Qualitative Information in Finance: Natural Language Processing and Information Extraction.

Marco Costantino, Russell J. Collingham, Richard G. Morgan
Laboratory for Natural Language Engineering
Department of Computer Science
University of Durham
Science Laboratories, South Road
DURHAM, DH1 3LE, U.K.
Tel. +44 191 374 2549, Fax. +44 191 374 2560
e-mail: marco.costantino@durham.ac.uk

July 9, 1996

Abstract

Quantitative data are today largely processed by computer programs based on traditional or artificial intelligence techniques including statistics, neural networks, genetic algorithms, etc.

Conversely, little progress has been made for the processing of *qualitative* information, which is mainly represented by financial articles from on-line news agencies or from financial newspapers. As a result, financial operators nowadays suffer from qualitative data-overload.

This article describes the importance of qualitative information in the financial operators' investment decision-making process and how natural language processing can be successfully used for processing and analysing such information. The paper focuses on information extraction which can identify specific kinds of information in a source article, producing a list of relevant templates or summaries of the original text. Information extraction can significantly reduce the qualitative data-overload from which traders, brokers, fund managers etc. suffer. The templates produced can be used for a "meta-analysis" of the effects of news on price behaviour. Finally, Natural Language Processing is briefly compared to other artificial intelligence techniques which are widely employed in finance: Neural Networks and Expert Systems.

1. Introduction

Every financial player in the market (fund managers, brokers, bank analysts, merchant banks etc.) with the exclusion of information providers has one single goal: maximise their investment by buying and selling securities at the right time.

The process that leads to investment decisions is the crucial aspect for the financial operators who need to support their decisions with the largest possible amount of relevant information, which can be *quantitative* or *qualitative*.

Quantitative information is information which can be easily expressed in numeric format (for example historical price data, prices of raw materials, data on inflation, currency exchanges etc.), whereas *qualitative* information cannot be accurately translated into numbers, for example a sentence such as "there are fears of an increase in the German interest rates". Quantitative and qualitative information can be either *historical* (e.g. time series of shares prices), or *real-time*, from on-line news agencies[1].

Financial operators nowadays have access to a huge and vast amount of information, provided by on-line news agencies, historical archives, government agencies and private organisations which collect, organise, process and distribute different types of data. Financial operators are usually connected in real-time to the news providers. This reduces drastically the lag between when the event happens and when the operator reads about it. The cost of this information is decreasing rapidly, making it more accessible to small or private investors.

Quantitative data, such as historical price databases or real-time price information are today largely processed by automatic computer programs based on various techniques - i.e.: MACD, regression, neural network, genetic algorithms, etc. One of the drawbacks of most of

[1]We here consider a situation in which the market is not perfect. In case of a perfect market situation, in fact, the operators would have free and complete access to the information needed which would allow them to perform correct decisions.

quantitative tools is that they process and produce *"numbers"* (quantitative data), which need to be interpreted by the financial operators, before they can be used.

A number of *qualitative tools*, able to deal with *qualitative information* has also been developed, such as *Expert Systems* and *Fuzzy Logic*.

Natural Language Processing and *information extraction* can be used as *financial qualitative* tools, to analyse *qualitative data* consisting of financial news articles from on-line news providers (e.g. *Dow Jones*) or from financial newspapers (e.g. *The Financial Times*) and produce relevant qualitative information. This information can be used to support and help the operators' decision-making process regarding investments in securities.

This work is organised as follows. In section 2 we analyse qualitative information and tools and their importance in finance. Section 3 describes how *natural language processing* and *information extraction* can be successfully used for processing such qualitative information. Finally, in section 4 we briefly compare *natural language processing* with other AI techniques currently used in finance.

2. Qualitative information and qualitative tools

Qualitative data are data which is difficult to be accurately expressed in a numeric format. For example, data regarding *rumours*, *fears*, *broker's recommendations, takeovers* etc. are qualitative data. A sentence such as:

"there are rumours of a possible takeover of Apple, the troubled computer manufacturer"

represents information which is extremely relevant to the financial operator, since, if the source of the information is reliable (e.g. a leading Merchant Bank) it is likely to cause a movement in the quotation of Apple's shares as well as those of the possible buyers.

This kind of news is extremely important because it affects the expectations of the operators regarding a specific share. The way in which the operators are influenced depends on how an operator perceives the news.

Operators are much more influenced by news reports than by analysts' forecasts or historical price analysis of share. When relevant news arrive, the price of shares is influenced immediately and operators base decisions on their personal experience and on other people's behaviour, rather than on forecasts produced by neural-networks financial forecasting system. Quantitative methods work fine and help the operators' decision-making process by suggesting a possible path of prices but, in the end, prices reflect the people's view of the news, consisting of quantitative and qualitative information.

The financial operators and information providers understood the importance of qualitative data as key-point in the trading decision-making process a long time ago. Therefore, financial operators receive in real-time news regarding companies (e.g. announcements, rumours, profit forecasts etc.), macroeconomics (e.g. movements of inflation rate, unemployment etc.), politics (e.g. changes in general macro economic policy of the government, tax policies etc.). They also have access to huge quantities of past information. Operators can make use of these information by quantifying them in probabilistic terms and use them for projecting the results of various strategies based on quantitative analysis. However, this quantification is not easy, since qualitative data is difficult to be accurately expressed in quantitative terms. Therefore, very few qualitative financial tools, able to make direct use of the qualitative information have been designed in the past.

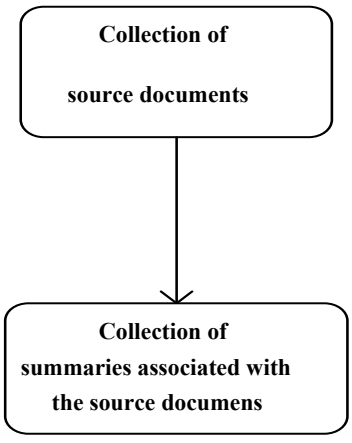
3. Natural Language Processing

Natural Language Processing belongs to the field of artificial intelligence. Its goal is to analyse input sentences in free natural language and store their meanings into an appropriate representation (e.g. semantic net). Once the information is available in the knowledge base, the relevant information can be easily retrieved in various ways and knowledge can be *inferred* by the system, using *inference* rules. An example of a large-scale system is the NLP System [Morgan *et al.*, 1996] under development at the University of Durham, UK.

3.1 Information Extraction

The goal of information extraction, which belongs to the field of Natural Language Processing, is to extract specific kinds of information from a source document. In other words, the input to the

system is a document (e.g. a newspaper article), while the output consists of a summary of the contents of the document (figure 2) . Information extraction is used to provide the financial operator with information which is a summary of the most important qualitative data supplied as input.



A possible source article for a financial information extraction can consist of the following text:

SCOTTS Inc. announced it will acquire Grace-Sierra Horticultural Products with 100 million dollars from a subsidiary of WR Grace, the specialty chemical group, and other investors.
Scotts said that after the deal, Grace-Sierra's business and operations would be combined with those of Scotts to form the world's largest turf and horticultural products company, with combined 1993 sales of nearly Dollars 600m.
Grace-Sierra manufactures and markets specialty fertilisers for nursery, golf course, greenhouse and consumer markets. WR Grace added that the deal included repayment of Grace-Sierra's indebtedness. The company added that the acquisition was expected to be financed through a combination of long-term sub-ordinated debt and bank borrowings.

An information extraction system will try to produce a summary of the original text according to a specification of the information to be searched defined during the design of the system. For example, a sensible summary of the article shown above could be:

```
Template: Takeover
     Company target: Grace-Sierra Horticultural Products.
     Company predator: SCOTTS Inc.
     Type of takeover: FRIENDLY
     Value: 100 million dollars.
     Attribution: SCOTTS Inc.
```

The summary presented above is called a *template* which is a structure with a predefined number of fields. A *template*, thus, is a schematisation of the contents of the source document and it is widely used in information extraction. The "fields" of the *template* are called "*slots*" which are filled with the information extracted from the source article. The slots of the template (called *takeover*) shown above are thus *Company target*, *Company predator*, *Type of takeover* and *Value*. The slots can be filled with information directly extracted from the text (e.g. *Marsam Pharmaceuticals Inc.* or with text "inferred" by the system for data ""hidden" the meaning of the

text, such as, in the above example, the *Type of takeover* slot, which contains "*FRIENDLY*", a term that cannot be found in the source document.

Templates can be of different kinds and can include different types of slots. More than one template can be produced for a source article. For example, for the article shown above another different template could be produced:

```
Template: Summary
    Companies: Grace-Sierra Horticultural Products.
               SCOTTS Inc.
    Values:    100 million dollars
               600 million dollars
```

As we have already mentioned, the definition of the templates is usually done during the design of the system and it assumes great importance since it will influence the overall performance of the system.

Two groups of articles can be processed by a financial information extraction system: documents from on-line services (e.g. *Dow Jones*, *Bloomberg* etc.) and from financial newspapers or magazines (e.g. *The Financial Times* etc.). The first tend to be rather summarised, while the latter tend to be much longer and include further analysis of the news and the automatic processing using an information extraction system can therefore be extremely useful. In figure 3 the same news reported from a newspaper (*The Financial Times*)and from an on-line news provider (*Dow Jones*) is shown.

A natural language processing financial tools based on information extraction can reduce the operator's qualitative *data overload*, by simplifying and reducing the amount of qualitative information that is needed to support the decision-making process.

Most of the information extraction systems have been developed and tested within government agencies or scientific environments. This has lead to very specialised systems able to work only in restricted situations and domains.

Information extraction is a relatively new field compared to other artificial intelligence techniques. The first systems were developed by the end of the 70s. Recently, the MUC conferences [ARP, 1994] and [ARP, 1996] provided the ideal framework for the development of systems based on new and more effective techniques.

As far as the financial domain is concerned, very few financial information extraction systems have been realised in the past. One of the few systems is ATRANS [Lytinen and Gershman, 1986], a system for extracting information from Telex messages regarding money transfers between banks. However, the system has been successful mainly because of the extremely limited domain and the reduced information to be extracted. The systems that competed in the MUC-5 competition [ARP, 1994] were also able to perform the extraction of information from financial articles. However, they were only able to extract information regarding Joint Ventures and, thus, work only in an extremely restricted subset of the financial domain.

A typical article from "The Financial Times"

Disney to buy Capital Cities
Tuesday August 1 1995
By Tony Jackson in New York
Walt Disney is to pay 19.1bn dollars for Capital Cities/ABC, owner of the ABC television network, creating the world's largest entertainment company. The deal is the second biggest takeover after that of RJR Nabisco by Kohlberg Kravis Roberts for about 25bn dollars in 1988.
In an agreed deal, Disney will pay one share plus 65 dollars for each Capital Cities/ABC share, valuing the latter at 124 dollars at yesterday's prices. The combined company will be called simply Walt Disney, and Mr Thomas Murphy, chairman of Capital Cities/ABC, will join the Disney board.
Mr Michael Eisner, Disney's chairman, claimed the synergies between the two companies were tremendous". He said: "Disney's intellectual property will appear on ABC's networks, and Disney's distribution systems will syndicate ABC's programmes."
Mr Warren Buffett, the billionaire portfolio investor whose company Berkshire Hathaway owns 12.9 per cent of Capital Cities/ABC, described the deal as "the marriage of the number one content company in the world with the number one distribution company".
The ABC television network rivals NBC for the top position among US networks, with about a 17 per cent market share. Capital Cities/ABC, which had revenues last year of 6.4bn dollars , also owns eight television stations, America's largest network of radio stations and an 80 per cent share of the leading cable television channel ESPN. It also publishes newspapers, books and magazines.
Walt Disney had been rumoured for some time to be interested in buying a TV network, but was thought to want CBS or NBC, either of which would cost 5bn dollars or less. It is much larger than Capital Cities/ABC, with a market value of 31bn dollars.
The deal allows Capital Cities/ ABC shareholders to take all their payment in either cash or stock, subject to availability. Mr Buffett said "the odds are extremely high that we will have a very large amount of Disney stock".

The same topic from an on-line news service (Dow-Vision)
Disney, Capital Cities -2-: Details On Merger DIS CCB
Source: Dow Jones News Service via DowVision
Date: Jul 31, 1995 Time: 8:04 am
BURBANK, Calif. -DJ- Walt Disney Co. (DIS) and Capital Cities ABC Inc. (CCB) agreed to merge in a transaction valued at about 19 billion dollars at current share prices. In a joint press release, the companies said Capital Cities/ABC shareholders will have the right to receive one Disney common share and 65 dollars in cash for each Cap Cities common share. The transaction has been approved by both companies' boards, the companies said.

Figure 3: The difference between a newspaper article and an on-line news

Tools based on information extraction are particularly useful for information providers who are extremely keen on trying to further classify, reduce and summarise huge amounts of financial information that they provide to their customers. For example, the *Dow-vision* news service, an Internet news service by *Dow Jones*, automatically provides some additional information linked to the articles available to the user in real time. However, such information is quite naive and simple, such as the market sector and the category to which the company

belongs. Such information is often obtained using simple pattern-matching techniques, rather than NLP techniques, or even by hand by the information provider's operators.

The financial information extraction system under development at the University of Durham is able to process source articles and produce relevant templates using deep natural language understanding techniques [Costantino *et al.*, 1996].

3.2 Information extraction and meta-analysis

Templates extracted from the source text can be improved by adding additional information:

- Knowledge which is not directly available in the text. For example, a text can contain information regarding a *takeover* even if the concept of *takeover* or *acquisition* is not present in the source article.
- The "way" in which a particular information is cited in the source article. A written text, in fact, incorporates the writer's perception of the event and the *way* in which it is written is likely to have some influence on the reader's opinion - view of the facts.
- How many times a particular news (topic) has been cited in the source article. This information is likely to be related to the importance of the news. A NLP information extraction system would perform such operation through a *semantic comparison* of the occurrences.

3.3 Forecasting with Natural Language Processing

Natural Language Processing, finally, can be used as a prediction - forecasting tool, suggesting to the financial operator final investment decisions (buy/sell decisions). We can summarise an investment decision-making process based on qualitative knowledge in the following steps: (a) the operator reads the news, (b) the operator gives his own interpretation to the news, (c) the operator compares and analyses the new information with the knowledge he already owns, (d) a final decision (buy/sell) is made.

Financial tools based on Natural Language Processing can potentially automate such processes. The system would process the new data, identify the relevant information, process it according to specific domain-knowledge and inference rules for that situation and present to the operator not only the summary / template, but also suggest an operative decision.

Figure 4 shows an example of use of a NLP System as prediction tool. The system analyses the source article and identifies the most relevant information which are displayed in the template. The user has immediate access to the information extracted which he can use to support his decision-making process. The NLP System also produces a suggested investment decision which, in case of the example in figure 4, would be to sell the shares.

3.4. Natural Language Processing, Expert Systems and Neural Networks

Natural Language Processing can be compared with other knowledge-based tools based on artificial intelligence techniques which have successfully applied to finance and, in particular two techniques: neural networks and expert systems.

A financial tool based on Natural Language Processing differs from a common Neural Network forecasting tool for two main reasons: the kind of input and the kind of output of the system. The input of a Neural Network is normally strictly numerical, for example time-series of share prices, while the input of a NLP System consists of news.

Neural Networks are normally used in the financial market for the prediction of future prices of shares. The output of a neural network is therefore strictly numeric and consists of forecasts of share prices. A NLP System, instead, does not produce any forecast of the price of shares. Its output consists of a *summarisation* of the original input which can potentially include a suggested decision for the user (e.g. figure 4).

Text from "The Times, 15th March 1996"

UNITED BISCUITS (Holdings), the McVitie's biscuits to KP snacks group, slumped to annual losses of more than 100 million dollars after suffering exceptional charges totaling more than 150 million dollars after a catalogue of disasters. Colin Short, chairman, said UB's board took last year's "tortuous" decision to retreat from the disastrous foray into America and Spain, but emphasized that prospects have been transformed by the disposal of poor-performing businesses.

NLP information extraction result:

```
Template: Dividend Announcement
      COMPANY_NAME: UNITED BISCUITS
      CATEGORY: yearly profits.
      VALUE: 100 million dollars losses.
      CHANGE TO LAST YEAR: unknown.
      COMMENT: due to exceptional charges.
```

NLP prediction tool suggestions:

```
sell UNITED BISCUITS shares
```

Figure 4: Example text of a possible NLP prediction tool.

NLP Systems can potentially be integrated with Neural Networks. The NLP System would process source articles, identifying the relevant quantitative information which could be supplied to the trading neural network for producing price forecasts. The input of the data to the Neural Network could be performed by the NLP System, rather than by the user. This can be extremely useful for large collections of source articles such as the Financial Times on CD-ROM. For example, the NLP System could potentially identify the market movements in the following article from *The Financial Times*:

TSB delivered the best performance among high street banks as the market responded strongly to whispers that it could be a takeover target for a domestic or overseas bank seeking to establish or increase a presence in UK retail banking. Yesterday the shares powered ahead to reach an all-time high of 259p before coming off the best level to finish a net 7 up at 257p.

The NLP System would extract the relevant information producing the following template which could be shown to the user and supplied as input to the Neural Network:

Template: Market Movement
Company Name: TSB
Type of securities: shares
Movement amount: 7p
Reason: reaction to takeover rumours

NLP Systems can also be potentially used with existing Expert Systems to produce final investment decisions. The user of Expert Systems is normally required to enter the information which will be used to generate the suggested investment decision. By integrating a NLP System and an Expert System, this phase could be eliminated. The NLP System would identify and extract the relevant information from the source articles. This information, for example the *dividend announcement* template shown in figure 4, could be supplied in the correct input format to the Expert System which would generate a final investment decision (figure 5). The NLP System could therefore substitute the user in the identification of the relevant information to supply to the Expert System.

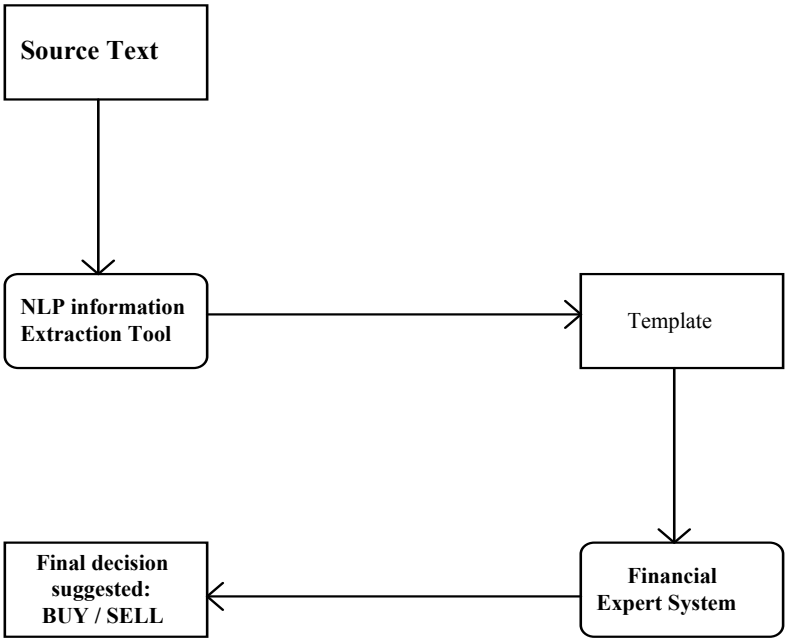


Figure 5: Natural Language Processing System as possible prediction tool.

4 Conclusions

In this paper we have analysed the importance of *qualitative* information in finance. Natural Language Processing and, in particular, information extraction, can be used for processing and analysing such information.

The user can process large quantities of articles obtaining the most relevant information (templates). Templates can be also filled with information which is not directly present in the source articles, but is *inferred* by the system. Information extraction can act as a "filter" for the news eliminating those that do not satisfy any of the extraction criteria. However, unlike information retrieval systems based on key-word searches, an information extraction system allows the extraction of specific information, such as takeoevers, mergers etc. Traditional key-word information retrieval engines are able to locate particular information in the source document, but this is limited to the search of specific words (e.g. takeovers) and a summary of

the contents of the original article is not provided. The information extracted can potentially be used in conjunction with existing financial tools, such as neural networks and expert systems, enhancing their functionalities.

The financial information extraction system under development at Durham University performs information extraction on articles from on-line agencies or newspapers producing templates according to the *financial activities approach* [Costantino *et al.*, 1996].

References

[ARP, 1994] ARPA, *Proceedings of the Fifth Message Understanding Conference*, Morgan Kaufmann Publishers, August 1994

[ARP, 1996] ARPA, *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann Publishers, March 1996

[Costantino *et al.*, 1996] M. Costantino, R. J. Collingham, R. G. Morgan, Information Extraction in the LOLITA System using Templates from Financial News Articles, in Proceedings ITI '96, June 1996, Pula.

[Lytinen and Gershman, 1986] S. Lytinen and A. Gershman, "ATRANS: Automatic Processing of Money Transfer Messages", in M. Kaufmann, editor, *9th International Joint Conference on Artificial Intelligence*, pages 821-825, 1986

[Morgan *et al.*, 1996] R. G. Morgan, R. Garigliano, P. Callaghan, S. Poria, M. H. Smith, A. Urbanowicz, R. J. Collingham, M. Costantino, C. Cooper and The LOLITA Group, "University of Durham: Description of the LOLITA System as used in MUC-6", in [ARP, 1996]