that can extract complex date expressions from natural language texts.


Figure 2.7: A local grammar for the identification of date expressions.

For example, some of the date expressions recognised by the above local grammar include:

- on January, on Monday
- in the year 1997
- on Monday the $1^{\text {st }}$ of January, on Monday the $1^{\text {st }}$ of January 2003
- on Tuesday January the $24^{\text {th }}$, on Tuesday January the $24^{\text {th }} 1995$

From Kartunen, Chanod, Grefenstette, and Schiller's (1996) point of view, finite state automaton used to encode phonological rewrite rules can be viewed as a special kind of regular expressions, but providing additional operators and new types of expressions. A regular expression (REGEX) is an expression that describes a set of strings in a concise manner, without having to list all elements. Four of the most common operations used to construct REGEX are concatenation, alternation, quantification and grouping. We will discuss each of them in turn.

First, concatenation simply means joining one or more strings/expressions together. For example, the expression "a" will identify the character, or the article $a$, and the expression " an " will identify the article an. Second, alternation is a vertical bar that separates alternatives. For example, the local grammar expression for speaking can be represented as "broadly |generally |roughly speaking". Third, quantification is the use of a quantifier that provides "counting". The most common quantifiers are + , * and ?. The plus sign means that there is at least one occurrence of the previous expression. For example, "bo+ks" will match boks, books, booooks, etc., but not bks. The asterisk indicates that there are 0 , or more than one occurrence of the previous expression. Compared with "bo+ks", "bo*ks" will match bks in addition to the patterns matched by "bo+ks". While for the question mark, it matches 0 or 1 occurrence of the previous expression. Unlike "bo+ks" and "bo*ks", "bo?ks" only matches boks or bks. Finally, grouping is used to limit the scope of alternation, or to group multiple characters into larger units such that quantifiers can be applied. Grouping is achieved by using parentheses around an expression. For instance, " $\mathrm{b}(\mathrm{o} \mid \mathrm{an})+\mathrm{ks} "$ will match boks, books, banks, bananks, boanks, booananks, banoks, etc. Note that " o " and "an" can appear simultaneously, in any order, and appear more than once, as illustrate by the tokens boanks, booananks, banoks. For more details of using REGEX, please refer to Friedl (2002).

Instead of using the FSA representation for local grammars, Karttunen et al. (1996:312) has illustrated how REGEX can be used to identify date expressions. They first started to define simple REGEX that can identify number, day of week, month, and so on, and subsequently concatenated them to form a more complex DateExpression REGEX:

```
1 To9 \(=[1|2| 3|4| 5|6| 7|8| 9]\)
0To9 \(=\) [\%0|1To9]
\(\mathrm{SP}=\) [","]
Day \(\quad=\) [Monday \(\mid\) Tuesday \(|\ldots . .\).\(| Saturday \mid\) Sunday \(]\)
Month \(=\) [January \(\mid\) February \(|\ldots . . .\).\(| November \mid\) December \(\mid\) Jan. \(\mid\) Feb. \(\mid \ldots . . .\). Nov. \(\mid\) Dec. \(]\)
Date \(=[1 \mathrm{To9}|[1 \mid 2] 0 \mathrm{To9}| 3[\% 0 \mid 1]]\)
Year \(=1 \mathrm{To} 0\) ( 0 To9 (0To9 (0To9)))
DateExpression = Day \(\mid\) (Day SP) Month " " Date (SP Year)
```

For example, item(s) within square brackets of the following sentence are the dates identified using Karttunen's DateExpression REGEX:

Today is [Tuesday, August 22, 2006] because yesterday was [Monday] and it was [August 21] so tomorrow must be [Wednesday, August 23].

From the above discussions, we can see that the use of local grammar, either represented by FSA, or by REGEX, provides an alternative for natural language processing and traditional information extraction. Hunston and Sinclair (2000) pointed out that even with the latest parsers, there will be segments of text (leftovers) that will never be adequately described - but with local grammar and finite state automata, such leftovers can be sufficiently handled. Gross (1993) extended the application of local grammars for articulating date, times and for financial reports. Barnbrook and Sinclair (1995) suggested that dictionary definitions are formatted according to local grammar patterns. Choi and Nam (1997) used much the same approach to identify Korean proper nouns, while Lee (1999) used local grammar for stemming Korean language. Ranchhod, Mota and Baptista (1999) have extended Gross' work into Portuguese. Mason (2004) has used local grammars 'discovering' verb patterns in English. Various studies have shown the benefits of using FSA for parsing as in Roche (1997), for the word disambiguation task (Roche, 1992; Silberztein, 2000; Carvalho, Mota and Ranchhod, 2002; Laporte and Monceaux, 2000), for Spanish-English and Spanish-German translations
(Casacuberta and Vidal, 2004), and for French- English machine translation (Kumar and Byme, 2003).

### 2.3 Sentiment Analysis

What is sentiment? The OED ${ }^{1}$ provides a variety of definitions for the term sentiment. Two of the definitions are:

- Sentiment is "an emotional thought expressed in literature or art."
- Sentiment is "what one feels with regard to something, mental attitude (of approval or disapproval, etc.,); an opinion or view as to what is right or agreeable."

In these two definitions, sentiment is used in identifying and extracting the feelings, or the opinions or views from statements made in the informative genre of texts like news reports - editorials, company reports, and so on. Note that the first definition, comprising the phrase "an emotional thought", relates to emotions, views or feelings expressed in imaginative texts like literary texts.

The use of sentiment analysis, or related terms like semantic orientation and polarity analysis, is now developing rapidly in two major areas: financial news analysis relating to financial instruments, and opinion analysis relating to goods and services (see Wiebe, Bruce, Bell, Martin and Wilson, 2001; Pang et al., 2002; Turney, 2002; Wilson, Wiebe and Hoffmann, 2005). Turney defines semantic orientation as "the evaluative character of a word. Positive semantic orientation indicates praise ("honest", "intrepid") and negative semantic orientation indicates criticism ("disturbing", "superfluous"). Semantic orientation varies in both direction (positive or negative) and degree (mild to strong)" ${ }^{2}$.

There are three major analysis systems used in financial trading: fundamental, technical and sentiment analysis, as defined in Table 2.9 below (taken from Wikipedia):

[^3]Table 2.9: Definitions of the three major analysis systems used in financial trading.

| Fundamental ${ }^{3}$ | Technical ${ }^{4}$ | Sentiment $^{5}$ |
| :--- | :--- | :--- |
| The analysis of financial | Study of the trading <br> statements, marketing and <br> production documents to <br> determine a security's | The linguistic analysis of <br> vistory (the price and <br> text to determine the mood <br> intrinsic value by studying its |
| type of traded security <br> economic well-being as <br> (stocks, commodities, etc.) <br> opposed to its price <br> to attempt to predict future | or emotional intent of the <br> text. |  |
| movements only. |  |  |

Fundamental Analysis deals with both numbers and texts; Technical Analysis deals with numbers solely, and Sentiment Analysis deals with Texts only, as illustrated in Table 2.10 below. All three analyses reflect on the same reality from three different perspectives. However, in this thesis, our focus is on spoken and written words expressed through informative texts, in other words, the financial news texts.

Table 2.10. Objects being analysed in the three major analysis systems used in financial trading.

|  | Analysis | Object | Medium |
| :---: | :---: | :--- | :--- |
| Factual | Fundamental | financial statements <br> marketing documents <br> production document | Text and numbers <br> Text <br> Text |
| Factual | Technical | price and volume movements <br> prediction system | Numbers <br> Numbers |
|  | Sentiment | reports about company <br> e-mails | Texts <br> Texts |
|  |  |  |  |

Financial markets are places where financial instruments are bought and sold. These instruments include shares, currencies and bonds. Some of these instruments are traded in millions, others in thousands and yet others in hundreds: the prices of instruments change frequently during a single trading session or over a longer trading horizon. The behaviour of financial markets is governed by the gains and losses of the investors in the markets.

[^4]In addition to the quantitative data related to trading volumes and price movements, financial traders rely on market sentiment. This sentiment is often expressed in news reports and editorials, and ranges from views about national economies to imminent takeovers, mergers and acquisitions, and from people leaving or joining an organisation to news about political and economic successes and failures. For instance, "the mood of a given investor or the overall investing public, can be either bullish or bearish." The metaphors, bullish and bearish, and more colourful slang, including the phrase dead cat bounce, to describe an upward movement of stock is much like a lifeless object moving merely because of the laws of gravity, shows a creative use of language in the specialist field of financial trading (Ahmad, 2002).

Financial reports, especially summaries of stock market behaviour on an hourly or daily basis, are reported in terms of stocks that rose most and stocks that fell most. The market movement is described in terms of metaphors related to market trends and cycles: uptrend, downtrend, boom-and-bust cycles, peaks and troughs (of cycles). The expression of optimism or pessimism relies on a choice of words whose meaning is generally understood. This is not to say that the words used in the expression of market sentiment have been standardised in much the same way that the terminology of science and technology is standardised. Rather, there is a general consensus on how to express optimism or pessimism about an instrument. Knowles (as cited by Salway and Ahmad, 1997:6) has identified a list of "health metaphors" in a 6 million word corpus of texts from the Financial Times.

Table 2.11: Health metaphors identified by Knowles (1996).

| abort | chronic | exposure | hurt | nurse | revive | surgery |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| addiction | clone | famine | immune | overweight | robust | symptom |
| ailing | collapse | fat | incubation | pain | rupture | syndrome |
| alive | complexion | fatal | indigestion | palatable | sanity | teething |
| anaemic | contagion | fatigue | infection | palliative | scar | temperature |
| anaethestic | convalesce | fever | inject | panic | shock | thin |
| anatomy | cripple | fit | injured | paralysis | sick | tired |
| appetite | cure | geriatric | life | patient | sleepy | transplant |
| atrophy | dehabilitating | haemmorhage | life-blood | pulse | slim | trauma |
| backbone | decline | hamstring | limp | rally | stagger | tumble |
| bill of health | depressed | handicap | medicine | recipe | starve | umbilical |
| bleed | depression | hangover | mid-life | recovery | sterilisation | vibrant |
| blood | diet | headache | miscarry | recuperate | stomach | viral |
| breath | disease | health | muscle | rehabilitation | strength | weak |
| bruise | endemic | healthy | myopic | relapse | stricken | weaken |
| casualty | epidemic | hunger | nerve | resuscitate | suffer | wound |
| choke | exhaust | hungry | nourish | revitalise | support system |  |

Selection of the health metaphors is subjective and depends on personal intuition, and also requires extensive knowledge of the specialist domain. It is possible to use the literature on buying and selling in semantic theory (see Jackendoff, 1991) as a framework for analysing the meaning of the news stories. The literature on natural language processing (Simmons, 1984) and on knowledge representation suggests that frame semantics have been used to build systems that can, in principle, analyse, extract and disseminate the meaning of a specialist news report. Frame semantics has a number of limitations, and a prominent one is the need for a lexicon that is rich and extensive in terms of meaningful data. Figure 2.8 below shows a typical news story published by Reuters in NewsML format in August 2006.
<?xml version=" 1.0 " encoding="iso-8859-1" ?>

- <newsitem itemid="185080" id="root" date="31 AUG 2006 10:16" xml:lang="en">
<headline>UPDATE 3-Rakuten threatens to sue tabloid, shares fall</headline>
- <text>
$<\mathrm{p}>$ (Updates with closing stock price, Shukan Shincho comment)</p>
$<p>$ By Jonathan Soble</p>
<p>TOKYO, Aug 31 (Reuters) - Japanese Internet firm Rakuten Inc. <4755.Q> threatened on Thursday to sue a tabloid magazine over a report suggesting that company executives had been questioned on suspicion of insider trading. $</ \mathrm{p}>$
<p>Shares in Rakuten, which operates Japan's biggest online shopping mall and other Internet businesses, fell 5 percent in an otherwise buoyant market. That brought the stock's decline to 13 percent since Wednesday, when rumours of the story's impending publication began to circulate in Tokyo trading rooms.</p>
$<\mathrm{p}>$ The sharp decline underscored investors' nervousness about potential investigations into Japanese Internet firms. A probe in January into alleged securities law violations by the founder of Rakuten rival Livedoor Co. helped snuff out a bull run in Japanese shares, in what came to be known as the Livedoor Shock.</p>
$<p>A$ Rakuten spokesman denied that the executives had been questioned. Tokyo police and prosecutors declined to comment on
the article, citing standard policy. $</ \mathrm{p}>$
<p>Japan's weekly tabloids have sometimes been ahead of more conservative newspapers in breaking major stories, but often
rely on anonymous sources with indirect knowledge of events and have a reputation for stretching the facts.</p>
< $\mathrm{p}>$ In the short term, however, "the damage from the rumour could continue to hurt the company's stock," he added.</p>
< p >Livedoor has seen its shares de-listed and its business shrink dramatically since Horie's arrest, a fate that has made some investors wary of other Internet start-ups.</p>
<p>Even before talk of the Shukan Shincho article began to circulate, Rakuten's shares had already fallen by half since prosecutors launched a probe of Livedoor on Jan. 16.</p>
<p>Rakuten's shares ended down 5.2 percent at 50,900 yen on Thursday after falling 8.5 percent the day before. Trading volume soared to 809,000 shares, six times the average for the last three months and nearly double the previous single-day record.</p>
<p>Mikitani, a former banker, has been Rakuten's chief executive and chairman since founding the firm in 1997 and has expanded the business steadily through acquisitions and alliances.</p>
$<p>$ Besides an Internet portal that directs users to some 54,000 online retailers, Rakuten operates a securities brokerage, reservation services for travel and entertainment, and a baseball team, the Rakuten Eagles.</p> <p>After the market closed, Rakuten announced that it would post an 18.6 billion yen ( $\$ 159$ million) special loss from the sale of its car financing and credit card unit to Orient Corp. $<8585$. T $>.</$ p>
$<$ text>
<topics>TEL PUB ASIA JP NEWS WWW RET CRIM FIN MRG LEN RTRS</topics> <newsitem>

Figure 2.8. A typical example of a news story from Reuters Newsfeed, published on 31 August 2006.

Some of the health metaphors identified by Knowles 10 years ago are still being used, these include: decline as in decline to 13 percent, sharp decline; hurt as in hurt the company's stock, and shock as in Livedoor Shock. However, other terms such as rise, up, fall (and its variants fell, fallen, falling), down, soared, loss, are also being used to express change in values of financial instruments; some of these usages can be found in the new story shown in Figure 2.8.

The words rise and fall have many senses and do cross grammatical categories; each could be a noun or a verb. However, financial report writers constrain the meaning by encoding the words within specific patterns - a local grammar. The verb rose has many
senses, for example, General Rose, colour of a rose, shares rose, but financial reporting has appropriated the word by co-locating its use with a cardinal number. Therefore, one can remove ambiguous patterns and focus on patterns such as bond rose 3.1 percent, fell 5 percent, falling 8.5 percent, down 5.2 percent which can be easily represented by a local grammar. The question is how can we identify such metaphorical words used in the financial news stories?

Some researchers approach this challenge by a priori lexicon of positive and negative words, or a priori set of grammatical patterns (as in Knowles, 1996; Surdeanu, Harabagiu, Williams and Aarseth, 2003; Chan and Lam, 2005; Tetlock, Mytal and Sofus, 2005) that indicate change in the value of a financial instrument - including metaphorical terms like above, below, $u p$ and down - and use them to 'represent' positive or negative news stories. Others use the frequency of collocational patterns for assigning a 'feel good or feel bad' score to the story (see, for example, DeGennaro and Shrieves, 1997; Wiebe et al., 2001; Koppel and Shtrimberg, 2004), or by employing different classifier algorithms coupled together by a voting scheme (see Das and Chen, 2006). Table 2.12 below shows such sentiment proxies - frequent metaphorical or literal keywords that can be used as placeholders for investor or trader sentiment; this seems to mirror Turney's work.

Table 2.12: Examples of the lexical content of a news story and the implied sentiment.

| Sentiment | Lexical Content |
| :--- | :--- |
| 'Good' news stories | appear to comprise collocates like revenues rose, share rose; |
| 'Bad' news stories | may contain profit warning, poor expectation; |
| 'Neutral' stories | usually contain collocates such as announces product, <br> alliance made; |

### 2.4 Sublanguages and Idiom Principle

Corpus analysis techniques have been developed to study languages at various levels of linguistic description, vocabulary, grammar, semantics and pragmatics, by relying almost exclusively on texts and speech produced by language users. Frequency analysis of linguistics tokens, includes, for example, words at the level of vocabulary and phrases and sentences at the level of grammar. Relations between the vocabulary and
grammatical levels for inferring meaning like collocation, are analysed. The results of the frequency counts are used to generate statistical metrics and quantitative results are produced, and linguistic hypotheses accepted and rejected using the metrics. Corpus analysis techniques work well with special languages.

A related notion to that of local grammar is that of collocation. Linguist John Firth's remarkable phrases, "You shall know a word by the company it keeps" (Firth 1957b), gives a clear idea that meaning of a word can be determined by looking at its distribution, he elaborates that:

Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night.
(Firth, 1957a:196)

Benson (1990) defined collocation as an arbitrary recurrent word combination. The Wikipedia defines collocation as "a sequence of words or terms which co-occur more often than would be expected by chance ${ }^{6 "}$. Clear gives a similar definition, "[...] a recurrent co-occurrence of words" (Clear, 1993:277). Sinclair not only defines collocation, but also describes the structure of it:

We may use the term node to refer to an item whose collocations we are studying, and we may define a span as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call collocates.
(Sinclair 1966:415)

For instance, if we want to study the collocations of the word shares, shares will be the 'node'. If we define the 'span' as four, then four lexical items on the left neighbours of shares and four lexical items on the right neighbours of shares will be considered to be the potential 'collocates' of shares. The identification of collocational patterns is carried

[^5]out by observing and counting the concordance output of the "node" words. Table 2.13 below shows some of the concordance output for fell:

Table 2.13: Sample of concordance output for fell.

| Left span |  |  |  |  | Right span |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| -4 | $\mathbf{- 3}$ | $\mathbf{- 2}$ | $\mathbf{- 1}$ | node | $\mathbf{+ 1}$ | $\mathbf{+ 2}$ | $\mathbf{+ 3}$ | $\mathbf{+ 4}$ |
| raw | material | and | fuel | fell | 0.1 | per | cent | in |
|  | second | section | price | fell | 3 | points | on | three |
| sales | of | new | homes | fell | 6.7 | percent | last | month |
| $[/ b]$ | [p] | Abbey | National | fell | 7 p | to | 592 p | as |
| price | of | new | properties | fell | by | 1 | $1 / 2$ | per |

An initial observation from the Table 2.13 is that the most significant collocates of fell are numbers.

Despite the discrepancy between the definitions of collocation, all of them focus on the co-occurrence of words, and such co-occurrence can be described by the idiom principle, as opposed to the open-choice principle governed by (universal) grammaticalness (Sinclair, 1991:109-115). The open-choice principle sees "language text as the result of a very large number of complex choices. At each point where a unit is completed [...], a large range of choice opens up and the only restraint is grammaticalness" (Sinclair, 1991:109). Sinclair rejects the open-choice principle as it does not account for the fact that linguistic choice is not random in the language:

It is clear that words do not occur at random in a text, and that the open-choice principle does not provide for substantial enough restraints on consecutive choices. We could not produce normal text simply by operating the open-choice principle.
(Sinclair 1991:110)

On the other hand, for the idiom principle,
a language user has available to him or her a large number of semi preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.
(Sinclair 1991:110)
these semi-preconstructed phrases, or collocation units are available for the language user to choose. After a choice has been made, "then all the slot-by-slot choices are massively reduced in scope or even, in some cases, pre-empted" (Sinclair, 1991:110).

In collocation, there are frequently used words and less frequent words that are simultaneously used. This collocation between high frequency words (node) and many lower frequency words is called "downward collocation" (ibid: 115). Similarly, the collocation between low frequency words (node) and many higher frequency words is called "upward collocation" (ibid: 116). For example, the frequently used "washed" may occur with many proteins and acids (Harris, 1991). Many financial instruments, for example, currencies and indices co-occur with the frequently used verbs fell and rose. In addition said can be used with a number of people and organisation names. While for Sinclair, he notes that the word back collocates with at, down, from, into, up, her, him, and so on, all of which occur more frequent than back itself. However, the grammatical nature of the collocates that the upward collocation and downward collocate attract is different:

There appears to be a systematic difference between upward and downward collocation. Upward collocation, of course, is the weaker pattern in statistical terms, and the words tend to be elements of grammatical frames, or superordinates. Downward collocation by contrast gives us a semantic analysis of a word.
(Sinclair 1991:116)

By looking at the grammatical categories of the collocates of back, Sinclair (1991:117118) notes that the upward collocates are mostly prepositions and pronouns, whilst downward collocates consist of a large number of verbs and nouns.

Generally, collocation strength is computed by the so-called "collocation strength" of a word pair. The pair may be neighbours, or may co-occur with other interspersing words: up to five left neighbours and five right neighbours of a high frequency word appear to be significant. Smadja (1994) defines the strength $k_{i}$ of the collocates $w_{i}$ of the high frequency word $w$ (node word, as referred by Sinclair) as:

$$
\begin{equation*}
k_{i}=\frac{f r e q_{i}-\bar{f}}{\sigma} \tag{4}
\end{equation*}
$$

where freq $_{i}$ is the frequency of the collocation of $w_{i}$ with the node $w ; \bar{f}$ is the average frequency and $\sigma$ is the standard deviation.

The collocation between non-neighbouring collocates may be just as useful between neighbouring collocates. According to Smadja (1994), the strength of their collocations can be statistically tested under certain statistical assumptions.

We have presented the key elements of collocation and its related terms: the notion of node and collocational span; the notion of upward and downward collocation; the strength of collocation; and collocation as an embodiment of the idiom principle (Sinclair 1991). In the next chapter, we will describe a method that can be used to extract information from arbitrary domains through collocation analysis.

### 2.5 Conclusion

In this chapter, we have reviewed some of the recent information extraction systems. Among these systems, there is considerable reliance on syntactic parsing where a large dictionary of commonly used words, together with a gazetteer and named entities descriptions, are used to extract information from text. There is also a move to generate a knowledge-base derived from the data by asserting the results of lexical syntactic analysis in terms of representation schema, whether it is a quasi-logical form as in LaSIE-II or a local grammar formalism for phrasal word expression as in Li et al. (2003).

Information extraction systems that based on a knowledge rich approach require extensive knowledge of the knowledge of the rules of the grammar of a given language, and also require an extensive and well prepared lexicon. Despite the fact that some of the information systems claim to be able to cope with multi-lingual texts, they work on the Romance languages only. Moving to orthographically different writing systems such as Chinese and Arabic, those information extraction systems will probably fail to work, or significant modification to the systems will be unavoidable.

We have also discussed notions of local grammar, and how it can be represented using FSA, or REGEX. Local grammar provides an alternative approach towards information extraction and natural language processing. The application of local grammar in tasks such as machine translation; word disambiguation; parsing; identifying and extracting patterns; and extracting proper nouns from natural language texts, have shown promising results.

Sentiment analysis and opinion analysis related to goods and services have also been discussed. Traditional approaches towards sentiment analysis often depend on a priori lexicon of positive and negative words, or a priori of a set of grammatical patterns. Building such lexical resources is either based on intuitions, or requires manually analysing a modest-size corpus and identifying relevant expressions. While for corpusbased approaches, annotated training data, or the use of a POS tagger, is required. Annotating individual examples may require less skill than linguistic analysis but the burden of annotating hundreds or even thousands of documents in order to collect a range of linguistic patterns still exists.

In summary, with a knowledge rich approach there is plentiful availability of grammatical resources, lexical resources and knowledge bases, and even lexical semantic resources like gazetteers, and it is possible to obtain a fairly good result on carefully selected training and testing samples. However, to deal with arbitrary texts in arbitrary languages, one has to start at least with a bottom up approach like that we replicated in our research for this thesis. Four of the properties of sublanguages, namely limited subject matter (1); lexical, semantic, and syntactic restrictions (2); deviant rules of grammar (3); and high frequency of certain constructions (4); as suggested by Lehrbeger, forms the basis of our algorithm. Property (1) ensures that certain terms will be domain specific. Properties (2), (3) and (4) suggest that the use of certain terms will be restricted and follow certain constructs, and such restrictions establish patterns of usage. Some of the highly frequent patterns can be identified through collocation analysis. Our approach, briefly, involves collecting a large corpora of texts used in a specialist domain, and then analysing that corpora of texts to create a lexicon, and through collocation and colligation constructing a
local grammar to find patterns of usage of critical lexical items which are specific to the domain.

Sinclair emphasis the importance of the idiom principle, and points out that "the principle of idiom is far more pervasive and elusive than we have allowed so far. It has been noted by many writers on language, but its importance has been largely neglected" (Sinclair, 1991:111). For us, our approach works closely with the idiom principle. In the next chapter, we will try to describe how our work is at variance with modern themes in information extraction in that it relies totally on our data - through collcational analysis to identify patterns of usage.

## Chapter 3

## 3 Method

### 3.1 Introduction

There is an inextricable link between economics and language (Rubinstein, 2000); this is, perhaps, true of all social sciences if we believe in linguistic philosophy. Applied linguists have examined the language of economics at various levels of description ranging from the lexical to the pragmatic. Paradigm shifts in economic theory have been studied in considerable detail (for example, by Henderson, Dudley-Evans and Backhouse, 2000). The notion of bounded rationality in economic and financial transactions has led some authors (like Markham, 2004) to argue that a gap 'persists between accounts of behaviour framed by rational theory [in economics] and experimental evidence of how people actually behave in bargaining situations'. This is a challenge 'to the epistemological foundations of mainstream narrative' (Mehta, 1993:85).

The automatic analysis of written texts has been used to assess a reviewer's attitude to a range of artefacts, including films and cars, banking institutions, and holiday destinations (see Turney, 2002). These analyses are based on methods of genre analysis and authorship attribution studies, and seek to unravel the 'semantic orientation' of reviews. Such techniques have been used in identifying abusive postings (or flames) in Internet newsgroups (Spertus, 1997).

Since the late 1970 's, investor and consumer behaviour and attitudes have been studied using questionnaire surveys for creating a confidence index. Yale's International Centre of Finance (now) publishes investor confidence indices covering, amongst others, attitudes to house prices, the possibility of market crash, and the 'confidence' in (the US) economy. Yale's results show a systematic and clear difference between attitudes of the individual investor and the institutional investor as measured by almost all the 15 or so
indices (Shiller, 1993, 2003 and Yale, 2004). The use of 'focus groups' and questionnaire-based techniques for assessing semantic orientation of human beings, and then publishing the result of such an analysis on the Internet, leads to the (de)construction of a social reality in cyberspace. The Internet helps to construct a 'geographically-dispersed' and 'temporally-malleable' reality with its concomitant scope and limitations as a qualitative research method in social sciences (see for instance, Markham, 2004).

Researchers in financial economics and empirical economics have explored the 'relationship between public information releases and private information arrival' on market volatility. DeGennaro and Shrieves (1997) found a positive correlation between market volatility and frequency of key words and proper nouns in Reuters Newswire headlines. Others have used a similar approach to identify movements in the market (Baestaens and Van den Bergh, 1995). In financial investing, sentiment analysis is based on the controversial notion that market information may not be wholly contained in the prices of goods in the market. Sentiments of buyers and sellers, based on their expectations, greed, fear, or hope, are equally important. A balanced approach to the study of change in the market is required, an approach that cherishes rational analysis, is facilitated by using statistical techniques and mathematical methods, and informed by an understanding that rational analysis is bounded since human behaviour can be irrational.

Sentiment analysis has started at Surrey since 1997. In two EU-IST projects (Analyst Control Environment 1997-1999 and Generic Information based Decision Assistant 2001-2003) the scope of sentiment analysis was extended to include the analysis of full news texts (Ahmad, Vrusias and Ledford, 2001; Gillam et al., 2002; Ahmad et al., 2003). Supported by ESPRIT and the EPSRC, Surrey pioneered (1989-96) a corpus-based approach to terminology extraction (Ahmad and Rogers, 2001) and ontology learning (Gillam et al., 2005). These approaches have been developed and tested in the software of the System Quirk framework for language analysis, developed in close collaboration with applied linguists and translators (Kugler, Ahmad and Thurmair, 1995).

### 3.2 Algorithm

Methods based on corpus linguistics and information extraction, and pioneered in System Quirk, have been used to identify sentiments as unambiguously as is possible from natural language texts. Methods have been developed to: (a) automatically identify the vocabulary used in sentiment-bearing phrases: key terms, movement verbs and orthographic and numerical cues in financial reports; and (b) automatically identify the local grammar used for expressing sentiment. The identification is made by corpus comparison methods. Once identified, the local grammar rules are used to extract and classify sentiment-bearing sentences from other specialist texts. These methods are used to extract patterns of language used in financial news that indicate changes in events or values of objects. We discovered a local grammar that governs the ordering of the keywords, verbs and the markers in sentiment-bearing sentences in the financial news. Such grammars have been found earlier where specific information is communicated: in scientific communications (Harris, 1991); in phrases for telling date and time (Gross, 1993); in phrases comprising name and designation/address (Traboulsi, Cheng and Ahmad, 2004).

We present a systematic method of finding such collocation. Starting from a corpus of specialist text, domain specific terms are automatically selected and their collocations are identified. By examining the collocations, a local grammar is then constructed. The local grammar is used either for building a linguistic description of the language governing specialist texts or for purposes of recognising patterns of local grammar usage in unseen texts. Such recognition is useful for information extraction tasks.

Collocation illustrates the idiom principle for Sinclair. He suggested words appear to be chosen in pairs or groups and these are not necessarily adjacent. According to Ahmad (2002), Sinclair made a key point in regard to different registers of language in which text is written applies very clearly to specialist domain. From Sinclair's point of view, once a choice of register has been made, "then all the slot-by-slot choices are massively reduced in scope or even, in some cases, pre-empted" (Sinclair, 1991:110). We will show, the use of "slot-by-slot choices" implemented in our algorithm, is what Sinclair would indeed regard as a "spectacular" manner. What interests us in his work on idiom principle is where Sinclair argued that many uses of words and phrases show a tendency to occur in a
certain semantic environment. The example he uses is the verb to happen which he suggested is associated to unpleasant things like accidents.

What we envisage is computing his so called upward (and downward) collocation patterns, which is the collocation of a less (or more) frequent token $a$ with a more (or less) frequent token, say $b$. This is a part of our algorithm that will be discussed next.

The algorithm we developed can be divided into 5 stages:

1. Select corpora
2. Generate wordlist and statistical information for each token
3. Extract candidate terms
4. Extract key collocates for each candidate term, and
5. Extract local grammar from collocates.

We will explain the process and functions performed in each stage in details in the following sections.

### 3.2.1 Select corpora

In the first stage, two corpora are selected, $C O R P U S_{C}$ and $C O R P U S_{R C} . C O R P U S_{C}$ can be any number of texts that one wants to be analysed. In our case, we have used the Reuters Corpus; comprising news texts produced in 2002 and containing 3.63 million words distributed over 9,063 texts. The average length of the news reports is 400 tokens. The subjects typically covered by Reuters Financial News Service are: Company Outlooks, Company Results, Economic Indicators, Funds and Initial Public Offering News. The news reports are first written by a news reporter in the field and then sub-edited at base. The news reports are in that sense jointly authored. The diachronic distribution of texts is shown in Table 3.1.

Table 3.1: Diachronic distribution of text in the Reuters Corpus.

|  | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| No. of stories | 778 | 682 | 572 | 650 | 631 | 589 | 962 | 658 | 887 | 1017 | 980 | 657 |
| No. of tokens | 323472 | 298202 | 246812 | 269743 | 278724 | 255094 | 401430 | 252072 | 326032 | 379056 | 402089 | 260216 |

$\operatorname{CORPUS}_{R C}$ can be any corpus that one wants to contrast with the $\operatorname{CORPUS}_{C}$. We have chosen the British National Corpus, comprising 100 -million tokens distributed over 4124 texts (Aston and Burnard, 1998). Figure 3.1 below shows the corpora selected in this stage.

| INPUT: $\mathrm{CORPUS}_{\text {c }}$ | /* a corpus of specialist texts comprising $\mathrm{N}_{\mathrm{C}}$ indivi |
| :---: | :---: |
| INPUT: CORPUS $_{\text {RC }}$ | ference language corpus comprising $\mathrm{N}_{\mathrm{RC}}$ individu |

INPUT: CORPUS $\quad / *$ a reference language corpus comprising $\mathrm{N}_{\mathrm{RC}}$ individual words */
Figure 3.1. Stage 1: Selecting input corpora.

### 3.2.2 Generate wordlist and statistical information for each token

After selecting the corpora, second stage of the algorithm will generate a wordlist of $\operatorname{CORPUS}_{\mathcal{C}}$, together with statistical information such as frequency, weirdness, frequency $z$-score and weirdness $z$-score for each token. These four statistical measures will be used to identify candidate terms from the $\operatorname{CORPUS}_{\mathcal{C}}$, which will be discussed in stage three. Furthermore, average frequency, standard deviation, average weirdness and weirdness standard deviation of all tokens in $\operatorname{CORPUS}_{C}$ are also calculated. In addition, numerals such as the cardinal number million, billion, one, two, three, etc. are replaced with the token "NOMBER".

Weirdness is the ratio of the relative frequency of token $\boldsymbol{w}^{k}$ in the $\operatorname{CORPUS}_{C}$ to the relative frequency of token $w^{k}$ in the $C O R P U S_{R C}$. A high weirdness ratio indicates significant use of the term in $\operatorname{CORPUS}_{C}$ as compared with the $\operatorname{CORPUS}_{R C}$, and this may be a clue that token $w^{k}$ is a candidate term. Weirdness is defined as:

$$
\begin{equation*}
\text { Weirdness }=\frac{f\left(w_{c}^{k}\right)}{N_{C}} / \frac{f\left(w_{R C}^{k}\right)}{N_{R C}} \tag{5}
\end{equation*}
$$

where
$f\left(w_{c}^{k}\right)=$ frequency of word $w^{k}$ in the CORPUS ${ }_{C}$;
$N_{c}=$ total number of words in the CORPUS $_{C}$;
$f\left(w_{R C}^{k}\right)=$ frequency of word $w^{k}$ in the CORPUS ${ }_{R C}$; and
$N_{R C}=$ total number of words in CORPUS $_{\text {RC }}$.

It is possible that token $\boldsymbol{w}^{k}$ occurs in $\operatorname{CORPUS}_{C}$ but does not occur in $\operatorname{CORPUS}_{R C}$, resulting an infinite weirdness value for token $\boldsymbol{w}^{k}$, thus generating a false clue. However, using the "Add-one" smoothing technique introduced by Gale and Church (1990), or the so-called smoothed weirdness suggested by Gillam et al. (2005b) will overcame this issue. In other words, if token $\boldsymbol{w}^{k}$ does not exist in $\operatorname{CORPUS}_{R C}$, simply assign 1 to $f\left(w^{k}{ }_{R C}\right)$.

The frequency or weirdness value of a token on its own does not tell much. For example, when one says high frequency, high weirdness, to what extent does "high" justify itself? A hundred, or a thousand? A value is meaningful only relative to the means of the sample or the population. Furthermore, another problem occurs when we want to compare scores measured with different units or on a different population. How do we compare the frequency of token x with the weirdness of token y ? Scores from different distributions, such as the ones in our algorithm, can be standardised for comparisons that take account of their respective distributions by transforming the scores into $z$-scores. $Z$ score is a statistical measure that indicates the relative importance of a specific data set (x) within a given data collection by comparing the data occurrence with the mean and standard deviation of that data collection. $Z$-score is defined as:

$$
\begin{equation*}
Z=\frac{X-\mu}{\sigma} \tag{6}
\end{equation*}
$$

where
$\mathrm{X}=$ frequency of the word occurrence in the data set,
$\mu=$ the mean or average frequency of all words in the text collection, and $\sigma=$ standard deviation of word frequencies.

Similarly, weirdness $z$-score is defined as:

$$
\begin{equation*}
\text { Weirdness z - score }=\frac{W\left(w_{c}^{k}\right)-\mu\left(W\left(w_{c}\right)\right)}{\sigma\left(W\left(w_{c}\right)\right)} \tag{7}
\end{equation*}
$$

where
$W\left(w_{c}^{k}\right)=$ Weirdness of word $w^{k}$ in the CORPUS ;
$\mu\left(W\left(w_{c}\right)\right)=$ average weirdness of all words in the CORPUS ${ }_{C}$;
$\sigma\left(W\left(w_{c}\right)\right)=$ standard deviation of weirdness of all words in the CORPUS ${ }_{c}$.
and Frequency z-score is defined as:

$$
\begin{equation*}
\text { Frequency z-score }=\frac{f\left(w_{c}^{k}\right)-\mu\left(f\left(w_{c}\right)\right)}{\sigma\left(f\left(w_{c}\right)\right)} \tag{8}
\end{equation*}
$$

where
$f\left(w_{c}^{k}\right)=$ frequency of word $w^{k}$ in the CORPUS $_{C}$;
$\mu\left(f\left(w_{c}\right)\right)=$ average frequency of all words in the CORPUS ${ }_{C}$;
$\sigma\left(f\left(w_{c}\right)\right)=$ standard deviation of frequency of all words in the CORPUS ${ }_{c}$.

A Text Analysis System is designed and implemented to perform such statistical analysis. Details of the system will be reported in next section. Statistical information for all the tokens in the corpus is passed to the next stage, in which candidate terms are extracted. A Summary for generating wordlist and statistical information for tokens in CORPUS ${ }_{C}$ is listed in Figure 3.2 below:

```
a. COMPUTE \(f\left(w_{c}\right) \quad l^{*}\) Frequency of all words, w in CORPUS \(_{c}{ }^{* /}\)
    Relative frequency \(f_{R}\left(w_{C}\right)=f\left(w_{C}\right) / N_{C}\)
b. COMPUTE \(f\left(w_{\mathrm{RC}}\right) \quad / *\) Frequency of all words, w in CORPUS \(_{\mathrm{Rc}} * /\)
    Relative frequency \(f_{R}\left(w_{R C}\right)=f\left(w_{R C}\right) / N_{R C}\)
c. COMPUTE \(W\left(w_{c}\right)=f_{R}\left(w_{c}\right) / f_{R}\left(w_{R C}\right) \quad / *\) weirdness ratio of all words, w in CORPUS \({ }_{c} * /\)
d. COMPUTE mean \(\mu\left(f\left(w_{c}\right)\right) \quad / *\) average frequency of all words, w in CORPUS \({ }_{c} * /\)
e. COMPUTE standard deviation \(\sigma\left(f\left(w_{c}\right)\right) \quad / *\) standard deviation of all words, w in CORPUS \(_{\mathrm{c}}{ }^{*} /\)
f. COMPUTE mean \(\mu\left(W\left(w_{c}\right)\right) \quad / *\) average weirdness of all words, w in CORPUS \({ }_{c}{ }^{* /}\)
g. COMPUTE standard deviation \(\sigma\left(W\left(w_{c}\right)\right) \quad / *\) standard deviation of weirdness of all words,
win CORPUS \({ }^{*}{ }^{*}\)
h. COMPUTE Z-Score \(Z\left(f\left(w_{c}\right)\right) \quad / *\) Frequency \(Z\)-Score of all words, win CORPUS \(_{c}{ }^{* /}\)
i. COMPUTE Z-Score \(Z\left(W_{C}\left(w_{c}\right)\right) \quad / *\) Weirdness Z-Score of all words, w in CORPUS \({ }_{c}\) */
```

Figure 3.2. Stage 2: Generate wordlist and statistical information for each token in CORPUS $_{\mathrm{C}}$.

### 3.2.3 Extract candidate terms

In this stage, an exclude list, $S_{E x C}$, is applied to remove unwanted tokens specified by the user. Any tokens that appear in the $S_{\text {Exc }}$ will be removed. Subsequently, by restricting threshold values of frequency and frequency $z$-score, tokens with frequency less than the $F_{\text {threshold }}$ will be filtered out, and the remaining ones are further filtered by the $\boldsymbol{Z F}_{\text {rhershold }}$ value, hence producing a set of statistically frequent tokens. Furthermore, restricting threshold values of weirdness and weirdness $z$-score, tokens with weirdness less than $W_{\text {threshold }}$ will be filtered out, and the remaining ones are further filtered by ZW Threshold value, hence producing a set of domain specific tokens. Any token filtered
out due to its presence in the exclude list, or has frequency $<\boldsymbol{F}_{\text {THRESHoLD }}$, or has weirdness < $W_{\text {THRESHoLD }}$ will not be included when calculating the frequency (weirdness) $z$-score. An overview of the steps used in this stage can be found in Figure 3.3 below:
a. Let $S_{W}$ and $S_{C A N}$ the sets of all words and candidate terms in CORPUS ${ }_{c}$;
$S_{E X C}$ be the set of exclude words;
$S_{C A N}=\{\phi\} ;$
$S_{W^{*}}=S_{W}-S_{E X C} ;$
b. CATEGORISE Words

FOREACH word $w^{l}$ in $S_{W^{*}}$
IF
$f\left(w^{\prime}\right)>=F_{\text {THRESHOLD }}$ AND . $\quad I^{*}$ Frequency threshold */
$\boldsymbol{W}\left(\boldsymbol{w}^{\prime}\right)>=\boldsymbol{W}_{\text {Threshold }}$ AND $\quad$ /* Weirdness threshold */
$\boldsymbol{Z}\left(f\left(w^{b}\right)\right)>=\boldsymbol{Z} \boldsymbol{F}_{\text {THRESHOLD }}$ AND $\quad / *$ Frequency Z-Score threshold */
$\boldsymbol{Z}\left(\boldsymbol{W}\left(\boldsymbol{w}^{\prime}\right)\right)>=\boldsymbol{Z} \boldsymbol{W}_{\text {THRESHOLD }} \quad{ }^{\prime *}$ Weirdness Z-Score threshold */
THEN
$S_{C A N}=S_{C A N} \cup\left\{w^{i}\right\}$
NEXT $w^{i}$;
Figure 3.3. Stage 3: Extract candidate terms from CORPUS ${ }_{c}$.

### 3.2.4 Extract key collocates for each candidate term

In this stage, a concordance for each term in $S_{\boldsymbol{C A N}}$ is generated. A concordance is a list of the candidate terms used in a corpus, with their immediate contexts occurring on the left and right. Below are the concordance samples of the candidate term shares:
between 1.673 billion pounds and 1.9 billion pounds 40 percent . wilson bowden 's results helped its

1100 gmt they were only up 0.5 percent . the
billion, have plummeted 95 percent this year. its jarvis is likely to remain in the spotlight after its
shares fell 4.3 percent in morning trade to 468 pence,
shares rise four percent to the day's high of
shares have outperformed the construction and building index by about 17
shares were down 4.35 percent at 2.20 euros in dublin in
shares fell 13 percent on thursday after fraud police arrested colin

From the concordance, a neighbourhood of $\boldsymbol{n}$ tokens around the candidate term shares is analysed. Normally, $\boldsymbol{n}=5$, which means 5 tokens on the left of shares and 5 tokens on the right. In other words, $n$, is what Sinclair referred to as the collocation span. Instead of manually browsing the concordance to identify significant collocates of shares, the concordance is analysed automatically using statistical measures suggested by Smadja (1994).

Strength, spread and peak strength for each token $\boldsymbol{x}_{\boldsymbol{i}}$ with a vicinity of $\boldsymbol{n}$ around the candidate term shares are then computed. The default threshold values $1,10,1$ are assigned to strength, spread and peak strength; and these values can be adjusted to suit individual use. Generally, a low threshold will yield results with high recall but lower precision. This statistical information will be used for identifying potential collocates of the candidate term. Table 3.2 below shows the 10 most frequent tokens around the candidate term shares, and the corresponding statistical information.

Table 3.2: Statistical information for the top 10 tokens that co-occur with the candidate term shares.

| $f_{\text {neighbourhood of shares }}$ |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $x^{\prime}$ | $f$ | -5 | -4 | -3 | , | 3 | 4 | 5 | sum | $\mu$ | $\sigma$ | spread | strength |
| nomber | 213292 | 709 | 889 | 520 | ... | 1449 | 1014 | 1430 | 9560 | 956.00 | 436.93 | 171813.60 | 62.00 |
| percent | 41091 | 118 | 120 | 181 | ... | 1083 | 898 | 435 | 2975 | 297.50 | 387.96 | 135458.25 | 19.22 |
| to | 100446 | 284 | 181 | 176 | ... | 251 | 594 | 265 | 2216 | 221.60 | 158.68 | 22661.84 | 14.29 |
| were | 11227 | 26 | 14 | 6 | $\cdots$ | 199 | 119 | 58 | 1733 | 173.30 | 402.52 | 145817.01 | 11.15 |
| at | 26824 | 137 | 54 | 121 | ... | 183 | 119 | 651 | 1556 | 155.60 | 185.11 | 30839.44 | 10.00 |
| have | 14289 | 26 | 26 | 52 | ... | 225 | 76 | 105 | 1478 | 147.80 | 286.52 | 73886.16 | 9.50 |
| and | 70898 | 133 | 112 | 125 | $\cdots$ | 170 | 159 | 127 | 1476 | 147.60 | 81.82 | 6024.44 | 9.48 |
| on | 41486 | 200 | 218 | 137 | ... | 115 | 159 | 197 | 1456 | 145.60 | 60.81 | 3327.64 | 9.35 |
| 's | 38772 | 87 | 105 | 43 | $\cdots$ | 116 | 142 | 71 | 1403 | 140.30 | 227.35 | 46519.21 | 9.01 |
| reuters | 11915 | 66 | 46 | 453 | $\cdots$ | 3 | 3 | 3 | 1028 | 102.80 | 168.32 | 25499.56 | 6.57 |

Strength, or z -score for the collocate $\left(w^{k}, x^{i}\right)$ is defined as:

$$
\begin{equation*}
\text { strength }=\frac{f_{i}-\mu}{\sigma} \tag{9}
\end{equation*}
$$

where
$f_{i}=$ sum of the frequency of $x^{i}$ among $n$ neighbourhood;
$\mu=$ average frequency of the words that appear within the neighbourhood of $w^{k}$;
and
$\sigma=$ standard deviation of frequency of the words that appear within the neighbourhood of $w^{k}$.

Token $x^{i}$ with strength $>=S T_{\text {rhesshoLD }}$ are selected, thus filtering out low frequency collocates. Apart from analysing the frequency of $x^{i}$, and its distribution, or spread, around $\boldsymbol{n}$ neighbourhood of the candidate term is also analysed. Spread is defined as:

$$
\begin{equation*}
\text { spread }=\frac{\sum_{j=-n}^{n}\left(f_{i}^{j}-\overline{f_{i}}\right)^{2}}{2 * n} \tag{10}
\end{equation*}
$$

where
$f_{i}^{j}=$ frequency of $x^{i}$ in position $j ;$
$\overline{f_{i}}=$ average frequency of $x^{i}$ among $n$ neighbourhood; and
$n=$ number of neighbourhood, normally $n=5$, and $n \neq 0$.

The spread value indicates how token $x^{i}$ is being used around the $n$ neighbourhood. If the spread is small, then $x^{i}$ can be used equally in almost any position around the candidate term shares. On the other hand, if spread is high, then $x^{i}$ can only be used in certain position(s) around the candidate term shares. Since $x^{i}$ can be used in certain position(s), peak strength was introduced to identify such position(s). Peak strength is defined as:

$$
\begin{equation*}
\text { peak strength }=\overline{f_{i}}+\left(k_{1} \times \sqrt{\text { spread }^{i}}\right) \tag{11}
\end{equation*}
$$

where
$\bar{f}_{i}=$ average frequency of $x^{i}$ among $n$ neighbourhood;
$k_{1}=$ threshold value specified by the user, normally $k_{1}=1$; and spread $d^{i}=$ spread for $x^{i}$.

With equations (9), (10) and (11), one of the collocate for shares is shares fallen, where fallen is 2 tokens on the right of shares - shares $X$ fallen. But what is $X$ ? To reveal $X$, one needs to start from the bi-gram of shares and select a statistically significant $x^{i}$ that occurs either at position -1 or +1 . These bi-grams now become a new set of candidate term. By repeating the strength, spread and peak strength analysis, one can generate a list of statistically significant tri-gram collocates. Following Smadja (1994), this process can be repeated to generate $n$-gram collocates until the conditions in $\left(\mathrm{C}_{1}\right),\left(\mathrm{C}_{2}\right)$ and $\left(\mathrm{C}_{3}\right)$ no longer hold.

$$
\left\{\begin{align*}
\text { strength }>=S T_{\text {THRESHOLD }} & \left(C_{1}\right)  \tag{12}\\
\text { spread }>=S P_{\text {THRESHOLD }} & \left(C_{2}\right) \\
\text { peak strength }>=\bar{f}_{i}+\left(k_{1} \times \sqrt{\text { spread }^{i}}\right) & \left(C_{3}\right)
\end{align*}\right\}
$$

where
$S T_{\text {THRESHOLD }}=$ strength threshold specified by the user, normally $S T_{\text {THRESHOLD }}=1$;
$S P_{\text {THRESHOLD }}=$ spread threshold specified by the user, normally $S P_{\text {THRESHOLD }}=10$;

In the above example of shares, one of its bi-gram collocates is shares have, and one of the trigram collocates for shares have is shares have fallen, revealing $X$ as have. A summary of extracting collocates for each of the candidate terms can be found in Figure 3.4 below:
a. Let $S_{W N}$ the set of all words that occur on the left and right neighbourhood of candidate terms;
$S_{\text {coll }}$ the set of all the collocates;
$S_{t m p C o l l}$ the set of the temporary collocates;
$S_{W N}=\{\phi\}_{;}$
$S_{\text {Coll }}=\{\phi\}$;
$S_{\text {tmpColl }}=\{\phi\}$;
$n=$ size of the neighbourhood, such that
$-n<=$ left neighbourhood $<=-1$;
$+1<=$ right neighbourhood $<=+n$;
$n \in Z^{+}$and $n \neq 0$; normally, $n=5$
b. FOREACH word $w^{i}$ in $S_{C A N}$
for ( $\mathrm{j}=-\boldsymbol{n} ; \mathrm{j}<-1 ; \mathrm{j}++$ )
COMPUTE $f\left(S_{W N}\right) \quad / *$ frequency of all words on the left neighbourhood of $w^{i} * /$
for $(\mathrm{j}=1 ; \mathrm{j}<\boldsymbol{n} ; \mathrm{j}++$ )
COMPUTE $f\left(S_{w N}\right) \quad / *$ frequency of all words on the right neighbourhood of $\boldsymbol{w}^{\boldsymbol{i}}$ */
$S_{W N}{ }^{*}=S_{W N}-S_{E X C}$
COMPUTE mean $\mu\left(\mathcal{f}\left(S_{W N^{*}}{ }^{*}\right)\right) \quad l^{*}$ average frequency of all words around $w^{i}{ }^{* /}$
COMPUTE standard deviation $\sigma\left(f\left(S_{W N^{*}}^{*}\right)\right) / *$ standard deviation of frequency of all words around $w^{t}{ }^{* /}$
FOREACH word $x^{l}$ in $S_{W N^{*}}{ }^{*}$
COMPUTE spread $\operatorname{sp}\left(x^{t}\right)$
/* spread of word $x^{1}$ */
COMPUTE strength $s t\left(x^{\prime}\right)$
/* strength of word $x^{L^{*}}$ //
COMPUTE peak strength $p s t\left(x^{\prime}\right)$
/* peak strength of word $x^{i}$ */
COMPUTE mean $\mu\left(x^{\prime}\right) \quad I^{*}$ average frequency of $x^{\prime}$ among $n$ positions */
IF
$\boldsymbol{s p}\left(x^{\prime}\right)>=\boldsymbol{S} \boldsymbol{P}_{\text {threshold }}$ AND $/ *$ spread threshold */
$s t\left(x^{\prime}\right)>=S T_{\text {THRESHOLD }}$ AND /* strength threshold */
$\operatorname{pst}\left(x^{\prime}\right)>=\boldsymbol{P S T}_{\text {THRESHOLD }} \quad \quad^{*}$ peak strength threshold */ THEN

IF
$f\left(x^{h-1}\right)>=\mu\left(x^{\prime}\right) \quad \quad / *$ frequency of $f\left(x^{\prime}\right)$ at position -1*/
THEN
THEN
$S_{\text {Coll }}=S_{\text {Coll }} \cup\left\{x^{I}+w^{l}\right\} ;$
$S_{\text {tmpColl }}=S_{\text {tmpColl }} \cup\left\{x^{l}+w^{l}\right\} ;$
IF
$f\left(x^{p+1}\right)>=\mu\left(x^{\prime}\right) \quad \quad / *$ frequency of $f\left(x^{l}\right)$ at position $+1 * /$
THEN
$S_{\text {Coll }}=S_{\text {Coll }} \cup\left\{w^{\prime}+x^{\prime}\right\} ;$
$S_{\text {timpColl }}=S_{\text {tmpColl }} \cup\left\{w^{i}+x^{i}\right\} ;$
NEXT $x^{\prime}$;
NEXT $\boldsymbol{\omega}^{l}$;
c. IF

RE-COLLOCATE
THEN
$S_{C A N}=S_{\text {tmp } C_{\text {oll }} ; ~}$
$S_{\text {tmpColl }}=\{\phi\} ;$
GOTO b;
Figure 3.4. Stage 4: Extract key collocates for each candidate term in CORPUS ${ }_{c}$.

So far, we have described our approach which can automatically identify domain specific candidate terms from arbitrary domains, and extract statistically significant collocates for each of the candidate terms. In the next section, we will discuss how to construct a local grammar that can be used to extract information unambiguously from the collocates identified by our algorithm.

### 3.2.5 Extract local grammar from collocates

Automatic acquisition of domain specific candidate terms and their statistically significant n -gram collocates from a corpus has been discussed in the previous 4 stages. The frequent collocates have an unambiguous interpretation and the avoidance of ambiguity is the cornerstone of modern information retrieval. Furthermore, it has been demonstrated the frequent collocates of collocates have still more unambiguous interpretation.

The ambiguities typically occur because a pivotal verb (or noun) in a sentence can be replaced by other verbs (or nouns). It is the special nature of financial news that restricts the use of such verbs (nouns) to a very small subset of such words in the language and thereby minimises ambiguity. This approach is contrary to the current paradigm of natural language processing that is grounded in universal grammar - where many words can be used interchangeably. The approach used in this thesis is called local grammar.

In the next section, an illustration of how the algorithm works will be given using the Reuters Corpus. Constructing local grammar from the collocation patterns will also be discussed.

### 3.3 Exemplar

In the first stage, two corpora are selected, $\operatorname{CORPUS}_{C}$ and $\operatorname{CORPUS}_{R C}$. In this case, CORPUS $_{C}$ is the Reuters Corpus comprising news texts produced in 2002 which contains 3.63 million words distributed over 9,063 texts (see Table 3.1 for details). We have
chosen the British National Corpus (comprising 100-million tokens distributed over 4124 texts) as $C O R P U S_{R C}$.

From Table 3.3 below, we can see that by using the combination of frequency and frequency $z$-score, 12 out of 20 of the tokens are common words like prepositions, the verb, to be, and conjunctions, which tell little about the finance market despite their high frequency. Such common words can be excluded in the analysis by specifying their existence in the exclude list, $S_{\text {EXC }}$ (see Appendix B for the default list of commonly excluded words used in the algorithm). If we use weirdness and weirdness $z$-score, the majority of selected tokens are proper nouns, which may give us some indication about the important/active players in the finance market, but little about events happening in the market.

Table 3.3: Filtering effect by using frequency and frequency z-score, or weirdness and weirdness $z$-score.

| Top 20 by $f\left(w^{i}\right)>=5$ and $Z\left(f\left(w^{i}\right)\right)>=2$ |  |  |  | Top 20 by $W\left(w^{i}\right)>=2$ and $Z\left(W\left(w^{i}\right)\right)>=0.02$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $w^{1}$ | F( $\mathbf{w}^{\mathbf{b}}$ ) | $w^{1}$ | f( $\mathbf{w}^{\mathbf{l}}$ ) | $w^{1}$ | $\mathrm{f}\left(\mathbf{w}^{\text {l }}\right.$ ) | $\mathbf{w}^{1}$ | $\mathrm{f}\left(\mathbf{w}^{\text {l }}\right.$ ) |
| to | 100446 | year | 20779 | vivendi | 1566 | glaxosmithkline | 501 |
| and | 70898 | has | 18117 | we're | 948 | midcap | 497 |
| said | 43413 | but | 15876 | mmo2 | 772 | diageo | 488 |
| on | 41486 | after | 14663 | easyjet | 743 | invensys | 365 |
| percent | 41091 | have | 14289 | cegetel | 686 | techmark | 365 |
| 's | 38772 | market | 13811 | astrazeneca | 646 | euros | 2010 |
| by | 27763 | would | 13607 | we've | 642 | gsk | 299 |
| at | 26824 | pounds | 13193 | ryanair | 567 | rusnak | 296 |
| was | 23370 | up | 13101 | corus | 550 | cgnu | 286 |
| from | 20921 | shares | 12718 | hbos | 535 | allfirst | 271 |

It is important to use the four threshold values - frequency, weirdness, frequency $z$-score and weirdness $z$-score together rather than using only the $z$-scores. Note that the $z$-score values are depended on the mean $(\mu)$ and standard deviation $(\sigma)$ of the sample, if we ignore the frequency and/or weirdness threshold values, $\mu$ and $\sigma$ will vary significantly, hence filtering out some of the domain specific terms. In this case, a set of statistically frequent yet domain specific tokens are selected, which forms the candidate terms, $S_{C A N}$, for CORPUS . And human intervention is kept to a minimum; more resources can be allocated to verify the quality of candidate terms generated, rather than manually analysing and selecting them. This may not be apparent when dealing with a small
collection of texts in a single domain. However, considering the vast number of readily available and accessible electronic texts, it is impractical to perform such analyses manually; not to mention the cost and the time required. Table 3.4 below shows the top 40 candidate terms generated using threshold values $5,2,2,-0.2$ for frequency, weirdness, frequency $z$-score and weirdness $z$-score respectively.

Table 3.4: Filtering effect by using the combination of frequency and frequency $z$-score, and weirdness and weirdness $\boldsymbol{z}$-score.

| Top 40 by $f\left(w^{i}\right)>=5$ and $Z\left(f\left(w^{i}\right)\right)>=2$ and $W\left(w^{i}\right)>=2$ and $Z\left(W\left(w^{i}\right)\right)>=0.2$ |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{w}^{\mathbf{1}}$ | $\mathrm{f}\left(\mathbf{w}^{\prime}\right)$ | $w^{1}$ | $\mathrm{f}\left(\mathbf{w}^{\mathbf{i}}\right)$ | $\mathbf{w}^{1}$ | $\mathrm{f}\left(\mathrm{w}^{\text {' }}\right.$ ) | $\mathbf{w}^{1}$ | $\mathbf{f}\left(\mathbf{w}^{\mathbf{l}}\right)$ |
| said | 43413 | last | 9343 | bst | 6123 | investors | 4747 |
| percent | 41091 | group | 9177 | points | 5674 | financial | 4692 |
| year | 20779 | down | 7592 | expected | 5550 | analysts | 4611 |
| after | 14663 | bank | 7507 | uk | 5422 | half | 4595 |
| market | 13811 | ftse | 7271 | pence | 5414 | while | 4566 |
| pounds | 13193 | index | 6923 | stock | 5348 | price | 4501 |
| shares | 12718 | sales | 6854 | share | 5195 | month | 4454 |
| reuters | 11915 | firm | 6703 | quarter | 5039 | prices | 4410 |
| london | 10989 | growth | 6459 | sector | 4905 | fell | 4282 |
| company | 9854 | business | 6335 | profits | 4784 | months | 4257 |

Recall the $S_{E X C}$ mentioned earlier, if we knew that certain words are of no interest to us despite the fact they passed the threshold values for frequency, weirdness, frequency $z$ score and weirdness $z$-score, such as reuters and bst (British Summer Time) shown in Table 3.4 above, we can include these two words in the $S_{E X C}$. In this case, they will be excluded in stage 3 of our algorithm. Consequently, any collocates that co-occur with reuters or bst, will also be excluded in the collocation analysis (stage 4). As a result, additional candidate terms, and their collocates, can be selected and extracted.

Going back to Table 3.4, among these candidate terms, there are: financial services providers like $\operatorname{bank}(s)$; financial instruments like share(s), stock(s); verbs such as down, growth (inflection of the verb grow), fell that express changes to ftse (index); price(s), sales, profits related to the financial instruments; and the changes are reported in percent, pounds, points, or pence. It is obvious these candidate terms indeed are specific to the finance domain, and the algorithm is able to identify them automatically. This set of candidate terms is then passed to the next stage, where the collocation analysis will be performed. For simplicity, only collocations for the candidate term fell are considered in this example.

Table 3.5: Collocations for the candidate term fell.
$\nabla$ fell_nomber
$\nabla$ fell_nomber_points
$\boldsymbol{C}$ average_fell_nomber_points
- industrial_average_fell_nomber_points
$\nabla$ index_fell_nomber_points
- composite_index_fell_nomber_points
- fell_nomber_percent
- airways_fell_nomber_percent
- alliance_fell_nomber_percent
- astrazeneca_fell_nomber_percent
- average_fell_nomber_percent
- barclays_fell_nomber_percent
- bp_fell_nomber_percent
- bt_fell_nomber_percent
- carton_fell_nomber_percent
- corus_fell_nomber_percent
- glaxosmithkline_fell_nomber_percent
- group_fell_nomber_percent
- hbos_fell_nomber_percent
- holdings_fell_nomber_percent
- hsbc_fell_nomber_percent

```
\nabla- fell
```

\nabla- fell
\nabla fell_by_nomber
\nabla fell_by_nomber
\nabla fell_by_nomber_percent
\nabla fell_by_nomber_percent
- sales_fell_by_nomber_percent

```
                - sales_fell_by_nomber_percent
```

- index_fell_nomber_percent
- items_fell_nomber_percert
- logica_fell_nomber_percent
- mmo2_fell_nomber_percent
- national_fell_nomber_percent
- output_fell_nomber_percent
- prices_fell_nomber_percent
- profit_fell_nomber_percent
- profits_fell_nomber_percent
- prudential_fell_nomber_percent
- reuters_fell_nomber_percent
- revenues_fell_nomber_percent
- sage_fell_nomber_percent
- sales_fell_nomber_percent
- share_fell_nomber_percent
- shares_fell_nomber_percent
- shell_fell_nomber_percent
- stock_fell_nomber_percent
- traffic_fell_nomber_percent
- tsb_fell_nomber_percent
- turnover_fell_nomber_percent
- vodafone_fell_nomber_percent
- wireless_fell_nomber_percent
$\nabla$ fellto
V fell_to_nomber
- fell_to_nomber_percent

Sinclair's example of back shows that its upward collocates are mostly prepositions and pronouns, whilst downward collocates consist of a large number of verbs and nouns. In our case, the situation is slightly different. It is true that the majority of upward collocates of fell are prepositions; and the downward collocates are mostly nouns, or company names, to be exact, and verbs. However, nouns also made a substantial occurrence within the upward collocates of fell, and there is no occurrence of pronouns. Figure 3.5 shows the most frequent upward and downward collocates of fell. A list of all the collocates that have strength $>=0.1$ can be found in Appendix $C$ (downward collocates) and Appendix D (upward collocates).


Figure 3.5: Upward and downward collocates of fell.

From the collocates of fell, the choice of lexical items being accepted on the right context of fell is very limited - an optional article (by,to) followed by nomber percent/points. The token nomber refers to numerals such as the cardinal number million, billion, one, two, three, and so on. However, for the left context of fell, lexical items belonging to the proper nouns (or coreference to the proper nouns) category are freely accepted. These collocation patterns of fell can be easily represented by the following local grammar:


Figure 3.6. Local grammar of the candidate term fell.

Similarly, if all the terms that were used to express changes to financial instruments can be identified, the above local grammar can be easily extended to accommodate such changes, and deployed to extract market sentiment from news stories efficiently.

### 3.4 Conclusion

In this chapter, we have presented the algorithm, and illustrated how it can be used to extract domain specific candidate terms, and generate statistically significant collocates. Compared with other researchers, our approach removes the burden of relying on gazetteers, annotating (manually) or POS tagging of the corpus. Moreover, minimum human intervention is required during the process. Riloff makes the claim that she creates dictionaries of case frame automatically, and suggests that while her system does in 5 hours, a manual analysis will take 1,500 hours - this may be due to human limitation on reading and understanding texts. The case frames are generated using a set of manually crafted grammar rules. So, Riloff's system may be domain-independent; but will require considerable resources for constructing grammars for languages other than English. To the best of my knowledge grammars of this type do not exist for Chinese and neither for Arabic. Moreover, despite the fact that both the AutoSlogs and the LOUELLA Parsing System automates a significant amount of processes involved in information extraction/retrieval, their systems rely heavily on the availability and performance of parsers. This will be the bottleneck that made their systems truly domain and language independent.

Traditional approaches to information extraction show a considerable reliance on POS tagger, syntactic parsing, or manually annotating a training corpus in order for the system to generate templates that can be used to extract information effectively. Such approaches not only require enormous efforts to fine-tune the systems to work on a designated domain, but also restrict the applicability of the systems across different domains. In this chapter, an alternative approach to information extraction using corpus linguistics and statistical techniques has been presented. Through frequency, weirdness and $z$-score analyses of specialist texts, domain specific terms can be automatically identified by the algorithm described, and the frequency of a lexical item is indeed related to the acceptability of that item in the language being studied, as stated by Quirk, Greenbaum, Leech and Svartvok (1985).

Firth's insight on collocation as an abstraction at the syntagmatic level corresponds to Sinclair's idiom principle. Using collocational analysis, collocates for each of the
candidate terms identified by the algorithm can be readily extracted. The syntagmatic choice (restrictions may be lexical, grammatical or semantic) of words that co-occur with these candidate terms is restricted, but with some degree of freedom in that certain parts of the construct can be varied. In the domain of finance, changes in values of financial instruments are expressed in a precise manner. Such expressions were revealed by the algorithm developed, and local grammars can be easily constructed based on the collocation patterms. This has removed the burden of relying on a priori lexicon, or $a$ priori grammars that are generally constructed based on introspection and intuition.

Smadja (1991:10-20) identifies five interesting properties of collocations:

1. Collocations are arbitrary, meaning that collocation is not simply based on syntactic and / or semantic constraints. This shares the same view as Sinclair's idiom principle.
2. Collocations are cohesive opaque lexical clusters, meaning that the presence of one or several words of a collocation often implies or suggests the rest of it. This is analogous to Firth's famous phrase, "You shall know a word by the company it keeps". For example, in the domain of finance, when we see the word percent, the most probable left collocate will be number percent, or riselfall number percent.
3. Collocations are domain dependent, denoting words in a particular domain may have different meaning than in general language, or in another domain. For example, "volatile memory" refers to the computer memory that requires power to maintain the stored information, and is specific to Computer Science. The word "volatile" and "memory" alone are used differently.
4. Collocations are recurrent, meaning that the co-occurrence of words is not by chance, but rather that they are repeated in a given context. This is reinforced by Sinclair, "[...] they can be important in the lexical structure of the language because of being frequently repeated" (Sinclair, 1991:170).
5. Collocations do not translate well across languages, meaning that collocation must be learned or translated as a whole, as its meaning may be different from the individual meanings of the words put together. The "Shepherd's pie" example given by Kavanagh (1996) explains this issue very clearly: "Shepherd's pie is a dish made of meat with mashed potatoes on top, and is neither a pie, nor a food
exclusive to Shepherds. To add to the confusion, shepherd's pie translates to a different collocation in French: paté chinois."

Our algorithm is in part with Smadja's observations, except for the issue that Collocations do not translate well across languages, which will be discussed in the next chapter.

Another point to note is that Smadja's approach relies on the manual selection of specific words for the collocation analysis to be carried out, and criteria for selecting these words are not clear. On the other hand, our approach has automated the selection of these words through frequency, weirdness and $z$-score analyses, and appears to work well. Moreover, our algorithm is domain independent, which means switching to arbitrary domains is feasible, and appears to be language independent.

In the next chapter, the Text Analysis System developed upon this algorithm which utilises statistical techniques such as weirdness, collocation, $z$-score, and so on, will be presented. The usability of the system as an aid to corpus linguistics and information extraction, especially multi-lingual information extraction, will also be evaluated.

## Chapter 4

## 4 Case Studies and Evaluation

In Chapter 3, we have described an algorithm that can automatically identify domain specific terms through contrastive analysis at the lexical level between a sample of texts from an arbitrary domain and general language sample like the British National Corpus. Statistically significant collocates of these candidate terms with certain grammatical constructions can be automatically extracted. In this chapter, we will first perform crosslingual, cross-domain case studies through the use of Text Analysis System - an implementation of the algorithm described earlier. The three case studies are Sentiment Analysis for Financial Texts in English; Sentiment Analysis for Financial Texts in Chinese; and Sentiment Analysis for Film Reviews. An evaluation of the system will be discussed next in section 4.2. And finally, a brief description of the Text Analysis System will be given in section 4.3.

### 4.1 Case Studies

The rise of sentiment analysis is linked to the argument that the role (and bounded rationality) of intuitive beliefs and choices is significant. Sentiment (or emotional) analysis plays a key role in the theories and observations of organisational behaviour and in the newly emerging hybrid disciplines of investor psychology and behavioural economics. These disciplines emphasise the role of qualitative information: the role of perceptions and expectations about the performance of economies, organisations and individuals as reported in news and company reports.

A considerable body of literature exists where econometricians have analysed the 'impact of news' on the prices, and indirectly on the order flows. The tradition here is to use methods and techniques of time-series analysis, more precisely generalised autoregressive
conditional heteroskedasticity (Robert Engel was awarded the 1973 Economic Nobel Prize for the GARCH), and to isolate the unexpected changes (shocks) that may occur in the price over a period of time. Some researchers pre-select keywords that indicate change in the value of a financial instrument - including metaphorical terms like above, below, up and down - and use them to 'represent' positive/negative news stories. Others use the frequency of collocational patterns for assigning a 'feel-good/bad' score to the story.

We will perform a corpus-based examination (as discussed in Chapter 3) of a specialist register of news reports, that is, the financial news reports, for examining how changes in the perceived values of financial instruments are reported. Firstly, a contrastive analysis at the lexical level between a sample of financial news texts and a general language sample like the British National Corpus will be carried out, thus highlighting the domain specific terms. Secondly, grammatical patterns/constructions associated with these domain specific terms that may be used more frequently in the specialist domain for reporting changes in financial instruments will be identified. Finally, a local grammar will be constructed out of these patterns for which the market sentiment can be extracted unambiguously.

To evaluate the effectiveness of our approach, experiments will be carried out on different datasets: UK English Financial news text published by Reuters, Chinese Financial news text published by Ming Pao (Hong Kong), as well as film reviews from the Internet Movie Database (IMDB) archive.

### 4.1.1 Sentiment Analysis for Financial Texts in English

We have exclusively used Reuters Financial News Service as the basis of our corpus. Reuters has privileged access to financial news released by a number of monetary authorities and has a long well-established reputation. Typically, a corpus is constructed from a diverse range of texts, especially a range of different sources. In that sense, our corpus is a digital library organised diachronically.

Our Reuters Corpus comprises 9,063 texts, in turn, comprising 3.63 million words of text, published during Jan - Dec 2002. The average length the news reports is 400 tokens. The subjects typically covered by Reuters Financial News Service are: Company Outlooks, Company Results, Economic Indicators, Funds and Initial Public Offering News. The news reports are first written by a news reporter in the field and then subedited at base. The news reports are in that sense jointly authored. The diachronic distribution of texts is as follows:

Table 4.1: Diachronic distribution of text in Reuters corpus

|  | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| No. of stories | 778 | 682 | 572 | 650 | 631 | 589 | 962 | 658 | 887 | 1017 | 980 | 657 |
| No. of tokens | 323472 | 298202 | 246812 | 269743 | 278724 | 255094 | 401430 | 252072 | 326032 | 379056 | 402089 | 260216 |

In this section, we will look at how the most frequent words are used in the description of the changes in financial markets, because the selection of the first 200 words we have produced comprises more than $56 \%$ of the total texts, and we have noticed that the keywords and prepositions are included in the 100 -most frequent list. We do not deny that there may be other important words which are less frequent which may contribute reporting of change. However, our focus is to see how we can derive some information about sentences describing the changes in market. We will investigate how the most frequent word, that is the most frequent open class word we found in our Reuters corpus, percent, actually behaves in relation to the other words using the method we have described in Chapter 3.

A frequency analysis of many general language corpora shows that the first 100 most frequent single tokens comprise just under half of a given text corpus and that an overwhelming number of tokens belong to so-called closed-class words: determiners, conjunctions, prepositions and certain verbs. There are only a few open class tokens, especially nouns or adjectives. In the British National Corpus, the three possible openclass tokens in the first 100 most frequent tokens are people, time, and way (see Table 4.2 for details). In the Lancaster-Oslo Bergen corpus the two open-class tokens were man and time. The first ten most frequent single tokens comprise around a quarter of the general language corpus and these are all short (fewer than 4 or five characters) closed class tokens. These observations confirm Zipf's Law (1945), that is, in a corpus of
natural language texts, the frequency of any word is roughly inversely proportional to its rank in an ordered frequency list.

Table 4.2: Distribution of the first 100 most frequent tokens in the BNC according to the cumulative frequency of ten tokens at a time.

| Rank | Token | Cumulative <br> Relative <br> Frequency | No. of <br> Open Class <br> Words |
| :--- | :--- | ---: | ---: |
| $1-10$ | the, of, and, to, a, in, that, it, is, was | $22.51 \%$ | 0 |
| $11-20$ | i, for, 's, on, you, he, be, with, as, by | $7.01 \%$ | 0 |
| $21-30$ | at, have, are, this, not, but, had, his, they, from | $4.53 \%$ | . |
| $31-40$ | she, which, or, we, an, n't, there, her, were, one | $3.42 \%$ | 0 |
| $41-50$ | do, been, all, their, has, would, will, what, if, can | $2.56 \%$ | 0 |
| $51-60$ | when, so, no, said, who, more, about, up, them, some | $1.98 \%$ | 0 |
| $61-70$ | could, him, into, its, then, two, out, time, like, only | $1.59 \%$ | 0 |
| $71-80$ | my, did, other, me, your, now, over, just, may, these | $1.37 \%$ | 2 |
| $81-90$ | new, also, people, any, know, very, see, first, well, after | $1.22 \%$ | 0 |
| $91-100$ | should, than, where, back, how, get, most, way, down, our | $1.02 \%$ | 3 |
|  |  | $47.20 \%$ | 1 |

The cumulative frequency of the first 100 most frequent tokens in our Reuters Financial Corpus is $47.47 \%$; slightly higher than that of the BNC. However, the key difference is in the appearance of 45 possible open-class tokens: the $10^{\text {th }}$ most frequent token is percent, followed by year $\left(22^{\text {th }}\right)$, million $\left(26^{\text {th }}\right)$, and market $\left(30^{\text {th }}\right)$; there are proper nouns including Reuters $\left(37^{\text {th }}\right)$, London $\left(42^{\text {th }}\right)$, FTSE $\left(61^{\text {th }}\right)$, and $($ the $) U K\left(80^{\text {th }}\right)$. The key tokens including share (including the plural shares $-34^{\text {th }}-$ and the singular share $-82^{\text {th }}$, and the synonym stock- $\left.83^{\text {th }}\right)$, company $\left(47^{\text {th }}\right)$, group $\left(48^{\text {th }}\right)$, investors, profits and analysts $\left(86^{\text {th }}-\right.$ $88^{\text {th }}$ ) make a significant appearance (see Table 4.3).

Table 4.3: Distribution of the first 100 most frequent tokens in the Reuters corpus according to the cumulative frequency of ten tokens at a time.

| Rank | Token | Cumulative <br> Relative <br> Frequency | No. of <br> Open Class <br> Words |
| :--- | :--- | ---: | ---: |
| $1-10$ | the, to, of, in, a, and, s, said, on, percent | $21.70 \%$ | 1 |
| $11-20$ | for, it, its, that, by, at, was, is, as, with | $7.30 \%$ | 0 |
| $21-30$ | from, year, has, but, be, million, which, after, have, market | $4.44 \%$ | 3 |
| $31-40$ | would, un, pounds, shares, will, are, Reuters, had, this, an | $3.37 \%$ | 4 |
| $41-50$ | were, London, not, u, he, we, company, group, last, billion | $2.71 \%$ | 4 |
| $51-60$ | bank, more, than, over, also, one, been, new, down, about | $2.08 \%$ | 4 |
| $61-70$ | FTSE, they, sales, index, first, firm, bst, growth, their, some | $1.81 \%$ | 7 |
| $71-80$ | or, business, there, two, points, could, expected, pence, quarter, UK | $1.55 \%$ | 7 |
| $81-90$ | out, share, stock, sector, financial, investors, profits, analysts, off, <br> three | $1.32 \%$ | 8 |
| $91-100$ | while, price, since, may, half, month, prices, months, if, into | $1.19 \%$ | 8 |
|  |  | TOTAL | $47.47 \%$ |

Another notable point about the 100 most frequent tokens in our corpus is the appearance of prepositions ( $u p$ and down) amongst the 100 high frequency tokens ( $32^{\text {nd }}$ and $59^{\text {th }}$ ) and the derived noun growth, $68^{\text {th }}$ (from grow). Together with keywords like profits $\left(87^{\text {th }}\right.$ ), these tokens might throw some light on the mood of the market.

The comparison of the frequency distribution of the first 100 tokens in the BNC and in our corpus helps us in identifying certain keywords used by the financial news writers/reporters. The 45 possible open class tokens are also potential keywords. Unlike, science and engineering texts, we also see a number of potential verbs, prepositions and adjectives. The second most frequent 100 words, although comprising much smaller proportions of texts in both the BNC (c. $6 \%$, Table 4.5) and our corpus (c. $9 \%$, Table 4.4), show the preponderance of more keywords, and, especially for us, more of the potential verbs that might have been used to describe changes in the market. In particular, we note the frequent use of fell $\left(105^{\text {th }}\right.$ most frequent with 3508 occurrences amounting to $0.11 \%$ of the total text corpus) and rose ( $110^{\text {th }}$ most frequent with 3450 occurrences amounting to $0.108 \%$ ). Indeed, if we add up all the inflections of rise and fall we obtain a frequency of $0.24 \%$ and $0.27 \%$ in our corpus.

When we compare the distribution of the potential movement tokens, rose and fell, in our corpus with that of the BNC, we find that rose, including its partial homonyms, for example, rose - a flower or a proper name, is used 10 times more frequently in our
corpus. The same is true of fell. The other inflected forms are used at least 5 times more frequently in our corpus than in the BNC.

Table 4.4: Distribution of tokens ranked between 101-200 in the Reuters corpus, grouped ten at a time.

| Rank | Token | Cumulative Relative Frequency | No. of Open Class Words |
| :---: | :---: | :---: | :---: |
| 101-110 | week, trading, around, profit, chief, fell, economic, no, interest, rose | 1.10\% | 8 |
| 111-120 | cut, higher, news, years, next, earnings, rates, european, rise, industry | 1.05\% | 8 |
| 121-130 | second, british, low, deal, world, government, time, end, back, investment | 1.00\% | 6 |
| 131-140 | added, executive, his, thursday, stocks, when, markets, results, other, biggest | 0.95\% | 5 |
| 141-150 | britain, still, banks, told, i, wednesday, all, economy, tuesday, before | 0.90\% | 5 |
| 151-160 | september, now, tax, five, rate, lower, companies, europe, says, services | 0.86\% | 6 |
| 161-170 | six, euro, early, strong, like, who, vodafone, fall, t , reported | 0.82\% | 5 |
| 171-180 | cash, any, gmt, telecoms, third, high, debt, earlier, bid, so | 0.77\% | 5 |
| 181-190 | analyst, statement, united, monday, four, country, data, top, oil, value | 0.72\% | 6 |
| 191-200 | under, our, just, friday, recovery, largest, states, further, hit, consumer | 0.69\% | 6 |
|  | TOTAL | 8.87\% | 60 |

Table 4.5: Distribution of tokens ranked between 101-200 in the BNC grouped ten at a time. Note that there are some abbreviations in this table, e.g. the title mr and erm (European exchange Rate Mechanism).

| Rank | Token | Cumulative <br> Relative <br> Frequency | No. of <br> Open Class <br> Words |
| :--- | :--- | ---: | ---: |
| $101-110$ | made, got, much, think, work, between, go, years, er, many | $0.91 \%$ | 3 |
| $111-120$ | 've, being, those, before, right, because, through, 're, yeah, good | $0.86 \%$ | 2 |
| $121-130$ | three, make, us, such, still, year, 'll, must, last, even | $0.75 \%$ | 2 |
| $131-140$ | take, own, too, off, here, come, both, does, say, oh | $0.70 \%$ | 1 |
| $141-150$ | used, 'd, going, 'm, erm, use, government, day, man, might | $0.64 \%$ | 4 |
| $151-160$ | same, under, yes, however, put, world, another, want, thought, while | $0.59 \%$ | 1 |
| $161-170$ | life, again, against, never, need, old, look, home, something, mr | $0.55 \%$ | 4 |
| $171-180$ | long, house, why, each, part, since, end, number, out_of, found | $0.50 \%$ | 4 |
| $181-190$ | place, different, went, little, really, ', came, left, children, local | $0.48 \%$ | 7 |
| $191-200$ | within, always, without, four, around, great, give, set, system, small | $0.45 \%$ | 4 |
|  |  | $6.43 \%$ | 32 |

When we consider the most frequent 200 words in our Reuters corpus of text (Table 4.3 \& Table 4.4); together, they compile $56.34 \%$ of the total number of tokens in the texts. It would be interesting to see how these three fifths of the texts, or rather the tokens in the three fifths of the texts, co-occur with each other.

So far, the selection of the candidate terms was based on frequency analysis and human introspection. The introspection part is time consuming and requires expert knowledge in the domain being studied. With the algorithm discussed in Chapter 3, the Text Analysis System is able to select domain specific candidate terms based on the threshold values of frequency, weirdness, frequency $z$-score and weirdness $z$-score specified automatically. Here, we set the threshold values as $5,2,2,-0.2$ for frequency, weirdness, frequency $z$-score and weirdness $z$-score respectively. These values depend on size of the corpus size and can be adjusted. Generally, the higher the threshold values, the less number of domain specific terms being selected from the corpus.

Note that the weirdness $z$-score threshold value is negative; this is expected as the number of proper nouns present in the Reuter Corpus is much higher than that of the BNC, thus a large weirdness standard deviation value. The system selected 115 candidate terms out of the 33,612 word vocabulary. Table 4.6 below shows the comparison between the two sets of candidate terms: one set was generated by the system (from the entire corpus) and the other set was selected by the user ( 121 from the top 200, as highlighted in Table 4.3 and Table 4.4). Terms in bold are common to both sets; terms underlined (17) are unique to the system generated set, while terms in italic (23) are unique to the user selected set. $81 \%$, or 98 of the system generated candidate terms are present in the user selected set, this shows that our algorithm performs pretty well at generating domain specific candidate terms from a corpus, without human intervention. This is important as corpora users can have a general overview of the corpus, without knowing, or just knowing a little about the corpus being investigated.

Table 4.6: Comparison between system generated candidate terms and user selected candidate terms, ranked by frequency and grouped ten at a time.

| Rank | System generated Candidate Terms | User selected Candidate Terms |
| :---: | :---: | :---: |
| 1-10 | said, percent, its, year, after, market, pounds, shares, Reuters, London | percent, year, million, market, up, pounds, shares, Reuters, London, company |
| 11-20 | u.s, company, last, group, down, bank, FTSE, index, sales, firm | group, billion, bank, more, over, one, new, down, FTSE, sales |
| 21-30 | growth, business, bst, points, expected, UK, pence, stock, share, quarter | index, first, firm, bst, growth, business, two, points, pence, quarter |
| 31-40 | sector, profits, investors, financial, analysts, half, while, price, since, month | UK, share, stock, sector, financial, investors, profits, analysts, three, price |
| 41-50 | prices, fell, months, around, week, trading, economic, profit, rose, chief | may, half, month, prices, months, week, trading, profit, chief, fell |
| 51-60 | higher, british, interest, rise, news, european, cut, next, rates, low | economic, interest, rose, cut, higher, news, years, earnings, rates, european |
| 61-70 | earnings, second, deal, industry, added, executive, investment, thursday, britain, markets | rise, industry, second, british, low, deal, world, government, time, end |
| 71-80 | stocks, tax, results, biggest, economy, lower, told, banks, says, wednesday | back, investment, executive, thursday, stocks, markets, results, biggest, britain, banks |
| 81-90 | strong, gmt, rate, tuesday, fall, september, companies, europe, early, services | wednesday, economy, tuesday, september, tax, five, rate, lower, companies, europe |
| 91-100 | telecoms, reported, euro, oil, cash, united, debt, monday, vodafone, earlier | services, six, euro, early, strong, vodafone, fall, reported, cash, gmt |
| 101-110 | statement, recovery, analyst, june, friday, data, country, top, july, hit | telecoms, third, high, debt, bid, analyst, statement, united, monday, four |
| 111-120 | states, largest, value, bid, street | country, data, top, oil, value, friday, recovery, largest, states, hit |
| 121-130 |  | consumer |

The most frequent candidate term that appears in both sets is percent. The word percent is very pervasive; it occurs more than 41,000 times, making it $1.1 \%$ of the total texts. It is a keyword for financial texts specifically, and scientific texts generally. Normally, changes are measured in percentage term. In scientific texts, the "\%" symbol is used to indicate percentage. While in financial texts; we have noted the preponderance of the word itself, and often related to changes in value. As percentage refers to change, we noted that the most significant collocates are the numbers which are found. How can we make such a claim? This is achieved by using collocation and the algorithm discussed in Chapter 3.

We took a neighbourhood of five around the candidate term percent, which means we look at the five words which were 'left neighbours' of the term and five 'right neighbours' of the term. By setting the threshold values of spread, strength and peak strength to $10,1,1$ (these values can be adjusted depending on the corpus size), we see
from Table 4.7 that the token occurring most frequently with percent in those ten different positions is the word nomber. Nomber itself occurs 213,292 times, the word nomber and percent co-occur with each other 62,932 times, of which, $64.7 \%$ of the time the word nomber occurs next to percent. This is closely followed by the word up. Up itself occurs 13,101 times, the word $u p$ and percent co-occur with each other 4503 times, of which, $63.8 \%$ of the time the word $u p$ occurs with percent - not next to percent, but a token away from percent. For instance, up 2 percent, up 2.2 percent. This is closely followed by 683 of its co-occurrence ( $2^{\text {nd }}$ peak) with the token percent, but with 2 tokens in-between (for example, up by 2 percent, up nearly 2 percent). Note that the right collocation percent up, percent $\mathrm{x} u p$, and so on have much lower frequency than the left ones. The words down, rose and fell share the same collocation patterns with $u$ p.

Table 4.7: Ten most frequent collocates of percent generated by the Text Analysis System.

|  | $f_{\text {neighbourhood of percent }}$ |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | f | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | Total |
| percent | 41091 |  |  |  |  |  |  |  |  |  |  |  |
| nomber | 213292 | 2291 | 1804 | 1118 | 306 | 40743 | 90 | 5805 | 3849 | 4268 | 2658 | 62932 |
| up | 13101 | 235 | 132 | 683 | 2876 | 1 | 89 | 68 | 115 | 178 | 126 | 4503 |
| year | 20779 | 266 | 571 | 179 | 1 | 0 | 169 | 630 | 727 | 833 | 512 | 3888 |
| down | 7592 | 265 | 117 | 371 | 2561 | 0 | 107 | 46 | 85 | 146 | 103 | 3801 |
| shares | 12718 | 435 | 898 | 1083 | 59 | 0 | 0 | 81 | 181 | 120 | 118 | 2975 |
| rose | 4041 | 88 | 67 | 319 | 2170 | 0 | 0 | 3 | 54 | 62 | 67 | 2830 |
| fell | 4282 | 117 | 97 | 280 | 1993 | 0 | 1 | 5 | 55 | 83 | 67 | 2698 |
| points | 5674 | 29 | 844 | 563 | 1 | 0 | 4 | 0 | 253 | 118 | 49 | 1861 |
| said | 43413 | 202 | 82 | 29 | 29 | 0 | 48 | 21 | 667 | 398 | 354 | 1830 |
| pence | 5414 | 66 | 110 | 109 | 2 | 0 | 0 | 2 | 1213 | 290 | 34 | 1826 |

This can be illustrated even more clearly by visualising collocates of percent around its left and right neighbours, as shown in Figure 4.1 below:


Figure 4.1: Left and right neighbours of collocate percent.

By further analysing the collocates of percent, percent indeed relates to changes in the stock market: - profit or loss of share(s), instrument(s), are reported as a percentage of the previous price. Table 4.8 shows the tri-gram collocates of percent that occur at least 100 times in the Reuters corpus.

Table 4.8: Tri-gram collocates of percent, with frequency $>=100$.

| Positive | $f_{\text {Coll }}$ | $f_{\text {Raw }}$ | Negative | $f_{\text {Coll }}$ | $f_{\text {Raw }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $u p$ nomber percent | 2876 (22\%) | 13101 | down nomber percent | 2561 (34\%) | 7592 |
| rose nomber percent | 2170 (54\%) | 4041 | fell nomber percent | 1993 (47\%) | 4282 |
| added nomber percent | 435 (12\%) | 3583 | lost nomber percent | 513 (33\%) | 1562 |
| climbed nomber percent | 432 (61\%) | 705 | dropped nomber percent | 383 (45\%) | 846 |
| gained nomber percent | 408 (57\%) | 720 | shed nomber percent | 323 (55\%) | 587 |
| jumped nomber percent | 392 (53\%) | 744 | slipped nomber percent | 292 (47\%) | 621 |
| rising nomber percent | 261 (22\%) | 1193 | dipped nomber percent | 289 (58\%) | 497 |
| grew nomber percent | 147 (29\%) | 504 | off nomber percent | 270 (6\%) | 4658 |
| gaining nomber percent | 109 (52\%) | 208 | fallen nomber percent | 158 (18\%) | 888 |
| surged nomber percent | 105 (33\%) | 316 | slid nomber percent | 155 (52\%) | 298 |
| bounced nomber percent | 102 (31\%) | 327 | falling nomber percent | 154 (14\%) | 1134 |
| rallied nomber percent | 100 (34\%) | 298 | slumped nomber percent | 133 (34\%) | 392 |
| nomber percent higher | 958 (24\%) | 3937 | tumbled nomber percent | 128 (37\%) | 344 |
| nomber percent gain | 224 (21\%) | 1057 | plunged nomber percent | 104 (33\%) | 316 |
| nomber percent increase | 213 (14\%) | 1568 | nomber percent lower | 645 (19\%) | 3327 |
| nomber percent jump | 154 (38\%) | 404 | nomber percent fall | 569 (18\%) | 3149 |
| nomber percent above | 103 (8\%) | 1266 | nomber percent drop | 369 (27\%) | 1376 |

The collocates can be categorised into two groups: positive and negative. Positive collocates refers to the upward movement of price (instrument), while negative collocates refers to the downward movement. Pertaining to their functions in the finance domain, we will refer them as sentiment words. These sentiment words can either appear on the left of percent - up nomber percent, down nomber percent, bounced nomber percent, slumped nomber percent, and so on; or, directly on the right of percent - nomber percent higher, nomber percent lower, nomber percent drop, nomber percent increase, and so on. Percentage value (\%) shown in Table 4.8 indicates the Confidence Level that a sentiment word is used in the sense that it reflects changes in the financial market precisely, rather than other senses. Confidence level (CF) is defined as:

$$
\begin{equation*}
C F^{i}=\frac{f^{i}{ }_{\text {Coll }}}{f_{\text {Raw }}^{i}} \tag{13}
\end{equation*}
$$

where
$f^{i}$ coll $=$ frequency of collocate for sentiment word $i$ in the training corpus; and
$f_{\text {Raw }}^{i}=$ absolute frequency of sentiment word $i$ in the training corpus.

Take the example of rose which has many senses as found in the Collins Bank of English:
access was willed, the barrier rose : the horror of the scene thus
On the far side pale grey cliffs rose, their tops level with the ground on grateful for peace, for what General Rose has done here. [p] My parents have You [f] acted. [f] A hopeless silence rose between them. Through the years his

However, in the financial domain, we are $54 \%$ confident that rose refers to the change in value of an instrument on the positive way, as identified by the collocate rose nomber percent. Other n-gram collocates such as up by nomber percent, fell over nomber percent, fell nearly nomber percent, rose just nomber percent, rose over nomber percent, variants of the tri-gram collocates also exist, but are less frequent.

From the collocates of percent generated by the Text Analysis System, a local grammar to extract the positive and negative market sentiment is derived, as shown in Figure 4.2 and Figure 4.3 respectively.


Figure 4.2: A local grammar to extract positive market sentiment in English.


Figure 4.3: A local grammar to extract negative market sentiment in English.

Using the local grammar derived above, market sentiment can be extracted efficiently without the need to use parser, POS tagger or gazetteer. Some of the sentiment bearing sentences extracted by the local grammar are listed below:

| , with food sales | up | 14 | percent | but bar sales flat |
| ---: | :--- | :--- | :--- | :--- |
| price ) has gone | up | by 20 | percent | because it is a |
| halifax, showing prices | up | nearly five | percent | on the month in |
| overnight that prices were | up | only 18 | percent | in the third quarter |
| enterprise shares | rose | 2.83 | percent | to 584 pence in |
| home loan repayments, | rose | to 2.3 | percent | in the year to |
| 152 pence, logica | rose | over eight | percent | and cmg climbed 6 |
| last week house prices | rose | only 1.4 | percent | last month-- |
| paring losses to trade | down | 9.5 | percent | at 105 pence. |
| 2001 , only slightly | down | from 33 | percent | percentin 1998. |
| the economy next year | down | to 2.25 | percent | from a previous estimate |
| shares in viacom closed | down | nearly three | percent | in new york at |
| budget airline easyjet | fell | 9.7 | pcrcent | , with investors ignoring |
| the next ten years | fell | to 51 | percent | from 54 percent in |
| but its shares | fell | nearly seven | percent | as investors worried about |
| before goodwill amortisation | fell | by 6.7 | percent | to 46.3 |

Through the case study of Sentiment Analysis for Financial Text in English, our approach shows promising results. The domain specific candidate terms are identified automatically, and their collocates indeed reflect changes in values of financial instruments. Local grammars constructed from the collocates can be used to track such changes unambiguously.

### 4.1.2 Cross-validation: Sentiment Analysis for Financial Texts in Chinese

In this study, we would like to investigate whether our algorithm will work on a different language - Chinese. Regarding the availability of corpora resources, there are fewer Chinese corpora available than is the case for English. We used the Ming Pao Financial news text ( 1.56 million words) published in Hong Kong between Aug 2000 and July 2001 in Big5 encoding. We will refer this as the Ming Pao Financial Corpus (MPFC). Word segmentation of the MPFC was done using the Chinese University of Hong Kong's Jansers system. We have compiled frequency lists from the TaBE (Localization for Taiwan and Big5 Encoding, see Hsiao et al. 2000) Project ( 54.2 million tokens) and the LDC Chinese Resources ${ }^{1}$ ( 4.9 million tokens), as a Chinese General Language Corpus ( $\mathrm{TaBE}+\mathrm{LDC}$ ) totalling 59.1 million tokens. These corpora, from which the frequency lists were obtained, appear to be representative.

[^6]In the $\mathrm{TaBE}+\mathrm{LDC}$ corpus，there are 37 possible open－class tokens in the first 100 most frequent tokens，mainly in the area of Computer Science．The $13^{\text {th }}$ most frequent token is綢（internet），followed by 首頁（homepage， $28^{\text {th }}$ ），軟盛（software， $31^{\text {th }}$ ），細路（network， $32^{\text {th }}$ ），服務（service， $37^{\text {th }}$ ），全球（global， $38^{\text {th }}$ ），系統（system， $39^{\text {th }}$ ），資訊（information， $41^{\text {th }}$ ），资訊網（information network， $42^{\text {th }}$ ），資料（information， $53^{\text {th }}$ ），蒙脑（computer， $71^{\text {th }}$ ），技術 （technique， $80^{\text {th }}$ ），研究（research， $98^{\text {th }}$ ）and 科舉技術（scientific technique， $99^{\text {th }}$ ），as shown in Table 4．9．

Table 4．9：Distribution of the first 100 most frequent tokens in the TaBE＋LDC corpus according to the cumulative frequency of ten tokens at a time．

| Rank | Token | Cumulative Relative Frequency | No．of Open Class Words |
| :---: | :---: | :---: | :---: |
| 1－10 | 的，之，在，及，是，與，一，回，我，了 | 6．79\％ | 0 |
| 11－20 | 上，有，綱，中，以，和，你，人，為，年 | 2．63\％ | 1 |
| 21－30 | 前，新開，或，台滴，不，第，下，首頁，大，他 | 2．03\％ | 3 |
| 31－40 | 乾睤，網路，新，也，您，所，服務，全球，系統，日 | 1．77\％ | 6 |
| 41－50 | 资訊，资㷖網，公司，要，而，就，對，國立，月，隆 | 1．59\％ | 5 |
| 51－60 | 三，號，资料，等，時，淂酹，傑，到，將，會 | 1．41\％ | 2 |
| 61－70 | 請，小，提供，區，站，名，於，者，文章，相關 | 1．34\％ | 5 |
| 71－80 | 媘，頁，行政院，本，可，二，由，我們，股份有限公司，技術 | 1．28\％ | 4 |
| 81－90 | 其，來，但，地，可以，用，後，這，交通大瞟，废商 | 1．15\％ | 2 |
| 91－100 | 一個，元，中央研究院，內容，同题，使用，計算中心，研究，科醇技術，國家科掌委員會 | 1．09\％ | 9 |
|  | TOTAL | 21．10\％ | 37 |

As TaBE forms majority of the corpus，it comes as no surprise that the proper nouns such as 行政院（Executive Yuan， $83^{\text {th }}$ ），交通大學（Chiao Tung University， $89^{\text {th }}$ ），中央研究院 （Academic Sinica， $93^{\text {th }}$ ），計算中心（Computing Centre， $97^{\text {th }}$ ），國家科學委員會（National Science Council， $100^{\text {th }}$ ）are all within 台溫（Taiwan， $24^{\text {th }}$ ）．Moreover，the first ten most frequent single tokens comprise about 7 percent of the general language corpus and these are all single character closed class tokens．

Table 4．10：Distribution of the first 100 most frequent tokens in the Ming Pao Financial Corpus according to the cumulative frequency of ten tokens at a time．

| Rank | Token | Cumulative Relative Frequency | $\qquad$ |
| :---: | :---: | :---: | :---: |
| 1－10 | 的，在，及，有，元，會，是，為，公司，將 | 9．91\％ | 2 |
| 11－20 | 但，市場，不，百分之，中，成，至，亦，年，大 | 4．68\％ | 3 |
| 21－30 | 於，而，表示，以，該，股，業務，美元，銀行，個 | 3．76\％ | 5 |
| 31－40 | 已，較，對，美國，與，他，後，新，經溶，報道 | 3．00\％ | 3 |
| 41－50 | 可，集回，兩，上，周，投资，每，多，日報，星島 | 2．60\％ | 5 |
| 51－60 | 息，高，則，報，低，今年，去年，最，股㑑，減 | 2．36\％ | 8 |
| 61－70 | 預期，升，盁利，指出，令，並，末，橃展，月，第 | 2．16\％ | 6 |
| 71－80 | 內，中圆，跌，角，由，認為，仍，收購，上市，湎長 | 2．00\％ | 7 |
| 81－90 | 昨日，仙，約，之，更，室訊，計䛚，日，時，股份 | 1．80\％ | 6 |
| 91－100 | 由於，有關，香港，向，內地，達，要，再，了，目前 | 1．59\％ | 4 |
|  |  | 33．87\％ | 49 |

The cumulative frequency of the first 100 most frequent tokens in the MPFC is $33.87 \%$ ； $12 \%$ higher than that of the TaBE＋LDC．Comparing the number of possible open－class words，the difference is $12: 49$ in the MPFC and 37 in the TaBE＋LDC；this is not as dramatic as in English： 45 in Reuters and 5 in BNC．

In MPFC，the $5^{\text {th }}$ most frequent token is $\boldsymbol{\pi}$（yuan，currency used in China），followed by 百分之（percent， $14^{\text {th }}$ ）。市場（market， $12^{\text {th }}$ ）and its constituents such as 公司（company， $9^{\text {th }}$ ），銀行（Bank， $29^{\text {th }}$ ），集国（Group， $42^{\text {th }}$ ），and 茞訊（telecom， $86^{\text {th }}$ ）；activities in the market like投资（investment， $46^{\text {th }}$ ），收䢂（takeover， $78^{\text {th }}$ ），上市（flotation， $79^{\text {th }}$ ）；年（year， $19^{\text {th }}$ ），周 （week， $45^{\text {th }}$ ），今年（this year， $56^{\text {th }}$ ），去年（last year， $57^{\text {th }}$ ），月（month， $69^{\text {th }}$ ），昨日（yesterday， $81^{\text {th }}$ ），and $日\left(\right.$ day， $88^{\text {th }}$ ）indicate when such activities happened．Proper nouns including美國（USA， $34^{\text {th }}$ ），星島（SingTao，a newspaper in Hong Kong， $50^{\text {th }}$ ），中國（China， $72^{\text {th }}$ ），香港（Hong Kong， $93^{\text {th }}$ ），and 內地（Mainland China， $95^{\text {th }}$ ）may give us clues to the location or organisations that such activities occurred．高（higher， $52^{\text {th }}$ ），低（lower， $54^{\text {th }}$ ），隇 （decrease， $60^{\text {th }}$ ），升（rise， $62^{\text {th }}$ ），跌（fall， $73^{\text {th }}$ ）and 增長（increase， $80^{\text {th }}$ ）reflect the change in息（interest， $51^{\text {th }}$ ），股（share， $26^{\text {th }}$ ）and its derivatives 股㑯（share price， $59^{\text {th }}$ ）and 股份
（share， $90^{\text {th }}$ ），which may lead to 盈利（profit， $63^{\text {th }}$ ）．Profits and loss are expressed in currencies such as $\bar{\pi}$（yuan），美元（US dollar， $28^{\text {th }}$ ），角（ 0.1 yuan， $74^{\text {th }}$ ）or 仙（cent， $82^{\text {th }}$ ）． All of these are 報道（reported， $40^{\text {th }}$ ）in 日報（daily newspaper， $50^{\text {th }}$ ）．

Similar to English，the comparison of the frequency distribution of the first 100 tokens in the TaBE＋LDC and in MPFC again helps us in identifying certain keywords used by the financial news writers and reporters．The 49 possible open class tokens are also potential keywords．The second most frequent 100 words，although comprising much smaller proportions of texts in both the TaBE＋LDC（c．7．7\％，Table 4．12）and MPFC（c． $11.2 \%$ ， Table 4．11），show the preponderance of additional keywords，such as 投資者（investor， $103^{\text {th }}$ ），哣券（Security， $116^{\text {th }}$ ），基金（fund， $129^{\text {th }}$ ），資金（capital， $137^{\text {th }}$ ），股東（shareholder， $145^{\text {th }}$ ），股橅（equity， $178^{\text {th }}$ ）；and especially for us，more of the potential verbs that might have been used to describe changes in the market．In particular，we note the frequent use of 上升（rise， $108^{\text {th }}$ most frequent token with 1963 occurrences amounting to $0.15 \%$ of the total text corpus）and $\mathrm{F}^{\text {跌（drop，}} 112^{\text {th }}$ most frequent token with 1875 occurrences amounting to $0.14 \%$ ）．

Table 4．11：Distribution of tokens ranked between 101－200 in the Ming Pao corpus according to the cumulative frequency of ten tokens at a time．

| Rank | Token |  | Cumulative <br> Relative <br> Frequency | No．of Open Class Words |
| :---: | :---: | :---: | :---: | :---: |
| 101－110 | 湓科，家，投资者，客戶，半，只，收市，上升，公布，可能 |  | 1．48\％ | 6 |
| 111－120 | 項目，下跌，又，其，不過，錐券，能，影䇾，堛加，主要 |  | 1．37\％ | 6 |
| 121－130 | 指，所，眼務，下，等，這，兌，若，基金，雖然 |  | 1．28\％ | 3 |
| 131－140 | 說，前，厘，港元，和黄，現時，資金，美，也，和 |  | 1．18\％ | 4 |
| 141－150 | 佔，無，人，相信，分析㗐，股東，部分，底，未來，行 |  | 1．10\％ | 5 |
| 151－160 | 包括，本，正，或，政雛，资迹，指數，提供，出現，其他 |  | 1．05\％ | 6 |
| 161－170 | 企業，因為，消息，日国，政府，移動，主席，便，收，董事 |  | 1．00\％ | 7 |
| 171－180 | 進行，日本，市，港股，交易，作，位，股盒，期，國鄚 |  | 0．96\％ | 7 |
| 181－190 | 業緁，就，推出，惧款，網絡，情況，出售，人士，季，紐約 |  | 0．91\％ | 9 |
| 191－200 | 才，銀，到，明年，可以，名，受，垻，䢒，全球 |  | 0．84\％ | 4 |
|  |  | TOTAL | 11．16\％ | 56 |

When we compare the distribution of the potential movement tokens，低（lower），減 （decrease），升（rise），跌（fall），增長（increase），上升（rise）and 下跌（fall），in MPFC with that of $\mathrm{TaBE}+\mathrm{LDC}$ ，we find that none of them exists in $\mathrm{TaBE}+\mathrm{LDC}$ ．The only exception is 高 （higher，or surname Gao），which exists in both corpora，but it is used at least 3 times more frequently in MFPC．

Table 4．12：Distribution of tokens ranked between 101－200 in the TaBE＋LDC corpus， grouped ten at a time．

| Rank | Token | Cumulative Relative Frequency | No．of Open Class Words |
| :---: | :---: | :---: | :---: |
| 101－110 | 资料中心，路，科技，企業，生品，並，教育，那，査詢，多 | 1．01\％ | 6 |
| 111－120 | 管理，都，還业，能，軍子，四，應，主，設備，中心 | 0．92\％ | 5 |
| 121－130 | 啊，已，陳，時間，簡介，好，留言，有限公司，錄，業 | 0．86\％ | 4 |
| 131－140 | 間，至，台北市，陜，弡展，日期，政府，文，活動，此 | 0．81\％ | 5 |
| 141－150 | 林，讓，行，國際，一頁，家，各，說，最，類 | 0．75\％ | 2 |
| 151－160 | 期，工商，高，如，内，會員，再，一策，主题，八十 | 0．72\％ | 4 |
| 161－170 | 五，郎，軍，法，被，如何，工作，支授，國，科 | 0．69\％ | 2 |
| 171－180 | 去，目前，美，維错，使，軍話，表，又，自己，學生 | 0．67\％ | 4 |
| 181－190 | 资源，对，世界，學，版，用戶，美娄，從，全，如果 | 0．64\％ | 4 |
| 191－200 | 名稱，組，給，個，天，得，八，她，華，其他 | 0．62\％ | 1 |
|  |  | 7．69\％ | 37 |

Consider the most frequent 200 words in MPFC（Table 4.10 \＆Table 4．11）；together they compile $45.02 \%$ of the total number of tokens in the texts．It would be interesting to see how half of the texts，or rather the tokens in half of the texts，co－occur with each other．

As is the case in English，we set the threshold values as 5，2，2，-0.2 for frequency， weirdness，frequency $z$－score and weirdness $z$－score respectively．Note that the weirdness $z$－score threshold value is again negative；this is expected as the number of proper nouns present in the MPFC is much higher than that of the $\mathrm{TaBE}+\mathrm{LDC}$ ，thus creating a large weirdness standard deviation value．The Text Analysis System selected 130 candidate terms out of the 28,455 vocabulary．Table 4.13 below shows the comparison between the two sets of candidate terms：one set was generated by the system（from the entire corpus）
and the other set was selected by the user（106 from the top 200，as highlighted in Table 4.10 \＆Table 4.11 above）．

Terms in bold are common to both sets；terms underlined（50）are unique to the system generated set，while terms in italic（25）are unique to the user selected set． $76 \%$ ，or 81 of the system generated candidate terms，are present in the user selected set，this shows that our algorithm performs pretty well at generating candidate terms from a corpus，without human intervention．This is important as corpora users can have a general overview of the corpus，without knowing，or knowing little，about the corpus being investigated． Note that the complete absence of numerals，such as the cardinal number 百（hundred），$\mp$ （thousand），億（billion），一（one），二（two），三（three），etc．in the system generated terms，is because they are being replaced with the token＂NOMBER＂during the pre－process．It is common for a single character in Chinese to represent one or more meanings．For example，息 has the meaning of interest，information，rest，and breath；and the contemporary meanings are 利息（interest， $379^{\text {th }}$ ），消息（information， $243^{\text {th }}$ ），休息（rest， $12272^{\text {th }}$ ），and 氣息（breath，not found in MPFC）．To reduce ambiguity，all single character tokens are filtered out by the Text Analysis System when generating candidate terms．

Table 4．13：Comparison between system generated candidate terms and user selected candidate terms，ranked by frequency and grouped ten at a time．

| Rank | System generated Candidate Terms | User selected Candidate Terms |
| :---: | :---: | :---: |
| 1－10 | 公司，市場，百分之，表示，業務，美元，銀行，美國，經渾，報道 | 元，公司，市場，百分之，年，表示，股，業務，美元，銀行 |
| 11－20 | 集圆，投資，日報，星岛，今年，去年，股便 ，预期，盈利，指出 | 美國，經济，報道，集回，周，投資，日報，星島，息，高 |
| 21－30 | 發展，中國，恐為，收購，上市，增長，昨日 ，冠訊，計量，股份 | 報，低，今年，去年，股偩，減，預期，升，盕利，指出， |
| 31－40 | 由於，有關，香港，內地，目前，㭆科，投資者，客戶，收市，上升 | 發展，月，中國，跌，角，嗐為，收購，上市，畦長，昨日 |
| 41－50 | 公布，可能，項目，下跌，不過，镫券，影響 ，場加，主要，基金 | 仙，軍訊，計堚，日，股分，有關，香港，內地，目前，盈科 |
| 51－60 | 雖然，港元，和黄，現時，资金，相信，分析員，股東，部分，未來 | 投资者，客戶，收市，上升，公布，項目，下跌，钲券，影算，增加 |
| 61.70 | 包括，臨羅，资產，指數，出現，其他，消息 ，因為，日圆，政府 | 主要，服務，兌，基金，厘，港元，和黄，资金，相信，分析员 |
| 71－80 | 移動，主席，萓事，進行，日本，港股，交易 ，股模，業紻，國際 | 股東，部分，未來，包括，覃羅，資產，指數，提供，出現，企業 |
| 81－90 | 推出，货款，網絡，情況，出售，人士，紐約 ，明年，持有，方面 | 消息，日回，政府，移動，主席，萓事，進行，日本，市，洪股 |
| 91－100 | 因此，收入，利率，成本，估計，總裁，大幅 ，合作，本港，升幅 | 交易，股栱，國際，業緀，推出，貸款，網絡，情況，出售，人士 |
| 101－110 | 成為，没有，收费，表現，其中，預測，賃格 ，網上，放緩，水平 | 季，紐約，明年，場，通，全球 |
| 111－120 | 旗下，機會，歐洲，股市，希望，至於，現金 ，合併，噱損，減少 |  |
| 121－130 | 亞洲，完成，另外，決定，透要，分別，透過 ，集資，鰦緬，以及 |  |

The most frequent candidate terms that appears in both sets are 公司（company），市場 （market）and 百分之（percent）．What interests us is the term 百分之．The word 百分之 is pervasive；it occurs 6,403 times，making it $0.5 \%$ of the total texts，and was used in the same manner as in English－its most significant collocates are the numbers．We will illustrate this by using collocation and the algorithm discussed in Chapter 3.

We took a neighbourhood of five around the candidate term 百分之，five left neighbours of the term and five right neighbours of the term．By setting the threshold values of spread，strength and peak strength to $10,1,1$（the same setting as the previous experiment），we see from Table 4.14 that the token that occurs most frequently with 百分 $\mathcal{Z}$ in those ten different positions is the word nomber．Nomber itself occurs 83,572 times；the word nomber and 百分之co－occur with each other 10,319 times； $62 \%$ of the times the word nomber occurs it is immediately next to 百分之．Closely followed is the word 升（rise）．升itself occurs 3，032 times and the word 升 and 百分之co－occur with each other 698 times，of which， $77.4 \%$ of the times the word 升 occurs with 百分之－directly to the left of 百分之．For instance，升百分之．This is followed by 95 of its co－occurrence （ $2^{\text {nd }}$ peak）with the token 百分之，but with 1 token in－between（for example，升了百分之，升至百分之）The words 坦長（increase）and 跌（fall）share the same collocation patterns with 升．

Table 4．14：Ten most frequent collocates of 百分之 generated by the Text Analysis System．

|  | $f_{\text {neighbourhood of }}$ 百分之 |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | f | －5 | －4 | －3 | －2 | －1 | 1 | 2 | 3 | 4 | 5 | Total |
| 百分之 | 6403 |  |  |  |  |  |  |  |  |  |  |  |
| nomber | 83572 | 534 | 581 | 481 | 232 | 0 | 6401 | 201 | 439 | 745 | 705 | 10319 |
| 的 | 46849 | 218 | 189 | 107 | 48 | 669 | 0 | 501 | 44 | 89 | 138 | 2003 |
| 至 | 6195 | 12 | 22 | 27 | 29 | 404 | 0 | 320 | 176 | 89 | 40 | 1119 |
| 升 | 3032 | 17 | 14 | 2 | 95 | 540 | 0 | 7 | 10 | 4 | 9 | 698 |
| 百分之 | 6403 | 111 | 95 | 135 | 0 | 0 | 0 | 0 | 134 | 96 | 111 | 682 |
| 較 | 4211 | 85 | 279 | 111 | 5 | 0 | 0 | 1 | 102 | 45 | 26 | 654 |
| 增長 | 2584 | 39 | 42 | 58 | 72 | 209 | 0 | 57 | 64 | 41 | 32 | 614 |
| 報 | 3267 | 49 | 11 | 1 | 12 | 46 | 0 | 1 | 233 | 229 | 23 | 605 |
| 為 | 7416 | 28 | 30 | 15 | 12 | 312 | 0 | 58 | 55 | 39 | 46 | 595 |
| 跌 | 2729 | 27 | 20 | 8 | 79 | 410 | 0 | 1 | 3 | 13 | 22 | 583 |

This can be illustrated more clearly by visualising collocates of 百分之 around its left and right neighbours，as shown in Figure 4.4 below：


Figure 4．4：Left and right neighbours of collocate 百分之．

By further analysing the collocates of 百分之，百分之 indeed relates to the change of stock market－where the profit or loss of share（s），instrument（s），are reported as a percentage of the previous price．Table 4.15 shows the tri－gram collocates of 百分之that occur at least 20 times in MPFC．

Table 4．15：Tri－gram collocates of 百分之，with frequency $>=\mathbf{2 0}$ ．

| Positive | $\boldsymbol{f}_{\text {Coll }}$ | $\boldsymbol{f}_{\text {Raw }}$ | Negative | $\boldsymbol{f}_{\text {Coll }}$ | $\boldsymbol{f}_{\text {Raw }}$ |
| :--- | ---: | ---: | :--- | ---: | ---: |
| 升百分之 nomber | $540(18 \%)$ | 3032 | 跌百分之 nomber | $410(15 \%)$ | 2729 |
| 上升百分之 nomber | $240(12 \%)$ | 1963 | 下跌百分之 nomber | $256(14 \%)$ | 1875 |
| 增長百分之 nomber | $209(8 \%)$ | 2584 | 減少百分之 nomber | $45(5 \%)$ | 880 |
| 增加百分之 nomber | $78(4 \%)$ | 1816 | 下降百分之 nomber | $39(8 \%)$ | 515 |
| 增百分之 nomber | $44(4 \%)$ | 1110 | 低於百分之 nomber | $30(7 \%)$ | 433 |
| 百分之 nomber 开幅 | $91(9 \%)$ | 986 | 跌幅百分之 nomber | $30(6 \%)$ | 593 |
| 百分之 nomber 增幅 | $43(12 \%)$ | 355 | 低百分之 nomber | $27(1 \%)$ | 3251 |
|  |  |  | 挫百分之 nomber | $21(9 \%)$ | 224 |
|  |  |  | 百分之 nomber 跌幅 | $158(18 \%)$ | 593 |

As in English，the collocates can be categorised into two groups：positive and negative． Positive collocates refers to the upward movement of price（instrument），while negative collocates refers to the downward movement．Pertaining to their functions in the finance domain，we will refer them as sentiment words．These sentiment words can either appear
on the left of 百分之一开百分之 nomber（rise percent nomber），增長百分之 nomber （increase percent nomber），下跌百分之 nomber（fall percent nomber），減少百分之 nomber （decrease percent nomber），and so on；or two words on the right of 百分之一百分之nomber升煏（percent nomber increase），百分之 nomber 跌幅（percent nomber decrease），and so on． Percentage values shown in Table 4.15 indicate the confidence level that a sentiment word is used in the sense of reflecting changes in the financial market precisely，rather than other senses．To take the example of $\boldsymbol{廾}$（sheng）；it has many senses as found in the Chinese English Dictionary（CEDICT）：

| 上升 | rise |
| :--- | :--- |
| 晉升 | promote to a higher position |
| 公升 | a liter |
| 升級 | to go up． |

However，in the financial domain，we are $18 \%$ confident that $\neq$ refers to the change in value of an instrument in the positive way，as identified by the collocate 升 百分之 nomber．Other n－gram collocates such as 跌近百分之 nomber（fall nearly percent nomber），升至百分之 nomber（rise to percent nomber），上升超過百分之 nomber（rise more than percent nomber），variants of the tri－gram collocates also exist，but less frequently．

From the collocates of percent generated by the Text Analysis System，a local grammar to extract the positive and negative market sentiment is derived，as shown in Figure 4.5 and Figure 4.6 respectively．


Figure 4.5: A local grammar to extract positive market sentiment in Chinese.


Figure 4.6: A local grammar to extract negative market sentiment in Chinese.

Using the local grammar derived above，market sentiment can be extracted efficiently in Chinese without the need to use a parser，POS tagger or gazetteer．Some of the sentiment bearing sentences extracted by the local grammar are shown below：

| ．昨日一口氣 ，yesterday readily | 上升 |  | 百分之 percent | 三點一，收市報一百一十六元 <br> 3．1，closed at 116 dollar |
| :---: | :---: | :---: | :---: | :---: |
| ．食品貨格也 | 上升 |  | 百分之 | 零點三，高於四月的 |
| ，food price also | rise |  | percent | 0．3，higher than April＇s |
| 出口也 | 上升 |  | 百分之 | ○點五，報八百九十六億五千萬美元 |
| export also | rise |  | percent | 0．5，at 89.65 billion USD |
| 每小時工資 | 上升 | 3 | 百分之 | 四點一。二月份的 |
| Hourly paid | rose |  | percent | 4．1．February |
| 經濟增長再 | 上升 | 至 | 百分之 | 三點四可見有相當的 |
| economic growth again | rose | to | percent | 3．4，shows considerable |
| 點，相等於 | 下跌 |  | 百分之 | 二點七七。上星期恆指仍然在 |
| point，equivalent to | fall |  | percent | 2．77．Last week，Hang Seng Index still |
| 杜粯斯指數於周一 | 下跌 |  | 百分之 | 一點六，收報一－四六二點 |
| on Monday，Dow Jones | fall |  | percent | 1．6，closed at 10462 points |
| 第二季明鄓 | 下跌 |  | 百分之 | 十七．而季内更因為 |
| Second season clearly | fall |  | percent | 17，within |
| ．較上星期 | 下跌 |  | 百分之 | 二，有分析員表示盛科 |
| ，compared with last week | fall |  | percent | 2，and analyst indicated PCCW |
| 納斯達克綜合指數更 | 下跌 | 3 | 百分之 | 三點四一，收報二千四百七十點五二 |
| Nasdaq Composite Index even | fell |  | percent | 3．41，closed at 2470.52 |
| 失業率仍 | 下跌 | 至 | 百分之 | 三點九。投資者已預期 |
| unemployment rate still | fall | to | percent | 3．9．Investors already expected |

So far，we have discussed two experiments－sentiment analysis in the domain of Finance written in two typologically different languages，English and Chinese．Our approach handled multi－lingual texts without any alternation to the algorithm．The local grammars constructed from the collocates generated by the Text Analysis System in the financial domain can be used to track changes in financial instruments precisely，and there are similarities between the usage of certain sentiment words．For example，in English， changes in values of the financial instruments are reported as：

| sentiment word | nomber | percent |
| :--- | :--- | :--- |
| rose | 2.83 | percent |

and in Chinese，they are reported as：

| 上升 | 百分之 | 三點一 |
| :--- | :--- | :--- |
| rose | percent | 3.1 |

We can see that the position of nomber and percent are interchanged while the position of sentiment word remains the same．Work by Ahmad et al．（2006）suggests that sentiment words in Arabic financial news texts operate in the same fashion，and Figure 4.7 below shows the local grammar used to extract market sentiment in Arabic．


Figure 4．7：A local grammar to extract market sentiment in Arabic．

Recall Smadja＇s observation that＂collocations do not translate well across languages＂，in our case，collocations translate well across languages in the financial domain，at least in three languages－English，Chinese and Arabic．

## 4．1．3 Sentiment Analysis for Film Reviews

The previous two experiments，sentiment analysis in the domain of Finance written in two typologically different languages，English and Chinese，show that the algorithm is able to cope with multi－lingual texts in the same domain．Apart from being employed to detect changes in the financial market，sentiment analysis has also been used to assess opinion about film，holiday resorts，cars，and consumer electronics．

We would like to investigate whether our algorithm will also work for domains other than the financial domain．In this experiment，we chose to work with movie reviews polarity dataset vl．1，which was made available by Pang，Lee and Vaithyanathan（2002）．The dataset contains 1386 movie reviews collected from the Internet Movie Database（IMDB） archive of the rec．arts．movies．reviews newsgroup，hereafter referred to as the IMDB corpus．The IMDB corpus is divided into two sets，a testing set containing 200 randomly selected reviews which will be used for evaluation purposes，and a training set comprising the remaining 1186 reviews（ 791,879 tokens in all）．

Table 4.16: Distribution of the first 100 most frequent tokens in the IMDB training corpus according to the cumulative frequency of ten tokens at a time.

| Rank | Token | Cumulative <br> Relative <br> Frequency | No. of <br> Open Class <br> Words |
| :--- | :--- | ---: | ---: |
| $1-10$ | the, a, and, of, to, is, in, 's, it, that | $22.77 \%$ | 0 |
| $11-20$ | with, as, for, film, his, this, he, but, by, on | $6.81 \%$ | 1 |
| $21-30$ | i, are, be, movie, who, at, an, not, from, has | $4.31 \%$ | 1 |
| $31-40$ | have, was, her, you, all, they, like, out, there, about | $3.09 \%$ | 1 |
| $41-50$ | so, more, up, or, she, which, what, when, some, their | $2.40 \%$ | 0 |
| $51-60$ | just, if, time, even, only, him, into, than, we, no | $1.88 \%$ | 1 |
| $61-70$ | its, good, most, can, will, story, been, much, also, other | $1.60 \%$ | 2 |
| $71-80$ | would, first, get, them, do, characters, character, very, well, see | $1.38 \%$ | 4 |
| $81-90$ | after, because, way, make, how, does, off, any, too, people | $1.17 \%$ | 2 |
| $91-100$ | life, films, really, plot, little, where, while, director, had, over | $1.09 \%$ | 4 |
|  |  |  | $46.50 \%$ |

The cumulative frequency of the first 100 most frequent tokens in the IMDB training corpus is $46.50 \%$, slightly higher than that of the BNC ( $45 \%$ ). Unlike in the domain of Finance, we see a much lower number of open-class tokens, 16 only, versus 45 in the Reuters corpus and 47 in MBFC. However, these 16 possible open-class tokens indeed highlight the characteristics of the IMDB training corpus: the $14^{\text {th }}$ most frequent token is film and its plural films $\left(92^{\text {th }}\right)$, followed by its synonym movie $\left(24^{\text {th }}\right.$ ); the tokens story $\left(66^{\text {th }}\right)$, characters $\left(76^{\text {th }}\right)$, character $\left(77^{\text {th }}\right)$ and plot $\left(94^{\text {th }}\right)$ tell us something about the film, and the token director $\left(98^{\text {th }}\right.$ ) tells us who directs the making (make, $84^{\text {th }}$ ) of the film. The token good $\left(62^{\text {th }}\right)$ may indicate the reviewer's opinion towards a film.

Compared with the Finance corpora, another notable point about the 100 most frequent tokens in this corpus is the significant appearance of 14 pronouns: it $\left(9^{\text {th }}\right)$ and its ( $61^{\text {th }}$ ); his $\left(15^{\text {th }}\right)$, he $\left(17^{\text {th }}\right)$ and $\operatorname{him}\left(56^{\text {th }}\right)$; $I\left(21^{\text {th }}\right)$, you $\left(34^{\text {th }}\right)$ and we $\left(59^{\text {th }}\right)$;her $\left(33^{\text {th }}\right)$ and she $\left(45^{\text {th }}\right)$; they $\left(36^{\text {th }}\right)$, their $\left(50^{\text {th }}\right)$, other $\left(70^{\text {th }}\right)$ and them $\left(74^{\text {th }}\right)$ are all amongst the 100 high frequency tokens. This is similar to the BNC, where 18 pronouns appeared within the 100 most frequent tokens.

Table 4.17: Distribution of tokens ranked between 101-200 in the IMDB training corpus, grouped ten at a time.

| Rank | Token | Cumulative <br> Relative <br> Frequency | No. of <br> Open Class <br> Words |
| :--- | :--- | ---: | ---: |
| $101-110$ | then, man, me, scene, never, movies, bad, these, could, my | $0.98 \%$ | 4 |
| $111-120$ | scenes, new, best, such, being, know, many, great, here, through | $0.90 \%$ | 4 |
| $121-130$ | were, us, love, made, action, another, john, now, big, back | $0.82 \%$ | 5 |
| $131-140$ | something, before, still, minutes, those, cast, go, end, makes, seems | $0.75 \%$ | 6 |
| $141-150$ | however, few, work, com, things, every, real, audience, better, <br> around | $0.69 \%$ | 6 |
| $151-160$ | down, role, same, since, actually, though, world, both, between, <br> going | $0.66 \%$ | 6 |
| $161-170$ | year, gets, long, years, may, think, last, old, should, take | $0.64 \%$ | 8 |
| $171-180$ | performance, comedy, enough, look, your, own, seen, fact, say, <br> why | $0.60 \%$ | 6 |
| $181-190$ | although, thing, come, funny, http, plays, right, almost, find, ever | $0.58 \%$ | $\mathbf{6}$ |
| $191-200$ | comes, played, directed, young, nothing, did, actors, course, part, <br> takes | $0.55 \%$ | 8 |
|  |  | TOTAL | $\mathbf{7 . 1 6 \%}$ |

The comparison of the frequency distribution of the first 100 tokens in the BNC and in IMDB training corpus helps us in identifying certain keywords used by the movie reviewers. The 16 possible open class tokens are also potential keywords. The second most frequent 100 words, though comprising much smaller proportions of texts in both the BNC (c. $6 \%$, Table 4.5) and our corpus (c. 7\%, Table 4.17), show the preponderance of additional keywords, and especially for us, more of the adjectives that might have been used to express the reviewer's opinion. In particular, we note the frequent use of bad $\left(107^{\text {th }}\right)$, best $\left(113^{\text {th }}\right)$, great $\left(118^{\text {th }}\right)$, better $\left(149^{\text {th }}\right)$ and funny $\left(184^{\text {th }}\right)$. If we add up all the inflections of good, we obtain a frequency of $0.33 \%$ in our corpus.

Consider the most frequent 200 words in our corpus of text (Table 4.16 \& Table 4.17): together they compile $53.66 \%$ of the total number of tokens in the texts. We were interested in seeing how half of the texts, or rather the tokens in half of the texts, cooccur with each other. Note that we have replaced all the words like don't, doesn't, can't, wasn't, which were used to express disagreement with "NEG", and all the numerals with "NOMBER".

Like the two previous experiments, we set the threshold values as $5,2,2,-0.2$ for frequency, weirdness, frequency $z$-score and weirdness $z$-score respectively. Note that
the weirdness $z$-score threshold value is again negative; this is expected as the number of proper nouns (movie titles, director names, actor names) present in the IMDB training corpus is much higher than that of the BNC, thus creating a large weirdness standard deviation value. The Text Analysis System selected 96 candidate terms out of the 32,654 word vocabulary. Table 4.18 below shows the comparison between the two sets of candidate terms: one set was generated by the system (from the entire corpus) and the other set was selected by the user ( 76 from the top 200, as highlighted in Table 4.16 \& Table 4.17). Terms in bold are common to both sets; terms underlined (56) are unique to the system generated set, while terms in italic (36) are unique to the user selected set. $52 \%$, or 40 of the system generated candidate terms are present in the user selected set, this shows that our algorithm performs quite well at generating candidate terms from a corpus, without human intervention.

Table 4.18: Comparison between system generated candidate terms and user selected candidate terms, ranked by frequency and grouped ten at a time.

| Rank | System generated Candidate Terms | User selected Candidate Terms |
| :---: | :---: | :---: |
| 1-10 | film, movie, even, good, story, characters, character, life, films, really | film, movie, like, time, good, story, first, characters, character, well |
| 11-20 | plot, little, director, scene, movies, bad, scenes, best, love, action | make, people, life, films, plot, director, man, scene, movies, bad |
| 21-30 | john, big, minutes, cast, makes, seems, few, com, real, audience | scenes, best, know, great, love, made, action, john, big, minutes |
| 31-40 | role, actually, gets, performance, comedy, funny, http, plays, ever, comes | cast, go, end, makes, seems, work, com, things, real, audience |
| 41-50 | played, directed, actors, course, takes, lot, script, screen, original, least | better, down, role, actually, though, world, going, year, gets, long |
| 51-60 | written, interesting, star, acting, rather, michael, picture, david, guy, american | years, think, last, old, take, performance, comedy, enough, seen, fact |
| 61-70 | pretty, fun, effects, www, seem, wife, starring, series, reviews, bit | say, thing, come, funny, http, plays, right, find, comes, played |
| 71-80 | instead, stars, based, screenplay, everything, watch, hollywood, trying, sex, music | directed, young, nothing, actors, course, takes |
| 81-90 | review, james, watching, running, dialogue, violence, rated, actor, becomes, everyone |  |
| 91-100 | gives, soon, moments, wants, turns, girl |  |

The coverage of system generated candidate terms may seems low when comparing with user selected ones. However, the system managed to pick up terms such as script, star, acting, fun, starring, stars, screenplay, American, Hollywood, which are characteristic of the IMDB corpus.

The most frequent candidate terms that appear within rank 1 to rank 10 in both sets are film, movie, good, story, characters and character. Since the whole IMDB corpus is about reviewers' opinions about film(s) and movie(s), we would like to use collocation and the algorithm discussed in Chapter 3 to examine how such opinions are expressed.

We took a neighbourhood of five around the candidate terms film and movie, which means we looked at the five left neighbours of the term and five right neighbours of the term. By setting the threshold values of spread, strength and peak strength to $10,0.5$, 0.5 , we see from Table 4.19 that the token occurring most frequently with film in those ten different positions is the word 's. However, what really interests us is the word $N E G$ and good. NEG itself occurs 3,650 times, the word $N E G$ and film co-occur with each other 246 times, of which, $24.4 \%$ of the time the word $N E G$ occurs next to film. For instances, film isn't, film doesn't, film wasn't. Good itself occurs 1,347 times and the word good and film co-occur with each other 119 times, of which, $29.4 \%$ of the time the word good occurs next to film.

Table 4.19: Ten most frequent collocates of film generated by the Text Analysis System.

| f - $\quad$ - $f_{\text {neighbourhood of film }}$ |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| film | 5506 |  |  |  |  |  |  |  |  |  |  |  |
| 's | 10866 | 68 | 74 | 35 | 60 | 68 | 348 | 10 | 70 | 53 | 51 | 837 |
| nomber | 10624 | 61 | 97 | 72 | 23 | 25 | 9 | 87 | 59 | 65 | 70 | 568 |
| but | 5052 | 28 | 21 | 22 | 44 | 0 | 11 | 79 | 22 | 23 | 40 | 290 |
| NEG | 3650 | 23 | 28 | 37 | 4 | 0 | 60 | 25 | 27 | 21 | 21 | 246 |
| about | 2106 | 16 | 19 | 8 | 61 | 0 | 43 | 7 | 12 | 16 | 15 | 197 |
| most | 1346 | 32 | 16 | 38 | 24 | 1 | 0 | 18 | 7 | 10 | 11 | 157 |
| first | 1150 | 15 | 14 | 8 | 13 | 70 | 1 | 8 | 12 | 4 | 5 | 150 |
| more | 2016 | 8 | 17 | 8 | 7 | 0 | 10 | 36 | 24 | 17 | 12 | 139 |
| good | 1347 | 7 | 16 | 8 | 6 | 35 | 1 | 16 | 11 | 10 | 9 | 119 |
| than | 1430 | 13 | 18 | 7 | 12 | 1 | 15 | 7 | 15 | 13 | 13 | 114 |

Movie, synonym of film, shares a similar collocation pattern, as shown in Table 4.20. The word NEG and movie co-occur with each other 173 times, of which, $23.1 \%$ of the time the word $N E G$ occurs next to film. Good itself occurs 1,347 times and the word good and film co-occur with each other 94 times, of which, $34 \%$ of the time the word good occurs next to film.

Table 4.20: Ten most frequent collocates of movie generated by the Text Analysis System.

|  | $f_{\text {neighbourhood of movie }}$ |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | f | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | Total |
| movie | 3510 |  |  |  |  |  |  |  |  |  |  |  |
| 's | 10866 | 47 | 33 | 22 | 16 | 30 | 144 | 19 | 48 | 35 | 29 | 423 |
| nomber | 10624 | 44 | 51 | 43 | 17 | 12 | 13 | 47 | 58 | 52 | 67 | 404 |
| NEG | 3650 | 19 | 12 | 26 | 2 | 0 | 40 | 13 | 20 | 21 | 20 | 173 |
| but | 5052 | 13 | 15 | 13 | 24 | 0 | 8 | 46 | 12 | 15 | 21 | 167 |
| about | 2106 | 10 | 7 | 6 | 36 | 0 | 43 | 22 | 12 | 16 | 12 | 164 |
| more | 2016 | 6 | 14 | 24 | 12 | 4 | 5 | 16 | 14 | 13 | 15 | 123 |
| good | 1347 | 5 | 14 | 6 | 9 | 32 | 0 | 2 | 6 | 8 | 12 | 94 |
| out | 2182 | 12 | 5 | 22 | 3 | 0 | 10 | 9 | 9 | 6 | 8 | 84 |
| just | 1632 | 10 | 8 | 8 | 2 | 0 | 12 | 16 | 11 | 11 | 5 | 83 |
| reviews | 325 | 0 | 0 | 1 | 28 | 0 | 47 | 3 | 0 | , | 0 | 79 |

Other less frequent collocates of film and movie that also indicate reviewer's opinions include bad, best, great, and so on, and they share the same collocation pattern with good. This can be illustrated more clearly by visualising the collocates of film around its left and right neighbours, as shown in Figure 4.8 below:


Figure 4.8: Left and right neighbours of collocate film.

As our algorithm is based on statistics, a larger corpus is preferred in order to produce statistically significant results. In this case, as the IMDB training corpus only has 0.79 million tokens, we need to analysis collocates of the all candidate terms generated by the Text Analysis System. Table 4.21 shows the $n$-gram collocates of the candidate terms that occur more than 15 times in the IMDB training corpus.

Table 4.21: $\mathbf{N}$-gram collocates expressing reviewer's sentiments, with frequency $\mathbf{> 1 6}$.

| Positive | $f_{\text {coll }}$ | Negative | $f_{\text {coll }}$ |
| :--- | ---: | :--- | ---: |
| even more | 100 | just $N E G$ | 102 |
| better than | 90 | NEG really | 81 |
| much better | 50 | $N E G$ seem | 75 |
| big screen | 47 | film NEG | 60 |
| good film | 35 | even worse | 45 |
| most interesting | 34 | movie $N E G$ | 40 |
| pretty good | 34 | too bad | 38 |
| good movie | 32 | worse than | 36 |
| more interesting | 30 | bad movie | 33 |
| good time | 29 | but i $N E G$ | 25 |
| even better | 27 | pretty bad | 24 |
| good job | 27 | still $N E G$ | 23 |
| fun to watch | 26 | bad movies | 19 |
| best thing | 19 | NEG funny | 19 |
| pretty well | 19 | any sense | 18 |
| worth watching | 19 | i $N E G$ even | 18 |
| best performance | 18 | better off | 17 |
| good idea | 18 | bad things | 16 |
| good thing | 18 | even less | 16 |
| great flick | 18 | hard time | 16 |
| well done | 18 |  |  |
| extremely well | 17 |  |  |
| okay movie | 17 |  |  |
| best films | 16 |  |  |
| great film |  |  |  |
| interesting character |  |  |  |

The collocates can be categorised into two groups: positive and negative. Positive collocates refers to the reviewer's agreement that a film is good, while negative collocates refers to a disagreement. Pertaining to their functions, we will refer to them as sentiment words, a majority of which are adjectives. These sentiment words can either appear directly to the left of the collocate - bad movie, bad film, good film, fun to watch, etc.; or directly on the right - too bad, pretty bad, pretty good, extremely well, and so on.

Using the collocates generated by the Text Analysis System, the reviewer's opinions can be extracted without the need to use parser, POS tagger or gazetteer. Below shows some of the sentiment bearing sentences extracted:

> avoid flaws : he has created a force behind one of the year's is considered one of the
> you can always count on a good direction by cameron, a but i found her more leonard and tarantino . perhaps the
> the cinematography is very hal holbrook and ed anser by telling them that they and pretentious . the acting is
> see it if you like to watch opaque . the whole experience story and those that do don't
> big dance number . this
> successful horror films ever is

### 4.2 Evaluation

### 4.2.1 The case for English financial news

The use of other statistical measures of the quantitative changes in the value of instrument(s) are important for us as we try to attempt to incorporate the use of sentiment indicators - or rather changes in sentiment - in an overall financial analysis framework. One such attempt has involved in helping the traders to correlate the quantitative signal, either in its raw form or the derived forms (return and volatility measures), with the movement of sentiment indicating phrases. In this section, we will use the local grammars extracted from the Reuters corpus (see Figure 4.2 and Figure 4.3 for details) to identify the market sentiments. We will create a time series of positive and negative sentiments separately and for each of the sentiments we have two time series: one positive (negative) time series based on the count of a predefined list of words relating to positive (negative) sentiments and the other where we only count a word as a sentiment word if it occurs in a local grammar pattern. We have the four time series: two 'raw' sentiment counts and two based on the filter provided by local grammar rules identified from a text corpus. These sentiment series are then correlated with the time series of various aggregates of share price values within two stock exchanges: the Financial Times Stock Exchange aggregates (FTSE 100 best companies aggregates and FTSE All share
index) for the London stock market; and the Dow Jones Industrial Average for the New York markets together with the National Association of Securities Dealers Automated Quotations system (NASDAQ), as well as the Standards and Poor ratings for US corporations (S\&P 500).

We have selected 163,879 news articles ( 64.03 million tokens) covering a period of 1 month from the Reuters Newsfeed, and details of the corpus is shown in Table 4.22 below:

Table 4.22: Distribution of news articles between 01 August and 31 August in the Reuters Newsfeed.

| Week | No. of News | Tokens |
| :---: | ---: | ---: |
| 01-04 Aug 2006 | 33,537 | 14.02 M |
| 07-11 Aug 2006 | 39,677 | 16.20 M |
| 14-18 Aug 2006 | 35,445 | 13.66 M |
| 21-25 Aug 2006 | 32,961 | 11.93 M |
| 28-31 Aug 2006 | 22,259 | 8.21 M |

The Reuters Newsfeed corpus is divided into five weekly intervals - with the weekends excluded as the stock markets were closed during such periods. With our algorithms, empirically determined positive and negative sentiments from each news article are extracted and ordered according to the timestamp provided by Reuters. The aggregated sentiments are then represented as a (quantitative) time series that were more familiar to the traders, and correlated with the financial indices. Note that the long term trend for all the financial indices are uptrend, meaning the closing value on the 31 August is higher than the opening value on 01 August. Table 4.23 below shows the correlation between five financial indices and all the news articles from Reuters Newsfeed in August 2006 (see Appendix E and Appendix F for details of the time series):

Table 4.23: Correlation between financial instruments and all the news articles from Reuters Newsfeed in August 2006 using RAW data.

|  | Time Series |  |  |  |
| :--- | ---: | ---: | ---: | ---: |
| Index | Raw +ve | Raw -ve | LG +ve | LG-ve |
| FTSE 100 | $-19 \%$ | $-39 \%$ | $8 \%$ | $-42 \%$ |
| FTSE ALL | $-29 \%$ | $-42 \%$ | $-1 \%$ | $-42 \%$ |
| Dow Jones | $-59 \%$ | $-68 \%$ | $-31 \%$ | $-35 \%$ |
| Nasdaq | $-57 \%$ | $-65 \%$ | $-30 \%$ | $-35 \%$ |
| S\&P 500 | $-56 \%$ | $-66 \%$ | $-29 \%$ | $-38 \%$ |

The Raw sentiment time series represent the sentiment generated using single words such as $u p$, down, rose, fell; whilst the $L G$ sentiments were identified using the local grammar patterns discussed earlier (Section 4.1.1). From Table 4.23, it seems that the financial indices and the news articles are highly anti-correlated to a certain extent, and the local grammar based time series produced a stronger correlation, as indicated by the FTSE 100 index.

We have computed simple correlations between raw numbers (time series) and the results are all anti-correlated (negative), except the correlation between FTSE 100 and LG +ve. In econometrics and financial studies instead of using a time series per se, researchers use, with the so-called return values; and in the literature on financial engineering, volatility is used. Volatility is defined as $^{2}$ :

$$
\begin{equation*}
\text { volatility }=s t d\left(\log \left(\frac{t_{Q}}{t-1_{Q}}\right)\right) \tag{14}
\end{equation*}
$$

where
$\boldsymbol{t}_{\boldsymbol{Q}}$ represents the financial instrument or sentiment value at time $\boldsymbol{t}$; $\boldsymbol{t}-1_{Q}$ represents the financial instrument or sentiment value at time $\boldsymbol{t}-1 ;$ and ( $\boldsymbol{t}_{\boldsymbol{Q}} / \boldsymbol{t}-\mathbf{1}_{Q}$ ) represents the Return value.

Table 4.24 below shows the correlations using the Return and Volatility values. It can be seen that the local grammar based time series project a much more accurate correlations using the volatility values. For instances, the FTSE indices correlate with the negative sentiments from the news articles very well $-84 \%$ of the time.

[^7]Table 4.24: Correlations between financial instruments and all the news articles from Reuters Newsfeed in August 2006 using Return and Volatility data.

|  | Time Series (Return) |  |  |  | Time Series (Volatility) |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Index | $\begin{gathered} \text { Raw } \\ \text { +ve } \end{gathered}$ | $\begin{aligned} & \text { Raw } \end{aligned}$ | $\begin{gathered} \text { LG } \\ +\mathbf{v e} \end{gathered}$ | $\begin{gathered} \mathrm{LG} \\ \text {-ve } \\ \hline \end{gathered}$ | $\begin{gathered} \text { Raw } \\ +\mathbf{v e} \end{gathered}$ | $\begin{gathered} \text { Raw } \\ \text {-ve } \end{gathered}$ | $\begin{aligned} & \text { LG } \\ & +\mathbf{v e} \end{aligned}$ | $\begin{gathered} \text { LG } \\ \text {-ve } \end{gathered}$ |
| FTSE 100 | -12\% | -34\% | -1\% | -58\% | -10\% | 30\% | -3\% | 84\% |
| FTSE ALL | -13\% | -36\% | -1\% | -59\% | -9\% | 31\% | -1\% | 84\% |
| Dow Jones | -5\% | -6\% | 13\% | 9\% | 25\% | -3\% | 31\% | -45\% |
| Nasdaq | 3\% | 1\% | 20\% | 9\% | -16\% | -17\% | -85\% | -68\% |
| S\&P 500 | -1\% | -4\% | 22\% | 13\% | -50\% | -50\% | 81\% | 89\% |

News articles from the Reuters Newsfeed in August 2006 are essentially a collection of news articles published by reporters/journalists around the world. We would like to investigate if there are any differences between the correlations generated using only country specific news articles, and using all news articles. Table 4.25 below shows the distribution of news articles for the U.S. and U.K. in August 2006.

Table 4.25: Distribution of the U.S. and U.K. news articles in Reuters Newsfeed between 01 August and 31 August.

|  | U.S. |  | U.K. |  |
| :---: | ---: | ---: | ---: | ---: |
| Week | No. of News | Tokens | No. of News | Tokens |
| 01-04 Aug 2006 | 16,027 | 6.60 M | 5,276 | 2.30 M |
| 07-11 Aug 2006 | 19,106 | 7.89 M | 5,731 | 2.21 M |
| 14-18 Aug 2006 | 16,887 | 6.72 M | 5,231 | 1.88 M |
| 21-25 Aug 2006 | 14,359 | 5.66 M | 5,400 | 2.02 M |
| 28-31 Aug 2006 | 9,226 | 3.58 M | 3,544 | 1.58 M |

We can see that the U.S. news articles alone account for nearly half (46\%) of the news articles published within August 2006, whilst only $16 \%$ for the U.K. news articles. Similarly, the correlations betweens the financial indices and the sentiments from country specific news articles are shown in Table 4.26 below:

Table 4.26: Correlations between financial instruments and country specific news articles from Reuters Newsfeed in August 2006.

| Index | Raw +ve | Raw -ve | LG +ve | LG -ve |
| :--- | ---: | ---: | ---: | ---: |
| FTSE 100 | $5 \%$ | $-41 \%$ | $25 \%$ | $-43 \%$ |
| FTSE ALL | $1 \%$ | $-39 \%$ | $19 \%$ | $-43 \%$ |
| Dow Jones | $-53 \%$ | $-38 \%$ | $-34 \%$ | $-22 \%$ |
| Nasdaq | $-53 \%$ | $-34 \%$ | $-35 \%$ | $-19 \%$ |
| S\&P 500 | $-55 \%$ | $-36 \%$ | $-36 \%$ | $-20 \%$ |

Compared with the results from Table 4.23, using country specific news articles indeed improve the correlation, especially for the U.K. financial market where both of the local grammar based time series show a stronger correlation. Whilst for the U.S. market, using country specific gave a much weaker correlation for some reasons. This is certainly an interesting result worth investigating further.

Similarly, an even better correlation was produced using the Return and Volatility values when compared with the results using all news articles. For instance, the local grammar based sentiment time series, both positive and negative ones, are correlated with the Dow Jones Industrial Average Index - $22 \%$ with the positive news and $60 \%$ with the negative news. On the other hand, the positive and negative sentiment time series from the news are anti-correlated with the movement of the FTSE financial indices - $88 \%$ of the time with the positive news; and $67 \%$ of the time with the negative news. This is shown clearly in Table 4.27 below:

Table 4.27: Correlations between financial instruments and country specific news articles from Reuters Newsfeed in August 2006 using Return and Volatility data.

|  | Time Series (Return) |  |  |  | Time Series (Volatility) |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Index | $\begin{gathered} \text { Raw } \\ \text { +ve } \end{gathered}$ | $\begin{aligned} & \text { Raw } \\ & \text {-ve } \end{aligned}$ | $\begin{gathered} \hline \mathbf{L G} \\ +\mathbf{v e} \end{gathered}$ | $\begin{gathered} \underline{\mathbf{L G}} \\ -\mathbf{v e} \end{gathered}$ | $\begin{gathered} \text { Raw } \\ +\mathbf{e v e} \end{gathered}$ | $\begin{gathered} \text { Raw } \\ \text {-ve } \end{gathered}$ | $\begin{gathered} \mathbf{L G} \\ +\mathbf{v e} \end{gathered}$ | $\begin{gathered} \text { LG } \\ -\mathbf{v e} \end{gathered}$ |
| FTSE 100 | 0\% | -52\% | 17\% | -56\% | -24\% | 93\% | -89\% | 67\% |
| FTSE ALL | 1\% | -51\% | 17\% | -58\% | -23\% | 93\% | -88\% | 68\% |
| Dow Jones | -2\% | 0\% | 6\% | 4\% | 19\% | -25\% | 22\% | -60\% |
| Nasdaq | -2\% | -7\% | 3\% | -3\% | -75\% | -79\% | -13\% | -51\% |
| S\&P 500 | -2\% | -5\% | 12\% | 1\% | 75\% | 75\% | -10\% | 58\% |

The correlation between the sentiment need to be interpreted very carefully especially since we have 'compressed' the time series: we have aggregated the sentiment values
over a day and then correlated with the value of share indices that day．The analysis of the movement of share indices is an art in itself and hence we will not attempt to draw any further conclusions．Suffice it to say that one when one sentiment time series（either positive or negative）correlates with a stock exchange index，then the other shows a weaker correlation or in many cases anti－correlates．

## 4．2．2 The case for Chinese financial news

In this section，we will use the local grammars extracted from the MPFC（see Figure 4.5 and Figure 4.6 for details）to identify the market sentiments and the correlations between various financial instruments from China，such as the Shanghai Index，the Shenzhen Index and the Hang Seng Index．However，a slight change to the local grammars is required as the symbol \％is preferentially used than the actual word 百分之（percent）．For example，the local grammar：

| sentiment word | percent | nomber |
| :--- | :--- | :--- |
| 上升 | 百分之 | 三點一 |
| rose | percent | 3.1 |

is modified to：

| sentiment word | nomber | Percent |
| :--- | :--- | :--- |
| 上升 | 三點－ | $\%$ |
| rose | 3.1 | $\%$ |

We have selected 597 Chinese financial news articles（293，933 tokens，segmented using the Standford Chinese Word Segmenter ${ }^{3}$ ）published between $02-27$ July 2007 from the Xinhuanet ${ }^{4}$ ，and details of the corpus is shown in Table 4.28 below：

Table 4．28：Distribution of news articles published by Xinhuanet between 02 July and 27 July 2007.

| Week | No．of News | Tokens |
| :---: | :---: | :---: |
| 02－06 July 2007 | 154 | 71,052 |
| 09－13 July 2007 | 178 | 99,100 |
| 16－20 July 2007 | 153 | 69,739 |
| 23－27 July 2007 | 112 | 54,042 |

[^8]Similar to the English financial news, the Xinhuanet corpus is divided into four weekly interval - with the weekends excluded as the stock markets were closed during such periods. Table 4.29 below shows the correlations between three financial indices and the news articles from Xinhuanet in July 2007 (see Appendix G for details of the time series):

Table 4.29: Correlations between financial instruments and news articles from Xinhuanet within July 2007.

| Index | Raw +ve | Raw -ve | LG +ve | LG-ve |
| :--- | ---: | ---: | ---: | ---: |
| Hang Seng | $26 \%$ | $2 \%$ | $-13 \%$ | $-63 \%$ |
| Shanghai | $31 \%$ | $-4 \%$ | $-11 \%$ | $-43 \%$ |
| Shenzhen | $54 \%$ | $-39 \%$ | $52 \%$ | $-12 \%$ |

From the above table, it seems that the financial indices and the news articles are highly anti-correlated, and the local grammar based time series produced a stronger correlation, as indicated by the Hang Seng and Shanghai indices. Again, if we use the return and volatility values, a better correlation is achieved, which is illustrated in Table 4.30 below:

Table 4.30: Correlations between financial instruments and news articles from Xinhuanet in July 2007 using Return and Volatility data.

|  | Time Series (Return) |  |  |  | Time Series (Volatility) |  |  |  |
| :--- | ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Index | Raw <br> +ve | Raw <br> -ve | LG <br> +ve | LG <br> -ve | Raw <br> +ve | Raw <br> -ve | LG <br> +ve | LG <br> -ve |
| Hang Seng | $26 \%$ | $2 \%$ | $-13 \%$ | $-63 \%$ | $54 \%$ | $-39 \%$ | $52 \%$ | $-12 \%$ |
| Shanghai | $13 \%$ | $5 \%$ | $20 \%$ | $-32 \%$ | $-25 \%$ | $-16 \%$ | $-50 \%$ | $62 \%$ |
| Shenzhen | $19 \%$ | $1 \%$ | $19 \%$ | $-35 \%$ | $-21 \%$ | $-20 \%$ | $-46 \%$ | $64 \%$ |

Note that the local grammar based sentiments correlate with the financial indices much better than the raw (single word) ones. However, both the Shanghai and Shenzhen financial indices exhibit an anti-correlation with the sentiment time series.

### 4.2.3 The case for film reviews

In this section, we will use the collocation patterns generated by the Text Analysis System from the IMDB training corpus to detect sentiments in some unseen film reviews

- the IMDB testing set. Reviews in the IMDB corpus have already been classified as positive and negative, based on the number of stars given by the reviewer. For example, with a five-star rating system, 4 stars and above are considered positive, while 2 stars and below are considered negative. For details of the classification, please refer to Pang, Lee and Vaithyanathan (2002). The IMDB testing set comprises 200 film reviews randomly selected from the polarity dataset v1.1: 100 from the positive reviews, and 100 from the negative reviews.

Table 4.31: Evaluation result for detecting sentiments in film reviews.

|  |  |  |  |  |  |  | F- |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | actual | correct | mrong | missing | precision | recall | Measure |
| Negative | 100 | 64 | 18 | 18 | $64.00 \%$ | $64.00 \%$ | $64.00 \%$ |
| Positive | 100 | 51 | 22 | 27 | $51.00 \%$ | $51.00 \%$ | $51.00 \%$ |
|  |  |  |  |  | $57.50 \%$ | $57.50 \%$ | $57.50 \%$ |

Using the local grammar constructed from the IMDB training corpus, an average of $57.5 \%$ precision and recall is achieved. Turney's approach reported an accuracy of 65.83\% on a 120 -review: 60 reviews for the movie The Matrix; 60 reviews for the movie Pearl Harbour. However, the IMDB testing set comprises a variety of film reviews, instead of two; furthermore, we noticed that the local grammar constructed did not identify any sentiments in certain reviews: 8 in the negative reviews, 13 in the positive reviews. If we ignore these reviews, we achieved an average of $64.09 \%$, which is comparable to Turney's results.

Table 4.32: Evaluation result for detecting sentiments in film reviews, after removing neutral reviews.

|  |  |  |  |  |  |  |  | F- <br>  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | actual | correct | wrong | missing | precision | recall | Measure |  |

When comparing our result with that of Pang et al. (2002) who used the same dataset, our results are better than the two graduate students who achieved $58 \%$ and $64 \%$ respectively; and with a much lower tie rates (where two sentiments are rated equally likely): $10 \%$ and $14 \%$ as opposed to $75 \%$ and $39 \%$. However, their hybrid machine learning approach has accuracy in the range of $73 \%$ and $83 \%$. One point to be noted in
their approach is that a significant amount of time is required to train the system, and their focus is on document classification rather than extracting sentiment bearing patterns.

Furthermore, our approach is not specific to film reviews, requires minimum human intervention, does not depend on POS tagging, and can be applied to other domains as long as sufficient training data exists. This has already been demonstrated in the finance domain (see Sections 4.2.1 \& 4.2.2). Comparing our approach to Turney's, his approach may not work, or enormous human effort is required for the identification of adjective and adverbs phrases, when switching to domains where POS tagging is not available; whereas, our approach is still valid, as shown in the case studies earlier.

### 4.3 System Description

The Text Analysis System has been developed using the Java programming language, third party Java libraries for processing documents in $X M L, H T M L, P D F, D O C, R T F$, XLS , PPT, and TXT format; as well as vljtable for representing the results in a style akin to a Microsoft Excel spreadsheet; and jfreechart for visualising the results graphically.


Figure 4.9. User interface of the Text Analysis System.

The nine buttons of the Text Analysis System (as shown in Figure 4.9 above) are designated for specific tasks:

1. Manage Corpus: Adds/Removes files to/from the Corpus
2. Save Results: Saves the Collocation results
3. Help: Provides Help about using the Text Analysis System
4. Wordlist: Produces a wordlist of the Corpus
5. KWIC: Produces concordance of the candidate terms
6. Collocation: Produces (bi-gram) collocates of the candidate terms
7. re-Collocate: Produces (n-gram) collocates of the candidate terms
8. Sentiment: Performs sentiment analysis of the corpus.
9. Options: Configure various parameters for the Text Analysis System

The three main components in the Text Analysis System are Wordlist, KWIC and Collocation (initiated by clicking buttons 4,5 and 6). The Wordlist component will produce a wordlist of a corpus, and select candidate terms from the corpus based on statistical values. A concordance for each candidate term is then produced by the KWIC component, and the Collocation component will extract collocates for each candidate terms consequently.

### 4.3.1 The Wordlist Component

The Wordlist component will produce a wordlist of the Corpus (selected files). By using statistics such as Z-Score and Weirdness, this component will be able to identify candidate terms automatically. The selection is based on the threshold values specified by the user.


Figure 4.10. Filtering and Sorting the wordlist.

Figure 4.10 shows the wordlist of a corpus. The three buttons (1,2 and 3) on the left hand side provides additional functions:

1. Search/Filter: Shows (default) or Hides the text-field(s) used for Searching/Filtering.
2. Toggle Scrollbar: Shows or Hides (default) the horizontal scrollbar.
3. Save: Saves the wordlist to a file.

The user is able to customise how the wordlist is being displayed. For instance, Figure 4.10 displays only words that end with se, and they are sorted in descending order based on frequency of the word. This is achieved by clicking the Table Header (A) and specifying constraints in the text-field (B) for each individual column. When clicking on the Table Header, the wordlist of the Corpus will be sorted accordingly; clicking on the same Table Header again will change the sorting order:


While for the constraint text-field beneath the Table Header, a Regular Expression (for textual data) or a Number (for numerical data) can be specified. For example, entering
the following regular expression under the token header will:
a[bcs]* $\quad \rightarrow$ display tokens that starts with $a$, followed by $b$ or $c$ or $s$ ([]), $b$ or $c$ or $s$ can appear 0 or more than once ( ${ }^{*}$ )
$\mathrm{a}[\mathrm{bcs}]+\quad \rightarrow$ display tokens that starts with $a$, followed by $b$ or $c$ or $s([])$, but $b$ or $c$ or $s$ must appear at least once $(+)$
r.*se\$ $\quad \rightarrow$ display tokens that starts with $r$, followed by anything (.) appearing 0 or more than once $\left(^{*}\right.$ ), and ends with (\$) se
and entering a Number under the freq header will:
$10 \rightarrow$ display tokens that have frequency $>=10$

### 4.3.2 The KWIC Component

The second component of the Text Analysis System is KWIC (Key Word In Context, or concordance). Clicking the KWIC button will list the candidate term(s), and clicking a candidate term will display its concordance on the right hand side.

Figure 4.11 below shows the concordance of the candidate term percent. Note that we can have multi-column sorting: first sorting by column -2 , and second sorting by column +1 , both in ascending order.


Figure 4.11. Concordance of the candidate term percent.

Similar to the Wordlist component, constraints can be specified in the text-field for each individual column to filter out irrelevant entries. Clicking the "Save" button will save the concordance result of the selected candidate term.

### 4.3.3 The Collocation Component

The third component of the Text Analysis System is Collocation, which is closely related to KWIC. Collocations are groups of words that frequently appear in the same context. For example, one of the potential collocates of percent (as shown in Figure 4.11) is NOMBER percent. By clicking the Collocations button, the user will see a list of collocates; left-clicking on a collocate will show its statistical information in a summarised view, while right-clicking on a collocate will show its statistical information in full details, as shown in Figure 4.12 below:


Figure 4.12. Statistical information for the collocate percent in summarised view.

The four buttons ( $1,2,3$ and 4 ) on the right hand side provides additional functions:

1. Search/Filter: Shows (default) or Hides the text-field(s) used for Searching/Filtering.
2. Toggle Scrollbar: Shows or Hides (default) the horizontal scrollbar.
3. KWIC: Shows concordance of the selected row.
4. Save: Saves the collocation results of the selected collocate to a file.

Apart from the Search/Filter, Toggle Scrollbar, KWIC and Save buttons, an additional Chart button will be shown when the user chooses to view the full statistical information of a collocate, as shown in Figure 4.13 below:


Figure 4.13. Statistical information for the collocate percent in full details.

The re-Collocate button is used to identify the $n$-gram collocates of the candidate terms. For example, from Figure 4.13, we can see that the most favourable collocate for percent is nomber percent. If we run the re-Collocate four times, we will get collocates such as nomber percent rise, over nomber percent, rose nomber percent, earnings rose nomber percent, and so on, as shown in Figure 4.14.


Figure 4.14. Collocates of the Candidate Term percent.

Collocates of all the candidate terms can be exported to a Resource Description Framework Schema (RDFS) file which can be manually edited, and verified by domain experts to confirm their validity.

### 4.4 Conclusion

In this chapter, we have presented the Text Analysis System and illustrated how it can be used to extract domain specific candidate terms, and generate statistically significant collocates. Compared with other researchers, our approach removes the burden of relying on gazetteers, annotating (manually) or POS tagging of the corpus. Moreover, minimum human intervention is required during the process. The same approach can be applied to arbitrary domains, and is able to cope with multi-lingual texts. This is illustrated by experiments in the domain of Finance, with financial news written in Chinese and English, and the domain of film reviews.

By constructing local grammar from the collocates generated by the system, we are able to extract information in an unambiguous manner. This has been illustrated by
identifying the market sentiments and tracking changes of instruments in the Finance domain, and recognising reviewer's opinions about films in movie reviews.

The use of local grammar to extract information in the financial domain forms the basis to develop the Sentiment and Time Series: Financial analysis system (SATISFI ${ }^{5}$ ), which is capable of performing time series analysis and text visualisations (see Gillam et al. 2002, Ahmad et al. 2004a). For example, sentiments extracted using local grammar are classified as positive or negative in relation to instruments and institutions that are also automatically identified. At the same time, time series of positive and negative sentiment are created, based on news arrival time, and visualised alongside other financial timeseries. Figure 4.15 below shows the correlation between news and FTSE 100 time series:


Figure 4.15: SATISFI in action: Correlation between news and instrument time series.

Local grammar based sentiments correlated well with the selected financial indices - FTSE, Dow Jones, Nasdaq and S\&P 500 - as illustrated using both the English and Chinese financial news articles. Financial traders have shown their interest in the SATISFI system as the system

[^9]could help them make better decisions. A description of the SATISFI system has been published in the "The Technical Analyst Magazine".

When using local grammar for recognising reviewer's opinions about films, our results are encouraging. We achieved an average of $64.09 \%$ precision and recall, which is comparable to Turney's results. When comparing our result with that of Pang et al., our results are better than the two graduate students who achieved $58 \%$ and $64 \%$ respectively; but lower than their hybrid machine learning approach which has accuracy in the range of $73 \%$ and $83 \%$. One point to be noted in both their approaches, their focus is on document classification rather than extracting individual sentiments within each film reviews. They also agreed that film reviews are difficult to classify, and commented that for film reviews, the overall sentiment is not necessarily the sum of the individuals - as they treat each individual sentiment equally. Clearly, this is not the case. For example, a "good film" is obviously not as good as a "great film". However, with our approach, we can solve the problem by assigning different weights to individual sentiments -1 for "good film", 1.5 for "great film", 2 for "best film", and so on. In this case, the overall sentiment for film reviews will be more reasonable.

Furthermore, platform independent implementation of the Text Analysis System and the support of Unicode enable the processing of text corpora from different languages, while using the same algorithm. This will allow corpus users around the globe to concentrate their efforts on interpreting results generated from the system, rather than spending enormous amounts of time and effort (re)implementing a system to cope with different languages.

## Chapter 5

## 5 Conclusions and Future Work

### 5.1 Conclusion

In this thesis, we have described various approaches towards corpus analysis and information extraction, and demonstrated how these techniques can be integrated to form a framework - the Text Analysis System - for analysing text corpora in arbitrary domains.

In the past, the use of gazetteers has played a major role in information extraction, particularly in the area of named entity recognition. Selecting domain specific terms to be included in the gazetteers is costly, requiring the knowledge of experts, not to mention the maintenance costs involved to keep such gazetteers up to date. POS tagging has been employed by a range of IE systems to perform such grammatical analysis as phrasal verbs identification in English, semantic orientation of reviews based on adjective/adverb phrases and ambiguity reduction. Prior training and configuring of the POS taggers from a manually annotated and tagged corpus is required. Parsing which often occur in conjunction with POS tagging, has normally been used to transform input text into a data structure - identifying boundaries of major noun or verb phrase constituents, and subsequently their syntactic and logical dependency relationships, for instance.

However, the techniques mentioned above are all domain dependent. If one moves to a different domain, the performance of IE systems using such techniques will be severely affected. In order to achieve a comparable performance, either the gazetteers for the new domain need to be constructed from scratch again or a manually annotated corpus for the new domain needs to be acquired and the POS tagger trained to adapt to the changes. As for the parser, if it is dependent on the POS tagger, reconfiguration is necessary. In some
cases, it may be difficult to obtain expert knowledge for the new domain. As both the POS tagger and parser rely on the quality and availability of annotated corpora, it will be expensive and difficult for the tagger (parser) to achieve a high performance in the new domain if such resources are scarce.

With the algorithm and the Text Analysis System we have developed, which utilises statistical techniques such as weirdness, collocation, z-score, and so on, constraints for switching to new domains are kept to a minimum. For instance, selecting domain specific terms, previously carried out by consulting thesauruses or experts, or based on intuitions, can now be done automatically by measuring the weirdness and $z$-scores of individual terms, and, consequently filtering out irrelevant terms based on the threshold values specified. In this case, even the less experienced user working for a new domain is able to perform the task. Note that the only constraint here is the presence of a reference corpus, or a wordlist representing the reference corpus. The reference corpus refers to a representative corpus for the language being studied, which may include certain aspects of the new domain. For example, the BNC corpus was used as the reference corpus when conducting sentiment analysis in the domain of finance and movie reviews, as illustrated in Chapter 4.

### 5.2 Main Contribution

We have developed an algorithm that can automatically identify domain specific terms and extract their statistically significant collocates from a corpus. Unlike other research, the algorithm is domain independent, and can be applied to arbitrary domains without any altemation. This is complimentary to Smadja's approach that relies on the manual selection of specific words for the collocation analysis to be carried out, and criteria for selecting these words are not clear. As the algorithm performs the analysis automatically, human intervention is kecp to a minimum and more resources can be allocated to analysing and refining the results generated.

Implementation of the algorithm - the.Text Analysis System - has been demonstrated with two different domains and two very different language systems. The system can aid corpus analysis, not only within arbitrary domains (finance and movie reviews in our
case), but also in relation to multi-lingual texts (Chinese and English). Some projects have claimed their system can perform multi-lingual information extraction. Their claim to be "multi-lingual" is only valid when referring to the alphabet in which the texts was written, as opposed to the actual writing system. Despite the fact that the system developed by the ECRAN project is able to process English, German and Italian texts, the underlying writing system of these three languages is still the same; that is, the Latin alphabet, or the Romance language. However, our Text Analysis System is truly multilingual, as illustrated in the case studies of analysing and extracting information from the orthographically different financial news texts written in English and Chinese.

The Text Analysis System was developed using the Java programming language, which ensures its portability and ability to operate on cross-platforms as the required Java Runtime Environment (JRE) now comes bundled with almost every operating system available. This promotes the software reusability - it is no longer necessary for linguists to redevelop a system due to the incompatibilities between different computing environments. Pre-processing, for example, converting documents in different formats to particular ones supported by individual IE systems, is no longer necessary in the Text Analysis System since document formats ranging from XML, HTML, PDF, DOC, RTF, $X L S, P P T$ to $T X T$ are directly supported. Considering the amount and varicty of electronic documents that can be easily accessed via the internet, a significant amount of time will be saved without involving a conversion between different document formats. Moreover, the Graphical User Interface representing results of the analysis in a style akin to Microsoft's Excel spreadsheet makes the Text Analysis System uscr friendly - no programming is involved, and the users can start analysing their data immediatcly.

The notion of local grammar and how it can be represented using FSA, or REGEX, has also been introduced. Local grammar provides an alternative approach towards information extraction and natural language processing. As noted by Hunston and Sinclair, such approach "would be more simple, more precise, and more useful than an analysis using general grammar" (Hunston and Sinclair, 2000:99). By merging individual local grammars, more complex tasks can be solved with minimum overheads. The application of local grammar in tasks such as machine translation; word
disambiguation; parsing; identifying and extracting patterns; and extracting proper nouns from natural language texts, have demonstrated this with promising results.

There are three major contributions I would like to report. My first contribution is sentiment analysis. Sentiment analysis is evolving at the branch of finance, particular behaviour of finance where the mood of the actors in the market, including the traders people who buy and sell stocks, institutions and regulatory bodies, is deem to have an effect on the state of the market.

In the past, there have been a number of studies on the effect of news on the market and they fall into two categories: (i) people actually analyse news stories by hand and correlated with the movements of the market (Ahmad, Cheng, Taskaya, Ahmad, Gillam, Manomaisupat, Traboulsi and Hippisley, 2003); (ii) there are a number of researchers who use a proxy for the news like timing of announcements of the news, or creating a new metric which is the linear combination of other stock market metrics to call them sentiment. My system was among the first which attempted to extract positive and negative words in financial news text. My system does not know whether the words extracted have a positive or negative sentiment, but what it does is to find sentiment bearing sentences with minimum ambiguity. This is done automatically by frequency analysis of single words and their statistically significant collocates in the domain. It turns out, fortuitously that the most statistically significant key words in financial texts have sentiment words as their key collocates - again established through statistically analysis. This study has been published in collaboration with other researchers and Ph.D students who were working with me at the time.

My second contribution lies in drawing up the basis of a contrastive analysis of how sentiment is expressed in different languages, especially languages which are typologically dissimilar. I think I am one of the first researchers who actually looked at Chinese text with the view of extracting sentiment bearing sentences in Chinese. What is particularly gratifying for me is the frequency based approach which I helped to develop appears to work just as well with the Chinese language. Much of the work in information extraction and in sentiment analysis is based on the analysis of English language text. I do understand the problem is quite complex and dealing with one language would not
leave much room to do work on other languages. Nevertheless, I think I have found a way of analysing texts, specially sentiment related texts, almost independent of the domain and, perhaps more importantly independent of the language.

My third contribution is in extracting the so-called local grammar expressions which help to distinguish between sentiment bearing sentences and those which do not. In information extraction, much of the effort is focused on the development of so-called case frames (Riloff 96), and Finite State Automata (Hobbs et al., 92). Riloff notes that earlier reported work, like Hobbs et al., the authors quite proudly present the number of case frames generated by their analysis. Indeed the case frames are generated automatically, but those case frames are based on manually crafted grammars specifically those grammars which can take into account the very complex role of verbs in English language. My system can also help in creating case frames but unlike other systems I do not rely on hand-crafted grammars, what I rely on is a training corpus.

### 5.3 Future Work

Our statistical approach towards corpus analysis and information extraction has significantly reduced the dependency on POS taggers and parsers that were used in various earlier IE systems, yet remains able to extract domain specific information unambiguously. This was achieved by constructing local grammars from the collocation patterns generated by the Text Analysis System. A cross-validation of our algorithm has been applied to multiple domains - finance and movie reviews; and the orthographically different languages - English and Chinese. The results are encouraging. However, our research raises several questions for future research.

First, there are other statistical techniques for corpus analysis and information extraction which should be explored and compared with the ones in the Text Analysis System. For instance, z -score statistics works very well in generating collocates from a large corpus, but may miss out less frequent yet important ones. This has been noted in the case study of movie reviews. Incorporating other statistical techniques such as log-likelihood ratio, mutual information or Pearson's $\chi$-square test, and allowing the user decide which one to use for generating collocates, should address this issue. Furthermore, it will be
interesting to carry out a more comprehensive study on the strengths and weaknesses of individual techniques with respect to the size of the corpora.

Second, advances in technology have greatly increased the amount and variety of electronic documents that can be freely accessed via the internet, downloaded and saved into local machines. For example, the gigaword corpora released by LDC includes Arabic, Chinese, English, French and Spanish, which covers the majority of languages being used. It would be interesting to explore how words are being used and related to each other in individual language and how words are being translated and used from one language to the other languages. For example, are the translated words being used in the same way as in the native language? Furthermore, if we consider the simplest task in corpus analysis - word counting (wordlist), processing the (complete) RCV1 corpus ( 169.9 million words in 806,791 texts) on a modern PC (a Dell PowerEdge 2650 with 1 GB memory and dual processors) takes 53300 seconds, or 14.8 hours, as reported by Ahmad, Gillam and Cheng (2005). Based on these findings, the estimated time to produce wordlist for one of the gigaword corpora will be 87.2 hours. If more complex tasks such as collocation, concordance are required, it could take at least a week to produce the results.

To expedite such computations, we can either use a more powerful machine, or carry out such tasks using Grid Computing. IBM defines Grid Computing as "the ability, using a set of open standards and protocols, to gain access to applications and data, processing power, storage capacity and a vast array of other computing resources over the Internet ..." Accessing more powerful machines is not always possible, but the availability of open standard protocols, resources and support from the Grid community make the choice of Grid Computing more appealing. Initial experiments using the grid's computing power to increase throughput have been conducted. We have created a 24 node grid infrastructure which can provide access to up to 64 processors simultaneously, in an attempt to support such analyses. Details of the Grid architecture can be found in Gillam, Ahmad and Dear (2005a). Repeating the same experiment using the RCV1 corpus, a wordlist can be produced within an hour ( 3572 seconds, we gain a throughput

[^10]increase by a factor of 15 ) by using 16 processors. When we used 64 processors, the time was reduced to half an hour ( 1683 seconds). Similarly, if we analyse the gigaword corpus using Grid, the time required to produce a wordlist will be reduced from 87.2 hours to 5.89 hours using 16 processors, or to 2.95 hours using 64 processors. It would be useful to produce a parallel implementation of the Text Analysis System to perform analyses in a Grid environment.

Third, the current implementation of the Text Analysis System was optimised for memory usage and intermediate results of the analysis were saved as Java Objects. If such objects could be organised and saved as a workplace such that the user can resume the previous analysis, view the result of the analysis at any time, or share the results of the analysis with other users, this would provide a very useful feature which would greatly enhance the system.

Finally, the use of local grammars to extract information unambiguously has shown promising results across multiple domains and multi lingual texts. However, constructing the local grammars from collocation patterns extracted by the Text Analysis System is currently done by hand. The use of Euclidean distance to measure the similarities between the generated collocation patterns which would enable the clustering of similar patterns could simplify the process of constructing a local grammar. This is an area to be explored further.

To sum up, with a knowledge rich approach there is plentiful availability of grammatical resources, lexical resources and knowledge bases, and even lexical semantic resources like gazetteers, and it is possible to obtain fairly good results on carefully selected training and testing samples. However, to deal with arbitrary texts in arbitrary languages, one has to start at least with a bottom up approach as we replicated in our research for this thesis. Our approach, briefly, sets about collecting a large corpora of texts used in a specialist domain, and then analyses that corpora of texts to create a lexicon automatically, and through collocation and colligation constructs a local grammar to find patterns of usage of critical lexical items which are specific to the domain. Implementation of the algorithm - the Text Analysis System - has been demonstrated with two different domains and two very different language systems. A cross-validation of our algorithm has been applied to multiple domains - finance and movie reviews; and the
orthographically different languages - English and Chinese. The results are encouraging.

Our preliminarily evaluation, based on the correlation between a time series of positive (negative) sentiment word (phrase) counts with a time series of indices produced by stock exchanges (Financial Times Stock Exchange, Dow Jones Industrial Average, Nasdaq, S\&P 500, Hang Seng Index, Shanghai Index, and Shenzhen Index) showed that when the positive (negative) sentiment series correlates with the stock exchange index, the negative (positive) shows a smaller degree of correlation and in many cases a degree of anti-correlation. Any interpretation of our result requires a careful econometrically well grounded analysis of the financial time series - this is beyond the scope of this work.

## Bibliography

Ahmad, K. (1995) Pragmatics of specialist terms: the acquisition and representation of terminology. In Steffens, P. (ed.), Machine Translation and the Lexicon, Berlin: Springer-Verlag, 51-76.

Ahmad, K. (2002) Events and the causes of events: the use of metaphor in financial texts. In Proceedings of the workshop at the International Conference on Terminology and Knowledge Engineering. Nancy, France.

Ahmad, K. and Rogers, M. (2001) Corpus linguistics and terminology extraction. In Wright, S. E. and Budin, G. (eds.), The Handbook of Terminology Management, Vol. 2 Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ahmad, K., Cheng, D., Taskaya, T., Ahmad, S., Gillam, L., Manomaisupat, P., Traboulsi, H. and Hippisely, A. (2003) The mood of the (financial) markets: In a corpus of words and of pictures. In Archer, D., Rayson, P., Wilson, A., and McEnery, T. (eds.) Proceedings of the Corpus Linguistics 2003 Conference.

Ahmad, K., Cheng, D. and Almas, Y. (2006) Multi-lingual sentiment analysis of financial news streams. In S. Cozzini, S. d'Addona and R. Mantegna (eds.), Proceeding of $1^{\text {st }}$ International Conference on Grid in Finance, Palermo, Italy.

Ahmad, K., Gillam, L. and Cheng, D. (2005) Textual and Quantitative Analysis: Towards a new, e-mediated Social Science. In Proceeding of the $1^{\text {st }}$ International Conference on e-Social Science, Manchester, United Kingdom.

Ahmad, K., Vrusias, B. and Ledford, A. (2001) Choosing Feature Sets for Training and Testing Self-Organising Maps: A Case Study. Neural Computing \& Applications. Volume 10, 56-66.

Ahmad, K., Taskaya-Temizel, T., Cheng, D., Ahmad, S., Gillam, L., Manomaisupat, P., Traboulsi, H. and Casey, M. (2004a) Fundamental Data to SATISFI the Chartist. The Technical Analyst Magazine, Vol. 1 (No. 3), 36-38.

Ahmad, K., Taskaya-Temizel, T., Cheng, D., Gillam, L., Ahmad, S., Traboulsi, H. and Nankervis, J. (2004b) Financial Information Grid - an ESRC e-Social Science Pilot. In Proceedings of the Third UK e-Science Programme All Hands Meeting (AHM 2004), Nottingham, United Kingdom.

Appelt, D. E. and Israel, D. J. (1999) Introduction to Information Extraction Technology. Tutorial for the $16^{\text {th }}$ International Joint Conferences on Artificial Intelligence (IJCAI99), Stockholm.

Aston, G. and Burnard, L. (1998) The BNC Handbook. Edinburgh: Edinburgh University Press.

Baestaens, D. J. E. and van den Bergh, W. M. (1995) The marginal contribution of news to the DEM/USD swap rate. Neural Network World, Vol. 5(4), 371-378.

Barnbrook, G. and Sinclair, J. (1995) Parsing CoBuild Entries. In Sinclair, J., Hoelter, M. and Peters, C. (eds.), The Languages of Definition: The Formalization of Dictionary Definitions for Natural Language Processing. Luxembourg: Office for Official Publications of the European Communities, 13-58.

Benson, M. (1990) Collocations and General-purpose Dictionaries. International Journal of Lexicography, 3: 23-34.

Breidt. E., Segond, F. and Valetto, G. (1994) Local Grammars for the Description of Multi-Word Lexemes and Their Automatic Recognition in Text. In Proceedings of Comlex-2380 - Papers in Computational Lexicography, Linguistics Institute, HAS, Budapest, 19-28.

Carvalho, P., Mota, C. and Ranchhod, E. (2002) Complex Lexical Units and Automata. In Mamede, N. and Ranchhod, E. (eds.), Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002), Berlin/Heidelberg: Springer-Verlag. Lecture Notes in Artificial Intelligence 2389, 229-238.

Casacuberta, F. and Vidal, E. (2004) Machine Translation with Inferred Stochastic FiniteState Transducers. Computational Linguist. 30, 2, 205-225.

Chan, K. and Lam, W. (2005) Extracting Causation Knowledge from Natural Language Texts. International Journal of Intelligent Systems, 20(3), 327-358.

Chen, K. J. and Liu, S. H. (1992) Word identification for Mandarin Chinese sentences. In Proceedings of the $14^{\text {th }}$ International Conference on Computational Linguistics (COLING), Nantes, France, 101-107.

Childs, L., Guthrie, L., and Sider, I. (1995) Lockheed Martin: LOUELLA PARSING, an NLToolset system for MUC-6. In Proceedings of the 6th Conference on Message Understanding (Columbia, Maryland, November 06-08, 1995).

Choi, K. and Nam, J. (1997) A Local-Grammar-based Approach to Recognizing of Proper Names in Korean Texts. In Zhou, J. and Church, K. (eds.), Proceedings of the Fifth Workshop on Very Large Corpora, ACL/Tsing-hua University/Hong-Kong University of Science and Technology, 273-288.

Clear, J. (1993) From Firth Principles. Computational Tools for the Study of Collocation. In Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.), Text and Technology. In Honour of John Sinclair. Philadelphia/Amsterdam: John Benjamins.

Das, S. R. and Chen, M. Y. (2006) Yahoo! for Amazon: Sentiment Extraction From Small Talk on the Web. Working Paper, Leavey School of Business, Santa Clara University, California. Available at: http://scumis.scu.edu/~srdas/chat.pdf, date accessed: 10 January 2007.

DeGennaro, R. and Shrieves, R. (1997) Public information releases, private informationarrival and volatility in the foreign exchange market. Journal of Empirical Finance, Vol. 4, 295-315.

Firth, J. R. (1957a) Modes of meaning. In Firth J. R. (ed.) Papers in Linguistics 19341951, London, 190-215.

Firth, J.R. (1957b) A synopsis of linguistic theory, 1930-1955. In Palmer, F. R. (ed.) (1968) Selected Papers of J. R. Firth 1952-1959. London:Longman

Friedl, J. (2002) Mastering Regular Expressions Second Edition. O'Reilly. ISBN 0-596-00289-0.

Gaizauskas, R., Greenwood, M. A., Hepple, M., Harkema, H., Saggion, H. and Sanka, A. (2005) The University of Sheffield's TREC 2005 Q\&A Experiments. In Proceedings of the 14th Text REtrieval Conference.

Gaizauskas, R., Greenwood, M. A., Hepple, M., Roberts, I. and Saggion, H. (2004) The University of Sheffield's TREC 2004 Q\&A Experiments. In Proceedings of the 13th Text REtrieval Conference.

Gale, W. and Church, K. W. (1990) What's wrong with adding one? IEEE Transactions on Acoustics, Speech and Signal Processing.

Gillam, L., Ahmad, K., Ahmad, S., Casey, M., Cheng, D., Taskaya, T., Oliviera, P. and Manomaisupat, P. (2002) Economic News and Stock Market Correlation: A Study of the UK Market. In Gillam, L. (ed.), Making money in the financial services industry, Proceedings of a workshop at the Terminology and Knowledge Engineering conference TKE 2002.

Gillam, L., Ahmad, K. and Dear, G. (2005a) Grid-enabling Social Scientists: some infrastructure issues. In Proceeding of the $1^{\text {st }}$ International Conference on e-Social Science, Manchester, United Kingdom.

Gillam, L., Tariq, M. and Ahmad, K. (2005b) Terminology and the Construction of Ontology. Terminology 11(1), 55-81

Grishman, R. (2002) Adaptive Information Extraction and Sub-language Analysis. In Proceedings of the $17^{\text {th }}$ International Joint Conferences on Artificial Intelligence (IJCAI-01) Workshop on Adaptive Text Extraction and Mining.

Gross, M. (1993) Local Grammars and their Representation by Finite Automata. In Hoey, M. (ed.), Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair. HarperCollins Publishers, 26-38.

Gross, M. (1997) The construction of local grammars. In Roche, E. and Schab`es, Y. (editors), Finite-State Language Processing, 329-354. Cambridge, MA, MIT Press.

Halliday, M. and Martin, J. (1993) Writing Science: Literacy and Discursive Power, London and Washington D.C: The Falmer Press.

Harris, Z. (1968) Mathematical Structures of Language. Wiley, New York.
Harris, Z. (1991) A Theory of Language and Information: A Mathematical Approach. Oxford: Clarendon Press.

Henderson, W., Dudley-Evans, T. and Backhouse, R. (2000) Economics and Language, London \& New York: Routledge.

Hirschman, L. and Sager, N. (1982) Automatic information formatting of a medical sublanguage. IN Kittredge, R. and Lehrberger, J. (eds.), Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin , New York : Walter de Gruyter, 27-80.

Hoey, M. (1991) Patterns of Lexis in Text. Oxford: Oxford University Press.
Hobbs, J. (1993) The Generic Information Extraction System. In Proceedings of the Fifth Message Understanding Conference (MUC-5), Baltimore, Maryland, 87-92.

Hobbs, J.; Appelt, D.; Bear, J.; Israel, D.; and Tyson, M. (1992) FASTUS: a system for extracting information from natural-language text. Technical Report Technical Note No. 519, SRI International Artificial Intelligence Center.

Hsiao, P. H., Hsieh, T. H., Tan, K. S., Tsai, C. H. and Yeh, W. (2000) Localization library for Taiwan and Big5 Encoding (libtabe). (Version 0.1-3) [Computer programming library]. Available at: http://sourceforge.net/projects/libtabe/, date accessed: 10 December 2005.

Hunston, S. and Sinclair, J. (2000) A Local Grammar of Evaluation. In Hunston, S. and Thompson, G. (eds), Evaluation in Text: Authorial Stance and the Construction of Discourse, Oxford University Press, 75-100.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. and Wilks, Y. (1998) University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In Proceedings of the Seventh Message Understanding Conferences (MUC-7).

Jackendoff, R. (1991) Semantic Structures. Cambridge (USA) \& London: The MIT Press.
Karttunen, L., Chanod, J. P., Grefenstette, G. and Schiller, A. (1996) Regular Expression for Language Engineering. Natural Language Engineering, 2(4): 305-328.

Kavanagh, J. (1996) The Text Analyzer: a Tool for Extracting Knowledge from Text. Master's of Computer Science Thesis, University of Ottawa, Canada. Available at: http://www.csi.uottawa.ca/~kavanagh/Thesis/index.html, date accessed: 5 December 2006.

Kendall, M. (1976) Time-Series. London: Charles Griffin and Company Ltd.

Kittredge, R. (1982) Sublanguages. American Journal of Computational Linguistics 8(2), 79-84.

Knowles, F. (1996) Lexicographical Aspects of Health Metaphors in Financial Text. In Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström and Catarina Röjder Papmehl (eds.), Euralex 1996, Proceedings Part II, 789-796.

Koppel, M and Shtrimberg, I. (2004) Good News or Bad News? Let the Market Decide. In AAAI Spring Symposium on Exploring Attitude and Affect in Text. Palo Alto: AAAI Press. 86-88.

Kugler, M., Ahmad, K. and Thurmair, G. (1995) Translator's Workbench: Tools and Terminology for Translation and Text Processing. Berlin: Springer-Verlag.

Kumar, S. and Byrne, W. (2003) A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, 63-70.

Laporte, É. and Monceaux, A. (2000) Elimination of lexical ambiguities by grammars: The ELAG system. Linguisticae Investigationes, 22:2, 341-367.

Lee, C. Y. (1999) Local Grammar-based Lexical Stemmer for Korean Language. In Proceedings of VEXTAL '99, Venezia, San Servolo, 139-144.

Li, W., Zhang, X., Niu, C., Jiang, Y. and Srihari, R. (2003) An Expert Lexicon Approach to Identifying English Phrasal Verbs. In Proceedings of the $41^{s t}$ Annual Meeting of the Association for Computational Linguistics, Japan, 513-520.

Markham, A. (2004). Internet communication as a tool for qualitative research. In Silverman, D. (ed.), Qualitative Research. Theory, Method and Practice, 95-124.

Mason, O. (2004) Automatic Processing of Local Grammar Patterns. In Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, University of Birmingham, 166-171.

Mehta, J. (1993) Meanings in the context of bargaining games: Narratives in opposition. In Henderson, W., Dudley-Evans, T. and Backhouse, R. (eds.), Economics and Language, 85-99.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs Up? Sentiment Classification using Machine Learning Techniques. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 79-86.

Pearson, J. (ed.) (1998) Terms in Context. John Benjamins Publishing Company, Amsterdam.

Peng, F. and Schuurmans, D. (2001) Self-supervised Chinese word segmentation. In F. Hoffman et al., (eds.), Advances in Intelligent Data Analysis, Proceedings 4th International Conference (IDA-01), 238-247.

Poibeau, T. (2000) A corpus-based approach to Information Extraction. Journal of Applied Systems Studies (JASS). Vol. 1, No. 2, 254-267.

Quirk, R., Greenbaum, S., Leech, G. and Svartvok, J. (1985) A comprehensive Grammar of the English Language. Longman.

Ranchhod, E., Mota, C. and Baptista, J. (1999) A Computational Lexicon of Portuguese for Automatic Text Parsing. In Proceedings of SIGLEX-99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL, College Park, Maryland, USA, 74-80.

Roche, E. (1992) Text disambiguation by finite-state automata: an algorithm and experiments on corpora. In COLING-92. Proceedings of the Conference, Nantes, 993997.

Roche, E. (1997) Parsing with Finite-State Transducers. In Roche, E. and Schabes, Y. (eds.), Finite-State Language Processing, MIT Press, Cambridge, 241-281.

Riloff, E. (1996) Automatically Generating Extraction Patterns from Untagged Text. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI96), The AAAI Press/MIT Press, 1044-1049.

Riloff, E. (1999) Information Extraction as a Stepping Stone toward Story Understanding. In A. Ram and K. Moorman (eds.), Understanding Language Understanding: Computational Models of Reading, MIT Press Language, Speech, And Communications Series. MIT Press, Cambridge, 435-460.

Rubinstein, A. (2000) Economics and Language: Five Essays. Cambridge: Cambridge University Press.

Salway, A. and Ahmad, K. (1997) The Role of Metaphors in Financial Texts. Available at: http://citeseer.ist.psu.edu/salway97role.html, date accessed: 20 December 2006.

Shiller, R. (1993) Measuring Asset Values for Cash Settlement in Derivative Markets: Hedonic Repeated Measures Indices and Perpetual Futures. Journal of Finance, Vol. 48 (3), 911-931.

Shiller R. J. (2003) The New Financial Order: Risk in the 21st Century. Princeton: Princeton University Press.

Silberztein, M. (2000) INTEX: An FST Toolbox. Theoretical Computer Science, Volume 231(1): 33-46.

Simmons, R. (1984) Computations from the English: A Procedural Logic Approach for Representing and Understanding English Texts. Englewood Cliffs, NJ: Prentice Hall.

Sinclair, J. (1966) Beginning the Study of Lexis. In Bazell, C.E., Catford, J.C., Halliday, M.A.K. and Robins, R.H. (eds.), In Memory of J.R. Firth. London: Longmans.

Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford University Press.
Smadja, F. (1991) Extracting collocations from text. An application: Language generation. Ph.D Thesis, Columbia University.

Smadja, F. (1994) Retrieving Collocations from Text: Xtract. In Armstrong, S. (ed.), Using Large Corpora. London: Mit Press.

Soderland, S. (1999) Learning Information Extraction Rules for Semi-structured and Free Text. Machine Learning 1:44.

Spertus, E. (1997) Smokey: Automatic recognition of hostile messages. In Proceedings of the Conference on Innovative Applications of Artificial Intelligence. Menlo Park, CA: AAAI Press, 1058-1065.

Srihari, R., Li, W., Niu, C. and Cornell, T. (2003) InfoXtract: A Customizable Intermediate Level Information Extraction Engine. In Proceedings of the HLTNAACL 2003 workshop on Software engineering and architecture of language technology systems, Vol. 8, 51-58.

Surdeanu, M., Harabagiu, S., Williams, J. and Aarseth, P. (2003) Using PredicateArgument Structures for Information Extraction. In Proceeding of the $41^{\text {st }}$ Annual Conference of the Association for Computational Linguistics (ACL'03), Sapporo, Japan, 8-15.

Tetlock, P., Mytal, S. and Sofus, M. (2005) More Than Words: Quantifying Language to Measure Firms' Fundamentals. Available at: http://www.mccombs.utexas.edu/faculty/Paul.Tetlock/papers/TSM More Than Wor ds 09 06.pdf, date accessed: 22 December 2006.

Traboulsi, H., Cheng, D. and Ahmad, K. (2004) Text Corpora, Local Grammars and Prediction. In Proceedings of 4th International Language Resource and Evaluation Conference (LREC) 2004, Vol. 3, 749-752.

Turney, P. (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceeding of the $40^{\text {th }}$ Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, 417-424.

Wiebe, J., Bruce, R., Bell, M., Martin, M. and Wilson, T. (2001) A Corpus Study of Evaluative and Speculative Language. In Proceeding of $2^{\text {nd }}$ ACL SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark, 1-10.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Conference on Human

Language Technology and Empirical Methods in Natural Language Processing. Vancouver, 347-354.

Yale (2004). Yale School of Management Stock Market Confidence Indexes ${ }^{\text {TM }}$. Available at: $\quad$ http://icf.som.yale.edu/financial data/confidence index/index.shtml, date accessed: 15 December 2004

Yao, Y. and Lua, K. T. (1998) Splitting-merging model of Chinese word tokenization and segmentation. Natural Language Engineering, 5: 309-324.

Yeh, C. L. and Lee, H. J. (1991) Rule-based word identification for Mandarin Chinese sentences: a unification approach. Computer Processing of Chinese \& Oriental Language 5(2): 171-184.

Zipf, G. K. (1935) The Psychobiology of Language, Houghton Mifflin, Boston.

## Appendix A - Publications

Ahmad, K., Bennet, C., Mountstephens, J., Cheng, D., Vassiliou, A. and Salway, A. (2005) Video Summarisation: Variation in Attention and Linguistic Description of Surveillance Videos. IEE International Symposium on Imaging for Crime Detection and Prevention.

Ahmad, K., Cheng, D. and Almas, Y. (2006) Multi-lingual sentiment analysis of financial news streams. In S. Cozzini, S. d'Addona and R. Mantegna (eds.), Proceeding of 1st International Conference on Grid in Finance, Palermo, Italy.

Ahmad, K., Cheng, D. and Traboulsi, H. (2003) Special Language and Local Grammar: Analysing Financial News Streams. The 14th European Symposium on Language for Special Purposes (LSP 2003), Surrey, United Kingdom.

Ahmad, K., Cheng, D., Taskaya, T., Ahmad, S., Gillam, L., Manomaisupat, P., Traboulsi, H. and Hippisely, A. (2003) The mood of the (financial) markets: In a corpus of words and of pictures. In Archer, D., Rayson, P., Wilson, A., and McEnery, T. (eds.) Proceedings of the Corpus Linguistics 2003 Conference.

Ahmad, K., Gillam, L. and Cheng, D. (2005) Society Grids. In Proceedings of the Fourth UK e-Science Programme All Hands Meeting (AHM 2005), Nottingham, United Kingdom.

Ahmad, K., Gillam, L. and Cheng, D. (2005) Textual and Quantitative Analysis: Towards a new, e-mediated Social Science. In Proceeding of the 1st International Conference on e-Social Science, Manchester, United Kingdom.

Ahmad, K., Taskaya-Temizel, T., Cheng, D., Ahmad, S., Gillam, L., Manomaisupat, P., Traboulsi, H. and Casey, M. (2004) Fundamental Data to Satisfi the Chartist. The Technical Analyst Magazine, Vol. 1 (No. 3), 36-38.

Ahmad, K., Taskaya-Temizel, T., Cheng, D., Gillam, L., Ahmad, S., Traboulsi, H. and Nankervis, J. (2004b) Financial Information Grid - an ESRC e-Social Science Pilot. In Proceedings of the Third UK e-Science Programme All Hands Meeting (AHM 2004), Nottingham, United Kingdom.

Gillam, L., Ahmad, K., Ahmad, S., Casey, M., Cheng, D., Taskaya, T., Fernandes de Oliviera, P. and Manomaisupat, P. (2002) Economic News and Stock Market Correlation: A Study of the UK Market. In Gillam, L. (ed.), Making money in the financial services industry, Proceedings of a workshop at the Terminology and Knowledge Engineering conference TKE 2002.

Hippisley, A., Cheng, D. and Ahmad, K. (2002) Chinese special languages and the notion of headedness. In Merja Koskela, C.L.M.N.a.N.P. (ed), Proceedings of the University of Vaasa, Porta Scientiae, vol. 1, 329-342.

Hippisley, A., Cheng, D. and Ahmad, K. (2005) The head-modifier principle and multilingual term extraction. Natural Language Engineering, Volume 11(2), 129-157, Cambridge University Press.

Hippisley, A., Tariq, M. and Cheng, D. (2001) Hierarchical data and the derivational relationship between words. In Bird, S., Buneman, P. and Mark Lieberman (eds.) Proceedings of the Institute for Research into Cognitive Sciences Workshop on Linguistic Databases, 125-133, Penn University.

Traboulsi, H., Cheng, D. and Ahmad, K. (2004) Text Corpora, Local Grammars and Prediction. In Proceedings of 4th International Language Resource and Evaluation Conference (LREC) 2004, Vol. 3, 749-752.

## Appendix B

List of common words (exclude list) used in the algorithm.

| u.s | xii | thu | gone | our | who |
| :--- | :--- | :--- | :--- | :--- | :--- |
| a | january | fri | goes | ourself | why |
| b | february | sat | got | pp | with |
| c | march | he | gotten | put | within |
| d | april | his | had | putting | without |
| e | may | him | he | saw | you |
| f | june | she | she | see | your |
| g | july | her | her | she | yourself |
| h | august | they | herself | since |  |
| k | september | them | him | sit |  |
| l | october | al | himself | sitting |  |
| m | november | all | his | so |  |
| n | december | am | however | some |  |
| o | monday | an | if | someone |  |
| p | tuesday | are | in | something |  |
| q | wednesday | as | into | such |  |
| r | thursday | be | inside | that |  |
| s | friday | because | is | the |  |
| t | saturday | been | it | their |  |
| u | sunday | before | its | then |  |
| v | jan | being | itself | there |  |
| w | feb | did | like | they |  |
| y | mar | do | me | this |  |
| z | apr | does | must | these |  |
| i | may | doing | my | those |  |
| ii | jun | dont | myself | unless |  |
| iii | jul | ed | neither | until |  |
| iv | aug | eds | never | us |  |
| v | sep | et | no | very |  |
| vi | oct | for | none | vol |  |
| vii | nov | gave | nor | we |  |
| viii | dec | give | not | what |  |
| ix | mon | tue | given | of | where |
| xi | wed | going | onto | whether |  |
|  |  |  |  | other | which |

## Appendix C

Downward collocates ( $\mathbf{2 , 5 2 0}$ in total) of fell sorted by strength. Note that only collocates with strength $>=0.1$ are listed.

| $f_{\text {neighbourhood of shares }}$ |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $x^{\prime}$ | $f$ | -5 | -4 | -3 | ... | 3 | 4 | 5 | sum | $\mu$ | $\sigma$ | spread | strength |
| back | 3680 | 1 | 0 | 0 | ... | 1 | 1 | 4 | 108 | 10.80 | 31.72 | 905.36 | 0.81 |
| sharply | 735 | 1 | 0 | 0 | ... | 0 | 0 | 0 | 105 | 10.50 | 27.69 | 689.85 | 0.78 |
| stocks | 3460 | 14 | 12 | 13 | ... | 0 | 3 | 2 | 103 | 10.30 | 16.01 | 230.61 | 0.76 |
| oil | 2823 | 7 | 21 | 5 | ... | 4 | 4 | 2 | 82 | 8.20 | 10.28 | 95.16 | 0.59 |
| profit | 4065 | 13 | 10 | 7 | ... | 2 | 0 | 0 | 80 | 8.00 | 13.79 | 171.20 | 0.57 |
| vodafone | 2771 | 4 | 1 | 8 | ... | 0 | 7 | 1 | 76 | 7.60 | 14.51 | 189.44 | 0.54 |
| services | 3075 | 3 | 14 | 35 | ... | 0 | 0 | 7 | 74 | 7.40 | 11.20 | 112.84 | 0.52 |
| british | 3927 | 10 | 11 | 11 | ... | 0 | 1 | 1 | 74 | 7.40 | 12.46 | 139.64 | 0.52 |
| news | 3890 | 18 | 8 | 4 | ... | 5 | 18 | 9 | 72 | 7.20 | 6.41 | 36.96 | 0.50 |
| revenues | 1498 | 11 | 4 | 1 | ... | 0 | 0 | 1 | 70 | 7.00 | 16.17 | 235.40 | 0.49 |
| mobile | 2181 | 7 | 34 | 16 | ... | 3 | 5 | 0 | 67 | 6.70 | 10.80 | 105.01 | 0.46 |
| average | 1930 | 3 | 5 | 13 | ... | 4 | 0 | 0 | 65 | 6.50 | 11.76 | 124.45 | 0.44 |
| bt | 1840 | 1 | 4 | 1 | ... | 0 | 16 | 1 | 64 | 6.40 | 9.92 | 88.64 | 0.44 |
| shell | 946 | 1 | 2 | 1 | ... | 0 | 20 | 0 | 62 | 6.20 | 11.49 | 118.76 | 0.42 |
| chip | 1645 | 10 | 18 | 4 | ... | 1 | 6 | 2 | 60 | 6.00 | 6.34 | 36.20 | 0.40 |
| giant | 1880 | 1 | 6 | 7 | ... | 0 | 0 | 1 | 60 | 6.00 | 13.60 | 166.40 | 0.40 |
| tax | 3373 | 4 | 11 | 8 | ... | 0 | 0 | 0 | 59 | 5.90 | 9.96 | 89.29 | 0.39 |
| lowest | 1025 | 0 | 0 | 0 | ... | 38 | 4 | 15 | 57 | 5.70 | 12.29 | 136.01 | 0.38 |
| nearly | 1095 | 2 | 0 | 0 | ... | 0 | 1 | 5 | 56 | 5.60 | 14.98 | 202.04 | 0.37 |
| output | 973 | 0 | 1 | 11 | ... | 0 | 0 | 0 | 55 | 5.50 | 9.59 | 82.85 | 0.36 |
| when | 3579 | 2 | 7 | 13 | ... | 1 | 6 | 2 | 55 | 5.50 | 5.87 | 31.05 | 0.36 |
| telecoms | 2979 | 13 | 15 | 10 | ... | 0 | 7 | 1 | 54 | 5.40 | 5.64 | 28.64 | 0.35 |
| retailer | 1643 | 1 | 10 | 16 | ... | 0 | 0 | 4 | 53 | 5.30 | 7.70 | 53.41 | 0.34 |
| nasdaq | 928 | 2 | 0 | 25 | ... | 0 | 0 | 1 | 53 | 5.30 | 8.65 | 67.41 | 0.34 |
| retail | 2163 | 1 | 14 | 14 | $\cdots$ | 0 | 4 | 2 | 53 | 5.30 | 7.12 | 45.61 | 0.34 |
| industrial | 1162 | 2 | 3 | 2 | ... | 0 | 2 | 0 | 53 | 5.30 | 13.65 | 167.61 | 0.34 |
| dow | 1098 | 1 | 37 | 2 | ... | 0 | 4 | 5 | 53 | 5.30 | 11.27 | 114.41 | 0.34 |
| crude | 722 | 1 | 3 | 11 | ... | 1 | 1 | 8 | 53 | 5.30 | 6.22 | 34.81 | 0.34 |
| royal | 2065 | 7 | 28 | 4 | ... | 0 | 12 | 1 | 52 | 5.20 | 8.97 | 72.36 | 0.33 |
| bp | 1010 | 0 | 2 | 7 | ... | 0 | 13 | 1 | 51 | 5.10 | 7.85 | 55.49 | 0.33 |
| short | 1285 | 1 | 2 | 0 | ... | 3 | 1 | 0 | 51 | 5.10 | 10.38 | 96.89 | 0.33 |
| phone | 1382 | 2 | 6 | 28 | ... | 0 | 5 | 5 | 51 | 5.10 | 8.43 | 63.89 | 0.33 |
| cents | 686 | 0 | 0 | 0 | $\cdots$ | 5 | 1 | 5 | 50 | 5.00 | 12.12 | 132.20 | 0.32 |
| hsbc | 861 | 5 | 1 | 0 | ... | 0 | 6 | 1 | 49 | 4.90 | 8.63 | 67.09 | 0.31 |
| maker | 1315 | 2 | 4 | 1 | ... | 0 | 0 | 0 | 49 | 4.90 | 11.74 | . 124.09 | 0.31 |
| items | 550 | 9 | 4 | 4 | $\cdots$ | 0 | 0 | 0 | 47 | 4.70 | 7.76 | 54.21 | 0.29 |
| early | 3084 | 7 | 4 | 1 | $\cdots$ | 6 | 8 | 6 | 46 | 4.60 | 4.55 | 18.64 | 0.28 |
| level | 1977 | 1 | 1 | 1 | ... | 0 | 31 | 12 | 46 | 4.60 | 9.98 | 89.64 | 0.28 |
| manufacturing | 1106 | 1 | 1 | 9 | ... | 0 | 0 | 0 | 44 | 4.40 | 7.26 | 47.44 | 0.27 |
| among | 1979 | 23 | 4 | 0 | ... | 0 | 17 | 0 | 44 | 4.40 | 8.44 | 64.04 | 0.27 |
| consumer | 2366 | 2 | 12 | 17 | ... | 1 | 0 | 1 | 43 | 4.30 | 6.27 | 35.41 | 0.26 |
| cable | 877 | 5 | 4 | 26 | ... | 0 | 7 | 1 | 43 | 4.30 | 8.04 | 58.21 | 0.26 |
| lloyds | 891 | 2 | 5 | 3 | $\cdots$ | 0 | 6 | 0 | 43 | 4.30 | 7.93 | 56.61 | 0.26 |


| low | 3799 | 4 | 4 | 0 | ... | 6 | 4 | 6 | 43 | 4.30 | 5.38 | 26.01 | 0.26 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| net | 1248 | 3 | 0 | 2 | ... | 0 | 0 | 1 | 42 | 4.20 | 11.22 | 113.36 | 0.25 |
| months | 4257 | 10 | 17 | 4 | ... | 0 | 0 | 1 | 42 | 4.20 | 5.65 | 28.76 | 0.25 |
| software | 903 | 8 | 7 | 19 | $\cdots$ | 0 | 0 | 2 | 41 | 4.10 | 6.03 | 32.69 | 0.24 |
| pre | 2121 | 1 | 32 | 0 | $\cdots$ | 1 | 0 | 6 | 41 | 4.10 | 9.97 | 89.49 | 0.24 |
| advertising | 1040 | 3 | 7 | 8 | $\ldots$ | 0 | 8 | 1 | 40 | 4.00 | 3.89 | 13.60 | 0.23 |
| astrazeneca | 646 | 0 | 0 | 3 | $\cdots$ | 0 | 3 | 0 | 40 | 4.00 | 9.92 | 88.60 | 0.23 |
| operator | 1364 | 10 | 2 | 9 | $\ldots$ | 0 | 0 | 0 | 40 | 4.00 | 6.55 | 38.60 | 0.23 |
| third | 2354 | 7 | 5 | 1 | $\cdots$ | 9 | 4 | 10 | 40 | 4.00 | 3.65 | 12.00 | 0.23 |
| mmo2 | 772 | 0 | 1 | 0 | $\ldots$ | 0 | 8 | 5 | 40 | 4.00 | 7.86 | 55.60 | 0.23 |
| per | 1983 | 4 | 1 | 1 | $\ldots$ | 1 | 3 | 1 | 40 | 4.00 | 5.91 | 31.40 | 0.23 |
| glaxosmithkline | 501 | 0 | 3 | 2 | $\cdots$ | 0 | 5 | 1 | 39 | 3.90 | 8.63 | 67.09 | 0.22 |
| granada | 922 | 5 | 3 | 4 | $\ldots$ | 1 | 3 | 9 | 39 | 3.90 | 3.14 | 8.89 | 0.22 |
| barclays | 1179 | 0 | 6 | 0 | $\ldots$ | 1 | 2 | 2 | 38 | 3.80 | 5.83 | 30.56 | 0.22 |
| exceptional | 599 | 4 | 7 | 0 | $\ldots$ | 0 | 0 | 0 | 38 | 3.80 | 8.50 | 64.96 | 0.22 |
| euros | 2010 | 3 | 1 | 1 | $\cdots$ | 6 | 7 | 12 | 37 | 3.70 | 4.11 | 15.21 | 0.21 |
| earnings | 3779 | 8 | 4 | 13 | ... | 3 | 2 | 0 | 37 | 3.70 | 4.14 | 15.41 | 0.21 |
| street | 2489 | 7 | 3 | 12 | $\ldots$ | 2 | 0 | 0 | 36 | 3.60 | 4.48 | 18.04 | 0.20 |
| investment | 3543 | 9 | 1 | 4 | ... | 1 | 4 | 1 | 36 | 3.60 | 3.69 | 12.24 | 0.20 |
| jones | 1075 | 0 | 1 | 34 | ... | 0 | 0 | 0 | 36 | 3.60 | 10.69 | 102.84 | 0.20 |
| much | 2371 | 1 | 0 | 0 | ... | 0 | 0 | 0 | 36 | 3.60 | 10.02 | 90.44 | 0.20 |
| both | 1909 | 1 | 3 | 0 | $\ldots$ | 1 | 5 | 0 | 36 | 3.60 | 6.43 | 37.24 | 0.20 |
| food | 1320 | 5 | 10 | 8 | ... | 0 | 0 | 11 | 35 | 3.50 | 4.58 | 18.85 | 0.19 |
| goodwill | 518 | 2 | 11 | 6 | $\ldots$ | 0 | 0 | 0 | 35 | 3.50 | 4.40 | 17.45 | 0.19 |
| markets | 3503 | 5 | 10 | 4 | ... | 0 | 1 | 0 | 35 | 3.50 | 4.62 | 19.25 | 0.19 |
| wireless | 693 | 1 | 1 | 5 | $\ldots$ | 0 | 0 | 0 | 35 | 3.50 | 6.87 | 42.45 | 0.19 |
| reported | 2928 | 3 | 2 | 5 | $\cdots$ | 0 | 2 | 17 | 34 | 3.40 | 5.17 | 24.04 | 0.18 |
| rate | 3174 | 1 | 0 | 2 | $\ldots$ | 0 | 9 | 1 | 34 | 3.40 | 4.67 | 19.64 | 0.18 |
| holdings | 684 | 1 | 2 | 8 | $\cdots$ | 0 | 0 | 1 | 34 | 3.40 | 6.98 | 43.84 | 0.18 |
| information | 761 | 7 | 11 | 13 | $\ldots$ | 0 | 0 | 0 | 33 | 3.30 | 5.10 | 23.41 | 0.17 |
| arm | 989 | 0 | 9 | 0 | $\cdots$ | 0 | 2 | 1 | 32 | 3.20 | 5.88 | 31.16 | 0.17 |
| trade | 2360 | 1 | 0 | 2 | ... | 11 | 4 | 11 | 32 | 3.20 | 4.29 | 16.56 | 0.17 |
| rival | 1690 | 3 | 6 | 1 | $\cdots$ | 0 | 7 | 6 | 32 | 3.20 | 3.49 | 10.96 | 0.17 |
| overall | 1102 | 10 | 9 | 9 | ... | 0 | 0 | 1 | 32 | 3.20 | 4.26 | 16.36 | 0.17 |
| traffic | 569 | 7 | 0 | 3 | $\ldots$ | 0 | 0 | 0 | 32 | 3.20 | 6.65 | 39.76 | 0.17 |
| revenue | 1284 | 3 | 6 | 4 | $\cdots$ | 0 | 0 | 0 | 31 | 3.10 | 5.67 | 28.89 | 0.16 |
| hit | 2569 | 3 | 3 | 0 | $\ldots$ | 1 | 17 | 3 | 31 | 3.10 | 5.07 | 23.09 | 0.16 |
| inflation | 2073 | 0 | 0 | 6 | ... | 0 | 0 | 0 | 31 | 3.10 | 5.90 | 31.29 | 0.16 |
| second | 3725 | 4 | 5 | 0 | $\cdots$ | 14 | 2 | 6 | 31 | 3.10 | 4.48 | 18.09 | 0.16 |
| composite | 368 | 0 | 0 | 0 | $\cdots$ | 0 | 0 | 0 | 31 | 3.10 | 6.87 | 42.49 | 0.16 |
| national | 1793 | 7 | 4 | 2 | ... | 0 | 0 | 1 | 31 | 3.10 | 3.98 | 14.29 | 0.16 |
| week | 4242 | 11 | 2 | 4 | $\cdots$ | 5 | 2 | 2 | 31 | 3.10 | 3.38 | 10.29 | 0.16 |
| carlton | 871 | 0 | 0 | 6 | ... | 0 | 3 | 0 | 31 | 3.10 | 5.17 | 24.09 | 0.16 |
| midcap | 497 | 12 | 6 | 3 | ... | 0 | 4 | 3 | 31 | 3.10 | 3.75 | 12.69 | 0.16 |
| insurer. | 663 | 17 | 2 | 0 | $\ldots$ | 0 | 0 | 0 | 31 | 3.10 | 5.95 | 31.89 | 0.16 |
| total | 1352 | 1 | 6 | 10 | $\cdots$ | 0 | 1 | 0 | 31 | 3.10 | 3.93 | 13.89 | 0.16 |
| technology | 1648 | 4 | 7 | 0 | ... | 1 | 0 | 3 | 30 | 3.00 | 4.83 | 21.00 | 0.15 |
| alliance | 726 | 3 | 0 | 5 | ... | 0 | 0 | 0 | 30 | 3.00 | 6.22 | 34.80 | 0.15 |
| equipment | 645 | 2 | 0 | 22 | ... | 0 | 0 | 1 | 30 | 3.00 | 6.77 | 41.20 | 0.15 |
| below | 1502 | 1 | 2 | 0 | ... | 2 | 0 | 0 | 30 | 3.00 | 7.09 | 45.20 | 0.15 |
| operating | 1429 | 4 | 2 | 0 | ... | 0 | 0 | 0 | 30 | 3.00 | 7.50 | 50.60 | 0.15 |
| around | 4249 | 2 | 6 | 0 | ... | 0 | 1 | 1 | 30 | 3.00 | 4.35 | 17.00 | 0.15 |
| amid | 982 | 0 | 0 | 0 | ... | 17 | 4 | 4 | 30 | 3.00 | 5.33 | 25.60 | 0.15 |
| trading | 4118 | 6 | 3 | 2 | ... | 5 | 3 | 7 | 29 | 2.90 | 2.42 | 5.29 | 0.14 |


| lost | 1562 | 7 | 0 | 0 | . | 0 | 0 | 22 | 29 | 2.90 | 7.06 | 44.89 | 0.14 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| logica | 498 | 0 | 1 | 5 | ... | 0 | 2 | 0 | 29 | 2.90 | 5.57 | 27.89 | 0.14 |
| slightly | 820 | 2 | 0 | 0 | ... | 0 | 0 | 0 | 29 | 2.90 | 5.63 | 28.49 | 0.14 |
| media | 1948 | 0 | 7 | 8 | ... | 1 | 5 | 2 | 28 | 2.80 | 3.01 | 8.16 | 0.13 |
| banking | 1783 | 4 | 6 | 6 | ... | 1 | 0 | 3 | 28 | 2.80 | 3.05 | 8.36 | 0.13 |
| each | 863 | 2 | 0 | 0 | ... | 1 | 5 | 1 | 28 | 2.80 | 5.55 | 27.76 | 0.13 |
| passenger | 478 | 1 | 2 | 1 | ... | 0 | 0 | 0 | 28 | 2.80 | 6.80 | 41.56 | 0.13 |
| materials | 208 | 1 | 5 | 11 | ... | 0 | 0 | 4 | 28 | 2.80 | 3.68 | 12.16 | 0.13 |
| prudential | 578 | 3 | 2 | 2 | ... | 0 | 2 | 2 | 28 | 2.80 | 4.76 | 20.36 | 0.13 |
| just | 2572 | 1 | 0 | 0 | ... | 3 | 0 | 14 | 27 | 2.70 | 4.88 | 21.41 | 0.12 |
| ons | 363 | 10 | 9 | 8 | ... | 0 | 0 | 0 | 27 | 2.70 | 4.37 | 17.21 | 0.12 |
| abbey | 1326 | 2 | 2 | 4 | ... | 0 | 4 | 1 | 27 | 2.70 | 3.92 | 13.81 | 0.12 |
| marconi | 755 | 0 | 0 | 4 | ... | 0 | 2 | 0 | 27 | 2.70 | 5.29 | 25.21 | 0.12 |
| tsb | 567 | 0 | 2 | 0 | ... | 0 | 0 | 4 | 27 | 2.70 | 6.57 | 38.81 | 0.12 |
| ago | 1753 | 1 | 14 | 3 | ... | 0 | 4 | 1 | 27 | 2.70 | 4.30 | 16.61 | 0.12 |
| employment | 390 | 4 | 0 | 7 | ... | 0 | 0 | 2 | 27 | 2.70 | 3.53 | 11.21 | 0.12 |
| drugs | 1000 | 1 | 6 | 8 | ... | 1 | 4 | 3 | 27 | 2.70 | 2.79 | 7.01 | 0.12 |
| designer | 286 | 7 | 0 | 13 | ... | 1 | 1 | 5 | 27 | 2.70 | 4.37 | 17.21 | 0.12 |
| international | 1309 | 1 | 2 | 6 | - | 0 | 1 | 0 | 27 | 2.70 | 3.62 | 11.81 | 0.12 |
| petrol | 241 | 0 | 3 | 0 | ... | 0 | 0 | 0 | 27 | 2.70 | 7.54 | 51.21 | 0.12 |
| tobacco | 611 | 1 | 0 | 8 | ... | 0 | 0 | 4 | 27 | 2.70 | 4.45 | 17.81 | 0.12 |
| debt | 2782 | 1 | 1 | 2 | ... | 2 | 1 | 0 | 27 | 2.70 | 5.42 | 26.41 | 0.12 |
| european | 3888 | 2 | 3 | 7 | ... | 2 | 0 | 4 | 27 | 2.70 | 3.16 | 9.01 | 0.12 |
| biggest | 3352 | 1 | 10 | 8 | ... | 0 | 0 | 3 | 26 | 2.60 | 3.69 | 12.24 | 0.12 |
| management | 2163 | 3 | 7 | 1 | ... | 0 | 0 | 0 | 26 | 2.60 | 4.01 | 14.44 | 0.12 |
| lower | 3327 | 7 | 10 | 5 | ... | 0 | 0 | 3 | 26 | 2.60 | 3.60 | 11.64 | 0.12 |
| dropped | 846 | 8 | 0 | 0 | - | 0 | 2 | 16 | 26 | 2.60 | 5.34 | 25.64 | 0.12 |
| sun | 762 | 0 | 1 | 1 | ... | 0 | 0 | 0 | 26 | 2.60 | 6.54 | 38.44 | 0.12 |
| benchmark | 738 | 1 | 4 | 5 | - | 0 | 0 | 5 | 26 | 2.60 | 3.66 | 12.04 | 0.12 |
| less | 1296 | 0 | 0 | 1 | ... | 0 | 0 | 1 | 26 | 2.60 | 5.50 | 27.24 | 0.12 |
| despite | 1660 | 0 | 0 | 0 | ... | 16 | 5 | 0 | 26 | 2.60 | 5.06 | 23.04 | 0.12 |
| losses | 1703 | 5 | 5 | 2 | ... | 1 | 7 | 2 | 26 | 2.60 | 2.50 | 5.64 | 0.12 |
| goods | 725 | 2 | 1 | 5 | ... | 0 | 0 | 0 | 25 | 2.50 | 3.54 | 11.25 | 0.11 |
| next | 3849 | 1 | 5 | 3 | $\cdots$ | 0 | 0 | 0 | 25 | 2.50 | 3.47 | 10.85 | 0.11 |
| wall | 2041 | 0 | 12 | 0 | ... | 0 | 0 | 2 | 25 | 2.50 | 4.35 | 17.05 | 0.11 |
| worries | 1057 | 0 | 1 | 0 | $\cdot$ | 8 | 7 | 5 | 25 | 2.50 | 3.21 | 9.25 | 0.11 |
| companies | 3136 | 2 | 7 | 0 | ... | 0 | 0 | 0 | 25 | 2.50 | 4.60 | 19.05 | 0.11 |
| systems | 633 | 2 | 3 | 6 | ... | 0 | 0 | 0 | 25 | 2.50 | 4.17 | 15.65 | 0.11 |
| volumes | 508 | 0 | 2 | 1 | $\cdots$ | 0 | 0 | 1 | 24 | 2.40 | 6.22 | 34.84 | 0.10 |
| unemployment | 620 | 0 | 1 | 0 | ... | 0 | 0 | 1 | 24 | 2.40 | 4.58 | 18.84 | 0.10 |
| airways | 695 | 0 | 3 | 1 | ... | 0 | 0 | 0 | 24 | 2.40 | 5.91 | 31.44 | 0.10 |
| almost | 888 | 0 | 0 | 0 | ... | 0 | 0 | 2 | 24 | 2.40 | 4.95 | 22.04 | 0.10 |
| hbos | 535 | 0 | 1 | 0 | ... | 0 | 8 | 0 | 24 | 2.40 | 3.72 | 12.44 | 0.10 |
| results | 3369 | 1 | 7 | 0 | $\ldots$ | 4 | 2 | 4 | 24 | 2.40 | 2.46 | 5.44 | 0.10 |

## Appendix D

## Upward collocates ( 65 in total) of fell sorted by strength. Note that only collocates with strength $>=0.1$ are listed.

| $f_{\text {neighbourhood of shares }}$ |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $x^{\prime}$ | $f$ | -5 | -4 | -3 | ... | 3 | 4 | 5 | sum | $\mu$ | $\sigma$ | spread | strength |
| nomber | 213292 | 212 | 137 | 93 | ... | 374 | 684 | 481 | 4968 | 496.80 | 659.23 | 391128.36 | 41.68 |
| percent | 41091 | 67 | 83 | 55 | ... | 280 | 97 | 117 | 2698 | 269.80 | 611.02 | 336005.56 | 22.59 |
| to | 100446 | 33 | 22 | 19 | ... | 656 | 117 | 83 | 1445 | 144.50 | 223.32 | 44885.05 | 12.05 |
| and | 70898 | 96 | 106 | 167 | $\cdots$ | 182 | 25 | 39 | 780 | 78.00 | 65.94 | 3913.20 | 6.46 |
| after | 14663 | 21 | 22 | 49 | $\cdots$ | 295 | 32 | 72 | 627 | 62.70 | 84.67 | 6452.41 | 5.17 |
| shares | 12718 | 45 | 47 | 72 | $\ldots$ | 0 | 6 | 11 | 592 | 59.20 | 120.54 | 13077.96 | 4.88 |
| on | 41486 | 67 | 40 | 30 | $\ldots$ | 144 | 39 | 49 | 534 | 53.40 | 47.72 | 2049.44 | 4.39 |
| from | 20921 | 13 | 16 | 3 | $\cdots$ | 68 | 50 | 116 | 294 | 29.40 | 37.90 | 1292.44 | 2.37 |
| index | 6923 | 12 | 24 | 24 | ... |  | 0 | 3 | 266 | 26.60 | 52.47 | 2477.44 | 2.13 |
| said | 43413 | 41 | 47 | 55 | $\ldots$ | 5 | 8 | 76 | 249 | 24.90 | 27.63 | 686.89 | 1.99 |
| by | 27763 | 13 | 3 | 7 | ... | 6 | 8 | 33 | 247 | 24.70 | 53.33 | 2559.61 | 1.97 |
| year | 20779 | 47 | 31 | 17 | ... | 21 | 19 | 66 | 245 | 24.50 | 19.74 | 350.85 | 1.96 |
| 's | 38772 | 56 | 40 | 51 | ... | 6 | 12 | 18 | 226 | 22.60 | 22.36 | 449.84 | 1.80 |
| group | 9177 | 11 | 9 | 26 | $\ldots$ | 2 | 1 | 24 | 211 | 21.10 | 27.00 | 656.09 | 1.67 |
| while | 4566 | 19 | 30 | 49 | $\ldots$ | 8 | 32 | 2 | 185 | 18.50 | 18.56 | 310.05 | 1.45 |
| pounds | 13193 | 20 | 14 | 7 | $\cdots$ | 11 | 106 | 15 | 176 | 17.60 | 31.86 | 913.44 | 1.38 |
| month | 4454 | 3 | 8 | 7 | ... | 5 | 59 | 72 | 171 | 17.10 | 25.85 | 601.29 | 1.33 |
| firm | 6703 | 9 | 13 | 48 | ... | 2 | 0 | 13 | 169 | 16.90 | 26.43 | 628.69 | 1.32 |
| sales | 6854 | 16 | 11 | 12 | ... | 2 | 1 | 5 | 167 | 16.70 | 26.49 | 631.41 | 1.30 |
| last | 9343 | 29 | 12 | 7 | ... | 30 | 36 | 14 | 166 | 16.60 | 11.64 | 122.04 | 1.29 |
| pence | 5414 | 7 | 7 | 4 | $\cdots$ | 14 | 1 | 120 | 166 | 16.60 | 36.64 | 1208.04 | 1.29 |
| but | 15876 | 17 | 22 | 51 | $\ldots$ | 7 | 10 | 25 | 165 | 16.50 | 14.22 | 182.05 | 1.28 |
| points | 5674 | 4 | 2 | 1 | $\ldots$ | 15 | 3 | 10 | 156 | 15.60 | 37.36 | 1256.24 | 1.21 |
| prices | 4410 | 4 | 6 | 8 | ... | 2 | 4 | 1 | 152 | 15.20 | 31.92 | 917.16 | 1.18 |
| company | 9854 | 8 | 24 | 42 | ... | 5 | 2 | 14 | 140 | 14.00 | 14.88 | 199.40 | 1.07 |
| or | 6028 | 1 | 3 | 1 | ... | 43 | 84 | 2 | 138 | 13.80 | 27.94 | 702.56 | 1.06 |
| also | 7749 | 3 | 3 | 2 | $\cdots$ | 3 | 1 | 5 | 113 | 11.30 | 29.45 | 780.61 | 0.85 |
| at | 26824 | 12 | 23 | 11 | $\ldots$ | 9 | 6 | 24 | 108 | 10.80 | 7.90 | 56.16 | 0.81 |
| sector | 4905 | 15 | 10 | 28 | ... | 0 | 2 | 0 | 101 | 10.10 | 12.44 | 139.29 | 0.75 |
| ftse | 7271 | 10 | 33 | 18 | $\ldots$ | 1 | 4 | 4 | 101 | 10.10 | 10.60 | 101.09 | 0.75 |
| market | 13811 | 22 | 2 | 9 | $\ldots$ | 7 | 15 | 12 | 99 | 9.90 | 7.82 | 55.09 | 0.73 |
| stock | 5348 | 13 | 6 | 7 | $\ldots$ | 0 | 0 | 1 | 98 | 9.80 | 15.20 | 207.96 | 0.72 |
| over | 7758 | 17 | 9 | 0 | $\cdots$ | 9 | 4 | 2 | 97 | 9.70 | 13.38 | 161.01 | 0.71 |
| quarter | 5039 | 8 | 27 | 14 | $\ldots$ | 4 | 5 | 14 | 96 | 9.60 | 8.13 | 59.44 | 0.70 |
| than | 7647 | 4 | 0 | 1 | $\ldots$ | 20 | 6 | 1 | 88 | 8.80 | 17.34 | 270.56 | 0.64 |
| profits | 4784 | 10 | 5 | 4 | ... | 0 | 1 | 2 | 79 | 7.90 | 16.49 | 244.69 | 0.56 |
| investors | 4747 | 3 | 1 | 1 | $\ldots$ | 5 | 46 | 12 | 79 | 7.90 | 14.11 | 179.29 | 0.56 |
| first | 6737 | 17 | 12 | 6 | $\ldots$ | 4 | 4 | 35 | 78 | 7.80 | 11.14 | 111.76 | 0.55 |
| more | 7942 | 2 | 2 | 0 | ... | 0 | 1 | 13 | 69 | 6.90 | 14.63 | 192.69 | 0.48 |
| share | 5195 | 9 | 3 | 1 | ... | 0 | 0 | 2 | 63 | 6.30 | 10.42 | 97.81 | 0.43 |
| reuters | 11915 | 4 | 25 | 4 | ... | 1 | 0 | 0 | 62 | 6.20 | 8.72 | 68.36 | 0.42 |


| bank | 7507 | 3 | 13 | 16 | $\cdots$ | 0 | 5 | 12 | 59 | 5.90 | 5.84 | 30.69 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| about | 7456 | 1 | 1 | 2 | $\cdots$ | 3 | 9 | 20 | 53 | 5.30 | 7.24 | 47.21 |
| was | 23370 | 6 | 5 | 0 | $\cdots$ | 0 | 5 | 35 | 53 | 5.30 | 10.71 | 103.21 |
| new | 7606 | 5 | 5 | 5 | $\cdots$ | 4 | 2 | 13 | 45 | 4.50 | 4.25 | 16.25 |
| price | 4501 | 3 | 3 | 18 | $\cdots$ | 0 | 1 | 2 | 44 | 4.40 | 5.76 | 29.84 |
| business | 6335 | 11 | 2 | 6 | $\cdots$ | 1 | 3 | 1 | 43 | 4.30 | 4.16 | 15.61 |
| uk | 5422 | 11 | 6 | 5 | $\cdots$ | 0 | 3 | 0 | 37 | 3.70 | 4.00 | 14.41 |
| growth | 6459 | 7 | 7 | 3 | $\cdots$ | 0 | 1 | 0 | 31 | 3.10 | 3.38 | 10.29 |
| half | 4595 | 9 | 2 | 11 | $\cdots$ | 0 | 1 | 1 | 31 | 3.10 | 4.23 | 16.09 |
| financial | 4692 | 6 | 4 | 2 | $\cdots$ | 4 | 3 | 7 | 29 | 2.90 | 2.38 | 5.16 |
| were | 11227 | 5 | 4 | 0 | $\cdots$ | 3 | 1 | 14 | 27 | 2.70 | 4.40 | 17.41 |
| london | 10989 | 2 | 4 | 3 | $\cdots$ | 3 | 0 | 2 | 25 | 2.50 | 1.84 | 3.16 |
| up | 13101 | 6 | 2 | 5 | $\cdots$ | 1 | 3 | 5 | 24 | 2.40 | 2.22 | 4.44 |

## Appendix E

Sentiments extracted from news articles in the Reuters Newsfeed during August 2006 together with the corresponding values of the FTSE financial indices in U.K.

|  |  |  | All News |  |  |  | UK News |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Date | $\begin{gathered} \text { FTSE } \\ 100 \end{gathered}$ | $\begin{aligned} & \text { FTSE } \\ & \text { ALL } \end{aligned}$ | Raw +ve | $\begin{aligned} & \text { Raw } \\ & \text {-ve } \end{aligned}$ | $\begin{aligned} & \text { LG } \\ & \text { +ve } \end{aligned}$ | $\begin{aligned} & \text { LG } \\ & \text {-ve } \end{aligned}$ | $\begin{aligned} & \text { Raw } \\ & \text { +ve } \end{aligned}$ | $\begin{gathered} \text { Raw } \\ \text {-ve } \end{gathered}$ | $\begin{aligned} & \text { LG } \\ & +\mathbf{v e} \end{aligned}$ | $\begin{aligned} & \text { LG } \\ & \text {-ve } \end{aligned}$ |
| 01/08/2006 | 5,880.80 | 2,983.52 | 37173 | 20192 | 1590 | 910 | 1876 | 607 | 170 | 76 |
| 02/08/2006 | 5,932.10 | 3,007.40 | 38872 | 19560 | 1783 | 618 | 1684 | 433 | 168 | 59 |
| 03/08/2006 | 5,838.40 | 2,961.64 | 39971 | 22873 | 1632 | 971 | 1779 | 694 | 161 | 121 |
| 04/08/2006 | 5,889.40 | 2,987.70 | 18961 | 10555 | 838 | 327 | 1438 | 425 | 122 | 39 |
| 07/08/2006 | 5,828.80 | 2,958.50 | 20243 | 12366 | 750 | 493 | 605 | 355 | 90 | 57 |
| 08/08/2006 | 5,818.10 | 2,954.32 | 34641 | 19930 | 1364 | 633 | 1465 | 655 | 122 | 84 |
| 09/08/2006 | 5,860.50 | 2,973.85 | 33052 | 19565 | 1193 | 728 | 1311 | 579 | 144 | 72 |
| 10/08/2006 | 5,823.40 | 2,954.94 | 34369 | 20331 | 1075 | 1034 | 1399 | 863 | 94 | 212 |
| 11/08/2006 | 5,820.10 | 2,953.31 | 23019 | 15191 | 1032 | 601 | 1092 | 548 | 143 | 64 |
| 14/08/2006 | 5,870.90 | 2,977.14 | 26263 | 16224 | 968 | 410 | 792 | 360 | 122 | 58 |
| 15/08/2006 | 5,897.90 | 2,991.00 | 28595 | 15996 | 1420 | 698 | 1054 | 522 | 152 | 87 |
| 16/08/2006 | 5,896.60 | 2,994.78 | 24029 | 12742 | 1360 | 479 | 1002 | 562 | 153 | 79 |
| 17/08/2006 | 5,900.40 | 2,998.34 | 24057 | 13290 | 1119 | 553 | 927 | 537 | 92 | 88 |
| 18/08/2006 | 5,903.40 | 3,001.06 | 16025 | 9783 | 898 | 487 | 839 | 389 | 101 | 54 |
| 21/08/2006 | 5,915.20 | 3,005.77 | 19601 | 11017 | 709 | 611 | 902 | 411 | 107 | 77 |
| 22/08/2006 | 5,902.60 | 3,000.22 | 22068 | 11157 | 1088 | 488 | 935 | 389 | 123 | 64 |
| 23/08/2006 | 5,860.00 | 2,980.55 | 19614 | 11260 | 752 | 684 | 878 | 501 | 95 | 85 |
| 24/08/2006 | 5,869.10 | 2,984.53 | 22890 | 12802 | 1047 | 768 | 1070 | 520 | 99 | 75 |
| 25/08/2006 | 5,878.60 | 2,990.21 | 15629 | 10114 | 830 | 450 | 1099 | 400 | 171 | 40 |
| 29/08/2006 | 5,888.30 | 2,996.09 | 22779 | 13030 | 990 | 567 | 1083 | 549 | 131 | 74 |
| 30/08/2006 | 5,929.30 | 3,015.89 | 23579 | 12876 | 1047 | 492 | 1465 | 527 | 147 | 60 |
| 31/08/2006 | 5,906.10 | 3,007.51 | 25033 | 13282 | 1288 | 578 | 1755 | 650 | 152 | 64 |

## Appendix F

Sentiments extracted from news articles in the Reuters Newsfeed during August 2006 together with the corresponding values of the financial indices in U.S.

|  |  |  |  | All News |  |  |  | US News |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Date | Dow Jones | Nasdaq | $\begin{gathered} \mathbf{S \& P} \\ 500 \end{gathered}$ | $\begin{aligned} & \text { Raw } \\ & +\mathbf{v e} \end{aligned}$ | $\begin{gathered} \text { Raw } \\ \text {-ve } \end{gathered}$ | $\begin{gathered} \text { LG } \\ +\mathbf{v e} \end{gathered}$ | $\begin{aligned} & \text { LG } \\ & \text {-ve } \end{aligned}$ | $\begin{gathered} \text { Raw } \\ \text { +ve } \end{gathered}$ | $\begin{gathered} \text { Raw } \\ \text {-ve } \end{gathered}$ | $\begin{gathered} \text { LG } \\ +\mathbf{v e} \end{gathered}$ | $\begin{aligned} & \text { LG } \\ & \text {-ve } \end{aligned}$ |
| 01/08/2006 | 11,125.73 | 2,061.99 | 1,270.92 | 37173 | 20192 | 1590 | 910 | 5072 | 2092 | 741 | 418 |
| 02/08/2006 | 11,199.92 | 2,078.81 | 1,277.41 | 38872 | 19560 | 1783 | 618 | 5298 | 1676 | 779 | 217 |
| 03/08/2006 | 11,242.59 | 2,092.34 | 1,280.27 | 39971 | 22873 | 1632 | 971 | 5767 | 2287 | 820 | 511 |
| 04/08/2006 | 11,240.35 | 2,085.05 | 1,279.36 | 18961 | 10555 | 838 | 327 | 1910 | 765 | 230 | 116 |
| 07/08/2006 | 11,219.38 | 2,072.50 | 1,275.77 | 20243 | 12366 | 750 | 493 | 3292 | 1568 | 413 | 278 |
| 08/08/2006 | 11,173.59 | 2,060.85 | 1,271.48 | 34641 | 19930 | 1364 | 633 | 5092 | 1920 | 645 | 345 |
| 09/08/2006 | 11,076.18 | 2,060.28 | 1,265.95 | 33052 | 19565 | 1193 | 728 | 4702 | 1779 | 516 | 335 |
| 10/08/2006 | 11,124.37 | 2,071.74 | 1,271.81 | 34369 | 20331 | 1075 | 1034 | 4444 | 2093 | 425 | 425 |
| 11/08/2006 | 11,088.02 | 2,057.71 | 1,266.74 | 23019 | 15191 | 1032 | 601 | 3103 | 1682 | 449 | 336 |
| 14/08/2006 | 11,097.87 | 2,069.04 | 1,268.21 | 26263 | 16224 | 968 | 410 | 4169 | 1659 | 557 | 241 |
| 15/08/2006 | 11,230.26 | 2,115.01 | 1,285.58 | 28595 | 15996 | 1420 | 698 | 4413 | 1777 | 787 | 406 |
| 16/08/2006 | 11,327.12 | 2,149.54 | 1,295.43 | 24029 | 12742 | 1360 | 479 | 3965 | 1764 | 668 | 271 |
| 17/08/2006 | 11,334.96 | 2,157.61 | 1,297.48 | 24057 | 13290 | 1119 | 553 | 3094 | 1665 | 463 | 338 |
| 18/08/2006 | 11,381.47 | 2,163.95 | 1,302.30 | 16025 | 9783 | 898 | 487 | 2455 | 1339 | 336 | 280 |
| 21/08/2006 | 11,345.04 | 2,147.75 | 1,297.52 | 19601 | 11017 | 709 | 611 | 2691 | 1341 | 304 | 332 |
| 22/08/2006 | 11,339.84 | 2,150.02 | 1,298.82 | 22068 | 11157 | 1088 | 488 | 3081 | 1413 | 454 | 262 |
| 23/08/2006 | 11,297.90 | 2,134.66 | 1,292.99 | 19614 | 11260 | 752 | 684 | 2632 | 1752 | 322 | 359 |
| 24/08/2006 | 11,304.46 | 2,137.11 | 1,296.06 | 22890 | 12802 | 1047 | 768 | 3623 | 1876 | 557 | 394 |
| 25/08/2006 | 11,284.05 | 2,140.29 | 1,295.09 | 15629 | 10114 | 830 | 450 | 2149 | 1095 | 297 | 223 |
| 28/08/2006 | 11,352.01 | 2,160.70 | 1,301.78 | 5192 | 2652 | 756 | 509 | 2594 | 1369 | 363 | 254 |
| 29/08/2006 | 11,369.94 | 2,172.30 | 1,304.28 | 22779 | 13030 | 990 | 567 | 2837 | 1716 | 325 | 314 |
| 30/08/2006 | 11,382.91 | 2,185.73 | 1,305.37 | 23579 | 12876 | 1047 | 492 | 3350 | 1443 | 389 | 200 |
| 31/08/2006 | 11,381.15 | 2,183.75 | 1,303.82 | 25033 | 13282 | 1288 | 578 | 3805 | 1650 | 593 | 343 |

## Appendix G

Sentiments extracted from the Xinhuanet news articles during July 2007 together with the corresponding values of three of the major financial indices in China.

| Date | Shenzhen | Shanghai | Hang Seng | Raw <br> +ve | Raw <br> -ve | LG <br> +ve | LG <br> -ve |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\mathbf{0 2 / 0 7 / 2 0 0 7}$ | $12,475.16$ | $3,836.29$ | $21,961.94$ | 417 | 127 | 19 | 7 |
| $\mathbf{0 3 / 0 7 / 2 0 0 7}$ | $12,761.06$ | $3,899.72$ | $22,151.14$ | 423 | 223 | 22 | $\mathbf{9}$ |
| $\mathbf{0 4 / 0 7 / 2 0 0 7}$ | $12,509.14$ | $3,816.17$ | $22,218.55$ | 635 | 183 | 28 | 5 |
| $\mathbf{0 5 / 0 7 / 2 0 0 7}$ | $11,783.58$ | $3,615.87$ | $22,252.99$ | 441 | 230 | 16 | 28 |
| $\mathbf{0 6 / 0 7 / 2 0 0 7}$ | $12,395.35$ | $3,781.35$ | $22,531.74$ | 281 | 188 | 17 | 30 |
| $\mathbf{0 9 / 0 7 / 2 0 0 7}$ | $12,846.41$ | $3,883.22$ | $22,817.43$ | 430 | 127 | 26 | 5 |
| $\mathbf{1 0 / 0 7 / 2 0 0 7}$ | $12,682.04$ | $3,853.02$ | $22,885.84$ | 627 | 182 | 28 | 15 |
| $\mathbf{1 1 / 0 7 / 2 0 0 7}$ | $12,785.67$ | $3,865.72$ | $22,607.02$ | 704 | 212 | 31 | 19 |
| $\mathbf{1 2 / 0 7 / 2 0 0 7}$ | $12,850.74$ | $3,915.99$ | $22,809.02$ | 453 | 316 | 37 | 34 |
| $\mathbf{1 3 / 0 7 / 2 0 0 7}$ | $12,816.42$ | $3,914.40$ | $23,099.29$ | 488 | 167 | 43 | 17 |
| $\mathbf{1 6 / 0 7 / 2 0 0 7}$ | $12,788.20$ | $3,905.30$ | $22,953.94$ | 603 | 200 | 81 | 11 |
| $\mathbf{1 7 / 0 7 / 2 0 0 7}$ | $12,759.98$ | $3,896.19$ | $23,057.30$ | 296 | 97 | 24 | 11 |
| $\mathbf{1 8 / 0 7 / 2 0 0 7}$ | $12,841.51$ | $3,930.06$ | $22,841.92$ | 476 | 126 | 65 | 13 |
| $\mathbf{1 9 / 0 7 / 2 0 0 7}$ | $12,790.94$ | $3,912.94$ | $23,016.20$ | 351 | 104 | 31 | 14 |
| $\mathbf{2 0 / 0 7 / 2 0 0 7}$ | $13,417.96$ | $4,058.85$ | $23,291.90$ | 765 | 145 | 61 | 16 |
| $\mathbf{2 3 / 0 7 / 2 0 0 7}$ | $14,139.27$ | $4,213.36$ | $23,365.56$ | 649 | 160 | 38 | 12 |
| $\mathbf{2 4 / 0 7 / 2 0 0 7}$ | $14,190.97$ | $4,210.33$ | $23,472.88$ | 549 | 139 | 23 | 6 |
| $\mathbf{2 5 / 0 7 / 2 0 0 7}$ | $14,403.08$ | $4,323.97$ | $23,362.18$ | 302 | 147 | 12 | 12 |
| $\mathbf{2 6 / 0 7 / 2 0 0 7}$ | $14,619.74$ | $4,346.46$ | $23,211.69$ | 191 | 88 | 11 | 2 |
| $\mathbf{2 7 / 0 7 / 2 0 0 7}$ | $14,614.10$ | $4,345.36$ | $22,570.41$ | 80 | 71 | 13 | 23 |


[^0]:    ${ }^{1}$ Investerwords.com - www.investorwords.com,date accessed: 10 February 2003
    ${ }^{2}$ Investerwords.com - www.investorwords.com,date accessed: 10 February 2003

[^1]:    ${ }^{3}$ Generic Information-based Decision Assistant (GIDA IST-2000-31123) is an EU-sponsored project, involving partners from Berlin, Madrid and London.

[^2]:    ${ }^{4}$ Investerwords.com - www.investorwords.com,date accessed: 10 February 2003

[^3]:    ${ }^{1}$ Oxford Dictionary Online, http://www.oed.com, date accessed: 10 December 2006
    ${ }^{2}$ Peter Turney, http://purl.org/peter.turney/ml text orientation apps.html, date accessed: 12 December 2006

[^4]:    ${ }^{3}$ Fundamental Analysis, http://en.wikipedia.org/wiki/Fundamental analysis, date accessed: 12 December 2006
    ${ }^{4}$ Technical Analysis, http://en.wikipedia.org/wiki/Technical analysis, date accessed: 12 December 2006
    ${ }^{5}$ Sentiment Analysis, http://en.wikipedia.org/wiki/Sentiment_analysis, date accessed: 12 December 2006

[^5]:    ${ }^{6}$ Collocation, http://en.wikipedia.org/wiki/Collocation, date accessed: 15 December 2006

[^6]:    ${ }^{1}$ http://projects.ldc. upenn.edu/Chinesel, date accessed: 20 January 2006.

[^7]:    ${ }^{2}$ Riskglossary, http://www.riskglossary.com/link/volatility.htm, date accessed: 12 August 2007

[^8]:    ${ }^{3}$ The Standford NLP Group，hitp：／／nlp．stanford．edu／software／segmenter．shtml，date accessed： 10 July 2007
    ${ }^{4}$ Xinhuanet，http：／／www，xinhuanet．com，date accessed： 12 July 2007.

[^9]:    ${ }^{5}$ The time series component of SATISFI is written by other colleagues in the GIDA project, and my main contribution were: 1) the Text Analysis Component that extracts sentiments from news articles, and represents them as sentiment time series; 2) System integration.

[^10]:    ${ }^{1}$ IBM Solutions Grid for Business Partners: Helping IBM Business Partners to Grid-enable applications for the next phase of e-business on demand, available at:
    http://www-304.ibm.com/jct09002c/isv/marketing/emerging/grid_wp.pdf, date accessed: 2 November 2006

