PROCEEDINGS OF SCIENCE

Multi-lingual Sentiment Analysis of Financial News Streams

Khurshid Ahmad*
Department of Computer Science, Trinity College
Dublin 2, EIRE
E-mail: kahmad@cs.tcd.ie

David Cheng
Department of Computing, University of Surrey
Guildford, Surrey, GU2 7XH, United Kingdom
E-mail: d.cheng@surrey.ac.uk

Yousif Almas
Department of Computing, University of Surrey
Guildford, Surrey, GU2 7XH, United Kingdom
E-mail: y.almas@surrey.ac.uk

Encouraged by the feasibility demonstration that a relatively low-cost grid environment can speed up the processing of continuous text streams of financial news in English, we have attempted to replicate our methods for automatic sentiment analysis in two major languages of the world – Arabic and Chinese. We show that our *local grammar* approach, developed on an archive of English (Indo-European language) texts, works equally on the typologically different Chinese (Sino-Asiatic) and Arabic (Semitic) languages.

Grid Technology for Financial Modeling and Simulation
February 3/4, 2006 – Palermo, (Italy)

* Speaker

1. Introduction

Literature on financial economics and sociology of financial markets suggests that 'the number of items of quantitative and qualitative information available to well-equipped actor is, in effect, infinite, yet the capacity of any agencement [humans, machines, algorithms, location,..] to apprehend and to interpret that data is finite' [1]. Both the qualitative and quantitative information is sent and received in different modalities – numerical, graphical and, increasingly in written text and in speech for instance. Much of the information about trading in a whole range of markets now literally *streams* in through trading terminals and through websites. One can receive raw numerical data about any financial instrument and *render* the data into a time series. The time series can, in principle, be then visualised and manipulated using grid-based systems that were initially developed for use in a number of scientific and engineering applications: The inherent parallel architecture of grid computing systems implies that the techniques developed for studying univariate and multivariate data on serial architectures can readily be ported over: discrete and continuous transformations ranging from Fourier transformations to wavelet analysis, and many varieties of auto-regressive models, can be parallelised with little or no effort.

Numerical data analysis is increasing, but indirectly, takes into account of data that may not have its origins in numerical data like the price and volume of a financial instrument: the key example here is the work of Granger and Engle [2, 3] on the volatility analysis of high-frequency financial data. The GARCH analysis focuses on the asymmetry in the distribution of innovations – the loosely called *random* component of volatility of a frequently-traded financial instrument. Engle (citing Nelson) has argued that 'volatility could respond asymmetrically to past forecast errors [....] negative returns seemed to be more important predictors of volatility than positive returns. Large price declines forecast greater volatility than similarly large price increases.' Engle has introduced the notion of the *news impact curve* [4] – negative news has longer lasting effect on financial instruments than the positive news.

But the term 'news' in 'news impact curve/analysis' is used in an esoteric and indirect manner – a variety of information proxies are used here: The proxies include timings of the announcements of various macro economic indicators: gross domestic product, non-farm payroll, retail sales and so on. Anderson *et al* have noted that the proxies can be rendered as a time series and correlated with the time series of cardinal values of prices, volumes and so on [5]. The authors have drawn six interesting conclusions based on their analysis: (i) news announcements matter and mostly have an immediate impact; (ii) the timing of the announcements matter; (iii) volatility adjusts to news gradually; (iv) 'pure announcement effects are present in volatility; (v) the effects of the announcements are generally 'asymmetric' in that the responses (or *innovations*) vary with the 'sign' of the news; and, (vi) the effect on traded volume persists longer

than on prices. The inference from the news impact curve have been analysed in a socio-behavioural context and the argument here is that sentiment may be expressed through action [1], in particular that

(a) panic buying and selling of financial instruments by the investors and traders, and

(b) the sometimes complacent attitude of the regulators, are good examples of economic, social and political action by individuals and groups.

It is not clear in the work of Engel and others whether the contents of a news report were analysed automatically or indeed analysed at all.

There are, however, a number of instance wherein researchers analyse the *content* of the news in addition to the timing of the news; the 'timing', in itself, is a difficult concept – is it the timing of the event being reported, the time when the news was written/read, the time at which the news is released, or is it the time when the reader receives and/or reads the news? The definition of 'news' notwithstanding a number of papers have appeared where the 'sentiment' – good news or bad news as articulated in a news report is correlated with the value of a given instrument: foreign exchange in DeGennaro and Shrieves [6] and company performance in Koppel and Shtrimberg [7] are a good example of sentiment analysis. These authors rely on their intuition about good/bad sentiment words and conduct a semi-manual analysis of texts. Given that the financial markets are continuously evolving with new instruments, new players and the expanding group of stakeholders, it appears that efforts should be made to create a framework where sentiments can be extracted without relying on the intuition of the analysts as to what is *good* or *bad* news. This is not an easy task and one can argue that it is an impossible task.

The sentiment is expressed in the news and views that emanate for and on behalf of the members in free natural language writing and speech excerpts. Over the years financial writing has matured and a mature writing style invariably develops its own widely used formats – comprising how information is ordered within a text such that readers with different information-extraction skills will be able to extract information with a minimal of effort. This requires minimizing ambiguity. The history of writing shows that this is achieved by not only choosing a small vocabulary but also the manner in which the vocabulary items are ordered in order to make up phrases and sentences [8]. This restriction on the language of a specialism has been exploited by us to find sentence templates that invariably contain sentiment bearing words in the context of financial instruments.

2. A framework for sentiment analysis

Consider texts written about financial markets in English. It is usually possible for a native speaker of English to instantly recognise that the text is in English – just by

looking at the relatively frequent use of the so-called grammatical words (loosely called stop-words also), *the, a, an, and, but, if, on, in* and so forth. Ordinary readers also recognise familiar aspects of the vocabulary of finance – *profit, loss, shares, stocks* and so forth. An expert in finance, whether native speaker or not, will recognise the fact that the text is about finance by instantly recognising keywords in finance and by relatively easily discerning the meaning of new words of the financial vocabulary. The native speaker is 'surprised' by the prolixity of the financial vocabulary.

We have developed a method for identifying the words that may surprise a native speaker by comparing the distribution of all the words in a collection of randomly sampled financial texts with that of the same words in a reference collection of texts. More prolific keywords in financial texts, the chances are that such a word will be less prolific in general language texts. Once we have identified keywords, based on a statistical criteria in our training collection of texts, then we look at the neighbourhood of these keywords; and, then look at the neighbourhood of the two word pair and so on. This neighbourhood, established on strict statistical criteria, yields information bearing sentences in the financial domain and, it turns out, sentences that typically carry sentiment information. These patterns are then used to build a finite state automaton. This automaton is then tested on an unseen set of texts – and the results vis-à-vis sentiment analysis are quite good. We have reported about our sentiment analysis framework elsewhere in detail [9]: The algorithm is summarised below –

I. Select training corpus (T_C) and a general language corpus (G_C);
II. Extract key words automatically:
    a. Identify candidate terms T_i by contrasting frequency (f_i) of T_i in T_C and G_C to obtain w_i;
    b. Apply statistical criteria (z-score for f_i and w_i);
III. Extract key collocates automatically;
IV. Extract local grammar using collocation and relevance feedback;
V. Assert the grammar as a finite state automaton.

The only non-automatic action required here is the choice of the training corpus (T_C) and the general language corpus (G_C).

2.1. English Sentiment Extraction

We have selected a training corpus of financial texts (Reuters RCV1 is available free for academic usage and comprises over 800, 000 texts, each containing between 200-400 words) and a corpus of general language (equally widely available *British National Corpus* – the *BNC*) to contrast the use of lexical items in the training corpus. The asymmetric choice of certain words in the training corpus, when compared with its use in everyday general language, typically indicates that asymmetrically chosen words are candidate terms. For instance, we found candidates *share* and *percent* in our training corpus. A small set of candidate terms are used to form compound words and the compound words show an asymmetry in the frequency of a set of preferential words

used in their vicinity: so *share* has preferential neighbours *share price* and *market share* and *percent* has companions *up, down, rise* and *fall*. We then go on to find complex collocation patterns that help in unambiguously detecting the sentiment words – this is revealed to us during the evaluation phase after processing 800,000 texts and comparing them with the 4,000 or so texts in the BNC (see Figure 1 below):

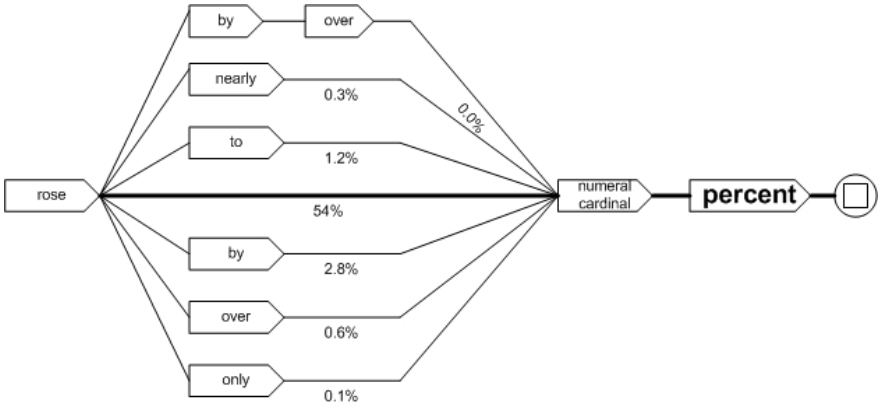

Figure 1: A local grammar for *percent* derived from a training corpus.

2.2. Arabic Sentiment Extraction

Arabic language has received limited attention and was until recently neglected in terms of the natural language processing research. Arabic is a Semitic language and its grammar and lexis is substantially different languages in, say, the Indo-European family (e.g. English, German, French, Hindi). The structure of Arabic words is more complex compared to English for instance. Arabic words are generated through a combination of a root (3 or more consonants) with a pattern, hence a root-and-pattern morphology [10][*].

In order to test the efficacy of our algorithm we need a general language corpus and a special language corpus: Our financial corpus comprised 8,815 texts and 1.48 million tokens published from March to August 2005 and posted by Reuters Arabic service. Unlike English though a general language corpus is not available to hand – as the business of creating 'representative corpora' of a language is difficult as demonstrated by the builders of the British National Corpus. We had to compile a general language corpus of Modern Standard Arabic (MSA). The MSA corpus comprises 2.6 million tokens in texts written between 1980 and 2005. The candidate 'terms' were extracted by comparing the asymmetric use of these words in the two corpora:

[*] Prefixes and suffixes can be attached to each word for delivering grammatical meanings and word can be inflected when refereeing to different numbers and genders. Vowels are omitted in Arabic texts unless there is a fear of ambiguity or for educational purposes. Average reader should be able to determine the vowels from the context. This omission causes many words with similar lexical structure to be different semantically causing ambiguity in parsing or pattern matching.

Word	Translit	Gloss	Freq	Relative Freq	Freq Z-score	Weird Z-score
المئة	al-meaa	percent	9594	0.0064	23.4	1.631
الدولار	al-dollar	the-dollar	4740	0.0031	11.53	0.301
بالمئة	bel-meaa	by-percent	4328	0.0029	10.52	5.218
مؤشر	moasher	index	3655	0.0024	8.887	0.260

Table 1: The distribution of candidate terms in our Arabic financial corpus.

The collocation analysis of the above words with others led us to our local grammar (Figure 2):

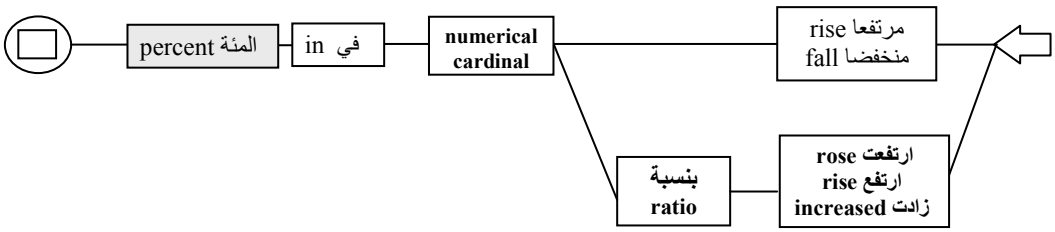

Figure 2: A local grammar for *almeaa* (المئة, *percent*) extracted automatically using our method.

2.3. Chinese Sentiment Extraction

For Chinese we used the Ming Bao Financial news text (1.56 million words) published in Hong Kong between Aug 2000 and July 2001. There are more readily available corpora in Chinese as compared to Arabic but still fewer Chinese corpora are available than is the case for English. We have instead used frequency lists from the TaBE (Localization for Taiwan and Big5 Encoding) Project and the LDC Chinese Resources [11, 12], comprising 5.96 million tokens. These corpora, from which the frequency lists were obtained, appear to be representative. Again, the candidate 'terms' were extracted by comparing the asymmetric use of these words in the two corpora (see Table 2). The most frequent term, like English and Arabic, was percent (百分之, *bai fen zhi*).

Word	Translit	Gloss	Freq	Relative Freq	Freq Z-score	Weird Z-score
報道	bao dao	report	3866	0.0025	4.23	38.37
百分之	bai fen zhi	percent	6403	0.0041	7.14	11.97
盈利	ying li	profit	2974	0.0019	3.20	6.36
收購	shou gou	acquisition	2673	0.0017	2.86	1.47
增長	zeng zhang	Increase	2584	0.0017	2.76	1.08

Table 2: The distribution of candidate terms in our Chinese financial corpus.

The collocation analysis led us to the following local grammar (Figure 3):

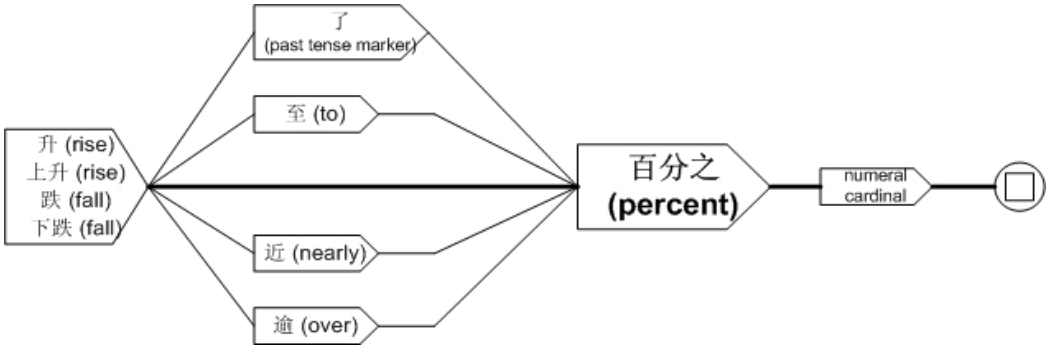

Figure 3: A local grammar for *percent* (*百分之, bai fen zhi*) extracted automatically from our method.

3. Grid-based Analysis

The above local grammar is used to extract 'true' sentiment from 'raw' sentiment – some sentiment words like *rose* or *fall* may be used as a name rather than as a verb. Our local grammar rejects such spurious use. The traffic volume is quite high – over a five day period Reuters supplied 15,668 financial news stories comprising 4,782,423 tokens – and requires a grid implementation to speed up the processing time of 1 Hour CPU time to 6 minutes on a 64 node processor. We had put together all the news produced within one hour – much the same as tick-data is compressed: Figure 4 below shows that in the 1st hour we had 9205 tokens in 35 news items (it was 01:00 hours on 3/1/2005) these contained 134 'raw' sentiment words and the filtration due to the local grammar led to 15 sentiment words. The peak in terms of number of news items supplied by Reuters was on the 61st hour – where 534 news items arrived within an hour comprising 233,181 tokens: 3085 raw sentiment words and 534 'actual' sentiment words!

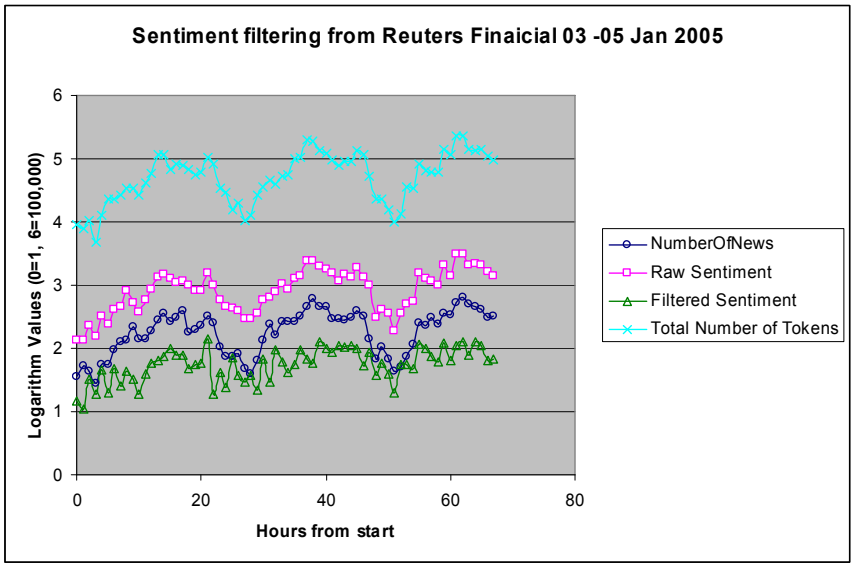

Figure 4: The 'filtration' of sentiments using the local grammar (Figure 1) for the period 03 Jan – 05 Jan 2005.

We are currently exploring how to *align* our series with the movement of financial instruments. We now discuss the effect of the use of the algorithm for extracting local grammar in Arabic and Chinese.

4. Conclusion and Future Work

The local grammar patterns in all three languages appear to extract most sentiment bearing phrases. This for us is a remarkable demonstration of a language of special purpose that transcends different language typologies (Indo-European → Sino-Asiatic → Semitic). We have conducted an initial manual evaluation and found the accuracy of extraction to range between 60-75%. More work is needed in this context.

The application of grid technology is essential due to the sheer volume of news; we have only used one news vendor at a time, but in the real world of financial trading more than one news source is used. Realistically a well-designed grid system can only cope with the volume. Once our extraction method is evaluated more robustly, then the use of such a technique in real time is even more dependent on a grid kind of environment.

References

[1] I. Hardie & D. MacKenzie, *An Economy of Calculation: Agencement and Distributed Cognition in a Hedge Fund* (2005) Available http://www.sps.ed.ac.uk/staff/mackenzie.html.
[2] C. W. J. Granger, *Time Series Analysis, Cointegration, and Applications, American Economic Review* (2004) 94(3), pp 421-425.
[3] R. Engle, *Risk and volatility: Econometric models and financial practice*, *American Economic Review* (2004) 94(3), pp 405-420.
[4] R. Engle & V. K. Ng, *Measuring and Testing the Impact of News on Volatility*, *Journal of Finance* (1993) 48(5), pp 1749—1777.
[5] T. G. Andersen, T. Bollerslev, F. X. Diebold & C. Vega, *Micro effects of macro announcements: Real time price discovery in foreign exchange*, NBER Working Paper 8959 (2002)
[6] R. DeGennaro & R. Shrieves, *Public information releases, private information arrival and volatility in the foreign exchange market*, *Journal of Empirical Finance* (1997) 4(4), pp 295–315.
[7] M. Koppel & I. Shtrimberg, *Good News or Bad News? Let the Market Decide*, in *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, Palo Alto: AAAI Press (2004) pp. 86-88.
[8] K. Ahmad, *Neologisms, Nonces and Word Formation*, in (Eds.) U. Heid, S. Evert, E. Lehmann & C. Rohrer, proceedings of *the 9th EURALEX Int. Congress*. Munich: Universitat Stuttgart (2000) Vol II, pp 711-730.
[9] K. Ahmad, L. Gillam & D.Cheng, Society *Grids*, In (Eds.) Simon Cox and David Walker. Proceedings of the *UK e-Science All Hands Meeting 2005.* 18-21 September, Nottingham UK. Swindon: EPSRC Sept (2005) pp 923-930.
[10] E. Badawi, M. Carter & A. Gully, *Modern Written Arabic - A Comprehensive Grammar*, London, Routledge (2004).
[11] P. H. Hsiao, T. H. Hsieh, K. S. Tan, C.-H. Tsai, & W. Yeh, *Localization library for Taiwan and BIG5 encoding (libtabe) (Version 0.1-3)* [Computer programming library], (2000) Available http://sourceforge.net/projects/libtabe/
[12] LDC Chinese Resources, Available http://projects.ldc.upenn.edu/Chinese/, accessed 20 Jan 2006.