IE-Expert: Integrating Natural Language Processing and Expert System Techniques For Real-Time Equity Derivatives Trading

Marco Costantino
Laboratory for Natural Language Engineering
Department of Computer Science
University of Durham, UK
Science Laboratories, South Road
Durham, DH1 3DL, UK
Tel. +44 181 932 2178, Fax. +44 181 932 2178
marco@afin.freeserve.co.uk

December 20, 1998

Abstract

Quantitative data are today largely analyzed by automatic computer programs based on traditional or artificial intelligent techniques, which provide traders with quantitative information that helps them hedge their risks. Qualitative data and, in particular, articles from on-line news agencies are instead not yet successfully processed. As a result, financial operators, notably traders, suffer from qualitative data-overload.

This paper describes how Natural Language Processing, Information Extraction and Expert Systems can be used for reducing the traders' qualitative information overload. In particular, the paper describes IE-Expert, an artificial intelligence system which is able to suggest investment decisions from qualitative information and to link this information to existing quantitative analysis.

1 Introduction: Quantitative and Qualitative information

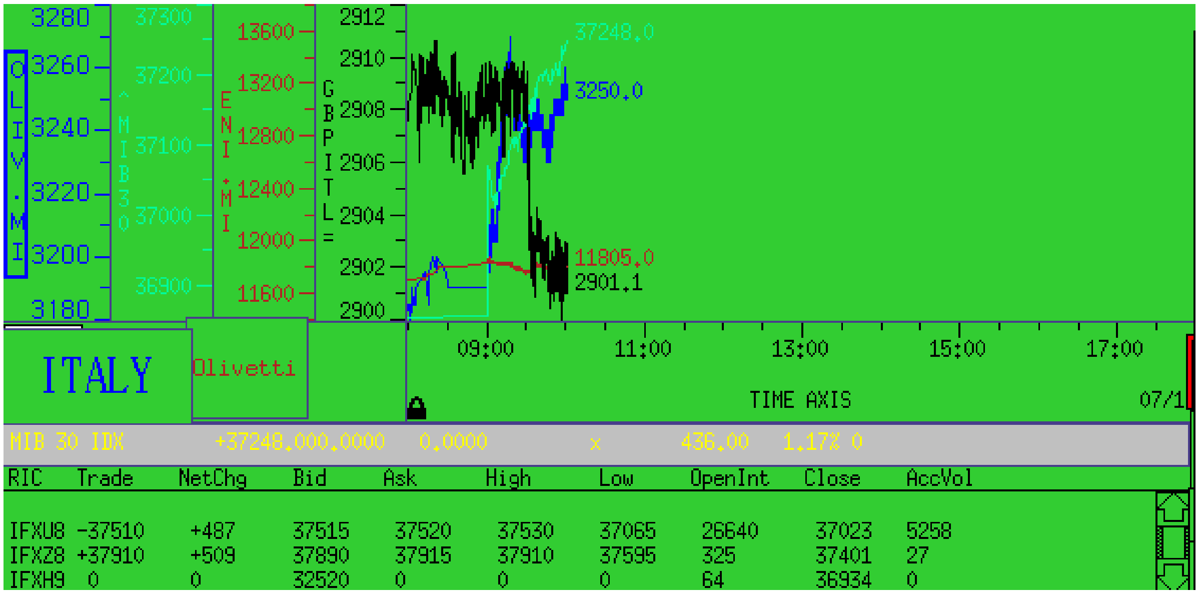
Equity derivatives traders[1] are have today access to a very large amount of information, both qualitative and quantitative, real-time and historical. Quantitative information consists of information which can be easily expressed in numbers (e.g. real-time prices from the major exchanges, volatility implied in exchange-traded option prices etc.[2]). Qualitative information is instead information which cannot be easily expressed in numeric format, for example a sentence such as *"there are fears of an increase in the German interest rates"* from a news article. Figure 1 and 2 show an example of quantitative information available to equity derivatives traders. This includes prices of financial instruments quoted on any exchange e.g. equities, derivatives, exchange-rates, currencies etc.

The real-time information regarding the current behavior of the market together with the risk-management information (prices and risks[3]) of the

[1] A *derivative* (or *derivative security*) is a financial instrument whose value depends on the values of other, more basic underlying variables [Hull, 1997]. *Equity derivatives* are *derivative instruments* based on an equity underlying, for example a stock, but also an equity index (e.g. FTSE-100) or an index future (e.g. the FTSE-100 index future). An *equity derivatives trader* is the financial operator which carries the risk of the derivatives position until expiry, minimising it by *hedging* the derivatives risk using other financial instruments.

[2] *Volatility* is a measure of how uncertain can be the future price movements of a stock an index or, in general, any financial instrument which can be used as an underlying for a derivative. A rough estimate of the volatility can be the standard deviation of the underlying prices over a specific number of days, normally the most recent 90 or 180 trading days. Obviously, the standard deviation can only be measured for past data and using this information for future predictions can be misleading. Equity derivatives traders, therefore, often use *implied volatility* for pricing options, which is the volatility "*priced into*" exchange-traded options available on the market. Implied volatility is therefore the "perception" of the market of the future behaviour of the volatility and can be measured using standard derivatives calculations (e.g. Black and Scholes) from option prices available in the market.

[3] The main risks (often called "*greeks*") associated to derivatives positions are *delta*, *theta*, *gamma*, *vega* and *rho*. *Delta* is the rate of change of the derivative price with respect to the price of the underlying asset [Hull, 1997], that is, the first derivation of the option. *Theta* is the rate of change of the value of the portfolio with respect to time with all else remaining the same [Hull, 1997]. The *gamma* of a derivative is the rate of change of the derivative's delta with respect to the price of the underlying asset [Hull, 1997], in other words the second derivation of the option. The *vega* of a derivative is the rate of change of the value of the derivative with respect to the volatility of the underlying asset [Hull, 1997]. Finally, *rho* is the rate of change of the value of the derivative with respect to



RIC	Trade	NetChg	Bid	Ask	High	Low	OpenInt	Close	AccVol
IFXU8	-37510	+487	37515	37520	37530	37065	26640	37023	5258
IFXZ8	+37910	+509	37890	37915	37910	37595	325	37401	27
IFXH9	0	0	32520	0	0	0	64	36934	0

Figure 1: Quantitative analysis available to equity derivatives traders

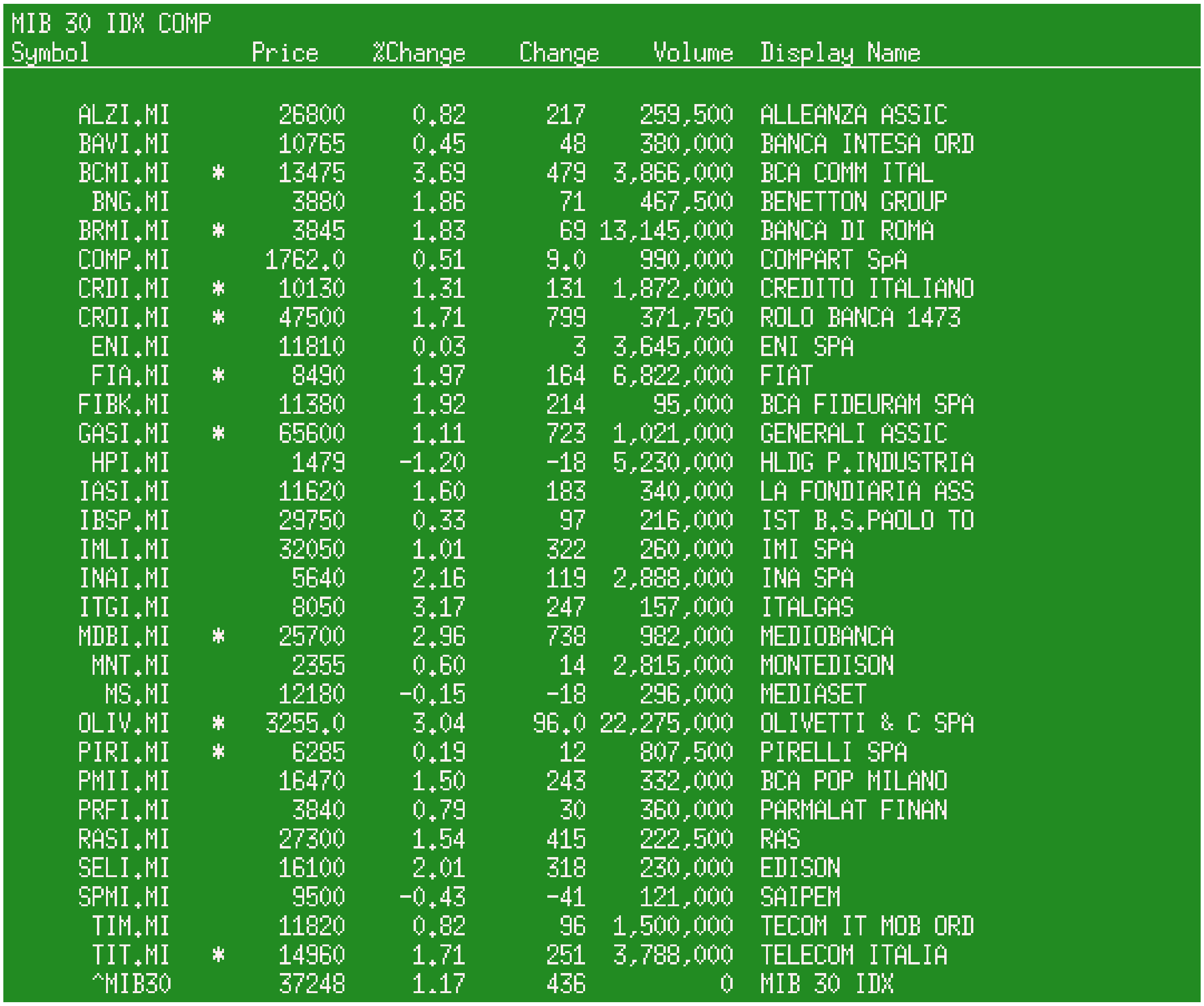
equity derivatives portfolios are used by traders to determine their trading strategies, aimed at maximizing their daily profits.

In order to successfully determine their trading and hedging strategies, however, traders must have a *view* of the market. Let's assume for example that the traders' portfolios are positively correlated to movements of the underlyings in the portfolio (positive delta and gamma of the portfolio). In this case, if the traders have a bearer view of the market[4] and therefore expect the price of the underlying to drop, they will try to hedge their risks by making their portfolios inversely correlated to movements of the underlying price (negative delta). Traders make then use of the quantitative and risk information available to determine the quantity and quality of the hedge to put in place. This could be done, for example, by shorting[5] a

the interest rate. A description of the most important models employed for option pricing can be found in [Hull, 1997]

[4] The term 'bearer' is commonly used in the financial environment when the financial operator expects the market to follow a downward trend. *'Bullish'* is its opposite and it is used then the operator expects the market to rise.

[5] The term "shorting" is commonly used in the financial environment to represent the operation of selling financial instruments, while "going long" represents buying financial instruments.



```
MIB 30 IDX COMP
Symbol              Price  %Change   Change     Volume  Display Name

     ALZI.MI        26800     0.82      217    259,500  ALLEANZA ASSIC
     BAVI.MI        10765     0.45       48    380,000  BANCA INTESA ORD
     BCMI.MI   *    13475     3.69      479  3,866,000  BCA COMM ITAL
      BNG.MI         3880     1.86       71    467,500  BENETTON GROUP
     BRMI.MI   *     3845     1.83       69 13,145,000  BANCA DI ROMA
     COMP.MI       1762.0     0.51      9.0    990,000  COMPART SpA
     CRDI.MI   *    10130     1.31      131  1,872,000  CREDITO ITALIANO
     CROI.MI   *    47500     1.71      799    371,750  ROLO BANCA 1473
      ENI.MI        11810     0.03        3  3,645,000  ENI SPA
      FIA.MI   *     8490     1.97      164  6,822,000  FIAT
     FIBK.MI        11380     1.92      214     95,000  BCA FIDEURAM SPA
     GASI.MI   *    65600     1.11      723  1,021,000  GENERALI ASSIC
      HPI.MI         1479    -1.20      -18  5,230,000  HLDG P.INDUSTRIA
     IASI.MI        11620     1.60      183    340,000  LA FONDIARIA ASS
     IBSP.MI        29750     0.33       97    216,000  IST B.S.PAOLO TO
     IMLI.MI        32050     1.01      322    260,000  IMI SPA
     INAI.MI         5640     2.16      119  2,888,000  INA SPA
     ITGI.MI         8050     3.17      247    157,000  ITALGAS
     MDBI.MI   *    25700     2.96      738    982,000  MEDIOBANCA
      MNT.MI         2355     0.60       14  2,815,000  MONTEDISON
       MS.MI        12180    -0.15      -18    296,000  MEDIASET
     OLIV.MI   *   3255.0     3.04     96.0 22,275,000  OLIVETTI & C SPA
     PIRI.MI   *     6285     0.19       12    807,500  PIRELLI SPA
     PMII.MI        16470     1.50      243    332,000  BCA POP MILANO
     PRFI.MI         3840     0.79       30    360,000  PARMALAT FINAN
     RASI.MI        27300     1.54      415    222,500  RAS
     SELI.MI        16100     2.01      318    230,000  EDISON
     SPMI.MI         9500    -0.43      -41    121,000  SAIPEM
      TIM.MI        11820     0.82       96  1,500,000  TECOM IT MOB ORD
      TIT.MI   *    14960     1.71      251  3,788,000  TELECOM ITALIA
      ^MIB30        37248     1.17      436          0  MIB 30 IDX
```

Figure 2: Quantitative analysis available to equity derivatives traders

number of futures contracts of the underlying where existing or a basket which successfully tracks the underlying when the future is not available[6].

As we can see, quantitative information helps traders understand the risks associated to their position but, in the end, they must take a view of the market, based on their personal qualitative judgment. Traders must therefore take into account the qualitative aspects of the market and need to refer to the current qualitative information available which can be grouped into two main categories:

- Analysis which is produced by internal analysts which includes forecasts for the main markets, indices and companies normally covering a period of one week.
- News from on-line news providers such as Dow Jones, Reuters, Bloomberg etc. which report the latest relevant news for the specific market or region. These news are the main source of qualitative information employed by traders to develop their view of the market.

Figure 3 shows an example of the real-time news available to equity derivatives traders.

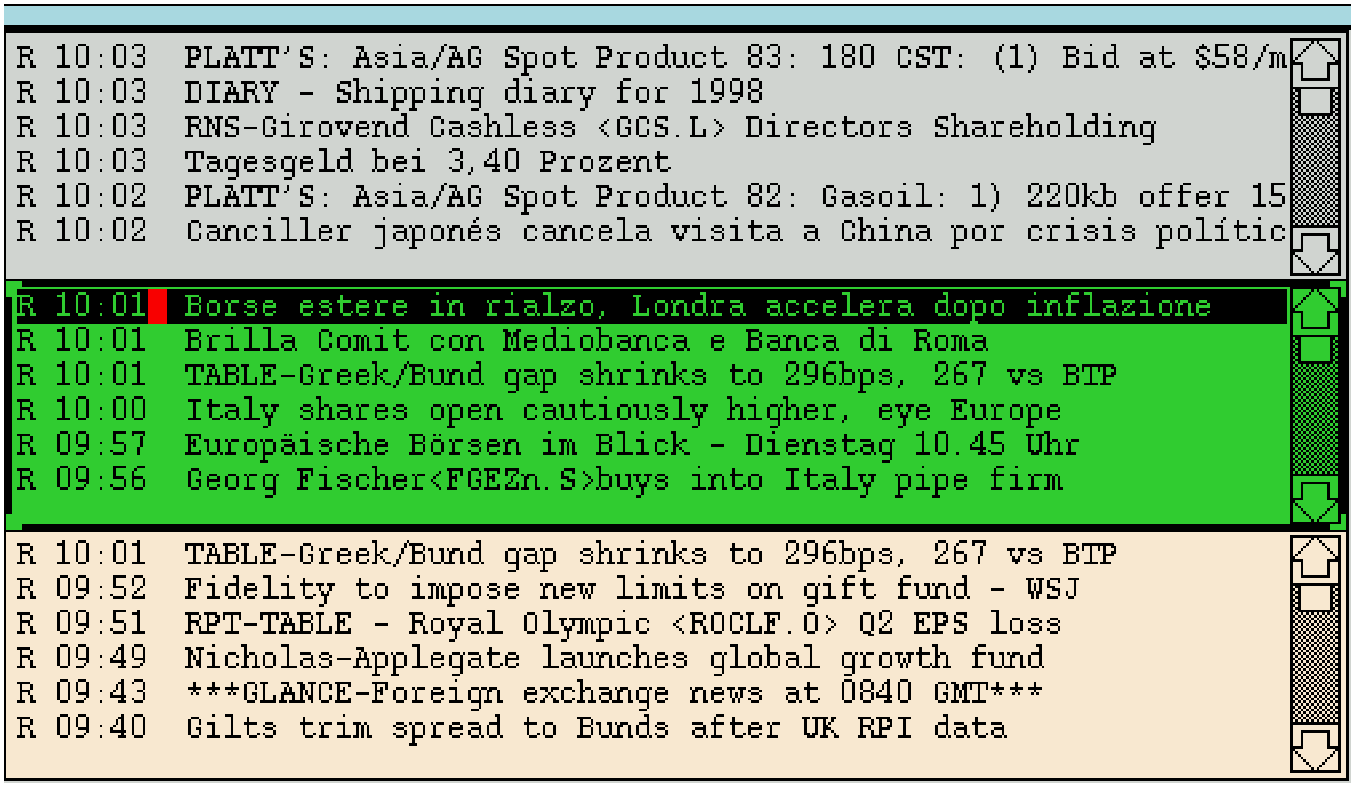
Real-time news are grouped by the information providers into specific categories, for example information for the Italian market, for global markets etc. The titles of the news scroll continuously on the screen and traders can click on a title displaying the full article. Figure 4 shows an article after clicking on the news' title.

A relevant number of articles are displayed every minute. The global-markets section of the market-sheet screen shown in figure 3, for example, tends to display an average of 5-7 articles a minute. Traders, often, are unable to capture and analyze this amount of information in such a short time and, therefore, the qualitative information is lost. Another important limitation of current systems is that there is no link between quantitative and qualitative information and traders must therefore carry out this further analysis.

In our view, the investment decision-making process of an equity derivative trader can therefore be summarized as following (figure 5).

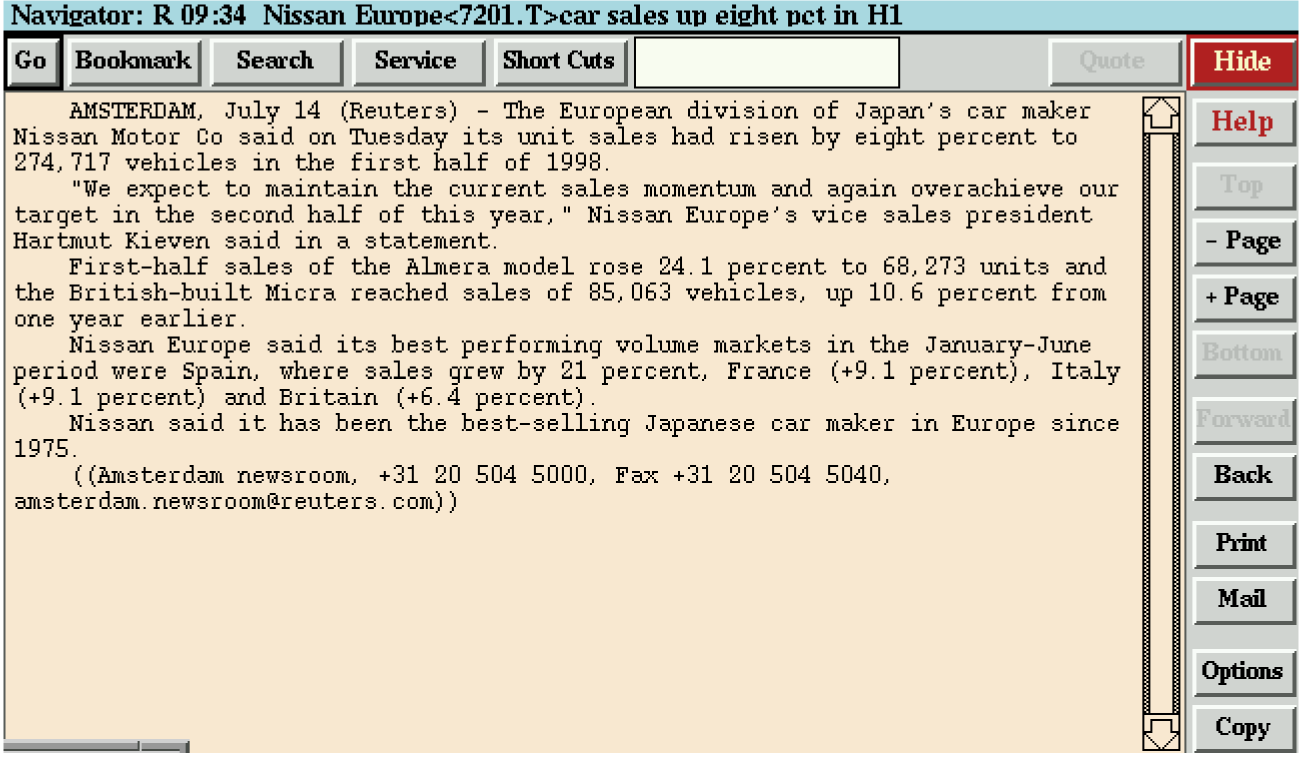
a) The trader assesses the global position and risks of the current portfolio using the quantitative information available; b) the trader takes a view

[6] If the original portfolio had, let's say, a *delta* of +0.5, selling one future contract will bring the new total delta to -0.5. At this point, for a decrease of the market price of the underlying by 1 per cent, the value of the trader's portfolio will increase by 0.5 per cent.



```
R 10:03   PLATT'S: Asia/AG Spot Product 83: 180 CST: (1) Bid at $58/m
R 10:03   DIARY - Shipping diary for 1998
R 10:03   RNS-Girovend Cashless <GCS.L> Directors Shareholding
R 10:03   Tagesgeld bei 3,40 Prozent
R 10:02   PLATT'S: Asia/AG Spot Product 82: Gasoil: 1) 220kb offer 15
R 10:02   Canciller japonés cancela visita a China por crisis polític

R 10:01   Borse estere in rialzo, Londra accelera dopo inflazione
R 10:01   Brilla Comit con Mediobanca e Banca di Roma
R 10:01   TABLE-Greek/Bund gap shrinks to 296bps, 267 vs BTP
R 10:00   Italy shares open cautiously higher, eye Europe
R 09:57   Europäische Börsen im Blick - Dienstag 10.45 Uhr
R 09:56   Georg Fischer<FGEZn.S>buys into Italy pipe firm

R 10:01   TABLE-Greek/Bund gap shrinks to 296bps, 267 vs BTP
R 09:52   Fidelity to impose new limits on gift fund - WSJ
R 09:51   RPT-TABLE - Royal Olympic <ROCLF.O> Q2 EPS loss
R 09:49   Nicholas-Applegate launches global growth fund
R 09:43   ***GLANCE-Foreign exchange news at 0840 GMT***
R 09:40   Gilts trim spread to Bunds after UK RPI data
```

Figure 3: The news available to equity derivatives traders.



```
    AMSTERDAM, July 14 (Reuters) - The European division of Japan's car maker
Nissan Motor Co said on Tuesday its unit sales had risen by eight percent to
274,717 vehicles in the first half of 1998.
    "We expect to maintain the current sales momentum and again overachieve our
target in the second half of this year," Nissan Europe's vice sales president
Hartmut Kieven said in a statement.
    First-half sales of the Almera model rose 24.1 percent to 68,273 units and
the British-built Micra reached sales of 85,063 vehicles, up 10.6 percent from
one year earlier.
    Nissan Europe said its best performing volume markets in the January-June
period were Spain, where sales grew by 21 percent, France (+9.1 percent), Italy
(+9.1 percent) and Britain (+6.4 percent).
    Nissan said it has been the best-selling Japanese car maker in Europe since
1975.
    ((Amsterdam newsroom, +31 20 504 5000, Fax +31 20 504 5040,
amsterdam.newsroom@reuters.com))
```

Figure 4: A financial news article on a trader's screen.

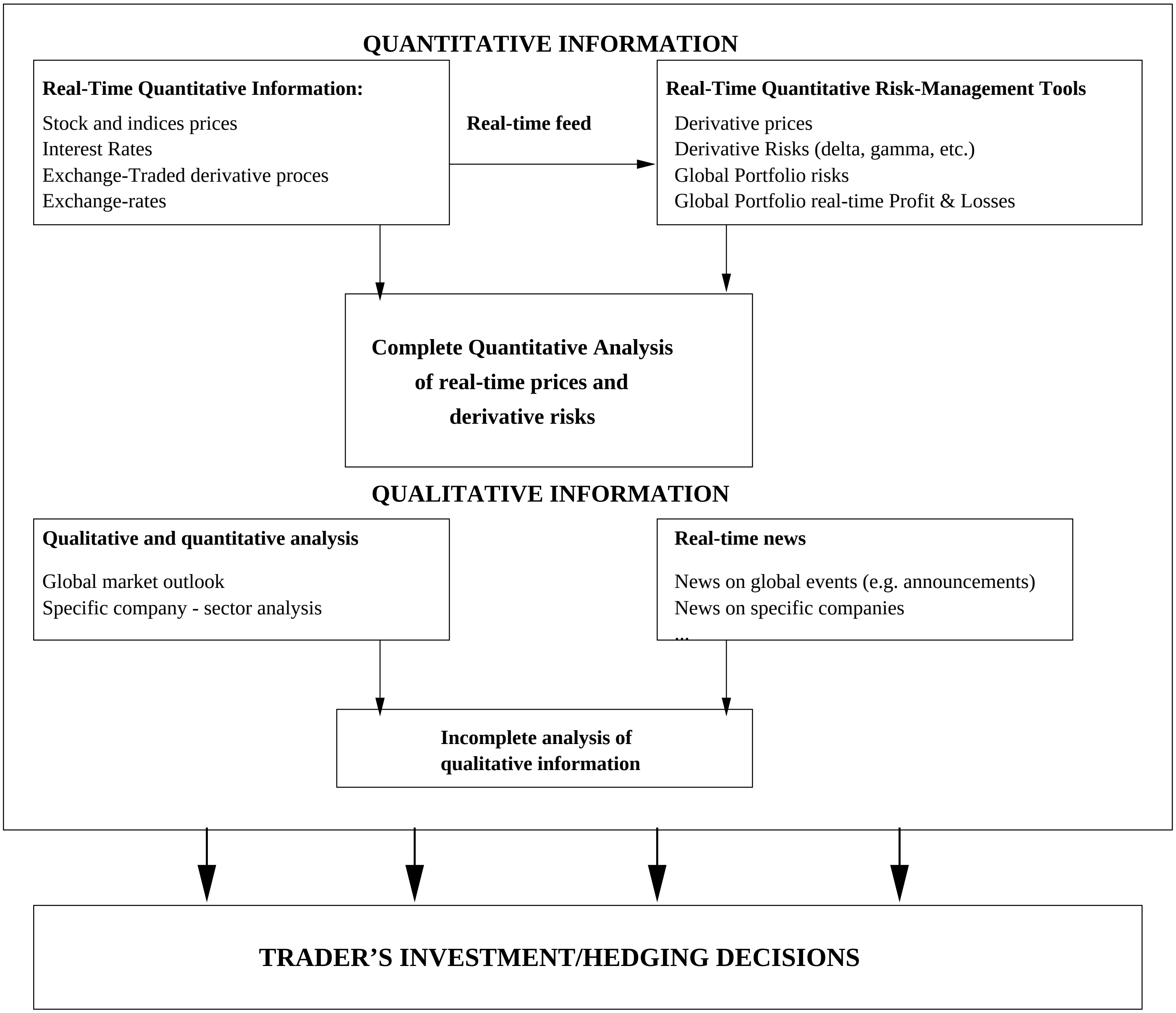

QUANTITATIVE INFORMATION
Real-Time Quantitative Information:
Stock and indices prices
Interest Rates
Exchange-Traded derivative proces
Exchange-rates
Real-time feed
Real-Time Quantitative Risk-Management Tools
Derivative prices
Derivative Risks (delta, gamma, etc.)
Global Portfolio risks
Global Portfolio real-time Profit & Losses
Complete Quantitative Analysis
of real-time prices and
derivative risks
QUALITATIVE INFORMATION
Qualitative and quantitative analysis
Global market outlook
Specific company - sector analysis
Real-time news
News on global events (e.g. announcements)
News on specific companies
...
Incomplete analysis of
qualitative information
TRADER’S INVESTMENT/HEDGING DECISIONS

Figure 5: The decision-making process of an equity-derivatives trader.

of the market, based on the current quantitative and qualitative information available; c) the trader decides the strategy to put in place using the quantitative and risk management information available.

In this paper we outline how natural language processing, information extraction and expert systems can be used to process the qualitative information available to the traders, suggesting possible investment decisions and providing a link between quantitative and qualitative information which is missing in today's financial tools. In particular, we focus on IE-Expert, a prototype system based on information extraction and expert systems technologies. Section 2 introduces Natural Language Processing, Information extraction and Expert Systems. In section 3 and 4 we introduce IE-Expert. Section 3 provides a general introduction to the system, while section 4 focuses on its implementation details. Finally, section 5 evaluates the results of the research.

2 Natural Language Processing, Information Extraction and Expert Systems

The goal of information extraction, which belongs to the field of Natural Language Processing, is to extract specific kinds of information from a source article [Riloff and Lehnert, 1994]. In other words, the input to the system is a collection of documents (e.g. a newspaper article), while the output is a representation of the relevant information from the source documents, according to specific extraction criteria. Using a information extraction system, instead, a "template" (summary) of the original article can be automatically created. The templates will contain the most important relevant information from the source article, while the non-relevant information will not be extracted.

Figure 6 shows a news-article processed by an information extraction system and the template generated by the system.

Information extraction is different from information retrieval. Information retrieval engines are able to locate the relevant documents within a collection, but they are unable to extract information from the relevant documents according to specific criteria. The power of an information extraction system compared to an information retrieval system is therefore in the ability to extract the relevant information in the articles according to specific extraction criteria and represent them in structures (templates), which information retrieval systems are unable to produce. A

Quarto Group, the USM-traded publishing and printing services company, announced that it is buying Front Line Art Publishing, the California-based publisher of art prints and posters, for up to Dollars 9m (Pounds 6m). An initial payment of Dollars 7m will be satisfied by Dollars 5.3m cash and a Dollars 1.7m loan note. There is a further performance-related payment of up to Dollars 2m. For the 1993 year Front Line made profits of Dollars 1.4m, excluding owner remuneration, on turnover of Dollars 5m. Net assets at December 31 were Dollars 1.6m.

Template extracted by the system:

```
Template: TAKEOVER
    COMPANY_TARGET:   Front Line Art Publishing
    COMPANY_PREDATOR: Quarto Group
    TYPE_TAKEOVER:    FRIENDLY
    VALUE:            9 million dollars
    ATTRIBUTION:      Quarto Group
```

Figure 6: A template extracted from a financial news article.

number of natural language processing and information extraction systems have been developed. However, most of these systems have been designed and tested within government agencies and the scientific community, and very few real applications have been commercially successful. A particularly interesting group of systems are those which participated in the MUC competitions [DAR, 1991], [DAR, 1992], [DAR, 1993], [DAR, 1995], a scientific competition for the evaluation of information extraction systems using standard evaluation measures within a specific domain. Among the best performing systems in the competitions are: the Hasten system [Krupka, 1995], the Shogun system [Jacobs *et al.*, 1993], the PLUM System [The PLUM System Group, 1993], the NYU system [Grishman, 1995a] and the LOLITA System [Morgan *et al.*, 1995].

Very few information extraction systems have been specifically designed for the financial domain. One of them is the Durham financial information extraction system, [Costantino, 1997, Costantino *et al.*, 1996] which is used in IE-Expert as information extraction engine. The system, which uses the LOLITA System as its NLP core [Morgan *et al.*, 1995, Garigliano *et al.*, 1993, Garigliano, 1995, Smith *et al.*, 1994], provides a set of pre-defined financial

Company related	**Company restructuring**	**General macroeconomics**
Merger	New product	Interest rates movements
Takeover	Joint venture	Currency movements
Flotation	Staff changes	General macroeconomics data (inflation, unemployment trade deficit)
New issue (shares, bonds etc.)	New factory	
Privatization		
Market movement		
Bankruptcy		
Broker's recommendations		
Taking a stake		
Dividend announcement		
Overseas listing		
Profit/sales forecasts		
Profits/sales results		
Directors' dealings		
Legal action		
Investigation		

Figure 7: The pre-defined financial templates available in the system

templates designed to capture the main financial events which can influence the securities prices. Figure 7 shows the full list of pre-defined templates available in the system, while figure 8 shows the definition of some of the templates.

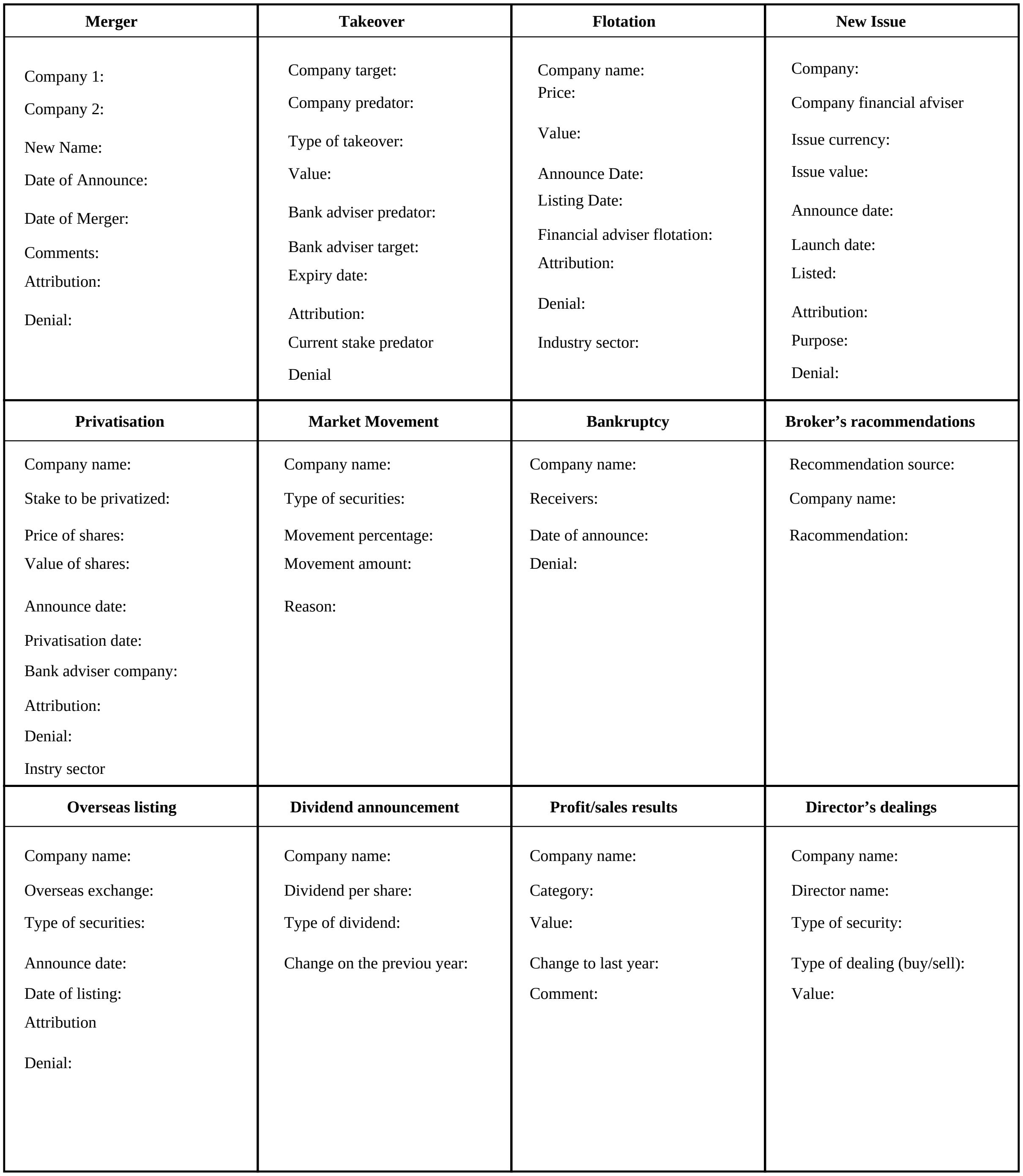
In addition, the system allows the user to define additional templates using a user-friendly natural language user-definable template interface (the user-definable template interface is discussed in more detail in section 4.2). For example, the takeover template shown in figure 9 could be defined by the user using the natural language sentences shown in figure 7.

An expert system can be defined as a "knowledge-based system that emulates expert thought to solve significant problems in a particular domain of expertise" [Zahedi, 1993].

The main characteristics of expert systems is that they are rule-based. This means that the expert system contains a predefined set of rules which is used for all decisions. The system uses the predefined rules to produce results by using inference rules which are coded into the system.

A generic expert system normally consists of two main modules: the knowledge base and the inference engine.

The knowledge base contains the system's knowledge regarding the spe-



Merger	Takeover	Flotation	New Issue
Company 1: Company 2: New Name: Date of Announce: Date of Merger: Comments: Attribution: Denial:	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: Company financial afviser Issue currency: Issue value: Announce date: Launch date: Listed: Attribution: Purpose: Denial:

Privatisation	Market Movement	Bankruptcy	Broker’s racommendations
Company name: Stake to be privatized: Price of shares: Value of shares: Announce date: Privatisation date: Bank adviser company: Attribution: Denial: Instry sector	Company name: Type of securities: Movement percentage: Movement amount: Reason:	Company name: Receivers: Date of announce: Denial:	Recommendation source: Company name: Racommendation:

Overseas listing	Dividend announcement	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 on the previou year:	Company name: Category: Value: Change to last year: Comment:	Company name: Director name: Type of security: Type of dealing (buy/sell): Value:

Figure 8: The most important templates definitions.

```
Template-name:       T=TAKEOVER
Variables:           V=COMPANY1 is an organization.
                     V=COMPANY2 is an organization
                     V=VALUE is money.

Template main-event: V=COMPANY acquired V=COMPANY2
                     V=COMPANY1 acquired V=COMPANY2 with V=VALUE
                     The acquisition of V=COMPANY2 by V=COMPANY1
                     The V=VALUE acquisition of V=COMPANY2 by
                     V=COMPANY1
                     V=COMPANY1 paid V=VALUE for V=COMPANY2.
                     V=COMPANY1 acquired a majority stake in
                     V=COMPANY2.
                     V=COMPANY1 took full control of V=COMPANY2.

Definition of slots:

S=COMPANY-PREDATOR:  V=COMPANY1
S=COMPANY-TARGET:    V=COMPANY2
S=TYPE_OF_TAKEOVER
HOSTILE:             T=TAKEOVER is hostile
FRIENDLY:            T=TAKEOVER is not hostile.
S=VALUE-TAKEOVER:    The cost of T=TAKEOVER
                     V=VALUE
S=ATTRIBUTION:       The person or organization who announced
                     T=TAKEOVER.
```

Figure 9: The definition of a user-definable template

cific area or domain for which it is designed to solve problems or make recommendations. For example, if the system has been designed for the financial domain, the knowledge base will contain rules specific for that area, for example for suggesting equity investment decisions. The knowledge base is coded into the system according to a specific notation. The main groups of notations are:

- Rules
- Predicates
- Semantic Networks
- Frames
- Objects

The inference engine processes and combines the facts related to the particular problem, case and question, using the relevant part of the knowledge-base. The selection of the appropriate data in the knowledge base is performed according to specific searching criteria. The way in which inference rules are written and applied to the information in the knowledge base varies greatly from system to system and can follow different paths.

The most important step for the development of expert systems is the acquisition of the domain specific knowledge, consisting of the methods that would be used by a domain expert for making appropriate decisions. This knowledge will normally consist of heuristics.

3 IE-Expert: Integrating Information Extraction and Expert Systems

This section describes how information extraction and expert systems can be integrated to successfully make use of the qualitative information available to traders in order to suggest possible investment and hedging decisions and link the results to the quantitative real-time information.

The qualitative information available to equity traders can be grouped in two different categories (see section 1): news articles from on-line news providers and in-house research material. IE-Expert is able to process these two categories of qualitative information and produce investment suggestions. In addition, it is able to provide a link between existing quantitative information and the investment decisions produced. The analysis is carried out in three main steps. The first step consists of the identification of relevant qualitative information from both real-time news and research material using the information extraction capabilities of IE-Expert. The second step consists of processing this information. Finally, the investment decision is shown to traders by linking it to existing quantitative information such as prices.

First of all, IE expert processes each of the incoming real-time news articles trying to identify any relevant information. The system processes all information using its information extraction capabilities. If no relevant information is found, the system skips the article and analyses the next incoming news. If any relevant information is found, a template is extracted according to the list of pre-defined templates shown in figures 7 and 9. Figure 6 shows a financial news article and the corresponding template extracted by the system.

At this point IE-Expert retrieves from the database any information available regarding the two companies involved in the takeover and the market sector (according to the relevant region) which they belong to. For example, the system could retrieve the following information:

```
Company:       Tele-Communications Inc.
Negative:      Market under-performer

Company:       BELL ATLANTIC
Positive:      Buy

Market sector: American Telecommunications
Positive:      Expanding rapidly.
```

Once the relevant qualitative information has been identified and processed from both sources, it is fed to the financial expert system, which processes it according to specific rules and suggests an investment decision. The expert system's knowledge consists of a set of investment decision rules which match the financial templates shown in figure 7 and 9, which represent the most likely causes of share prices changes.

From the information shown above, the expert system would produce the following investment suggestions:

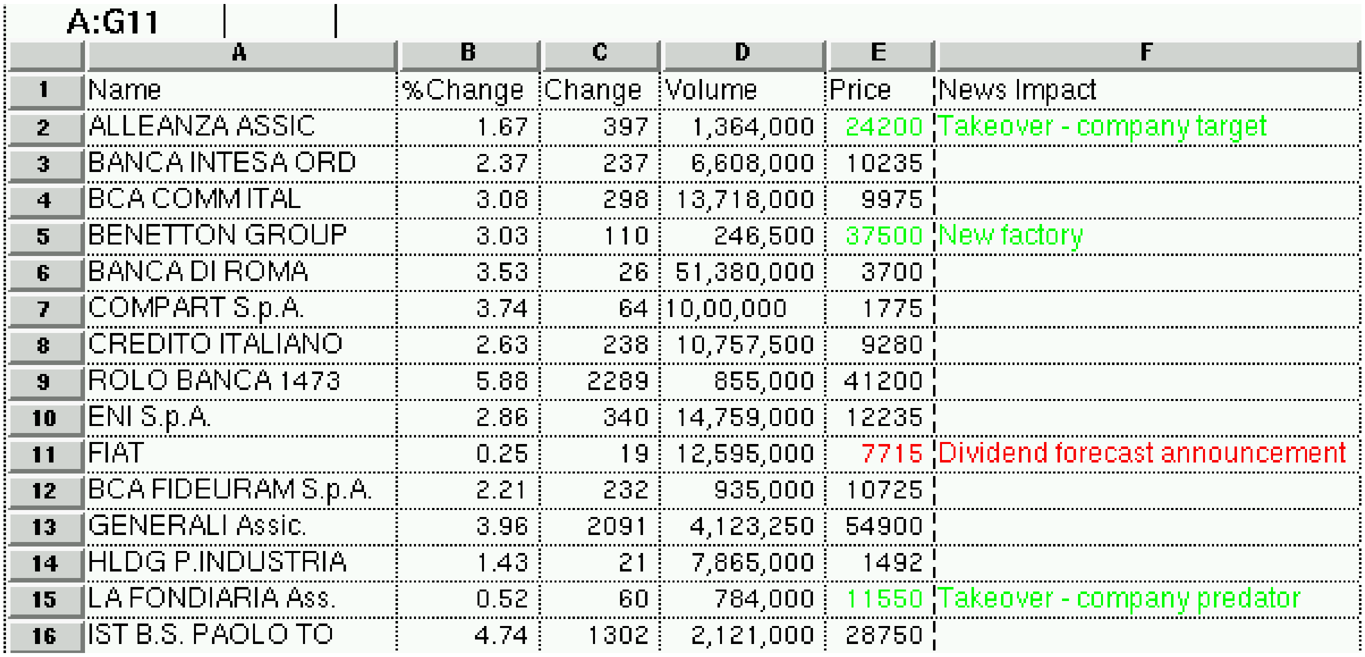
```
1) BELL ATLANTIC:
     Market Sector:   positive (expanding rapidly)
     Company:         positive (buy)

     Financial event: positive (takeover, company_predator)

     Investment decision suggested: share price likely to
     rise - buy

2) Tele-Communications Inc.

     Market Sector:   positive (expanding rapidly)
     Company:         negative (market under-performer)

     Financial event: positive (takeover, company_target)

     Investment decision suggested: share price likely to
     rise - buy
```

The expert system suggested a likely positive impact of the takeover event for both companies and that, as a consequence, the share price of the two companies is likely to increase.

News can also refer to the market as a whole and, in this case, the investment decision suggested by the expert system will be linked to the relevant market index (e.g. the SP 500 index). This is because news regarding macro-economical data such as a reduction in the level of unemployment, generally affect the market globally.

The last step of the system is to display the investment decisions suggested by the expert system in real time to traders. This is done using a specifically designed spreadsheet which reports the live quantitative information together with the information produced by the expert system. Figure 10 shows an example spreadsheet for part of the companies belonging to the MIB30 index.

The color of the column "price" will change whenever a relevant news is processed by IE-Expert and an investment decision is suggested. The color will be red for events with negative impact on the share price, green, for events with positive impact and blue for events with no impact on the share price. Although currently not possible, in the next version of the prototype the trader will be able to click on the financial event and display the relevant



	A	B	C	D	E	F
1	Name	%Change	Change	Volume	Price	News Impact
2	ALLEANZA ASSIC	1.67	397	1,364,000	24200	Takeover - company target
3	BANCA INTESA ORD	2.37	237	6,608,000	10235	
4	BCA COMM ITAL	3.08	298	13,718,000	9975	
5	BENETTON GROUP	3.03	110	246,500	37500	New factory
6	BANCA DI ROMA	3.53	26	51,380,000	3700	
7	COMPART S.p.A.	3.74	64	10,00,000	1775	
8	CREDITO ITALIANO	2.63	238	10,757,500	9280	
9	ROLO BANCA 1473	5.88	2289	855,000	41200	
10	ENI S.p.A.	2.86	340	14,759,000	12235	
11	FIAT	0.25	19	12,595,000	7715	Dividend forecast announcement
12	BCA FIDEURAM S.p.A.	2.21	232	935,000	10725	
13	GENERALI Assic.	3.96	2091	4,123,250	54900	
14	HLDG P.INDUSTRIA	1.43	21	7,865,000	1492	
15	LA FONDIARIA Ass.	0.52	60	784,000	11550	Takeover - company predator
16	IST B.S. PAOLO TO	4.74	1302	2,121,000	28750	

Figure 10: Merging real-time quantitative and qualitative information

article and market analysis which originated the event, to better understand its scope and impact on the share price.

IE-Expert helps therefore traders overcome their qualitative data-overload and link the quantitive and qualitative information together, which allows them to quicker define their current view of the market for the next investment and hedging decisions.

Although IE-Expert includes a set of pre-defined templates and expert system rules, it has been designed as a fully customizable system. This is because different traders and financial institutions might have different views and trading strategies.

The system is customizable at two different levels: the financial templates and the expert systems rules. New templates can be easily added to the system using a specific user-definable template interface. The user interface allows new users to define new templates using sentences in natural language using specific formal elements (e.g. the takeover template definition shown in figure 9). The user can also customize the expert system providing the rules for processing the extracted templates and the research analysis available. However, in the first version of the prototype the rules for the Expert System are currently hard-coded in the source code. A project is currently in progress for the development of a user-friendly interface for inputting new financial rules.

4 The implementation of IE-Expert

IE-Expert is based on two main components. The first component is the information extraction engine which identifies and extracts the relevant information from the incoming real-time financial news. The second component is the expert system, which is used to process the templates extracted by the information extraction component and the market analysis to produce investment suggestions. In addition, a postgres database stores the market data associated to a specific company or sector and an Applix spreadsheet is used to display the results. The architecture of the system is shown in figure 11.

4.1 The NLP information extraction component

The information extraction component is based on the Durham financial information extraction system, under development at the University of Durham, UK [Costantino, 1997].

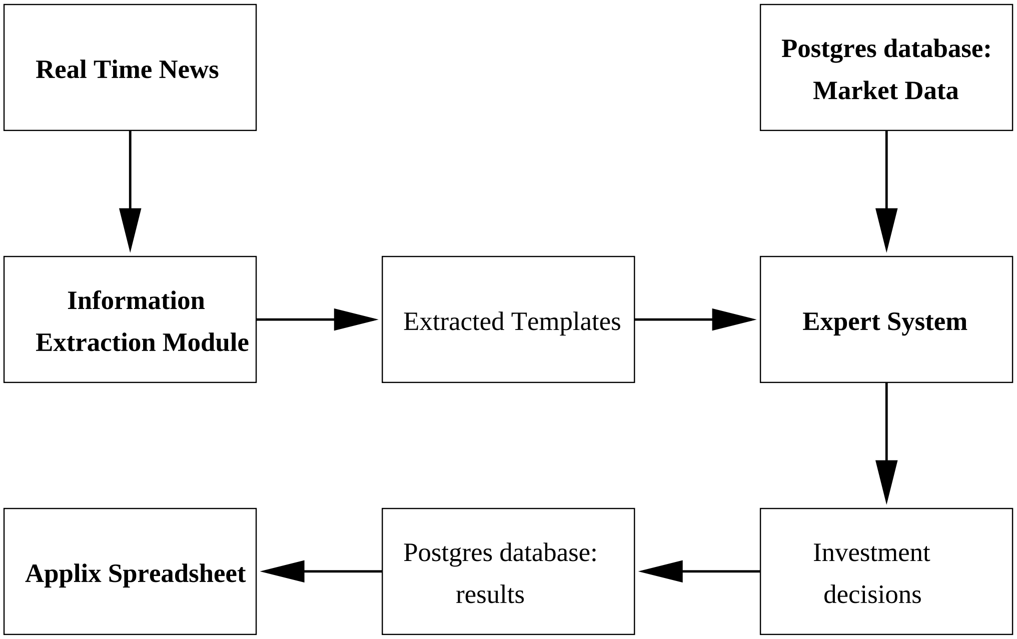

Real Time News
Postgres database: Market Data
Information Extraction Module
Extracted Templates
Expert System
Applix Spreadsheet
Postgres database: results
Investment decisions

Figure 11: The architecture of IE-Expert

The basic task of the natural language processing system is to process the input text and produce a representation of its meaning. This representation is then stored in an appropriate knowledge-base and can then be used for various different tasks and to generate natural language text. The core of the system is a large (over 100,000 nodes) semantic network, which consists of a hierarchy of nodes connected with arcs. The nodes represent entities (a company) and events (e.g. The company made a takeover). Each node is associated to specific control variables which are used to specify the type and properties of each node. Some of the control variables are as follows:

- **Rank.** This control gives the nodes quantification, i.e. individual, named individual, universal, existential, bounded existential etc. For example, the node "Robert" in the sentence "Robert owns a motorbike" is a named-individual.
- **Type.** This control is very similar to grammatical qualifications and comprises: entity, relation, typeless, event, fact, greeting etc. For example, the sentence "Robert owns a motorbike" is a fact.
- **Family** This control groups the nodes into semantic "families" which share specific properties, e.g.: living, animal, human, man-made, abstract, location, organization, human-organization etc. For example, the node "Robert" belongs to the family "human".

Source articles are processed by the system through four hierarchical modules: morphology, parsing, semantics and pragmatics (figure 12):

- The **morphology** module splits the input text into words and smaller units and produces for each word a list of possible meanings together with their syntactic and semantic categories. The input is then supplied to the parser.
- The **parser** module performs a full grammatical analysis of the source sentence recognizing the role of each of the words in the sentence, for example subject, object, verb and adjective. At this stage, the meaning of each of the words in the sentence is not yet determined, and will be resolved by the subsequent modules of the analysis.
- The **semantic analysis** module associates each of the words with their appropriate meanings and maps them onto the system's internal representation in a format compatible with the semantic network.

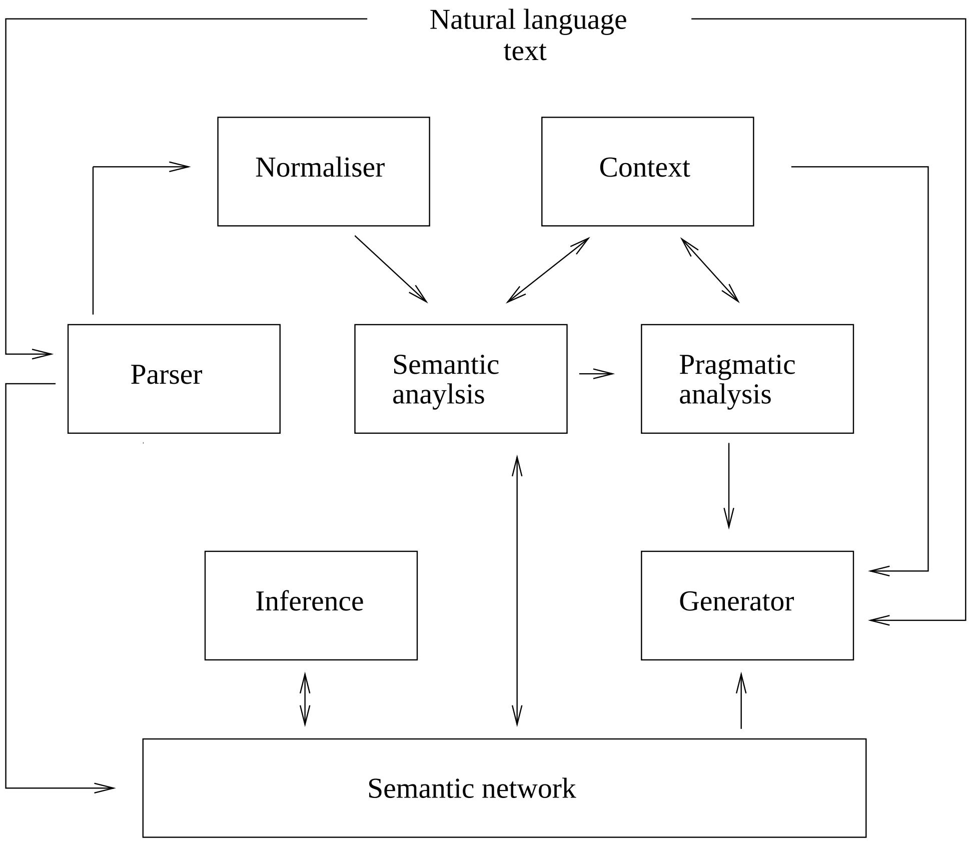


Figure 12: The Durham NLP System's core.

- The **pragmatic analysis** module performs the disambiguation of the meanings introduced by the semantic analysis module and type checking.

At the end of the analysis process the new knowledge is stored in the semantic network. To produce the templates, the new knowledge obtained from the analysis of the source articles is matched against the templates definitions defined using the user-definable template interface, which is discussed in more detail in the next section and the final templates are extracted from the source texts.

4.2 The User-Definable Template Interface

One of the main criticisms that can be made to many of the existing information extraction systems is that users cannot configure the systems to produce results (templates) which differ from those already available in the system. The *templates* are usually coded within the system and the user cannot modify the existing templates or add new ones without having to directly alter the system's code.

The *Hasten* system, which successfully participated in the MUC-6 competition [Krupka, 1995] is one of the few systems which shows a partially-customisable environment. The interface is based on *example-patterns* corresponding to relevant fragments of source texts which can be entered by the user and will be used for producing the templates. Although the interface presents the advantage of allowing the users' definition of the slots, two main problems arise in the definition of a new templates: the template is still coded in the system; the user is required to enter a considerable number of example patterns for the definition of each slot.

The Durham financial user-definable template interface has been designed to allow the end-user to enter new template definitions using natural language sentences using specific formal elements[7]. A generic template such as the takeover template shown in figure 9 can be represented in the system with the following key elements:

1. the **template-name** which uniquely identifies the template among the others in the collection;

[7] In a paper of this size it is impossible to fully describe all the technical details of the Durham user-definable template interface. Further information on this topic can be found in [Costantino, 1997].

2. the **main-events** of the template, which represent the conditions under which the template has to be instantiated by the system;
3. the **slot-names** which uniquely identify each of the slots in the template;
4. the **slot-rules** which are used by the system to identify the relevant information for each of the slots.

The user-definable template interface will therefore need to allow the user to define these elements. In the Durham financial user-definable template interface the users can define new templates using the following formal elements, which have been designed to reduce the amount of possible ambiguities in the template definitions without reducing the user's expressive power. The formal elements are:

- the **name of the template**, which distinguishes the template among the other templates in the system;
- the **template variables**, which identify the elements of the main-events which will be later used in the definition of the slot-rules.
- the **slot-names**, which identify the specific template's slots and can be used in the definition of other slot rules to refer to the information contained in the previous slots.

The user can enter a new template definition using sentences in natural language. For example, the user could define the template main-event as follows (see figure 9):

```
V=COMPANY1 acquired V=COMPANY2.
V=COMPANY1 acquired V=COMPANY2 with V=VALUE.
The acquisition of V=COMPANY2 by V=COMPANY1.
The V=VALUE acquisition of V=COMPANY2 by V=COMPANY1.
V=COMPANY1 paid V=VALUE for V=COMPANY2.
V=COMPANY1 acquired a majority stake in V=COMPANY2.
V=COMPANY1 took full control of V=COMPANY2.
```

where the template variables have been previously defined as

```
Variables:             V=COMPANY1 is a company.
                       V=COMPANY2 is a company.
                       V=VALUE is money.
```

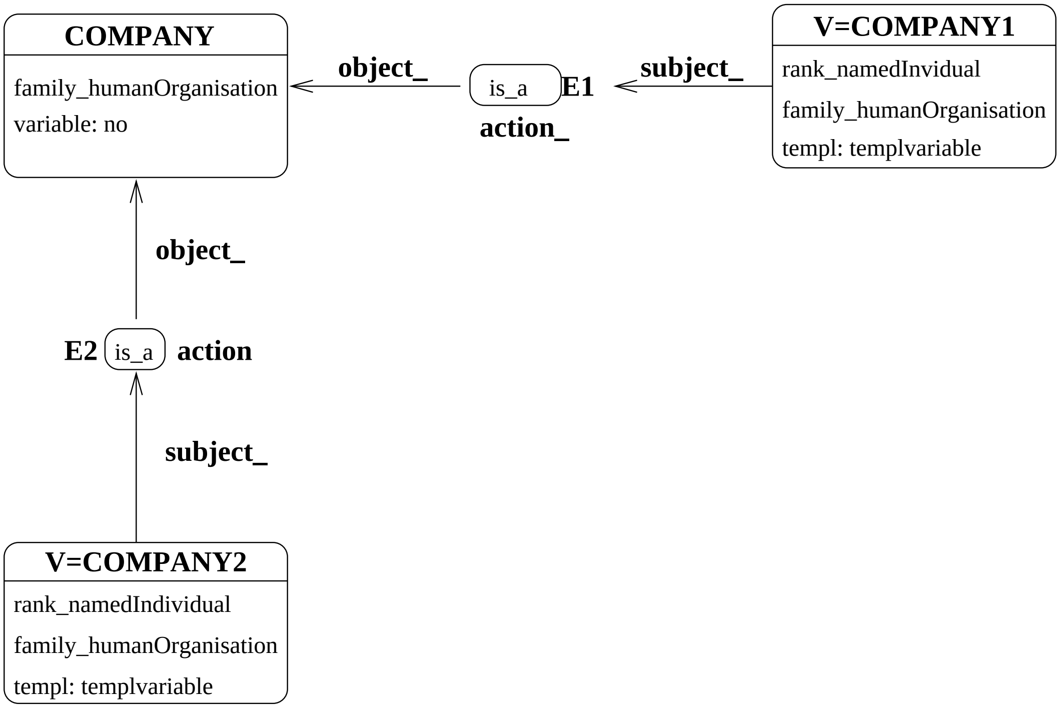


Figure 13: The processing of a the variable "V=COMPANY1 is a company."

Once the templates have been defined by the user using natural language sentences, they are processed by the natural language processing system and stored in the semantic network. The next step is to match the template definition against the set of source articles to extract the final templates.

The first step taken by the user-definable interface is to process the template definitions supplied by the user ("*ExtractionNeeds*"). This corresponds to the operation "*Customise (ExtractionNeed)*" of the TIPSTER phase II document [Grishman, 1995b]. The *template-name*, the *variables* (figure 13), the *main-conditions* (figure 14) and the *slot rules definitions* are processed and stored in the semantic network.

The inference system will then try to match these questions against the new information acquired from the processing of a source article. For example, for the main-event shown in figure 14:

```
V=COMPANY1 acquired V=COMPANY2
```

the inference system will recognize that an event such as:

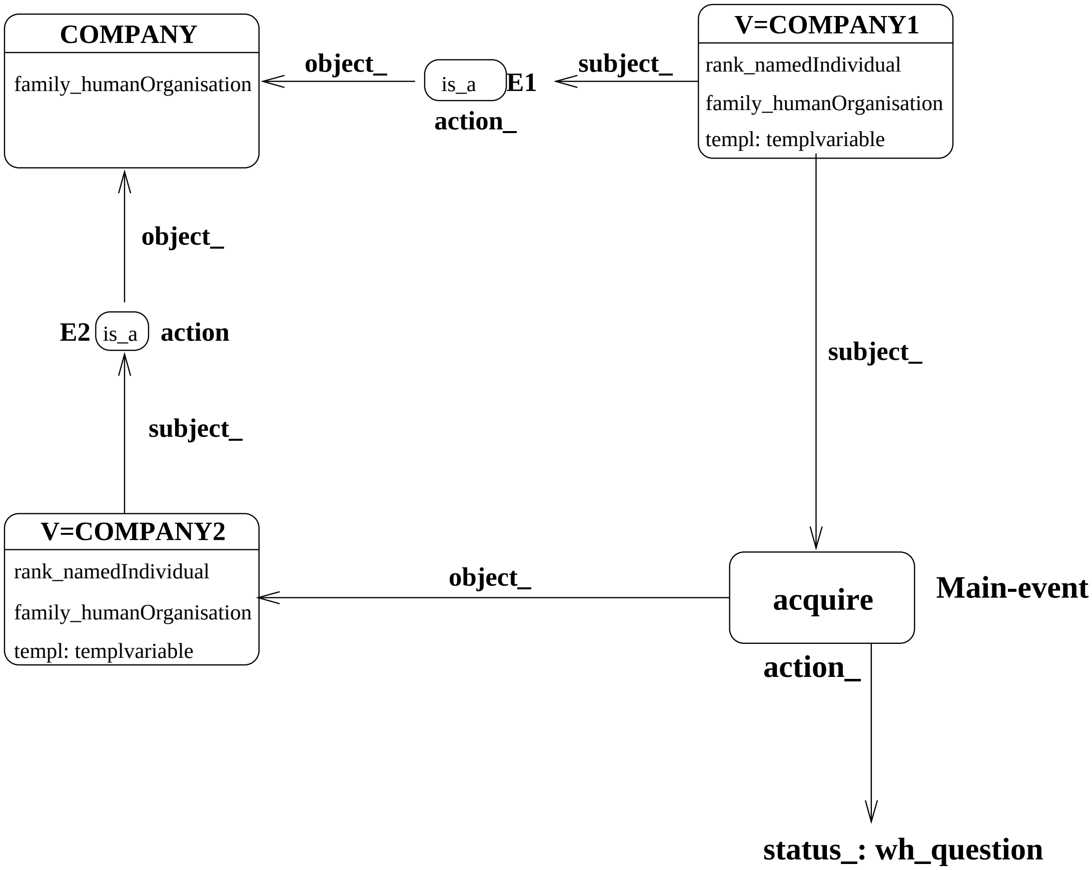


Figure 14: The processing of the main-event "V=COMPANY1 acquired V=COMPANY2.

```
Fiat purchased Renault
```

is a relevant one, because of the fact that the action is compatible with "acquire" and the subject and object can be matched against the variables "V=COMPANY1" and "V=COMPANY2".

Figure 15 shows the representation of the main-event and the candidate event "Fiat bought Renault". The inference system tries to match each of the components of the candidate event onto the main-event.

The inference system will therefore look for an event which satisfies the following condition:

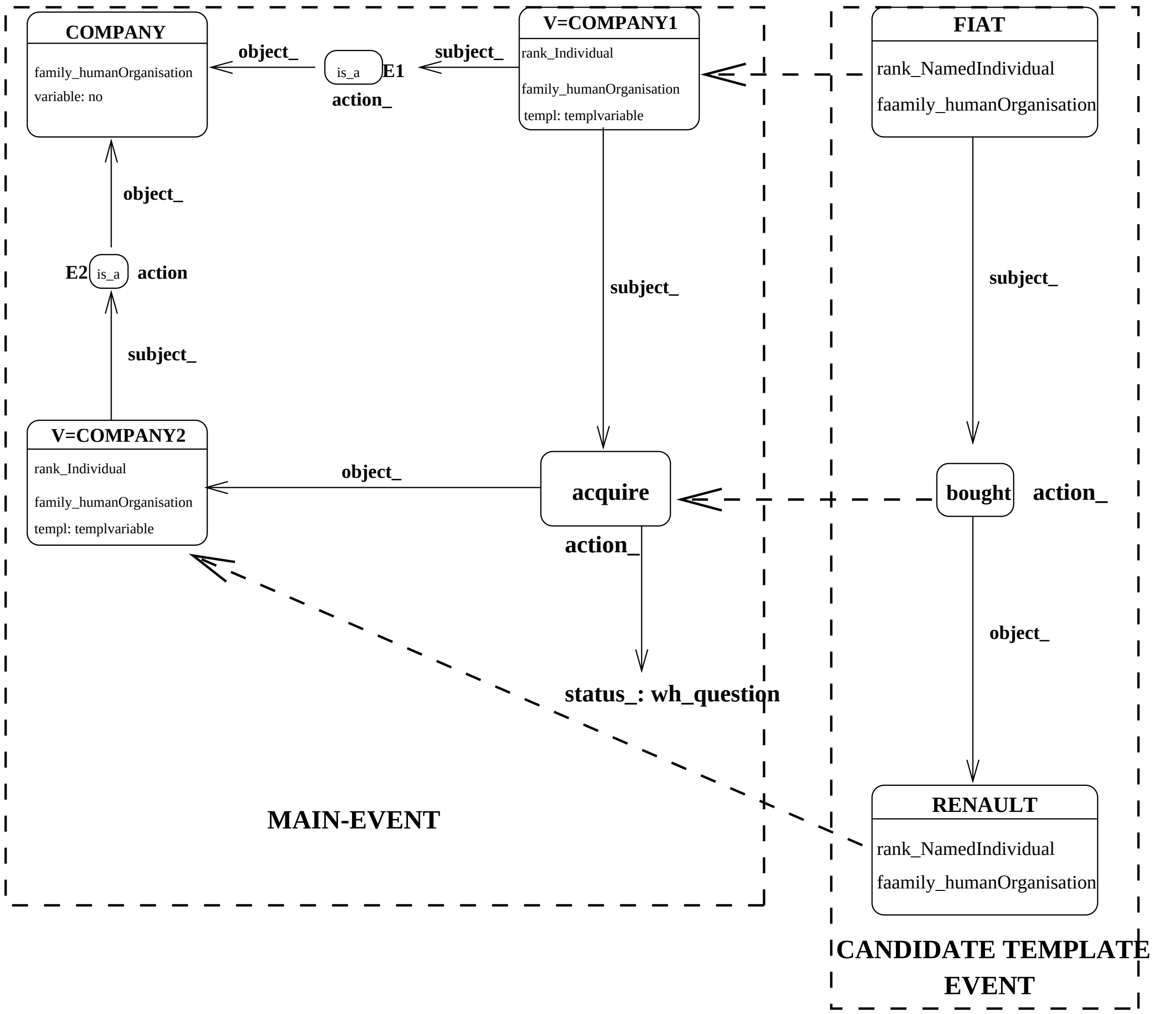


Figure 15: Identification of candidate main-events by the inference system.

$\exists$ *V=COMPANY1, V=COMPANY2. Acquire(V=COMPANY1,V=COMPANY2)*

Once the candidate events have been identified, these can be used by the inference system for searching for concepts which match the slot rules.

Inference and the variables

The variables are filled in by the inference system as part of the processing of the main-events. Therefore, specific calls to the inference system for locating information which corresponds to the variables are not necessary.

Inference and the slots

The slot-rules definitions entered by the user can be subdivided into two different categories:

- rules which refer only to a specific variable used in the main-event, for example:

  ```
  S=VALUE-TAKEOVER:       V=VALUE
  ```

  This kind of slots is filled with the concepts which have already been identified for the specific variable.

- rules which refer to specific variables, the template-name or other slot-names but adding additional conditions, for example:

  ```
  S=VALUE-TAKEOVER:       the cost of the T=TAKEOVER
  ```

  In this case, the inference system will be called again and will look for any event or entity which matches the slot-rules.

Figure 16 shows the takeover template extracted from a source financial article. The template has been produced using the takeover template definition shown in figure 9.

4.3 The Expert System component

Once the templates have been produced, the system retrieves any associated market data information from the database which is currently based on a

Reuters Holdings yesterday announced that it acquired Teknekron Software Systems for 125.1 million dollars cash. Teknekron, a software supplier and systems integrator based in Palo Alto, California with a workforce of 200, had turnover last year of 38.7 million dollars and pre-tax profits of 8.2 million dollars. Net assets at the end of 1992 were 3.6 million dollars. Reuters has 212,000 information outlets worldwide, including 350 of the latest digital Triarch systems. Under the deal, which has to clear both the US and UK regulatory authorities, Teknekron will retain operational control of the company. Two non-executive directors from Reuters will join the Teknekron board. Teknekron's management will also benefit from a stock appreciation plan, similar to a share option scheme.

Template produced by the LOLITA system:

```
<T=TAKEOVER> :=
    S=TYPE-TAKEOVER: FRIENDLY
    S=VALUE: "Million 125.1 dollar cash. "
    S=ATTRIBUTION: "Reuters Holdings. "
    S=COMPANY_TARGET: "Teknekron Software Systems. "
    S=COMPANY_PREDATOR: "Reuters Holdings. "
```

Figure 16: An example takeover template produced by the user-definable template interface using the takeover template definition shown in figure 9.

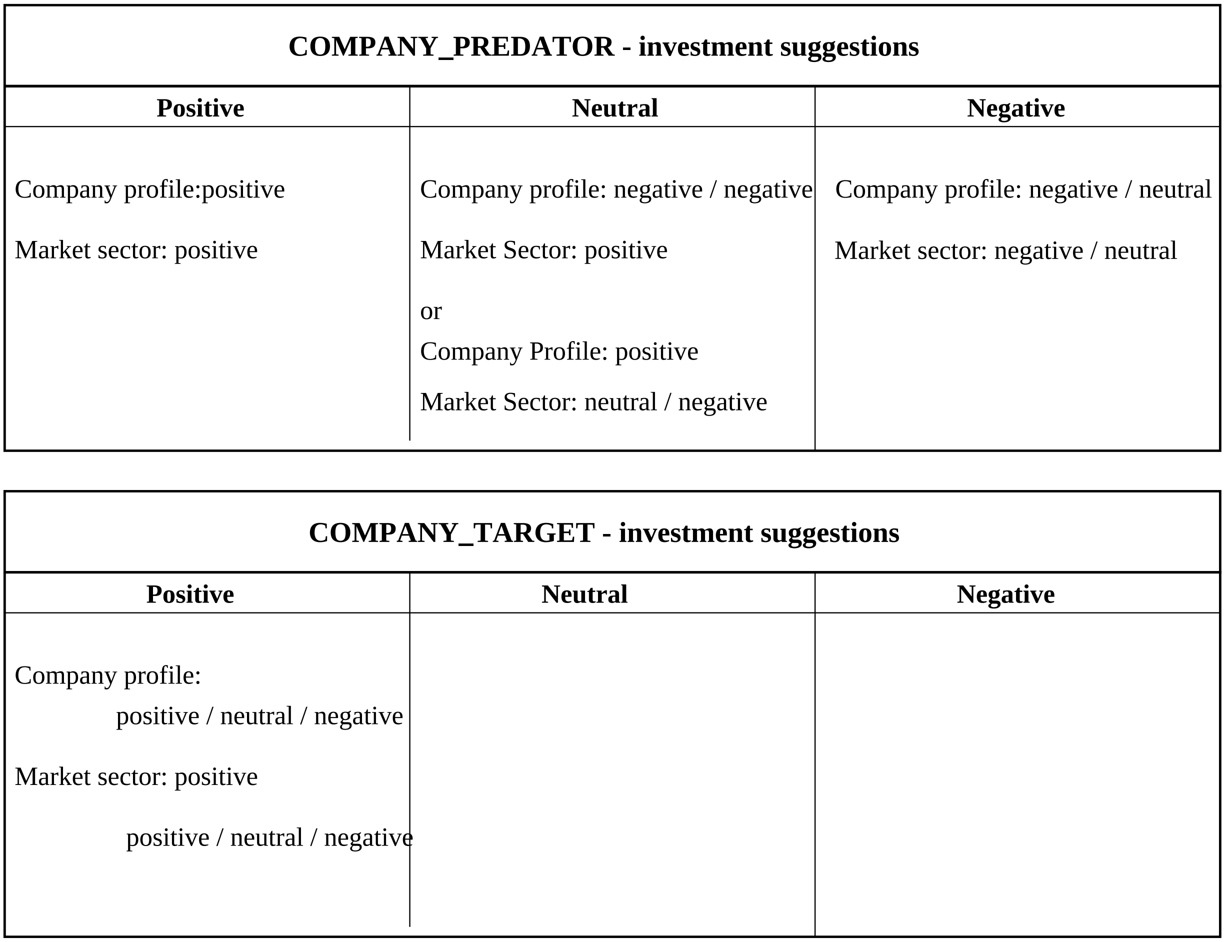
postgres server. The text of the template and the market data information are subsequently fed to the Expert System.

The expert system currently employed is relatively straightforward. The templates definitions and the market data available are parsed and stored in memory. At this stage, the information is processed using a set of rules corresponding to each of the pre-defined templates available in the information extraction module. The rule corresponding to the template is matched against the new information and an investment decision is produced. The expert system rules are represented as a table of True/False conditions which are matched against the slots of the template produced and of the associated market-data.

The current prototype of IE-Expert is based on a restricted number of rules, which have only been defined for the takeover template. Figure 17 shows the expert system rules associated to the takeover template. In addition, the initial version of the prototype requires the rules to be directly hard-coded in the system, making it impossible for end-users to easily add new rules to the expert system or update existing ones. Further work is currently being carried out firstly to increase the number of rules currently in the system and secondly to design and develop a user-friendly interface which allows the user to enter new rules directly. In fact, while the system currently allows the user to enter new template definition using the user-definable template interface (see section 4.2), the user is unable to enter new rules for the expert system, considerably limiting the global performance of the system. Further work is also being carried out to increase the number of standard templates directly available in the system and the corresponding expert system rules, the first one being the *merger* template, which has already been defined and it is shown in figure 18.

The investment decision produced by the expert system, together with the associated template and market information is subsequently stored in a database. Finally, an Applix Spreadsheet is used to retrieve the information from the database and display the results.

The information extraction system is written in the functional language Haskell and C language. The expert system is written in C. A postgres database is used for storing the market data information and the investment decisions and templates produced by the system. The information is displayed using an Applix spreadsheet which accesses directly the Postgres database. The system has been mainly written in C rather than in other languages for performance reasons. This choice, however, does not potentially prevent the system from being used within the Internet. The Applix



COMPANY_PREDATOR - investment suggestions		
Positive	**Neutral**	**Negative**
Company profile:positive Market sector: positive	Company profile: negative / negative Market Sector: positive or Company Profile: positive Market Sector: neutral / negative	Company profile: negative / neutral Market sector: negative / neutral

COMPANY_TARGET - investment suggestions		
Positive	**Neutral**	**Negative**
Company profile: positive / neutral / negative Market sector: positive positive / neutral / negative		

Figure 17: The expert systems rules for the takeover event.

```
Template_Name:          T=MERGER

Variables:              V=COMPANY1 is a company.
                        V=COMPANY2 is a company.
                        V=COMPANY3 is a company.

Template main-events:   V=COMPANY1 merged with V=COMPANY2 creating V=COMPANY3
                        V=COMPANY1 merged with V=COMPANY2

Definition of the slots:
S=FIRST-COMPANY:        V=COMPANY1
S=SECOND-COMPANY:       V=COMPANY2
S=NEW-NAME:             V=COMPANY3
S=DATE-OF-ANNOUNCE:     the date when T=MERGER is announced
S=DATE-OF-MERGER:       the date when T=MERGER takes place
S=ATTRIBUTION:          the person that announced T=MERGER
                        the company that announced T=MERGER
```

Figure 18: The merger template

spreadsheet, in fact, could be easily substituted with web pages interfaced directly to the core system and the interface to the core system could be written using Java. The system currently runs on a Sun SPARCstation with 80MB of RAM. However, it can easily be adapted for use within other Unix environments.

5 Evaluation and results

The evaluation of the results was carried out focusing on the performance of the information extraction module, which is essential for the system's success. This is because if any relevant information is missed or or non-relevant information is mistakenly extracted, the investment suggestions produced by the expert system could be misleading.

The performance of the information extraction module was evaluated scoring the results of the information extracted for the user-defined takeover template shown in figure 9 from an evaluation set of 55 financial articles (25 relevant takeover articles and 30 non-relevant financial articles)[8]. Figure 19 shows a relevant takeover article from the evaluation set.

Cowie Group, the car leasing and motor trading company, yesterday announced a big expansion of its bus operations with the 29.9 million pounds acquisition of Leaside Bus Company, the subsidiary of London Regional Transport (LRT). The deal, involving a 25.5 million pounds cash payment and 4.4 million pounds to settle intra-group loans, will enlarge Cowie's bus fleet from 128 vehicles to more than 600 and is expected to lead to a fourfold sales increase.

'We paid slightly more than we wanted to, but it was worth it for the enormous growth that it promises,' said Mr Gordon Hodgson, chief executive. The acquisition follows four months of talks between LRT and Cowie, which has been seeking a larger stake in the London bus network for more than two years.

At present, the group's bus and coach operations are dominated by Grey-Green - acquired 14 years ago - which serves 13 bus routes in London and employs 450 drivers. Leaside, by comparison, has a work force of about 1,800 and operates 28 routes.

Mr Hodgson, who is meeting Leaside managers today, said he was determined to introduce private sector efficiency to the business, which last year made profits of just 607,000 pounds on turnover of 43 million pounds. In the same period, Grey-Green made profits of 1.6 million pounds on sales of 14.4 million pounds. Cowie shares fell 3 1/2 p to 218 1/2 p yesterday - a new low for the year.

Figure 19: A relevant article of the evaluation set.

The scores have been computed using the MUC-6 scoring program which was released to the developers of the MUC-6 systems [Chinchor and Dungca, 1995]. The template definition of the scoring program was changed to the user-defined takeover template in place of the original MUC-6 templates, while the evaluation measures and criteria of the scorer were not modified. The scoring program matched the templates produced by the system for each article against the corresponding key templates producing a summary reporting the *precision*, *recall* and the combined *'F' measure*[9]. Figure 20

[8] A complete discussion of the methodologies for evaluating information extractions systems is beyond the scope of this paper. More information regarding this topic can be found in [Chinchor and Dungca, 1995, Callaghan, 1998]

[9] *Precision*, *recall* and *'F' measure* are standard measures employed in information retrieval and information extraction to evaluate the performance of a system. *Precision* can be thought of as the ratio of the number of relevant documents retrieved to the total number of documents retrieved [Rijsbergen, 1979]. The MUC *precision* measure was

shows the overall results for the 55 articles of the evaluation set. The final results showed that the system's overall performance measures were:

```
                                         P&R      2P&R      P&2R
F-MEASURES                               51.03    57.41     45.93\

OVERALL PRECISION:                       63%
OVERALL RECALL:                          43%
```

The overall figure (51%) is rather high. The precision (63%) is significantly higher than the recall (43%). The high performance of the information extraction module should allow the expert system to produce correct investment suggestions.

We believe that evaluating the system using 55 financial articles can provide us with a good indication of the system's performance. However, further evaluation experiments are currently being carried out to complete these results. Rather than evaluating the system using additional takeover articles, we are currently focusing on the evaluation of the merger template (figure 18) using a different set of financial articles. This would provide us with an indication of the performance of the system in domains other

adapted for information extraction systems:

$$precision = \frac{correct + (partial\ .\ 0.5)}{number\ of\ actual\ answers}$$

Recall is the ratio between the number of relevant documents retrieved and the total number of relevant documents (both retrieved and not retrieved) [Rijsbergen, 1979]. The MUC *recall* measure was adapted for information extraction systems:

$$recall = \frac{correct + (partial\ .\ 0.5)}{possible}$$

Finally, the *'F' measure* represents a way to combine the precision and recall measures into a unique value and was first introduced by van Rijsbergen [Rijsbergen, 1979]. The *'F' measure*, as combination of precision and recall, gives a values that falls between them. The β parameter in the *'F' measure* represents the relative importance given to recall over precision and in the case recall and precision are of equal weight, β assumes value 1.0. The *'F' measure* presents a higher value if precision and recall are more at the center of the recall-precision graph than if they are at the extremes of it. For example, if a system has precision and recall both of 50 per cent, the *'F' measure* will be higher than a system that has recall of 20 per cent and precision of 80 per cent. This is also because the aim of the formula is to direct developers towards an improvement of both recall and precision.

$$F\ measure = \frac{(\beta^2 + 1.0).P.R}{(\beta^2.P) + R}$$

```
Report for the standard takeover templates finalEval2:

 * * * TOTAL SLOT SCORES * * *
------------------------+-------------+--------------+------------------------
SLOT            POS   ACT| COR PAR INC | MIS  SPU  NON| REC PRE UND OVG ERR SUB
------------------------+-------------+--------------+------------------------
takeover         36    38|  28   0    0|   8   10    0|  78  74  22  26  39   0
  companytar     36    21|  12   0    5|  19    4    0|  33  57  53  19  70  29
  companypre     35    31|  18   0    3|  14   10    0|  51  58  40  32  60  14
  typetakeov     36    38|  28   0    0|   8   10    0|  78  74  22  26  39   0
  value          28     7|   3   0    1|  24    3    1|  11  43  86  43  90  25
  badviserpr      0     0|   0   0    0|   0    0    0|   0   0   0   0   0   0
  badviserta      0     0|   0   0    0|   0    0    0|   0   0   0   0   0   0
  expirydate      0     0|   0   0    0|   0    0    0|   0   0   0   0   0   0
  attrib          9     2|   1   0    1|   7    0    0|  11  50  78   0  89  50
  currentsta      0     0|   0   0    0|   0    0    0|   0   0   0   0   0   0
  denial          0     0|   0   0    0|   0    0    0|   0   0   0   0   0   0
------------------------+-------------+--------------+------------------------
ALL OBJECTS     144    99|  62   0   10|  72   27    1|  43  63  50  27  64  14
------------------------+-------------+--------------+------------------------
                                                   P&R      2P&R       P&2R
F-MEASURES                                       51.03     57.41      45.93
```

Figure 20: The final score report for the user-defined takeover financial template.

than the takeover template. We are also currently establishing the real-time capabilities of the system, measuring the average time for processing news articles and the correctness of the investment decisions suggested by the system, comparing them with investment decisions suggested by academic and financial experts.

6 Conclusions

In this paper we have shown how natural language processing, information extraction and expert systems can be used in finance. Information extraction and expert systems can be combined to process an incoming stream of news from on-line news providers, companies and market data to produce investment suggestions. The results can be subsequently linked to the existing real-time quantitative information. This allows traders overcome their qualitative-data overload and better define their trading and hedging strategies.

References

[Callaghan, 1998] P. C. Callaghan, *An Evaluation of LOLITA and Related Natural Language Processing Systems*, PhD thesis, University of Durham, Department of Computer Science, Laboratory for Natural Language Engineering, 1998.

[Chinchor and Dungca, 1995] N. Chinchor and G. Dungca, "The Scoring Method for MUC-6", in *Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann, November 1995.

[Costantino *et al.*, 1996] M. Costantino, R. J. Collingham, and R. G. Morgan, "Qualitative Information in Finance: Natural Language Processing and Information Extraction", *NeuroVe$t Journal*, 4 No.6, November 1996.

[Costantino, 1997] M. Costantino, *Financial Information Extraction using pre-defined and user-definable templates in the LOLITA System*, PhD thesis, University of Durham, Department of Computer Science, Laboratory for Natural Language Engineering, November 1997.

[DAR, 1991] DARPA, *Proceedings of the Third Message Understanding Conference (MUC-3)*, Morgan Kaufmann Publishers, May 1991.

[DAR, 1992] DARPA, *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, Morgan Kaufmann Publishers, June 1992.

[DAR, 1993] DARPA, *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, Morgan Kaufmann Publishers, August 1993.

[DAR, 1995] DARPA, *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann Publishers, November 1995.

[Garigliano *et al.*, 1993] R. Garigliano, R. G. Morgan, and M. H. Smith, "The LOLITA System as a Contents Scanning Tool", in *Avignon '93*, 1993.

[Garigliano, 1995] R. Garigliano, "Editorial", *Natural Language Engineering*, 1, March 1995.

[Grishman, 1995a] R. Grishman, "The NYU System for MUC-6 or Where's the Syntax?", in *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann Publishers, November 1995.

[Grishman, 1995b] R. Grishman, "Tipster Phase II Architecture Design Document (Tinman Architecture)", 1995.

[Hull, 1997] J. Hull, *Options, Futures and other Derivative Securities*, Prentice Hall, 1997.

[Jacobs *et al.*, 1993] P. S. Jacobs, G. Krupka, L. Rau, M. L. Mauldin, T. Mitamura, T. Kitani, I. Sider, and L. Childs, "GE-CMU: description of the Shogun system used for MUC-5", in *Fifth Messages Understanding Conference (MUC-5)*, Morgan Kaufmann, August 1993.

[Krupka, 1995] G. R. Krupka, "SRA: Description of the SRA System as Used for MUC-6", in *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann Publishers, November 1995.

[Morgan *et al.*, 1995] R. Morgan, R. Garigliano, P. Callaghan, S. Poria, M. Smith, A. Urbanowicz, R. Collingham, M. Costantino, C. Cooper, and The LOLITA Group, "University of Durham: Description of the LOLITA System as used in MUC-6", in *Sixth Messages Understanding Conference (MUC-6)*, Morgan Kaufmann, November 1995.

[Rijsbergen, 1979] C. J. V. Rijsbergen, *Information Retrieval 2nd Edition*, Butterworths, 1979.

[Riloff and Lehnert, 1994] E. Riloff and W. Lehnert, "Information Extraction as a Basis for High-Precision Text Classification", *ACM Transactions on Information Systems*, 12 No.3:296–333, 1994.

[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.

[The PLUM System Group, 1993] The PLUM System Group, "BBN: Description of the PLUM System as Used for MUC-5", in *Fifth Messages Understanding Conference (MUC-5)*, Morgan Kaufmann, August 1993.

[Zahedi, 1993] F. Zahedi, *Intelligent Systems for Business: Expert Systems with Neural Networks*, Wadsworth Publishing Company Belmont, California, 1993.