Natural Language Technology for Information Integration in Business Intelligence

Diana Maynard[1] and Horacio Saggion[1] and Milena Yankova[21] and Kalina Bontcheva[1] and Wim Peters[1]

[1] Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP
United Kingdom
{diana,saggion,wim,milena,kalina}@dcs.shef.ac.uk
[2] Onotext Lab, Sirma Group Corp.
135 Tazrigradsko Chassee, Fl.5
1784 Sofia, Bulgaria
milena@sirma.bg

Abstract. Business intelligence requires the collecting and merging of information from many different sources, both structured and unstructured, in order to analyse for example financial risk, operational risk factors, follow trends and perform credit risk management. While traditional data mining tools make use of numerical data and cannot easily be applied to knowledge extracted from free text, traditional information extraction is either not adapted for the financial domain, or does not address the issue of information integration: the merging of information from different kinds of sources. We describe here the development of a system for content mining using *domain ontologies*, which enables the extraction of relevant information to be fed into models for analysis of *financial and operational risk* and other business intelligence applications such as *company intelligence*, by means of the XBRL standard. The results so far are of extremely high quality, due to the implementation of primarily high-precision rules.
Keywords: Information Extraction; Ontology; Business Intelligence; Natural Language Processing; Information Fusion

1 Introduction

Business intelligence requires the collecting and merging of information from many different sources, both structured and unstructured, in order to analyse for example financial risk, operational risk factors, follow trends and perform credit risk management. The information is published either by the companies themselves on their web sites (e.g. balance sheets, company reports), by financial newspapers, specialised directories (e.g. Yahoo! Company and Fund Index[3]),

[3] http://biz.yahoo.com/i/

governamental bodies, etc. The analytical techniques frequently applied in business intelligence, however, have been largely developed for dealing with numerical data so, unsurprisingly, the industry has started to struggle with making use of this distributed and unstructured information. For example, Ellingsworth and Sullivan [8] found that traditional analytic techniques to understand trends in insurance claims could not help because the information was not fully described by structured data.

One solution to this problem is to apply text processing and Natural Language Processing (NLP) techniques to unstructured sources in order to transform them into structured representations suitable for such analysis. Information Extraction (IE) is a key NLP technology which automatically extracts specific types of information from text to create records in a database or populate knowledge bases, for example. One typical scenario for information extraction in the business domain is the case of insurance companies tracking information about ship sinkings around the globe [21]. Without an IE system, company analysts would have to read hundreds of textual reports and manually dig out that information. Another typical IE scenario is the extraction of information about joint ventures or other types of commercial company agreements from unstructured documents [2, 12]. This kind of information can help identify not only information about who is doing business with whom, but also market trends, such as which world regions or markets are being targeted by companies.

One additional problem with business information is that even in cases where the information is structured (e.g. balance sheets), it is not necessarily represented in a way machines can understand - and this is particularly true with legacy systems and documentation. One response to this problem has been the development of the emerging standard XBRL (eXtensible Business Reporting Language)[4]. XBRL is a universal XML-based specification for business information, providing both public and private companies with an effective way to prepare and distribute various business reports using the Internet in a cost effective and universal manner [18]. Structured data such as that from company balance sheets and tabular reports can be mapped into XBRL using automatic processes [10]. But when the information is unstructured, then NLP and text mining techniques are of paramount importance.

In this paper, we report on our work on information extraction for business intelligence in the context of the EU Musing project[5]. We are working with domain ontologies which represent our understanding of the domain of application and which capture the experts' knowledge. Ontologies contain concepts arranged in class/sub-class hierarchies (e.g. a bank is a financial institution), relations between concepts (e.g. a bank has a manager), and properties (e.g. a company has only one CEO). We have developed different applications in the business domain targeting real business scenarios defined by real users in the areas of financial risk management, internationalisation, and IT operational risk - ontologies are being developed for each of the scenarios. We focus here on ap-

[4] http://www.xbrl.org
[5] IST-2004-027097 http://www.musing.eu

plications for extracting information from company profiles and country/region for the developing internationalisation applications but we will also briefly describe techniques used in other scenarios. One key aspect of our work is the development of *ontology-based information extraction systems*[6] which are being developed using robust and adaptable tools from the GATE architecture [5]. A second key aspect of our work is a framework for *merging* information across different sources which also uses a domain ontology. The ontology acts as bridge between the text and a knowledge base, which in turn feeds reasoning systems or provides information to end users.

The following section describes the approach to text analysis we have adopted, while Section 3 describes the information extraction system in more detail. Section 4 describes the evaluation, and in Section 5 we compare our approach with previous work in the fields. Finally we discuss some related work and future directions in the last section.

2 Information Extraction

Information extraction (IE) is a technology for automatically extracting specific types of information from text [11]. The information to be extracted or the concepts to be targeted by the IE system are predefined in knowledge resources such as a domain ontology or templates. These concepts are elucidated by domain experts or can be automatically learnt (at least partially) from domain-specific texts. In the business domain, an information extraction template for joint ventures might be made up of the following key variables or concepts which need to be instantiated from text: partners (e.g. companies), nationalities, type of contractual form (e.g. alliance or joint venture), name of the contractual form, business sector, date of constitution of the alliance, etc.

Once target concepts, relations, and attributes have been defined for each domain, the information extraction system can be developed so that new documents can be semantically annotated by identifying instances of those concepts, attributes and relations. For example, company names can be identified in a number of ways such as gazetteer lookup, regular expression matching, or a combination of techniques. Relations between entities in text can be identified from syntactic relations found in parse trees or predicate-argument structures obtained from semantic analysis. The instances identified in text can then be mapped to the domain ontology, stored in a database, or used as semantic indexes for further processing (e.g. searching, reasoning). Some instances in the text may be already known to the system, while others may never have been encountered before: this is one of the key features of the IE technology.

We focus on an information extraction task which targets different domain ontologies. Ontology-based information extraction is a task which consists of finding in a text instances of concepts and relations between them as expressed in an ontology. This process is domain-specific and is carried out with a domain

[6] Musing ontologies extend the Proton Ontology http://proton.semanticweb.org.

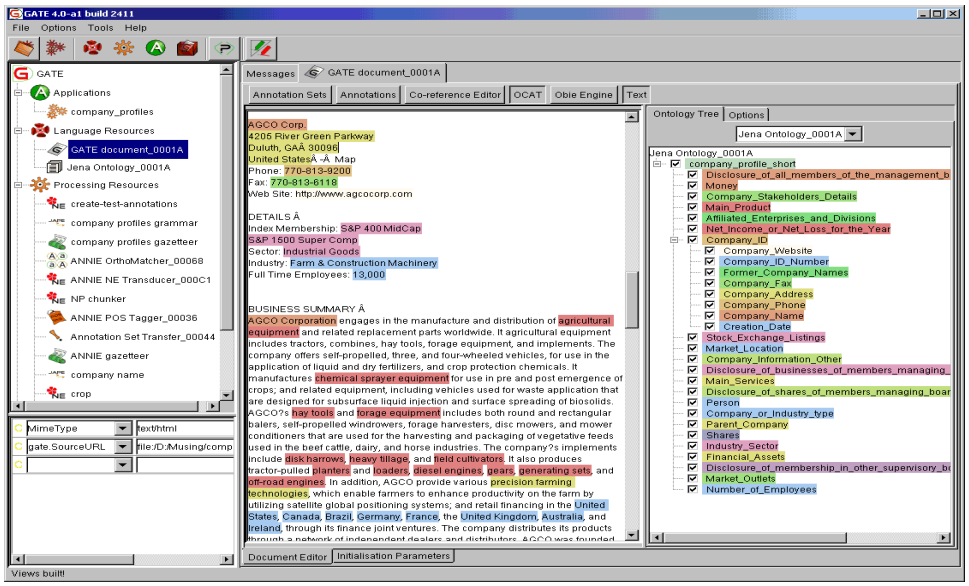

Fig. 1. GATE Development Environment and Text Automatically Annotated with Ontological Classes

ontology over texts which belong to that domain. Figure 1 shows our development environment and a text which has been automatically annotated with respect to an ontology for company information.

2.1 Data Sources and Ontology-based Annotation Tool

When developing an information extraction system, it is essential to have textual documents where the key domain concepts have been identified, so that a language engineer can create accurate information extraction rules. In addition to data provided by different partners in the project[7], a number of on-line data sources for business intelligence (e.g. Yahoo! Finance, World Bank, CIA Fact Book) have been targeted in order to boost system accuracy. We rely on the Ontology-based Corpus Annotation Tool (OCAT), a GATE plugin which uses one or more ontologies for annotation of concepts/classes. The required ontology can be selected from a pull-down list of available ontologies which are loaded into the system. GATE currently provides support for ontologies in both OWL and RDF. The current version of the tool supports only annotation with information about the ontology class, however future work will include the annotation of relations from the ontology. Ontology-based annotations in the text can be viewed by selecting the desired classes in the ontology tree.

[7] The European Business Register (EBR), Belgium and Verband der Vereine Creditreform e.V. (VVC), Germany are members of the Musing consortium.

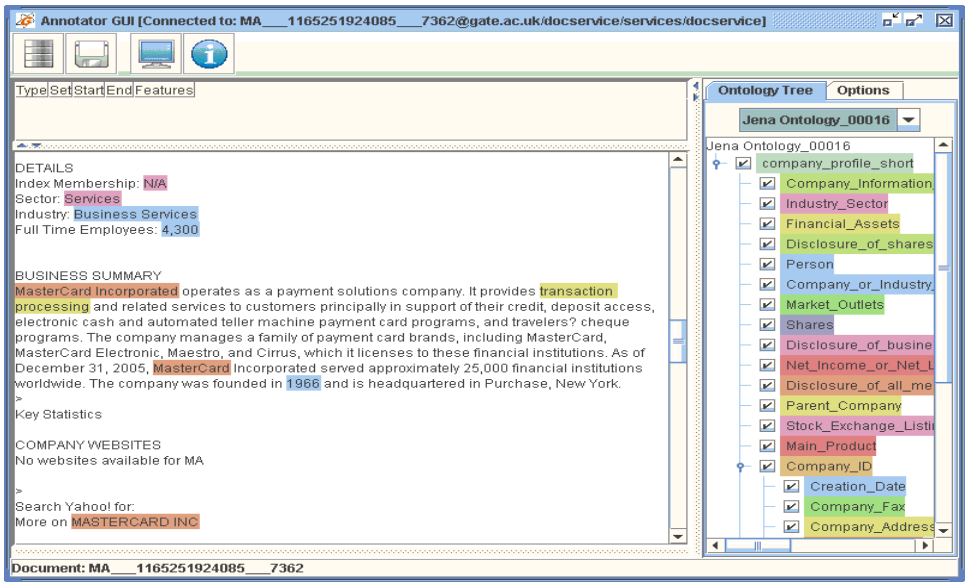

Fig. 2. Document Service for Ontology-based Annotation

We have developed a Web service which allows the user to *annotate* texts with ontological information over the Web (Figure 2). First, a set of documents (corpus) is annotated with key information using an initial information extraction system. This information may only be partially correct, so the user uses a corpus annotation tool to edit the annotations proposed by the system. The human annotations are then fed back to the system and developer to create a more accurate information extraction system, either by re-defining new rules or by machine learning. Once the system achieves the desired performance, the development cycle comes to an end and the system can be deployed by the final user.

2.2 Natural Language Processing Tools

We have developed our information extraction system using GATE. While GATE comes with a default information extraction system called ANNIE [16], it is only partially relevant to the business domain. The ANNIE system identifies generic concepts such as persons, locations, organization, dates, etc., so we had to develop new rules or adapt rules for our applications. The tools available in GATE to perform text analysis consist of: a document structure analyser which parses different input files into GATE documents; a tokeniser which identifies different types of words; a sentence splitter which segments the document into sentences; a part-of-speech tagger which associates POS tags to words and symbols; a morphological analyser which produces a root and affix for each word in the doc-

ument; a named entity recognition sub-system composed of a gazetteer lookup component and a rule-based pattern matching engine; and a coreference resolution algorithm. Other components which are sometimes necessary, depending on the text and task, are parsers which associate syntactic and semantic structures with sentences. For the work reported here, we have mainly adapted the named entity recognition components and developed a conceptual mapping to map concepts identified by our system into the ontologies of the application domains. The named entity system in GATE is a rule-based system developed using a pattern-matching engine called JAPE [6] which is ontologically aware; making the mapping of entities into ontological classes possible during entity recognition.

The ease of adaptation of the core ANNIE system to new applications depends on many factors: language, annotation types to be recognised, document type, level of structure in the text, and level of accuracy required (tradeoff between precision and recall). ANNIE does not deal with ontologies, however, so an ontology-based IE application requires a lot more initial adaptation than just the recognition of new entity types, for example. For more information about the adaptation process in general, we refer the reader to [13, 15]; for another example of adaptation to ontologies, see [17].

2.3 Merging Information across Different Sources

One of the fundamental problems one has to address with the proliferation of information is the identification and merging of ontological instances extracted from multiple sources. In the Semantic Web community, this problem is known as ontology population. An example of this is presented in Figure 3, where three texts refer to the same company Alcoa, using different expressions "ALCOA", "Alcoa Inc." and "Alcoa". It is important to identify the three instances as the same company because of the complementary information they bring (note that the interlinking or coreference between entities in the same text is solved by our coreference resolution mechanism). While one text provides information about the company profile (e.g. address, management), a second text provides information about where the company has business (e.g. 8 plants in the UK), the third one provides relevant financial information (e.g. share prices). The merging of these complementary sources provide a clearer picture about the company for BI purposes.

In the work presented here, merging and interlinking between pieces of information are carried out in an identity resolution framework which provides a generic solution to the merging problem. The framework is based on an ontology of the domain and a knowledge base containing known instances. For each new ontological instance discovered by the extraction process, the resolution process operates in four stages. First, a set of possible candidates is retrieved from the knowledge base (e.g. instances with the same class information). Second, evidence is collected from each of the candidate instances (e.g. attributes and values stored in the knowledge base). Third, a decision is made based on the similarity between the new instance and the instances retrieved from the database.

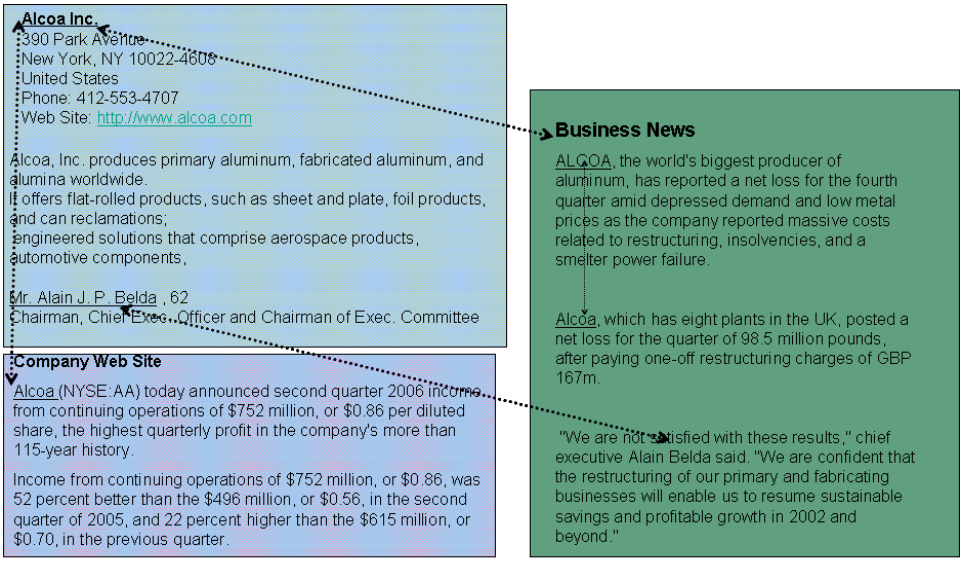

Fig. 3. Related information from multiple sources

The decision is based on a set of rules defined by the domain expert which are used to compute a similarity score between the new instance and each candidate (these rules may for example check name aliases; or similarity between values for similar attributes). Finally, the new instances and their attributes are asserted in the knowledge base. The framework uses the KIM [19] semantic repository implemented in OWLIM/Sesame.

3 MUSING Information Extraction System

In our framework, the documents to be analysed are first loaded into GATE and undergo document format analysis, which enables the documents to be processed by the application. Document structure analysis is then carried out in order to identify the layout. This consists of pre-processing modules such as tokenisation and sentence splitting. For example, a special splitting module is run in order to identify each row in a table in documents such as balance sheets. Then the information extraction system is run and the information is identified as annotations on the document. Finally, this information is mapped into XBRL and the appropriate ontology.

Because the system needs to take into account information from different kinds of sources, different applications are needed which may use slightly different sets of components. Not only do gazetteer lists and grammar rules differ for different kinds of concepts, but also pre-processing may differ to take into account

different structural information. For example, some web pages may contain a lot of extraneous information that should not be processed. Web pages in particular often contain information which is useful to the human user looking for other sources of information, such as information about other countries when looking at information for a specific country. These are often in the form of tables or drop-down boxes. Such information is very useful to a visual user but can be very misleading to a system which cannot distinguish the relative importance of information in different kinds of formats. We use some of GATE's processing resources to help us detect such information and ignore what is unimportant.

In the following sections we describe 3 applications for identifying and extracting information relevant for business intelligence from 3 kinds of domain-specific unstructured text: company profiles, country profiles, and balance sheets.

3.1 Information Extraction from Company Profiles

Structured information from company profiles needs to be extracted in order to be able to feed this data into statistical models of financial risk assessment or investment, e.g. assessment of the creditworthiness of a company. In addition, such information is necessary for providing services to companies who are looking for commercial partners working in the same sector in a different country, e.g. all software companies in Russia. The information from country profiles is therefore also needed as input. For example, if the system extracts the fact that Russia's investment Fitch rating is BBB+, increased from BBB, then the risk assessment model can take this into account and correspondingly revise risk downwards. One prototype we are developing is an International Enterprise Intelligence application whose objective is to provide customers with up-to-date and correct information about companies, mined from many different sources such as web pages, financial news, and structured data sources. A set of company profiles has been downloaded from Yahoo! and the most relevant concepts to extract have been identified in the ontology. Each concept is extracted along with the relevant information, for example the concept "number of employees" is associated with a feature and a value, such as "Number=2000".

Table 1 presents some examples of key concepts which, according to our users, need to be extracted from text for each company. The company domain specific ontology which extends the Proton model contains at the present time 24 concepts and 38 properties.

3.2 Country and Region Information Extraction

Our country/region profiles application enables us to extract general information about countries/regions from unstructured text. A set of country profiles has been downloaded from the CIA World Factbook[8], and a list of concepts to be extracted has been identified from the domain ontology. The following concepts have been extracted so far: country name; population; surface area; official

[8] https://www.cia.gov/cia/publications/factbook/index.html

Address Data	Company Data	Financial Data
Name of Company	Branch	Turnover
Telephone	Main Activities	Number of Employees
Postcode	Import/Export Activities	Turnover per Employee
Country	Legal Form	Shareholders
E-Mail	Managerial head	Related persons
...	...	...

Table 1. Relevant Concepts for Company Information

language; currency; exchange rate; foreign debt; unemployment rate; GDP; and foreign investments. Each concept is extracted with features and values depicting the information associated with it. In the case where we wish to extract information for multiple years (for example if we want to extract the exchange rate for the last 3 years), we extract separate features and values for each year, such as:

```
<Exchange>
<rate="afghanis per US dollar", amount2003=49, amount2004=48,
amount2005=541>
</Exchange>
```

There still remain some further concepts wich require a deeper level of analysis such as ratings, sustainability and vulnerability, which can be quite vague and hard to define in free text. The extracted concepts will be used in a Musing specific internationalisation application which will help companies or businesses searching for appropriate regions for internationalisation of their businesses.

3.3 Extracting Information from Financial Statements

While balance sheets and other financial statements contain both structured (tables) and unstructured information (explanatory notes), these statements are only currently available in documents in pdf, tiff, or similar binary formats which are difficult to process automatically. When a bank needs financial information about a company, a balance sheet would be requested and then analysed by a human analyst, who would typically re-enter all the information of the balance sheet in the bank system to produce a structured file before credit rating can be performed. This is a very tedious and error-prone practice. As an additional disadvantage, it is currently impossible for a bank to automatically obtain key information (relevant for our users) from a balance sheet such as *what were the net assets of the company in the 31 December 2001?* or *what is the purchase plan of the company?*: the analyst has to dig into the files in order to find the appropriate answers to these key questions. Some answers are found in free text descriptions in balance sheets, but this information is currently inaccessible to models of risk or the company's creditworthiness. The latter is required, for

example, by the Basel II accord which lays down guidelines for matters such as how much capital a bank or financial institution needs to keep in reserve to cover all its current lending. There are various methods of calculating the bank's expected loss/unexpected loss with differing degrees of complexity.

Our information extraction application over balance sheets aims to identify all specific financial information such as details of fixed assets, profits, goodwill (reflecting good relationship of a business enterprise with its customers) etc. from the files. The application identifies the structure of the balance sheets using patterns developed in JAPE, and maps each line of the balance sheet into the appropriate XBRL concept - as specified by an FRM expert. Another important aspect of this work is the identification of explanatory notes in the balance sheets as well as any concepts related to the financial risk management described by the domain experts in the ontology which curretly contains 45 concepts.

We have developed an application in GATE that extracts such information from company balance sheets in PDF format – some balance sheets are also available in other formats such as TIFF files or HTML pages. One of the problems of PDF files is that it is very difficult to extract information that is in tabular form. One solution is to first convert the PDF directly into a more easily processable format such as HTML, XML or XBRL. Alternatively, we can process the application directly as a PDF file in GATE, making use of GATE's language processing capabilities and the JAPE pattern-matching language [6] to identify things like column headings and separate rows. It is important to note that because the original documents are in PDF, the spatial/graphical structure of the document is not fully preserved and this will have consequences for extraction. For example, the numbers in each line are associated with particular dates which are given once at the top of the balance sheet. Some numbers appear to be totals but this is not explicitly mentioned, so analysis has to be performed on such figures based on positional information, and the meaning made clear.

Once the PDF file is loaded into GATE, the Balance Sheet application identifies each row in the table, using a specially modified version of the ANNIE sentence splitter which identifies each row as a separate sentence. Usually in balance sheets each column is headed by a date (usually a year), i.e. information in each column represents the information for that date. A JAPE grammar first identifies a line of date information in the table, e.g. 2001, 2002 etc., and then stores this information as annotations on the document as a whole (e.g. that the first column represents 2001, the second column represents 2002, etc.). Then various grammars look for the row entries in the table, for example identifying labels such as "Fixed Asset". For each concept, features and values are added to the annotation representing the amount and year. One annotation is thus produced for each row in the table, with the following information:

- year (e.g. year=2005)
- amount value (e.g. value=73,000)
- positive or negative (e.g. type=negative)
- string of the asset (e.g. string=Total Current Liabilities)

Negative values are sometimes displayed by a number in round brackets. A special grammar rule identifies these as negative. Our current work is looking at extensions to work with other document formats. Next stages in the process are to link the concepts denoting the entries in the table with concepts in the ontology, and to transform the final annotations (currently in XML) into XBRL, performed in collaboration with our financial partners.

4 Evaluation

Evaluation is an essential component of any information extraction application. Our quantitative evaluation compares annotations produced by the automatic system with annotations produced by human experts (known as key or gold standard annotations). We make use of traditional metrics used in information extraction [4]: precision, recall, and F-measure. Precision measures the number of correctly identified items as a percentage of the number of items identified. It measures how many of the items that the system identified were actually correct, regardless of whether it also failed to retrieve correct items. The higher the precision, the better the system is at ensuring that what is identified is correct. Recall measures the number of correctly identified items as a percentage of the total number of correct items measuring how many of the items that should have been identified actually were identified. The higher the recall rate, the better the system is at not missing correct items. The F-measure [20] is often used in conjunction with Precision and Recall, as a weighted average of the two – usually an application requires a balance between Precision and Recall. For the application on extraction of company information from different textual sources, we have obtained very encouraging results. An expert manually annotated the texts (using the tool described in Section 2.1) and we compared the results of the system annotations against this gold standard set. The results for each type as well as the totals are shown in Table 2.

For comparison purposes, our generic IE system ANNIE which identifies classical types of information such as People, Location, Organization, etc. has levels of precision of 93.5%, recall of 92.3%, and F-measure of 92.9% on general news texts.

The other two applications also show very encouraging results, although they require more work to complete the extraction of all relevant concepts.

5 Related Work

In a pure information extraction context in the business domain, JV-FASTUS [2] developed for the Message Understanding Conferences performed shallow and robust text analysis using a set of finite state transducers. For joint ventures the system achieved recall levels of 34%, precision levels of 62%, and combined F-score of around 45%. As with other systems in the MUC context, FASTUS targeted a template and not a domain ontology. Our work is different from traditional approaches to extraction not only because of the complexity of the domain

Concept	Precision	Recall	F-Measure
Company Address	100.00	66.70	80.00
Company Fax	100.00	100.00	100.00
Company Name	88.90	80.00	84.20
Company Phone	100.90	100.00	100.00
Company Website	50.00	70.00	58.30
Company or Industry Type	60.00	75.00	66.70
Creation Date	100.00	100.00	100.00
Industry Sector	60.00	100.00	75.00
Market Outlets	85.00	94.40	89.50
Market Location	69.60	94.10	80.00
Number 0f Employees	100.00	100.00	100.00
Stock Exchange Listings	100.00	100.00	100.00
Total	85.60	93.60	84.00

Table 2. Evaluation of company profiles application

but also because we address the problem of merging information extracted from different sources.

h-TechSight [17] is a system which also uses GATE (amongst other tools) to detect changes and trends in business information and to monitor markets. It uses semantically-enhanced information extraction and information retrieval tools to identify important concepts with respect to an ontology, and to track changes over time. This enables companies to keep an eye on competitors' products in the news and in company reports etc., and enables job seekers and job providers to monitor changes in the employment market (for example, required skills, salaries payable, locations of jobs, trends in company hiring policies, etc.). This system differs from MUSING in that the information acquired is only related to a quite shallow and simple ontology with a few fairly fixed concepts. The information discovery module realised in GATE is part of a much larger knowledge portal combining a number of different tools. It acts as a very good starting point or baseline from which to continue.

Information extraction is also used in the MBOI tool [9] for discovering business opportunities on the internet. The main aim is to help users to decide about which company tenders require further investigation. This enables the user to perform precise querying over named entities recognised by the system. Similarly the LIXTO tool is used for web data extraction for business intelligence [3], for example to acquire sales price information from online sales sites. However, this requires a semi-structured data source which is not always available or sufficient for the kind of financial information we are concerned with.

Ahmad et al. [1] have developed a system for analysing sentiment in business and financial news streams, using term recognition and collocation extraction techniques. The idea behind this is that positive and negative sentiments expressed in news can often make or break people, companies and even governments, creating effects such as economic bubbles through the power of financial

journalism. While this work does not directly address the problem we have in mind, the sentiment research supports the underlying theory about the importance of extracting such information from free text.

None of the systems above deals specifically with extracting information useful for financial business intelligence, and although there are systems which do so [7], they do not deal adequately with gathering information from unstructured text and the problem of merging information from different data sources or using an ontology to assist these processes.

6 Conclusions and Further Work

We have described the design and implementation of a system for knowledge extraction in business intelligence. The aim is to extract relevant information from a number of sources including the Web in order to build up a financial picture of a particular company for applications in financial risk management and internationalisation. Our system targets an ontology of the application domains containing the most relevant domain concepts and relations. The system produces annotations which will be used to populate a knowledge base or semantic repository with the assistance of a multi-source merging mechanism. The identification and extraction of such information has been largely implemented, and this paper describes the design approach to these tasks. Work will continue on refining this work and on the merging process which will follow. So far the actual extraction is of extremely high quality and there are few errors. Our future work on extraction will concentrate on different text types such as business reports and company web sites. As a continuation of our work on evaluation, we shall be looking at an evaluation metric specifically adapted to ontology-based information extraction, such as [14], since this will give us a more informed and practical result (giving credit for answers which are closely linked in the ontology to the correct answer).

7 Acknowledgements

This work is partially supported by the EU-funded MUSING project (IST-2004-027097).

References

1. Khurshid Ahmad, Lee Gillam, and David Cheng. Sentiments on a grid: Analysis of streaming news and views. In *5th Language Resources and Evaluation Conference*, 2006.
2. D.E. Appelt, J.R. Hobbs, J. Bear, D. Israel, M. Kameyama, and M. Tyson. Description of the JV-FASTUS system as used for MUC-5. In *Proceedings of the Fourth Message Understanding Conference MUC-5*, pages 221–235. Morgan Kaufmann, California, 1993.

3. R. Baumgartner, O. Frlich, G. Gottlob, P. Harz, M. Herzog, and P. Lehmann. Web data extraction for business intelligence: the lixto approach. In *Proc. of BTW 2005*, 2005.
4. Nancy Chinchor. Muc-4 evaluation metrics. In *Proceedings of the Fourth Message Understanding Conference*, pages 22–29, 1992.
5. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.
6. H. Cunningham, D. Maynard, and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, November 2000.
7. T. Declerck and H. Krieger. Translating XBRL into Description Logic: an approach using Protege, Sesame and OWL. In *Proceedings of Business Information Systems (BIS)*, Klagenfurt, Germany, 2006.
8. Marty Ellingsworth and Dan Sullivan. Text mining improves business intelligence and predictive modeling in insurance. *DM Review Magazine*, 2003.
9. J.-Y. Nie F. Paradis and A. Tajarobi. Discovery of business opportunities on the internet with information extraction. In *Workshop on Multi-Agent Information Retrieval and Recommender Systems (IJCAI)*, pages 47–54, Edinburgh, Scotland, 2005.
10. Franseco Fornasari, Alessandro Tommasi, Cesare Zavattari, Roberto Gagliardi, Thierry Declerck, and Michele Nannipieri. Xbrl web-based business intelligence services. In Paul Cunningham and Miriam Cunningham, editors, *Innovation and the Knowledge Economy: Issues, Applications, Case Studies. Proceedings of eChallenge 2005*. IOS Press, 2005.
11. R. Gaizauskas and Y. Wilks. Information Extraction: Beyond Document Retrieval. *Journal of Documentation*, 54(1):70–105, 1998.
12. P.S. Jacobs and L.F. Rau. Scisor: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, 1990.
13. D. Maynard, K. Bontcheva, and H. Cunningham. Towards a semantic extraction of Named Entities. In *Recent Advances in Natural Language Processing*, Bulgaria, 2003.
14. D. Maynard, W. Peters, and Y. Li. Metrics for evaluation of ontology-based information extraction. In *WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON)*, Edinburgh, Scotland, 2006.
15. D. Maynard, V. Tablan, K. Bontcheva, and H. Cunningham. Rapid customisation of an Information Extraction system for surprise languages. *Special issue of ACM Transactions on Asian Language Information Processing: Rapid Development of Language Capabilities: The Surprise Languages*, 2003.
16. D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. Named Entity Recognition from Diverse Text Types. In *Recent Advances in Natural Language Processing 2001 Conference*, pages 257–274, Tzigov Chark, Bulgaria, 2001. `http://gate.ac.uk/sale/ranlp2001/maynard-etal.pdf`.
17. D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Apects of the Semantic Web")*, Heraklion, Crete, 2005.
18. J. Montes. Consumer entertainment software - industry trends. In Brian Stanford-Smith and Enrica Chozza, editors, *E-Work and E-Commerce*, pages –7. IOS Press, Amsterdam, 2001.

19. B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM – Semantic Annotation Platform. *Natural Language Engineering*, 2004.
20. C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
21. Yorick Wilks and Roberta Catizone. Can We Make Information Extraction More Adaptive? In *M. Pazienza (ed.) Proceedings of the SCIE99 Workshop*, pages 1–16, Rome, Italy, 1999.