NATURAL LANGUAGE PROCESSING IN FINANCE: AUTOMATIC EXTRACTION OF INFORMATION FROM FINANCIAL NEWS ARTICLES.

Costantino M., Collingham R.J., Morgan R.G.
Laboratory for Natural Language Engineering
Department of Computer Science
University of Durham, U.K.
marco.costantino@durham.ac.uk

6 October 1995

Summary

Subject Areas: Information extraction in finance, Natural Language Processing

Word Count: 4,000 approx.

NATURAL LANGUAGE PROCESSING IN FINANCE: AUTOMATIC EXTRACTION OF INFORMATION FROM FINANCIAL NEWS ARTICLES.

Summary

Subject Areas: Information extraction in finance, Natural Language Processing

Word Count: 4,000 approx.

1. INTRODUCTION

The players of the financial market, such as fund managers, brokers, bank analysts, etc. nowadays have access to an extremely large amount of data, both quantitative and qualitative, real-time or historical, and can use this information to support their decision-making process.

Quantitative data, such as historical price databases or real-time price information are largely processed by automatic computer programs often based on artificial intelligence techniques, for the production of quantitative analysis, such as historical price analysis, price forecasts or technical analysis of price behaviour.

On the contrary, few progress has been done for the processing of qualitative data, which consists mainly of financial news articles either from financial newspapers, e.g. *The Financial Times*, or from on-line news services, such as *Dow Jones* or *Bloomberg*. As a result, the players of the financial market are overloaded with qualitative information that is potentially extremely useful but, due to lack of time, often ignored.

A way of solving this problem is to produce a short-summary (template) of the financial news articles according to specific criteria, rather than presenting the whole article, eliminating the information that is considered not to be relevant. The availability of summaries of the original articles, rather than the full articles, leads to a marked reduction of the time needed by the operators in the market for analyzing the information and, thus, can help reducing the data overload. The process of summarising news articles is called *information extraction from texts* and belongs to the field of *Natural Language Processing*.

Most of the information extraction systems have been developed and tested within government agencies and scientific environments. In addition, few systems successfully tackled information extraction in finance.

This article describes information extraction applied to finance and focuses on the financial information extraction system under development at the University of Durham (UK), which 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. The system, unlike many others developed in the past, has been designed in close contract with financial experts for working in real situations.

2. INFORMATION EXTRACTION AND N.L.P.

The goal of information extraction is to extract specific kinds of information from a source document (Riloff and Lehnert, 1994). In other words, the input to the system is a document (e.g. a newspaper article), while the output consists of a summary of the contents of the document (see figure 1). For example, the source article could consist of the following text:

FLORHAM PARK, N.J. (AP) – Generic drug maker Schein Pharmaceutical Inc. will acquire Marsam Pharmaceuticals Inc. for 240 million dollars, the two companies said.

The agreement calls for Schein to acquire all stock outstanding of Marsam at about 21 dollars a share. In May, Marsam, which makes injectable drug products, disclosed it had received unsolicited takeover offers in the range of 19 dollars a share. On Friday, Marsam shares closed at 19.3125 dollars, down 6.25 cents, in Nasdaq Stock Market trading.

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

```
Template: Takeover
       Company target: Marsam Pharmaceuticals Inc.
       Company predator: Schein Pharmaceutical Inc.
       Type of takeover: FRIENDLY
       Value: 240 million dollars
```

The summary presented above is called a *template* which is a structure with a predefined number of slots. A *template*, thus, is a schematisation of the contents of the source document and it is widely used in information extraction. The "fields" of the *template* are called "*slots*" which are filled with the information extracted from the source article. The slots of the template (called "takeover") shown above are thus *Company target*, *Company predator*, *Type of takeover* and *Value*. The slots can be filled with information directly extracted from the text (e.g. *Marsam Pharmaceuticals Inc.*) or with text "inferred" by the system for data "hidden" in the meaning of the text, such as, in the above example, the *Type of takeover* slot, which contains "*FRIENDLY*", a term that cannot be found in the source document.

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

```
Template: Summary
        Companies:  Marsam Pharmaceuticals Inc.
                    Schein Pharmaceutical Inc.
        Values:     240 million dollars
                    21 dollars a share
                    19 dollars a share
                    19.3125 dollars
                    6.25 cents
```

and

```
Template: Takeover
       Company target: Marsam Pharmaceuticals Inc.
       Company predator: Schein Pharmaceutical Inc.
       Type of takeover: FRIENDLY
       Value: 240 million dollars
```

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

Two groups of articles can be processed by a financial information extraction system: documents from on-line services, such as *Dow Jones* or Bloomberg and from financial newspapers or magazines, such as *The Financial Times*, *The Wall Street Journal* or *The Economist*. Articles from on-line news services tend to be already rather summarised and, therefore, further processing is not usually needed. Articles from newspapers tend to be much longer and include further analysis of the news, such as comments and experts' interpretation about the facts. On large collection of such articles, the identification of the information needed by a human becomes a difficult and long task and, therefore, automatic processing using financial information extraction systems can be extremely useful. In figure 2 the same news reported from a newspaper (*The Financial Times*) and from an on-line news provider (*Dow Jones*) is shown.

Most of the information extraction systems have been developed and tested within government agencies or scientific environments. This has lead to very specialized systems able to work only in restricted situations and domains. Information extraction is a relatively new field compared to other artificial intelligence techniques. The first systems such as, for example, SAM (Schank and Riesbeck, 1981), PAM (Schank and Riesbeck, 1981) and FRUMP (Tait, 1982) were developed by the end of the 70s and were based on the script-frames approach. More recently, the MUC conferences (DAR, 1994) (*Message Understanding Conference* sponsored by ARPA - Advanced Research Projects Agency, U.S.A.) have shown great vitality in the field and evident improvement in the technology.

As far as the financial domain is concerned, very few financial information extraction systems have been realised in the past. One of the few systems is ATRANS (Lytinen and Gershman, 1986), a system for

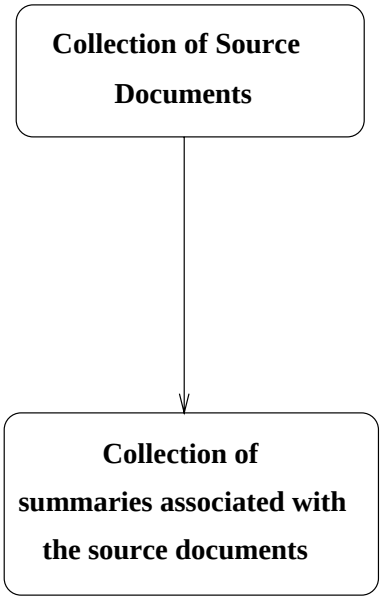

Figure 1: The behavior of a generic information extraction system.

extracting information from telex messages regarding money transfers between banks. However, the system has been successful mainly because of the extremely limited domain and the reduced information to be extracted. The systems that competed in the MUC-5 competition (DAR, 1994) were also able to perform the extraction of information from financial articles. However, they were only able to extract information regarding Joint Ventures and, thus, work in an extremely restricted subset of the financial domain.

3. FINANCIAL INFORMATION EXTRACTION AT DURHAM UNIVERSITY

Two main aspects have to be defined in the design of a financial information system.

- The kind of source articles, either from newspapers or from on-line news providers.
- The information to be extracted from the source documents, which will constitute the output of the system. Financial articles represent in fact an extremely broad domain, including different kinds of news: financial, economical, political etc. Therefore, the identification of an unique template able to summarise all the possible financial articles is extremely difficult, if not impossible. The best solution is thus to design more than one template.

The target source documents that have been chosen for the financial information extraction system under development at the University of Durham are articles from newspapers. However, the system is also able to successfully process articles from on-line services. This choice has been made taking into consideration the fact that articles from on-line news services are already rather summarised and further processing is often unnecessary (see figure 2).

As far as the output of the system is concerned, the system has been designed following the *"financial activities approach"* for the definition of the different templates to be extracted from the source documents.

3.1. The financial activities approach

The financial activities approach is based on the identification of specific *financial activities*. A *financial activity* is here defined as that potentially able to influence the decisions of the players of the market (brokers, investors, analysts etc.) regarding securities issued by companies. A finite number of relevant *financial*

A typical article from "The Financial Times"

Disney to buy Capital Cities
Tuesday August 1 1995
By Tony Jackson in New York

Walt Disney is to pay 19.1bn dollars for Capital Cities/ABC, owner of the ABC television network, creating the world's largest entertainment company. The deal is the second biggest takeover after that of RJR Nabisco by Kohlberg Kravis Roberts for about 25bn dollars in 1988.

In an agreed deal, Disney will pay one share plus 65 dollars for each Capital Cities/ABC share, valuing the latter at 124 dollars at yesterday's prices. The combined company will be called simply Walt Disney, and Mr Thomas Murphy, chairman of Capital Cities/ABC, will join the Disney board.

Mr Michael Eisner, Disney's chairman, claimed the synergies between the two companies were tremendous". He said: "Disney's intellectual property will appear on ABC's networks, and Disney's distribution systems will syndicate ABC's programmes."

Mr Warren Buffett, the billionaire portfolio investor whose company Berkshire Hathaway owns 12.9 per cent of Capital Cities/ABC, described the deal as "the marriage of the number one content company in the world with the number one distribution company".

Mr Buffett, who was instrumental in the merger of Capital Cities and ABC in 1986, bought shares in the company at the time for 17.25 dollars each, less than one seventh of the bid price. His holding is now worth 2.5bn dollars. He said yesterday: "I do not have blanket enthusiasm for all mergers. But this deal makes more sense than any deal I've ever seen, with the possible exception of Capital Cities and ABC."

The ABC television network rivals NBC for the top position among US networks, with about a 17 per cent market share. Capital Cities/ABC, which had revenues last year of 6.4bn dollars , also owns eight television stations, America's largest network of radio stations and an 80 per cent share of the leading cable television channel ESPN. It also publishes newspapers, books and magazines.

Walt Disney had been rumoured for some time to be interested in buying a TV network, but was thought to want CBS or NBC, either of which would cost 5bn dollars or less. It is much larger than Capital Cities/ABC, with a market value of 31bn dollars.

Mr Eisner, who worked as a programming executive with ABC throughout the 1970s, said discussions with ABC had gone on intermittently since he joined Disney in 1984. However, the deal had eventually been put together very quickly, he said. He had encountered Mr Buffett by chance while in Sun Valley, Idaho. When he raised the possibility of a merger, Mr Buffett took him to see Mr Murphy. "Eight days later, here we are," Mr Eisner said.

Mr Eisner said the takeover should not affect the joint venture struck last November between ABC and DreamWorks, the partnership between the Hollywood trio of Mr Steven Spielberg, Mr Jeffrey Katzenberg and Mr David Geffen. The venture aims to provide television programmes.

He said: "We want the best shows on this network that it's possible to get. If DreamWorks can come up [with them], I'm thrilled." Mr Katzenberg resigned as Disney's production chief last summer after a highly public row over promotion.

The deal allows Capital Cities/ ABC shareholders to take all their payment in either cash or stock, subject to availability. Mr Buffett said "the odds are extremely high that we will have a very large amount of Disney stock".

The same topic from an on-line news service (Dow-Vision)

Disney, Capital Cities -2-: Details On Merger DIS CCB
Source: Dow Jones News Service via DowVision Date: Jul 31, 1995 Time: 8:04 am

BURBANK, Calif. -DJ- Walt Disney Co. (DIS) and Capital Cities ABC Inc. (CCB) agreed to merge in a transaction valued at about 19 billion dollars at current share prices.

In a joint press release, the companies said Capital Cities/ABC shareholders will have the right to receive one Disney common share and 65 dollars in cash for each Cap Cities common share.

The transaction has been approved by both companies' boards, the companies said.

Figure 2: The difference between a newspaper article and an on-line news provider.

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

Figure 3: The three groups of financial activities

activities is identifiable in the financial market and can be grouped into three different categories (see figure 3).

- **Company related activities** which are those 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. This group includes, for example: merger, takeover, flotation, privatisation etc.
- **Company restructuring activities** which are activities related to changes in the productive structure of the company and include, for example, *new product*, *joint venture*, *staff changes* etc.
- **General macroeconomic activities** which include general macroeconomics news that can affect the prices of the securities quoted in the stock exchange and comprises, for example, interest rates movements, currency movements etc.

A specific template is associated to each of the *financial activities*. For example, the "takeover" financial activity is associated with the *takeover template* which is composed by 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*.

We believe that the *financial activities* identified represent the main information that brokers, investors, analysts etc. may want to extract from a source document and that represent and effective partitioning of the broad financial domain. In figure 3 the complete list of the *financial activities* identified is shown, while in figure 4 some of the specific templates associated to the first group of financial activities are listed.

3.2. Features of the system

As already pointed out earlier, the target documents for the system consist of of large collections of articles from financial newspapers, typically *The Financial Times* on CD-ROM.

The way in which the system works from the user's point of view is very similar to the way in which a generic information extraction system works (see section 2). However, the source articles are processed following a *two-stage* strategy.

First of all, the system processes the articles and identifies the list of relevant *financial activities* (see figure 3) in the source article. Basically, 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 merger(s)            found
2 market movement(s)   found
```

the user will draw the conclusion that the main *topic* contained in the article is about a *merger* and, probably, the two *market movements* are likely to be caused by the *merger*. At this stage, the user can decide that he is interested in more detail about the news or he is not interested in knowing more about the *merger*. 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 *merger* template, which will produce a filled template as, for example:

```
Template: merger
       Company 1: KNP
       Company 2: Buhrmann-Tetterode
       Company 3: VRG
       New name:  still unnamed
       Date of announce: 30/11/1992
       ...
```

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 collection of documents, the user might be interested only in particular topics and, just, decide to skip all the others.

An example of a full processing of an article is given in figure 5, while in figure 6 the schematization of the same process is given.

3.3. Implementation details

The system is written in the functional programming language Haskell (currently about 45,000 lines of code, corresponding to about 450,000 lines of code in a traditional programming language) and based on a large, WordNet-compatible semantic network (over 100,000 nodes) similar to a conceptual graph (Sowa, 1984). The system runs on a Sun SparcStation with 80 Mbytes of RAM. However, it can easily be adapted for usewithin other Unix environments.

3.4. Conclusions

In this paper we have shown how natural language processing and, more specifically, information extraction can successfully be used in finance and tried to give an overview of the financial information extraction system under development at the University of Durham.

The use of an information extraction system can be particularly interesting for any of the players of the financial market who have to deal with the increasing quantity of information avaiable nowadays. By processing large quantities of articles, the user can obtain the list of the financial activities (or topics) contained in the article and can access to its summary without having to read the whole article. THe financial information extraction sytem can thus act as a "filter" for the news, eliminating those that do not satisfy any of the extraction criteria. However, unlike information retrieval systems based on key-word searches, an information extraction system allows the extraction of specific information, such as takeovers,

mergers etc. Traditional key-word information retrieval engines are able to locate particular information in the source document. However, this is limited to the search of specific words (e.g. takeovers) and does not provide any templates able to summarise the main events in the article. The power of an information extraction system compared to a common key-words information retrival system is thus in the ability of *summarising* the contents of the articles according to predefined structures (templates), which traditional technologies are not able to produce.

In a paper of this size it is impossible to provide details of every aspect 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.

4. GLOSSARY OF TERMS

Natural Language Processing: a source text is processed and its meaning stored into a knowledge base. Once the meaning has been stored, it can be retrieved or manipulated in different ways.

Inference: inferece "is taken as the process by which a listener or reader draws out those parts of the meaning of a piece of natural language which are not manifest in the text itself " (Tait, 1982).

Information extraction: the process for which a source document is processed and different kinds of information are extracted, according to extraction criteria. Information extraction is different from Information Retrieval, which only *locates* the relevant information but it is not able to perform any "reasoning" on the data or to extract complex information from the articles.

Template: a structure with a predefined number of slots, one for each type of information to be extracted from a text (Riloff and Lehnert, 1994). A template can be seen as a *topic* (e.g. a *takeover*) associated to a number of slots which represent the information that has to be extracted from the source document for that particular topic (e.g. *company predator*, *company target*, etc.).

Slot: the fields associated to a particular template (topic). The slots are filled according to the information contained in the source documents and, organised in templates, represent the typical output of an information extraction system.

REFERENCES

N. Chichor. 1993. Muc-5 evaluation metrics. In *Fifth Messages Understanding Conference (MUC-5)*, August.

DARPA. 1994. *Proceedings of the Fifth Message Understanding Conference*. Morgan Kaufmann Publishers, August.

S.L. Lytinen and A. Gershman. 1986. Atrans: Automatic processing of money transfer messages. In Morgan Kaufmann, editor, *9th International Joint Concerence on Artificial Intelligence*, pages 821–825.

E. Riloff and W. Lehnert. 1994. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, 12 No.3:296–333.

R.C. Schank and C.K. Riesbeck. 1981. *Inside Computer Understanding*. Lawrence Erlbaum Associated.

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

J.I. Tait. 1982. *Automatic Summarising of English Texts, PhD Thesis*. University of Cambridge.

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 4: Specific templates associated with the company related financial activities

Source: The Wall Street Journal (full text) via DowVision Date: Jul 31, 1995 Time: 11:53 pm
FLORHAM PARK, N.J. (AP) – Generic drug maker Schein Pharmaceutical Inc. will acquire Marsam Pharmaceuticals Inc. for 240 million dollars, the two companies said.
The agreement calls for Schein to acquire all stock outstanding of Marsam at about 21 dollars a share. In May, Marsam, which makes injectable drug products, disclosed it had received unsolicited takeover offers in the range of 19 dollars a share. On Friday, Marsam shares closed at 19.3125 dollars, down 6.25 cents, in Nasdaq Stock Market trading.
Marvin Samson, chief executive officer of Marsam, and Agvar Chemicals Inc., which together hold about 28 per cent of Marsam shares outstanding, have agreed to grant Schein an option to purchase their shares at 21 dollars each.
Mr. Samson would continue in his position at Marsam.
Marsam said its board approved the offer, which is expected to commence Friday and expire at 12:01 a.m. EDT Sept. 2.
Marsam, based in Cherry Hill, N.J., develops, makes and markets multisource injectable drug products for hospital, institutional and home markets. It employs about 200 people.
Closely held Schein makes more than 400 pharmaceutical products in nearly every therapeutic category. It employs more than 1,600 people.

The financial activities found in the article are firstly displayed obtaining, in this case:

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

1 takeover(s) found
```

In case the user requests the display of the specific template, this is computed and displayed by the system obtaining, in this case:

```
Template: Takeover
        Company target: Marsam Pharmaceuticals Inc.
        Company predator: Schein Pharmaceutical Inc.
        Type of takeover: FRIENDLY
        Value: 240 million dollars
        Bank adviser predator:
        Bank adviser target:
        Expiry date: 12:01 a.m. EDT Sept. 2
        Attribution: Marvin Samson, chief executive officer of Marsam
        Current stake of the predator:
        Denial:
```

Figure 5: The processing of a typical Wall Street Journal article.

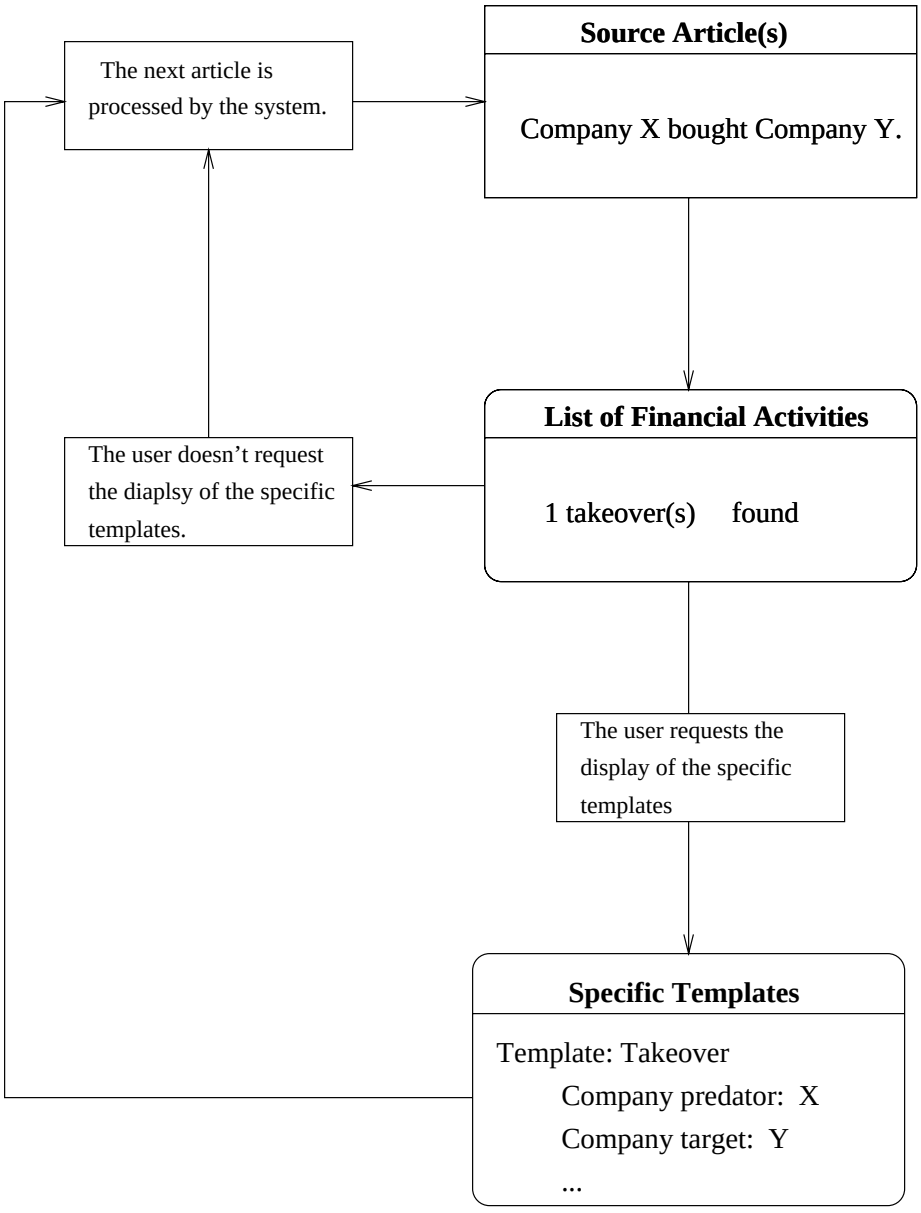

Figure 6: Schematization of the processing of an article.