Financial Information Extraction at the University of Durham

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

Abstract

This article describes the financial information extraction system under development at the University of Durham. Differently from many others developed in the past, the system has been designed for use in real situations and to alleviate the "data overload" from which traders, brokers, fund managers etc. suffer nowadays. The system is based on the financial activities approach, for the identification of the relevant templates to be extracted from the source articles. The goal of the system is to summarise financial news (either from newspapers or on-line services) producing specific templates associated to the various financial activities. The templates produced can be successfully used for a "meta-analysis" of the news on price behaviour. The system uses natural language processing techniques developed at Durham University which are based on *deep* natural language processing techniques, as opposed to pattern-matching or statistics.

1. Introduction

Few successful financial information extraction systems have been developed in the past. Moreover, most systems were developed and tested within government agencies and scientific environments. This has led to very specialised systems, able to work only in limited domains. The financial information extraction system under development at Durham University has been designed in close contact with financial experts and to work in real situations. The system is able to process news articles (either from newspapers or on-line news services) and produce a summary (*template*) of the most relevant information identified in the source articles. The templates are created according to the "financial activities" approach, which identifies a finite number of financial activities which are associated with a corresponding template. In addition, templates can be linked to other templates or to other sources of information external to the source articles.

The financial information extraction system is based on the natural language rocessing system under development at the University of Durham, which uccessfully participated in the MUC-6 information extraction competition [Morgan *et al.*, 1996]. The system uses *deep* natural language processing techniques rather than pattern-matching or statistics in order to capture and extract the meaning of the text.

The *templates* produced from the source articles can be used for reducing the qualitative *data-overload* suffered nowadays by the financial operators and can also be used for a "meta-analysis" of the effects of news on price behaviour.

2. The Financial Activities approach

Two main decisions have to be taken in the design of a successful financial information extraction system:

- The **type of source articles** which can either be from on-line news agencies (e.g. Dow Jones or Bloomberg) or from financial newspapers (e.g. The Financial Times).

- The **information to be extracted**. This aspect is crucial: if the wrong information is extracted, the system would probably be useless for the user. This choice is usually made during the general design of the system and will influence all the stages of its development.

The target source texts for the financial information extraction system under development at the University of Durham consists of articles from newspapers (e.g. *The Financial Times*) or (*The Wall Street Journal*). This choice has been made taking into consideration the fact that articles from on-line news agencies are already rather summarised and, therefore, do not need further processing. However, the system can be succesfully used to process such articles.

The definition of the information to be extracted (templates) has been done according to the *financial activities approach* [Costantino *et al*., 1996b]. This approach is based on the identification of various *financial activities* which can be defined as an event which is likely to influence the price of shares and, therefore, influence the decision-making process of the players of the market regarding these securities. The *financial activities* correspond therefore to the information that a financial operator would consider important for his decision making process. A finite number of financial activities can be identified in the financial market and can be subdivided into three different groups (figure 1):

- **Company related activities** which are activities related to the "life" of the company, changes in its status, in the ownership of the company, the number and ownership of its shares etc. These activities are likely to have a strong and direct influence on the prices of the shares of the same company (or to related companies), rather than a generalized change in the current trend and overall performance of the stock exchange. This category includes the following activities: takeover, merger, flotation, new issue (of shares), privatization, market movement, bankruptcy, broker's recommendations, taking a stake, dividend announcement, overseas listing, profit / sales forecasts, profits / sales results, directors' dealings, legal action and investigation.

- **Company restructuring activities** which are related to changes in the key elements of the productive structure of the company. Information about new products or lines of products, joint ventures, staff changes etc. are likely to have an immediate impact on the company's share price and might have some effect also on share prices of related companies, such as suppliers, customers, distributors, etc. Four activities belong to this group: new product (or line of products), joint venture, staff changes and new factory.

- **General macroeconomic activities**. This group includes activities which are likely to have an effect on the stock exchange market and to confirm the current trend or produce an inversion. All the main macroeconomic indexes, such as inflation, interest rate etc. belong to this category. The main activities belonging to this group are: interest rate movements, currency movements, inflation rate changes, unemployment rate, trade deficit, industrial production and deficit of the public sector.

Each *financial activity* is associated with a specific template. For example, the takeover template is composed of the following slots: *company target*, *company predator*, *type of takeover*, *value*, *bank adviser predator*, *bank adviser target*, *expiry date*, *attribution current stake of the predator*, *denial*, which can be filled or not depending on the information contained in the source article. Some of the templates associated with the financial activities are shown in figure 2.

Company Related	**Company Restructuring**	**General Macroeconomic**
Takeover	New product	Interest rate movements
Merger	Joint venture	Currency movements
Flotation	Staff changes	Inflation rate changes
New Issue (shares, bonds etc.)	New factory	Unemployment rate changes
Privatisation		Trade deficit changes
Market movement		Industrial production changes
Bankruptcy		Deficit of the public sector
Broker's racommendations		
Taking a stake		
Dividend announcement		
Overseas listing		
Profit / sales forecasts		
Profits / sales results		
Director's dealings		
Legal action		
Investigation		

Figure 1: the three groups of financial activities

It is our belief that the identification of the relevant templates based on the *financial activities approach* represents an effective partitioning of broad financial domain and the information. The templates identified represent the *qualitative information* that the financial operator takes into account in his decision-making process.

A *template* interface which allows the user to add new templates by using natural language sentences describing the information to be extracted is currently under development. This should ensure the maximum degree of freedom for the financial operator. A complete set of *financial templates* associated with the *financial activities* are already provided but the system can be personalised by adding new templates.

3. Architecture of the system

The financial information extraction system is based on the Natural Language Processing system under development at the University of Durham [Morgan *et al.*, 1996], which represents the core of the application and has been under development for the last nine years. Currently, around 20 researchers work on various aspects of the system, as part of the Durham's Laboratory for Natural Language Engineering. The system succesfully participated in the MUC-6 competition (Message Understanding Conference, U.S.A.), the most important competition for information extraction systems.

The system is based on "deep" natural language processing analysis of the input text, rather than on statistical, probabilistical or pattern-matching techniques which differs from traditional systems.

The system is built around a large (over 100,000 nodes), WordNet-compatible [Miller, 1990] semantic network, called *SemNet* which constitutes the knowledge-base of the system and is similar to a conceptual graph [Sowa, 1984]. The semantic network consists of a hierarchy of nodes connected with links. The nodes represent entities (e.g. *Ferrari*) or events (e.g. *The company made a takeover*). Control variables are used to indicate basic information about a node, such as its type (e.g. event, entity etc.), its family (e.g. human, inanimate, organisation), its lexical type (eg. noun, preposition, adverb) etc. Some of the most important control variables are:

Merger	Takeover	Flotation	New Issue

Company 1: Company 2: New Name: Date of Announce: Date of Merger: Comments: Attribution: Date	Company target: Company predator: Type of takeover: Value: Bank adviser predator: Bank adviser target: Expiry date: Attribution: Current stake predator: Denial:	Company name: Price: Value: Announce Date: Listing Date: Financial adviser flotation: Attribution: Denial: Industry sector:	Company name: Company fin. adviser: Issue currency: Issue value: Announce date: Launch date: Listed: Attribution: Purpose: Denial
Privatisation	**Market Movement**	**Bankruptcy**	**Broker's reccomendat.**
Company name: Stake to be privatised: Price of shares: Value of shares: Announce date: Privatisation date: Attribution: Denial: Industry sector:	Company name: Type of securities: Movement percentage: Movement amount: Reason:	Company name: Receivers: Date of announce: Denial:	Recommendation source: Company name: Recommendations:
Overseas listing	**Dividend annoucement**	**Profit/sales results**	**Director's dealings**
Company name: Overseas exchange: Type of securities: Announce date: Date of listing: Attribution: Denial	Company name: Dividend per share: Type of dividend: Change to previous year:	Company name: Category: Value: Change to last year: Comment:	Company name: Director name: Type of security: Type of dealing (buy/sell): Value:

Figure 2: Specific templates associated with the company related financial activities

- **Rank**. This control gives the nodes quantification, e.g. individual (*the loss Company XY made in the first quarter of '94*), universal (*every loss*), generic (*losses*, or *some losses*), named individual (*Ferrari*) etc.

- **Type**. This control value is similar to grammatical qualifications with few exceptions and additions: entity, relation, event, fact, greeting etc.

- **Family**. This control is used to group nodes into semantic "families", e.g. living, animal, human, man-made, abstract, location, organization, human-organization etc.

The semantic network contains an elaborate "knowledge-base" which includes *linguistic knowledge*, domain-specific knowledge (e.g. for a takeover) and can be dynamically expanded during the use of the system (e.g. with the processing of a source article) or using a Natural Language Interface.

The input natural language text is processed by various jerarchical modules and the result of the analysis stored in the semantic network. The main processing phases are: *morphology*, *parsing*, *semantics* and *pragmatics* (figure 3).

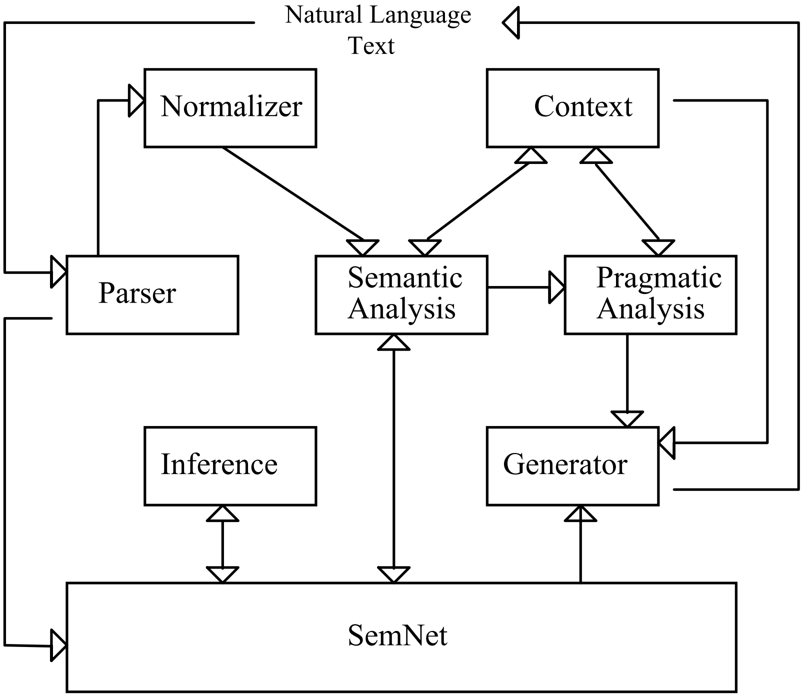

Figure 3: The Durham NLP System's core

- **Morphology**. This phase is responsible for splitting the input text words and smaller units and producing for each word a list of all possible meanings of that word combined with their syntactic (noun, verb, etc.) and semantic categories. The input is then supplied to the parser.

- **Parsing**. The task of the parser is to determine the syntactic information that is contained in a text. In other words, the parser performs a full grammatical analysis of the input text, recognising the role of each word in a sentence (e.g. subject, verb, adjective, object). At this stage, the meaning of the words in a sentence can still be ambiguous.

- **Semantic analysis**. The task of the semantic analysis is to associate the words with the appropriate meaning(s) and to map them onto the system's internal representation.

- **Pragmatic analysis**. Finally, this module provides disambiguation of meaning of words and type checking. Lexical ambiguities (e.g. different meanings of the same word) and anaphora are resolved using a series of preference heuristics, taking into account the *topic* of the current ext and the information in the *context*. In a financial article regarding a takeover, or example, the meaning of *to buy* will be recognised as being:

  ```
  to buy -> To acquire, To purchase.
  ```

  rather than:

  ```
  To buy -> To pay, To corrupt, To bribe.
  ```

At this stage the information is stored in the semantic network (the knowledge-base of the system) and can be later retrieved, producing output natural language text by using the *generator* [Smith *et al.*, 1994]. The *inference engine* can be used to perform various kinds of reasoning (e.g. analogy, inheriance etc.) on the data stored in the network [Long and Garigliano, 1994][1]

4 The financial information extraction module

The financial information extraction application is built as a module of the natural language processing system (figure 4).

[1] In a paper of this size it is impossible to describe the single modules in sufficient detail. A complete description of the architecture of the system can be found in [Morgan *et al*,. 1995].

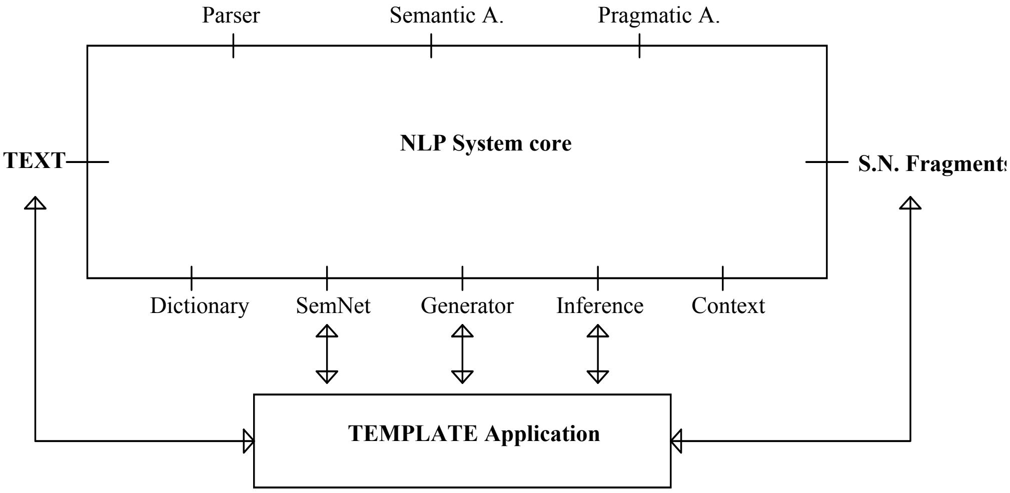

Figure 4: The financial information extraction module.

The input source text is therefore processed by the NLP system first and stored into *SemNet* (the knowledge-base). The task of the financial application is then to retrieve the information needed from the semantic network using the *inference engine* and to produce output in English using the *generator*. The information is retrieved according to the rules associated to each of the templates. Each template is defined as a *template main-event* and a set of *slot-rules*, which direct the search for the appropriate information in the semantic network. Five different types of slots can be used in the definition of a template:

- **Concept slot**. This slot is the most general slot usable in the definition of a template. The rule associated with the slot identifies the relevant concept in the semantic network which is passed to the generator obtaining the corresponding English text.

- **String slot**. This slot produces the output directly from a list of predefined strings and not using the English generator. It is mainly useful to produce a slot with a predefined number of alternatives which may not be present in the original text.

- **Net slot**. The slot fill-in rules are retrieved from the semantic network, allowing the creation of user-defined templates or rules by updating portions of semantic network.

- **Text Reference Slot**. The output is produced using fragments of the original text where possible and the generator if a semantic network's concept does not correspond to any fragments of the original text (e.g. when using inference functions). The slot is used when the exact copy of the original text is needed (e.g. the name of a company).

- **Template Reference Slot**. This slot is used to create a link to another template. The output of the slot will consist of a pointer to the new template. The template reference slots provide the basic mechanism for handling **hyper-templates** and for linking other sources of information to the template.

The slot-rules are built by checking the control variables associated with the nodes and by using the inference functions available in the core system [Costantino *et al.*, 1996a]. The *template main-event* is used to decide *wheter* the template itself has to be built. For example, a *takeover* template is filled when a *takeover event* is found in the semantic network updated with the information from the source article acording to the following main-event rules:

- the event that can be generalised to the concept of *takeover*, *acquisition*, *purchase* and other relevant concepts. For example: "*The acquisition of X by Y*", where X and Y are companies;

- the event that has a *takeover-action*, e.g. *to buy, to purchase, to take-over* etc. and the object a company. For example: "*X has acquired Y for 100 million dollars*" where Y is a company.

If such an event is found in the semantic-network, this means that the source article contained information about a takeover and all the other slots can be filled according to the fill-in rules. For example, the *company predator* slot is filled with the *subject* of the takeover main-event. The *company target*, instead, will be the *object* of the event, while the *takeover value* will be filled with the *instrument* of the takeover (the sum of money that the predator had to pay to acquire the target). The slots can be filled with information which is directly present in the source text or is *inferred* by the system. For example, the slot *type of takeover* of the takeover template is filled with information which cannot always be found in the source article. The slot, in fact, is filled with "FRIENDLY" if the company target agrees in principle to the takeover, while "HOSTILE" is used if the company target *does not want* to be acquired.

In figure 5, an example of how an article is processed and a template is filled is shown. More than one template can be built for a source article, e.g. a *takeover* template and a *market movement* template, where the text reports, for example, a change in the share prices caused by the takeover.

An important aspect in the development of the financial application is the domain-specific knowledge. Financial articles are based on highly technical and specific knowledge and lexicon. Sentences which would normally have a certain meaning in the normal text might present a totally different one in a financial context. Domain-specific knowlege consists of *semantic* and *pragmatic* rules which are used by the system to correctly *understand* and choose the meaning of a particular sentence in the financial context. For example, the takeover template is filled according to the following domain-specific rules which have been identified:

- if X takes full control of Y, this implies a takeover;
- X buys a majority stake in Y, this implies a takeover;
- X buys a 51 (or over) per cent of Y, this implies a takeover;
- X pays M for Y and Y is a company, this implies the takeover of Y by X.

In a paper of this size it is impossible to describe all aspects of the system. Additional details about the architecture of the system can be found in [Morgan *et al.*, 1996] while in [Costantino *et al.*, 1996a] the financial information extraction module is analysed in more detail.

Source article from *The Financial Times*

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.

Stages of the analysis:

1. The system processes the article and stores its contents in the semantic network;

2. The template application looks for events which satisfy the main-event takeover condition and identifies relevant information. For the takeover template, the following sentence is relevant:
   ```
   SCOTTS Inc. announced it will acquire Grace-Sierra Horticultural Products
   ```

3. Once a relevant main-event is found, the slots are filled according to their rules and connected to the main-event, obtaining the final template:
   ```
   Template: TAKEOVER
   ```

- from the object of the main event: *SCOTTS Inc. announced it will acquire **Grace-Sierra Horticultural Products***
  ```
  COMPANY_TARGET: Grace-Sierra Horticultural Products
  ```
- from the subject of the main event: ***SCOTTS Inc.** announced it will acquire Grace-Sierra Horticultural Products*
  ```
  COMPANY_PREDATOR: SCOTTS Inc.
  ```
- the instrument of the takeover main event: *SCOTTS Inc. announced it will acquire Grace-Sierra Horticultural Products **with 100 million dollars***
  ```
  VALUE: 100 million dollars.
  ```
- the slot is filled with "FRIENDLY" because the company target agrees, in principle, to the takeover and is not hostile to it. This information is not present in the source article, but *inferred* by the system.
  ```
  TYPE_OF_TAKEOVER: FRIENDLY
  ```
- the company/person who announced the takeover: ***SCOTTS Inc.** announced it will acquire Grace-Sierra Horticultural Products*
  ```
  ATTRIBUTION: SCOTTS Inc.
  ```

4. The template is shown to the user in the final format:
   ```
       Template: TAKEOVER
              COMPANY_TARGET: Grace-Sierra Horticultural Products
              COMPANY_PREDATOR: SCOTTS Inc.
              TYPE_TAKEOVER: FRIENDLY
              VALUE: 100 million dollars.
              ATTRIBUTION: SCOTTS Inc.
   ```

Figure 5: Analysis of an article by the system.

4.1. Hyper-templates and other sources of information

Hyper templates are structures whose slots can refer to other templates, thus creating a linked *chain* of templates [Costantino *et al.*, 1996a]. The mechanism allows the maximum degree of flexibility for the financial aplication. Hyper-templates can be used in different ways. For example, we can consider the case in which a *takeover* template is linked to an *organisation* template, connected to the company_pretador or company_target slots and containing additional details about the company. The user, in fact, after reading the *takeover* template, might be interested at that point in knowing more information about the company, thus accessing the *organisation* template (figure 6). The hyper-template mechanism is potentially usable for linking different kinds of information, not necessarily extracted from the source text, such as company databases or historical share prices.

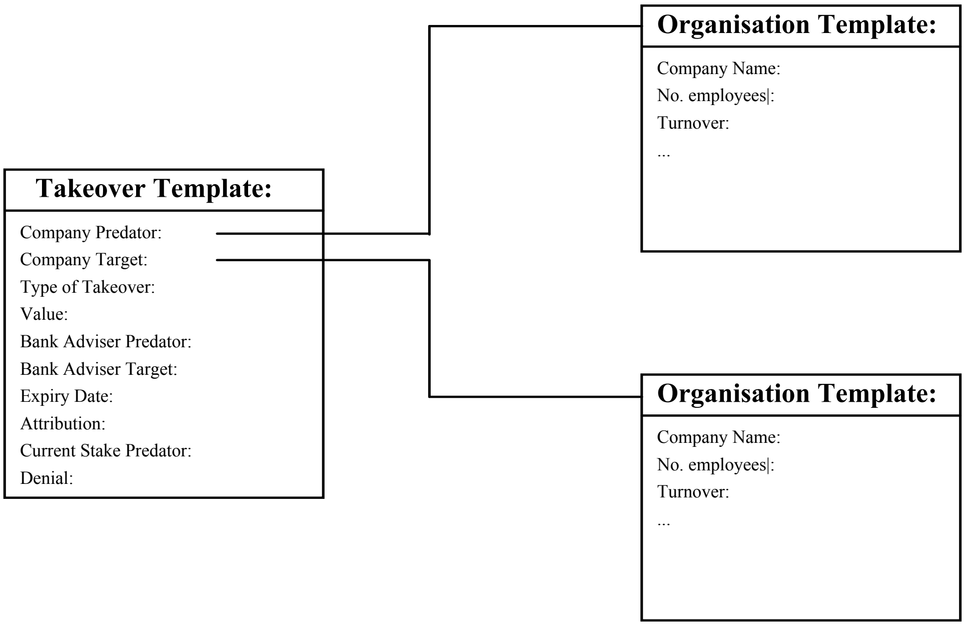

Figure 6: Hyper-Templates

5. Interaction with the system

The system processes the articles following a *two-stage* strategy. First of all, the system identifies the list of relevant *financial activities* (see figure 1) in the source article. This tells the user the main "topics" contained in the article. If, for example, the processing of an article produces:

```
Article N.1 - Financial activities found:

1 takeover(s)              found
2 market movement(s)       found
```

the user will draw the conclusion that the main *topic* contained in the article is about a *takeover* and, probably, the two *market movements* are likely to be caused by the *takeover*. At this stage, the user can decide that he is interested or not in more detail about the news. In the first case, he can request to the system the display of the full template associated to the financial activity. He can, for example, request the display of the *takeover* template, which will produce a filled template as, for example:

```
Template: TAKEOVER
     COMPANY_TARGET: Drakes Office Systems.
     COMPANY_PREDATOR: Filofax Group.
     TYPE_TAKEOVER: FRIENDLY
     VALUE: 3 million dollars.
     ATTRIBUTION: Filofax Group.
     ...
```

The user could in the same way request the display of the *market movement* templates. In case the user is not interested in these topics, he can just skip to the processing of the next article, avoiding the display of the full templates.

The two-stage processing has been chosen because of the fact that, on large collections of documents, the user might be interested only in particular topics and decide to skip all the others.

It is our belief that the financial operators can directly benefit from the use of the financial information extraction system for two main reasons:

- the system can drastically reduce the qualitative *data-overload* suffered by most financial operators. By processing larget quantities of articles, the user can obtain the list of the financial activities (or topics) contained in each of the articles and can access to their summaries without having to read the whole article. Moreover, irrelevant news, that do not satisfy any of the financial activities conditions or those added by the user are eliminated from the input data.

- the templates can be used for a "meta-analysis" of the news on price behaviour, by processing articles from large collections (e.g. *The Financial Times* on CD-ROM). The operator, in fact, could use the qualitative information to "*explain*" the behaviour of a share price in a certain time. For example, the operator could be interested in knowing why the price of a particular share share quickly increased. Using the financial information extraction system, he could link the increase with the fact that the company announced a takeover, reading the *takeover* template of the source article.

6. Implementation

The system is written in the functional language Haskell and the C Language (currently about 45,000 lines of code, corresponding to about 450,000 lines of code in a traditional programming language) and runs on a Sun SparcStation with 80Mbytes of RAM and running under Solaris 2.4, but can be easily ported to other Unix implementations. A parallel version of the system can be used to process more than one article at a time on multi-processors machines. Not all of the templates associated with the financial activities have been implemented yet. The testing of the system is carried out on a set of 80 financial articles from *The Financial Times* and evaluating the results using the MUC-6 scoring tools [ARP, 1996].

7. Conclusion

In this paper we have tried to give an overview of the financial information extraction system under development at the University of Durham. First results show that the system can be used successfully to process financial *qualitative* data and reduce the operators' qualitative *data overload*. The *financial activities approach* ensures that the information extracted is relevant for supporting the operators' investment decision-making process. The interface for allowing user-definable templates using natural language sentences currently under development should ensure the maximum degree of flexibility of the system.

Because of its commercial value, the system is not publicly available. However, we are keen to give demonstrations of the system and serious enquiries should be addressed to the authors.

References

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

[Costantino *et al.*, 1996a] M. Costantino, R. J. Collingham, and R. G. Morgan, "Information Extraction in the LOLITA System using Templates from Financial News Articles", in *Information Technology Interfaces '96*, University of Zagreb, June 1996.

[Costantino *et al.*, 1996b] M. Costantino, R. J. Collingham and R. G. Morgan, "Natural Language Processing in Finance", *The Magazine of Artificial Intelligence in Finance*, 2 No.4, 1996

[Long and Garigliano, 1994] D. Long and R. Garigliano, *Reasoning by Analogy and Causaility, a model and application,* Ellis Horwood, 1994.

[Miller, 1990] G. Miller, "Wordnet: An online lexical database", *International Journal of Lexicography*, 3 No.4, 1990.

[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]

[Smith *et al.*, 1994] M. H. Smith, R. Garigliano, and R. G. Morgan, "Generation in the LOLITA system: An Engineering Approach", in *7th International NL Generation Workshop*, June 1994.

[Sowa, 1984] J.F. Sowa, *Conceptual Structures, information processing in mind and machine*, Addison-Wesley, 1984.