Ontology-based Information Extraction for Business Intelligence

Horacio Saggion and Adam Funk and Diana Maynard and Kalina Bontcheva

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP
United Kingdom
{saggion,adam,diana,kalina}@dcs.shef.ac.uk

Abstract. Business Intelligence (BI) requires the acquisition and aggregation of key pieces of knowledge from multiple sources in order to provide valuable information to customers or feed statistical BI models and tools. The massive amount of information available to business analysts makes information extraction and other natural language processing tools key enablers for the acquisition and use of that semantic information. We describe the application of ontology-based extraction and merging in the context of a practical e-business application for the EU MUSING Project where the goal is to gather international company intelligence and country/region information. The results of our experiments so far are very promising and we are now in the process of building a complete end-to-end solution.

Keywords: Ontology-based Information Extraction; Business Intelligence; Cross-source Entity Coreference.

1 Introduction

Business intelligence (BI) can be defined as the process of finding, gathering, aggregating, and analysing information for decision making (See [9] for example). Semantic technologies of the type advocated by Semantic Web [6] are being applied for BI in the context of the EU MUSING[1] Project. MUSING is developing a new generation of BI tools and modules based on semantic-based knowledge and natural language processing (NLP) technology to mitigate the efforts involved in gathering, merging, and analysing information.

Information Extraction (IE) is a key NLP technology for automatically extracting specific types of information from text or other sources to create records in a database or populate knowledge bases, for example. Without an IE system, business analysts carrying out BI activities would have to read hundreds of textual reports, web sites, and tabular data to manually dig out the necessary information to feed BI models and tools.

[1] MUlti-industry, Semantic-based next generation business INtelliGence

The road infrastructure in Argentina is **excellent**, even in remote areas. This is in sharp contrast to Brazil and, to a lesser extent, Chile. The transportation and communication infrastructures seem more than adequate to allow quick price discovery and easy communication between the processors and farmers for sample test results and other marketing matters.
The forest area in India extended to about **75 million hectares**, which in terms of geographical area is approximately 22 per cent of the total land. Out of this, 9.5 million hectares is fallow and 7 million hectares is under shrub formation. Thus, an actual forest area is less than 17 per cent. *The total area under forest in Tamil Nadu is* **21,072 sq.km.** of which 17,264 sq.km. is reserved forest and 3,808 sq.kms is reserved land. This constitutes 16 per cent of the total geographical area of the State.
Political stability in India is **threatened** *by the Kashmir dispute and other internal issues.*
The population in India as of March 2001 stood at **1,027,015,247 persons**. With this, India became only the second country in the world after China to cross the one billionmark. India's population rose by 21.34% between 1991 - 2001. The sex ratio (i.e., number of females per thousand males) of population was 933, rising from 927 as at the 1991 Census. Total literacy rate in India was returned as 65.38%.

Fig. 1. Multiple Textual Sources of Information for Internationalisation Applications

Here, we concentrate on the application of *Ontology-based Information Extraction* (OBIE) in the context of Business Intelligence. OBIE is the process of identifying in text or other sources relevant concepts, properties, and relations expressed in an ontology. We are working with *domain ontologies* which represent the domain of application and which capture the experts' knowledge. Ontologies contain concepts arranged in class/sub-class hierarchies (e.g. a joint venture is a type of company agreement), relations between concepts (e.g., a joint venture has a starting date), and properties (e.g., a joint venture has only one starting date). An ontology we are working with is being developed for a e-business application in the internationalisation domain[2] where the objective is to model information about companies, countries, and regions. The ontologies are developed with the help of domain experts. These experts have identified that in a domain such as that of joint ventures, relevant concepts are: companies, nationalities, type of contractual form, date of constitution of the alliance, etc.

We have developed robust and adaptable technology for the extraction of relevant semantic information (expressed in the ontology) to be used in business intelligence processes in the following areas: financial risk management, internationalisation, and IT operational risk management. Specific applications in these areas are: credit risk assessment, international company intelligence, country or region selection, risk identification and mapping.

All these applications require the *extraction and merging* of information from a number of trusted but diverse data sources (e.g. database, financial reports, news reports, company web sites) which represents a challenge for any information extraction system.

We focus here on IE for the development of internationalisation applications. We rely on robust and adaptable tools from the GATE architecture [10].

[2] Internationalisation is the process that allows an enterprise to evolve its business from a local to an international dimension. In the MUSING context this involves for example the acquisition of information about international partnerships, contracts, investments, etc.

Extraction of information for internationalisation applications consists on the identification of all mentions of concepts, instances, and properties in text or other sources (e.g. multimedia material such as tables and images). Some examples are presented in Figures 1. Information in these sources is required for internationalisation applications dealing with, for example, companies desiring to take their business abroad and interested in knowing the best places to invest. These applications usually require the gathering of information on *economic indicators* such as the population of a particular country or region (e.g. the market potential), the status of the transport infrastructure, the literacy rate, the political situation, criminality indices or whether a region is prone to particular natural disasters. Both quantitative (e.g. numeric) and qualitative (e.g. categorical) information is necessary. Our IE components also extract information from images which is carried out by components which operate on the output of an OCR system.

Once the information has been gathered from different sources, the ontology has to be populated with all mentions found in text, in order to do so, the system has to decide if two mentions in different sources refer to the same entity in the real world. The populated ontology (or knowledge base(KB)) is queried by the different applications in the MUSING Project, the semantic information from the KB is used as valuable information for customers or applied to statistical models of decision making. Because some applications require perfect output from the IE system, a user verifies the extraction results.

In this paper we describe the development of an Ontology-based information extraction for business intelligence in the context of internationalisation applications. The paper is structured as follows: In the next Section, we describe the MUSING project with respect to the information extraction task. Section 3 describes our approach to cross-source entity identification for Ontology population. Adaptation of our NLP technology for internationalisation applications is described in Section 4. Section 5 reports on related work on business intelligence and ontology-based extraction. Finally, in Section 6, we present our conclusions.

2 MUSING Information Extraction Technology

In Figure 2 we present the MUSING IE architecture. A number of data sources for information extraction have been identified and documents and multimedia material collected and stored in the MUSING document repository. In addition to data provided by different partners in the project, a number of on-line data sources for business intelligence (e.g., Yahoo! Finance, World Bank, CIA Fact Book) are being targeted.

Documents are then processed by an Ontology-based annotation tool which automatically detects information specified in a domain ontology. The ontology has been developed through interaction with MUSING domain experts and implemented in OWL [11] expending the PROTON Upper Ontology[3].

[3] http://proton.semanticweb.org/

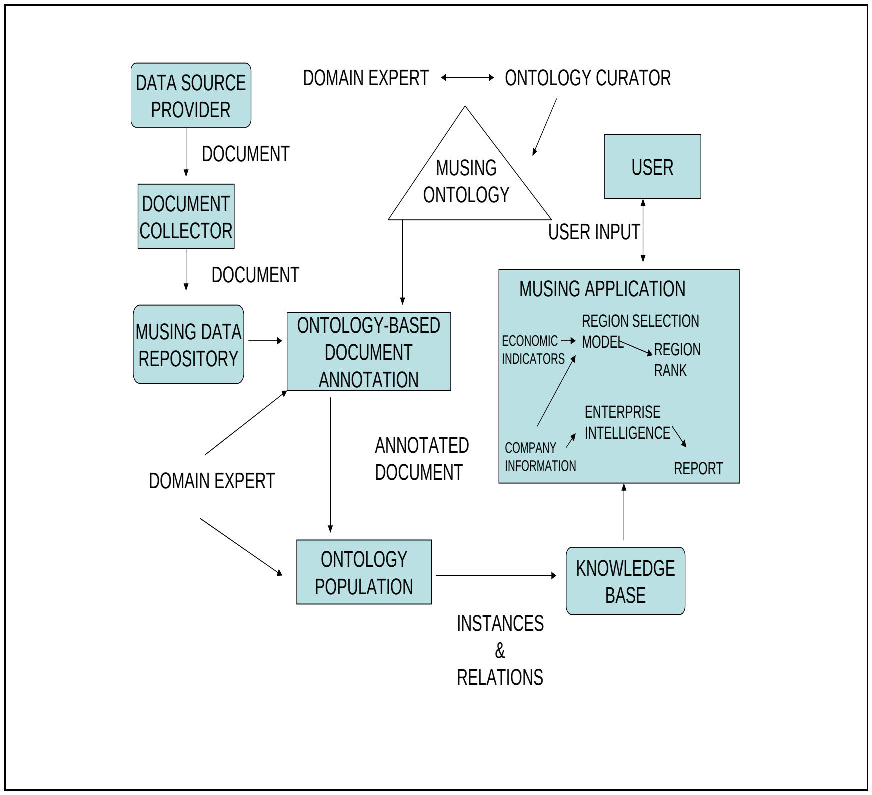


Fig. 2. Ontology-based Extraction Architecture in MUSING

A collaborative annotation tool developed by the University of Sheffield as part of the EU Neon Project[4] has been adapted to the MUSING ontologies in order to allow users not only to annotate documents from scratch but also to correct the results of the automatic semantic annotation process.

The current version of the tool supports both annotation of ontology class as well as relations. The tool is being used by experts to identify qualitative information in text which may include complicated statements about inflation, growth, reliability, etc. expressed in phrases, sentences or even full paragraphs.

Once documents have been automatically annotated, their annotations are analysed by an ontology population mechanism in charge of creating instances and relations for knowledge base population. Tuples in the knowledge base are used in different Musing applications. One such application is providing up-to-date information about companies (e.g. for identifying possible business partners) and another application is providing ranked lists of countries/regions for companies interested in investing into new country/regions.

[4] http://www.neon-project.org

2.1 GATE Processing Tools

We have developed our information extraction system using the General Architecture for Text Engineering (GATE). GATE is a framework for the development and deployment of language processing technology in large scale [10]. It provides three types of resources: Language Resources (LRs) which collectively refer to data; Processing Resources (PRs) which are used to refer to algorithms; and Visualisation Resources (VRs) which represent visualisation and editing components. GATE can be used to process documents in different formats including plain text, HTML, XML, RTF, and SGML. When a document is loaded or opened in GATE, a document structure analyser is called upon which is in charge of creating a GATE document, a LR which will contain the text of the original document and one or more sets of annotations, one of which will contain the document mark-ups (for example HTML). Annotations are generally updated by PRs during text analysis - but they can also be created during annotation editing in the GATE GUI. Each annotation belongs to an annotation set and has a type, a pair of offsets (the span of text one wants to annotate), and a set of features and values that are used to encode the information.

A key element in the annotations is the encoding of ontological information - our applications create *Mention* annotations which make reference to the target ontology as well as the ontological concept a string of text refers to.

While GATE comes with a default information extraction system called ANNIE [18], it is only partially relevant to the business domain. The ANNIE system identifies generic concepts such as person names, locations, organisation, dates, etc. Therefore we had to develop new rules or adapt rules for our applications.

For the work reported here, we have carried out adaptation of the GATE named entity recognition components because most target entities are not covered by ANNIE. We have also developed a conceptual mapping module to map concepts identified by our system into the ontologies of the application domains. Future versions of our system will apply machine learning techniques incorporated into the GATE framework.

3 Cluster-based Cross-document Entity Coreference

In a scenario such as the MUSING one where information is extracted from many sources, one has to deal with the problem of identifying whether two business entities in two different sources refer to the same individual in the real world. A problem known as ontology population [1] in the Semantic Web community. Solving this problem is extremely important in order to create an accurate picture of individuals as well as organisations. In fact, in business, the reputation of a particular company may depend on the reputation of its board of directors, and therefore bad news about a company director may influence a company's performance.

An example of this is presented in Figure 3, where a number of sources contain references to the same person name "Dale Merritt". Knowing if this particular person has criminal charges is important and may well influence a decision such as participating in a commercial agreement involving such person.

Dale Merritt, CEO of DotVentures, LLC has just signed on for a five week pilot season, with guests who specialize in domain investing and world wide web real estate, on the live talk radio show, “Domain Investing”...
DotVentures, an Internet domain name investing company, today announced a partnership between their company and Internet marketing software company, SearchMarketing. President of DotVentures, **Mr. Merritt** says, ”DotVentures and SearchMarketing have integrated ...
Dale Merritt, who was charged by the Federal Trade Commission along with his firm, Showcase Distributing, Inc., of Phoenix, Arizona, as part of a nationwide crackdown on business opportunity fraud, ..
Haystack Ranch Events... June 30 - July 1, 2007... Registration Contact: E-mail Brenda or **Dale Merritt** or phone 303.681.2098 (Brenda or **Dale Merritt**)

Fig. 3. Cross-source Coreference Problem Instance

We have applied text mining techniques to the cross-source coreference problem focusing on the problem of person name coreference. The context we are carrying out this piece of research is one where information about a particular entity is required and a set of documents are retrieved from data sources based on the entity description (e.g. person name). The task at hand consist on identifying what sets of documents refer to the same entity in the real world. Once this has been done, the entity can be assigned a unique identifier in the knowledge base.

As past and recent research [4, 20], we have addressed the problem as a document clustering problem. We have implemented an agglomerative clustering algorithm. The input to the algorithm is a set of vectors representations (e.g. terms and weights) which are extracted from the annotated documents. We have experimented with two types of data representation which are derived from the annotation of documents using our NLP tools. One representation is based on a bag-of-words approach while the other uses specific types of semantic information extracted from the annotated documents [21]. Terms are either words from the documents or named entities in the targeted ontology (PROTON Upper).

When clustering starts, there are as many clusters as input documents; as the algorithm proceeds clusters are merged until a certain termination condition is reached. The algorithm computes the similarity between vector representations in order to decide whether or not to merge two clusters. The similarity metric we use is the cosine of the angle between two vectors. This metric gives value one for identical vectors and zero for vectors which are orthogonal (non related). Various options have been implemented in order to measure how close two clusters are, one metric we have used is the following: the similarity between two clusters (sim_C) is equivalent to the “document” similarity (sim_D) between the two more similar documents in the two clusters; the following formula is used:

$$\mathrm{sim}_C(\mathrm{C}_1,\mathrm{C}_2) =$$

$$\max_{d_i \in C_1; d_j \in C_2} \mathrm{sim}_D(\mathrm{d}_i,\mathrm{d}_j)$$

Where C_k are clusters, d_l are document representations (e.g., vectors), and sim_D is the cosine metric.

If this similarity is greater than a threshold – experimentally obtained – the two clusters are merged together. At each iteration the most similar pair of clusters is merged. If this similarity is less than a certain threshold the algorithm stops. In order to identify the optimal threshold we have experimented with training data. The threshold was selected in order to obtain optimal performance in the training data.

In order to test the success of the implemented techniques, we have carried out a number of experiments with test data from the SemEval 2007 evaluation on People Web Search task [2]. In this evaluation, systems receive a name and a set of documents containing the name, the system has to decide how many different individuals are there, and what documents correspond to what individuals (e.g. clustering). Evaluation of the task is carried out using standard clustering evaluation measures of "purity" and "inverse purity" [13], and the harmonic mean of purity and inverse purity: F-score. Our algorithm is competitive when compared with the best system in that evaluation. One particular configuration of our system which uses specific types of semantic information obtained a (micro-averaged) F-score of 78% (same performance as the best SemEval 2007 system [7]) and a macro-average F-score one point more than the best system.

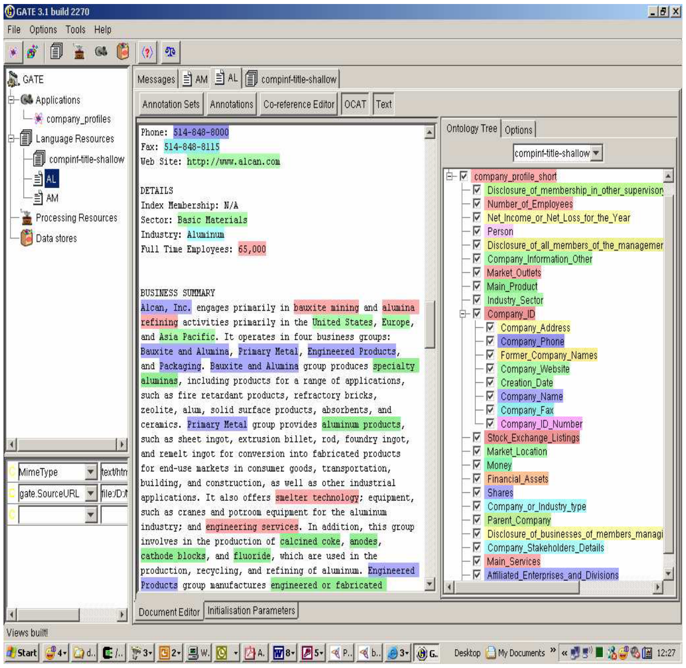
4 Information Extraction for Internationalisation Applications

In order to support intelligence gathering for BI we use a number of sources of information for developing internationalisation application: one is a set of company profiles that we have mined from Yahoo! Finance, another source is a set of around 100 company web sites, yet another source is company reports and newspaper articles provided by our partners in the project.

In order to collect available information for companies we have automatically gathered the main page of the company web site, and crawled pages containing contact information and company activities. We followed page links which contain certain keywords such as “contact us”, “about us”, etc. in the **href** attribute of an html anchor link or the text surrounding the anchor. For Yahoo! Finance documents we have developed a script which crawls the information for each company based on their company symbol.

4.1 Company Intelligence

One prototype we are developing is an International Enterprise Intelligence application the objective of which is to provide customers with up-to-date and correct information about companies. The information is mined from many different sources such as web pages, financial news, and structured data sources which after annotation and merged is stored in knowledge base tuples. Among other concepts to be targeted by the application are the company name, its main activities, its number of employees, its board of directors, etc.



```
Phone: 514-848-8000
Fax: 514-848-8115
Web Site: http://www.alcan.com

DETAILS
Index Membership: N/A
Sector: Basic Materials
Industry: Aluminum
Full Time Employees: 65,000


BUSINESS SUMMARY
Alcan, Inc. engages primarily in bauxite mining and alumina
refining activities primarily in the United States, Europe,
and Asia Pacific. It operates in four business groups:
Bauxite and Alumina, Primary Metal, Engineered Products,
and Packaging. Bauxite and Alumina group produces specialty
aluminas, including products for a range of applications,
such as fire retardant products, refractory bricks,
zeolite, alum, solid surface products, absorbents, and
ceramics. Primary Metal group provides aluminum products,
such as sheet ingot, extrusion billet, rod, foundry ingot,
and remelt ingot for conversion into fabricated products
for end-use markets in consumer goods, transportation,
building, and construction, as well as other industrial
applications. It also offers smelter technology; equipment,
such as cranes and potroom equipment for the aluminum
industry; and engineering services. In addition, this group
involves in the production of calcined coke, anodes,
cathode blocks, and fluoride, which are used in the
production, recycling, and refining of aluminum. Engineered
Products group manufactures engineered or fabricated
```

Fig. 4. OBIE for International Company Intelligence

The application consists of standard GATE components together with new linguistic and named entity recognition processors map concepts to ontological classes. Figure 4 shows the automatic annotation of concepts in text. The OCAT tool is used to display the link between ontology and annotated text. The result of the annotation is further analysed by the ontology population module responsible for knowledge base population. KB tuples returned in answer to a user request are used to display information in a Web-based user interface (e.g. Yellow Page style). We have carried out evaluation of this application using traditional IE metrics [8, 22]: precision, recall, and F-measure. An expert manually annotated 5 documents and we compared the results of the system annotations against this gold standard set. The overall performance of the system was an F-score of 84% (for details of this evaluation see [17]) which is acceptable for system deployment.

4.2 Country/Region Intelligence

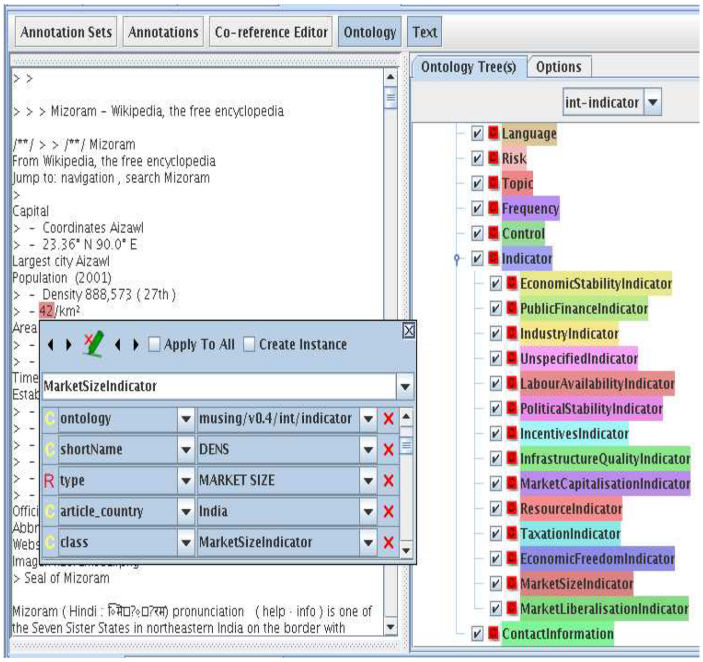
The second application we present here aims at extracting relevant information about countries and region in the globe. Sources of information used in this application are country profiles and statistics from various sources (e.g. World Bank, Monetary Fund). For system development and testing we have collected a corpus of documents, we have used crawling scripts that target specific web sites (e.g., BBC, Wikipedia, CIA World Fact Book) which contain the information required by the application. Concepts we target are: country name; official language; currency; exchange rate; foreign debt; unemployment rate; GDP; and foreign investments. In addition to these common types of information, more specific are economic indicators and indices such as the mortality index, region area, population, education, etc. In Figure 5 we show a screen-shot of a document annotated with semantic information for this particular application. The Figure also shows the ontology being used which contains among others some relevant economic indicators.

For each identified concept, features and values are stored in the annotation. The resulting information is used to feed a statistical model of country/region selection [15] which using both information about a country and a company (Section 4.1) decides which regions in the globe are more suitable to undertake business in.

The application's output can be seen in Figure 6. It shows a ranking of Indian regions most promising for investment. In addition to the ranking, the application indicates which model variables have mainly contributed to the obtained ranking.

The application was developed using standard as well as adapted GATE processing resources. A number of domain specific gazetteer lists were developed. One set of gazetteer lists is in charge of helping identify text types targeted by the application, another set identifies names of places associated with countries targeted by the application and helps associate capital cities with regions or countries for example. Named entity recognizers target the specific concepts or indicators required by the application and map them to the ontology. The rules for some particular types of text are highly accurate.

In Table 1 we present the performance of the extraction system in terms of precision, recall, and f-score. This is an evaluation of the extraction of 6 key economic indicators from 34 semi-structured sources about Indian regions in Wikipedia. The economic indicators targeted by the application are density of population (DENS), region surface (SURF), employment rate (EMP), literacy rate (LTR), literacy rate male (LRM), literacy rate female (LRF). The overall performance of the application is an F-score of 81%. Note that because country/region information changes periodically, this application has to be run whenever new documents are available, thus ensuring that the value of the indicators are up-to-date.



```
> >

> > > Mizoram - Wikipedia, the free encyclopedia

/**/ > > /**/ Mizoram
From Wikipedia, the free encyclopedia
Jump to: navigation , search Mizoram
>
Capital
> - Coordinates Aizawl
> - 23.36° N 90.0° E
Largest city Aizawl
Population (2001)
> - Density 888,573 ( 27th )
> - 42/km²
Area
> -
> -
Time
Estab
> -
> -
> -
> -
> -
> -
Offici
Abbr
Webs
Imag
> Seal of Mizoram

Mizoram ( Hindi : ि◌मो□?◌□?रम) pronunciation ( help · info ) is one of
the Seven Sister States in northeastern India on the border with
```

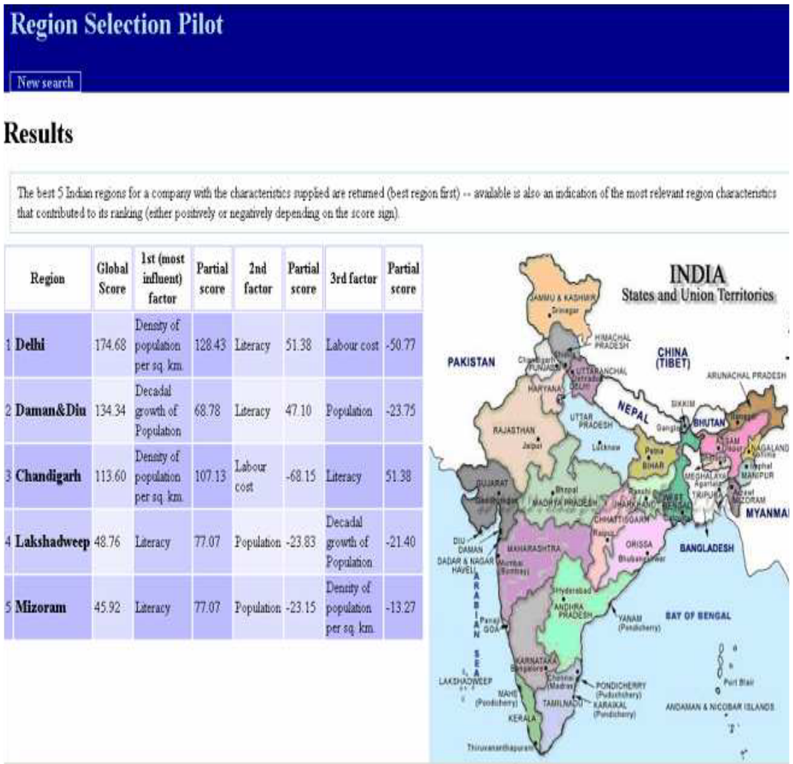
	MarketSizeIndicator	
C	ontology	musing/v0.4/int/indicator
C	shortName	DENS
R	type	MARKET SIZE
C	article_country	India
C	class	MarketSizeIndicator

Fig. 5. Output of the Automatic Annotation Process

5 Related Work

Information extraction in the business domain is not new, [23] developed a machine learning approach to identify patterns for the identification of *corporate management changes* in text, which is relevant in the context of BI. Such system should be able to identify positions in an organisation which are changing hands as well as who are the actors involved in the changes. While succession management is not an specific focus of MUSING we are dealing with a variety of applications in BI. In addition we are dealing with the whole development cycle from the creation of patterns to the extraction and mapping of the information to the ontology.

h-TechSight [19] is a system which also uses GATE 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 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.



Region Selection Pilot

Results

The best 5 Indian regions for a company with the characteristics supplied are returned (best region first) -- available is also an indication of the most relevant region characteristics that contributed to its ranking (either positively or negatively depending on the score sign).

Region	Global Score	1st (most influent) factor	Partial score	2nd factor	Partial score	3rd factor	Partial score
1 **Delhi**	174.68	Density of population per sq. km.	128.43	Literacy	51.38	Labour cost	-50.77
2 **Daman&Diu**	134.34	Decadal growth of Population	68.78	Literacy	47.10	Population	-23.75
3 **Chandigarh**	113.60	Density of population per sq. km.	107.13	Labour cost	-68.15	Literacy	51.38
4 **Lakshadweep**	48.76	Literacy	77.07	Population	-23.83	Decadal growth of Population	-21.40
5 **Mizoram**	45.92	Literacy	77.07	Population	-23.15	Density of population per sq. km.	-13.27

Fig. 6. Output of the Region Selection Application

Information extraction is also used in the MBOI tool [12] 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 [5], 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.

EBiZPort [16] is a portal for information gathering in BI. The tool incorporates a meta-search process to leverage different information sources also addressing the merging problem (but not at the entity level as in our case). The tool incorporates summarization, classification, and visualisation techniques. In this approach, it is still up to the user to find the relevant information in the

Concept	Precision	Recall	F-score
DENS	92%	68%	80%
SURF	100%	94%	97%
EMP	50%	100%	75%
LRT	88%	41%	64%
LRM	100%	29%	64%
LRF	100%	38%	69%
Total	94%	67%	81%

Table 1. Quantitative Evaluation of Region/Country Intelligence Application

mass of documents returned by the system. We go beyond this by providing extraction of relevant concepts to feed BI models.

Similar to our approach to instance merging is [3] where the problem of instance unification for author names is addresses. They mine information from the Web for authors including full name, personal page, and co-citation information to compute the similarity between two person names. Similarity is based on a formula which combines numeric features with appropriate weights experimentally obtained.

6 Conclusions and Further Work

Business Intelligence requires business analysts to gather, merge, and analyse considerable amounts of information in multiple formats and from heterogeneous sources. Information extraction technology is a key enabler to identify in text key pieces of information to be used in BI tools. Clustering techniques are powerful tools to merge information across different sources.

We have described how available and robust information extraction technology is being adapted to create an Ontology-based information extraction system in the context of the MUSING project. The system produces ontological annotations which are transformed into tuples for ontology population. The system already extracts and merges information from various sources and for specific applications in financial risk management and internationalisation. Applications are being created which use the valuable information in the knowledge based to perform reasoning or provide valuable information to customers. Performance measured through quantitative evaluation in both extraction and cross-source coreference look promising.

While MUSING is an ongoing project, we have already developed robust technology for deploying BI applications. The evaluation presented here is mainly quantitative, in the future the applications will be evaluated in terms of usability and user satisfaction.

Our current work is exploring the extraction of information from business graphics and tabular data which is based on the use of flexible gazetteer lookup procedures available in GATE which are being applied to OCR analysis of images. The output of the OCR analysis is being corrected by the exploitation of collateral information found around the graphics. This methodology has already prove useful, with improvements of around 3% over extraction from OCR alone.

Our future work will further improve our extraction tools incorporating Machine Learning capabilities into the extraction system [14], this will ensure that scalability is properly addressed in the extraction process. Our work on merging or ontology population will be extended to cover other semantic categories including locations, organisations, and specific business events (e.g., joint ventures).

Acknowledgements

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

References

1. H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. O'Hara, and N. Shadbolt. Managing Reference: Ensuring Referential Integrity of Ontologies for the Semantic Web. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 317–334, Siguenza, Spain, 2002.
2. J. Artiles, J. Gonzalo, and S. Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for Web People Search Task. In *Proceedings of Semeval 2007, Association for Computational Linguistics*, 2007.
3. Niraj Aswani, Kalina Bontcheva, and Hamish Cunningham. Mining information for instance unification. In *5th International Semantic Web Conference (ISWC2006)*, Athens, Georgia, 2006.
4. A. Bagga and B. Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85, 1998.
5. 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.
6. K. Bontcheva and H. Cunningham. The semantic web: A new opportunity and challenge for human language technology. In H. Cunningham, Y. Ding, and A. Kiryakov, editors, *Proceedings of Workshop on Human Language Technology for the Semantic Web and Web Services,* 2^{nd} *International Semantic Web Conference*, Sanibel Island, Florida, Oct 2003. `http://www.gate.ac.uk/sale/iswc03/iswc03.pdf`.
7. Y. Chen and J.H. Martin. Cu-comsem: Exploring rich features for unsupervised web personal named disambiguation. In *Proceedings of SemEval 2007, Assocciation for Computational Linguistics*, pages 125–128, 2007.
8. Nancy Chinchor. Muc-4 evaluation metrics. In *Proceedings of the Fourth Message Understanding Conference*, pages 22–29, 1992.

9. W Chung, H. Chen, and Nunamaker Jr. J.F. Business Intelligence Explorer: A Knowledge Map Framework for Discovering Business Intelligence on the Web. In *Hawaii International Conference on System Sciences*, Los Alamitos, CA, USA, 2003. IEEE Computer Society.
10. 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.
11. M. Dean, G. Schreiber, S. Bechhofer, Frank van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL web ontology language reference. W3C recommendation, W3C, Feb 2004. http://www.w3.org/TR/owl-ref/.
12. 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.
13. A. Hotho, S. Staab, and G. Stumme. WordNet improves text document clustering. In *Proc. of the SIGIR 2003 Semantic Web Workshop*, 2003.
14. Y. Li, K. Bontcheva, and H. Cunningham. An SVM Based Learning Algorithm for Information Extraction. Machine Learning Workshop, Sheffield, 2004. `http://gate.ac.uk/sale/ml-ws04/mlw2004.pdf`.
15. A. Majocchi and R. Strange. The FDI Location Decision: does Liberalisation Matter? *Transactional Corporation Review*, 2007. To Appear.
16. A. Marshall, D. McDonald, H. Chen, and W. Chung. EBizPort: Collecting and Analysing Business Intelligence Iformation. *Journal of the American Society for Information Science and Technology*, 55(10):873–891, 2004.
17. D. Maynard, H. Saggion, M. Yankova, K. Bontcheva, and W. Peters. natural language technology for information integration in business intelligence. In W. Abramowicz, editor, *10th International Conference on Business Information Systems*, Poland, 25-27 April 2007. `http://gate.ac.uk/sale/bis07/musing-bis07-final.pdf`.
18. 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.
19. 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.
20. X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Personal name resolution crossover documents by a semantics-based approach. *IEICE Trans. Inf. & Syst.*, Feb 2006, 2006.
21. H. Saggion. Shef: Semantic tagging and summarization techniques applied to cross-document coreference. In *Proceedings of SemEval 2007, Assocciation for Computational Linguistics*, pages 292–295, 2007.
22. C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
23. Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Unsupervised Discovery of Scenario-level Patterns for Information Extraction. In *Proceedings of ANLP-NAACL'00*, Seattle, WA, 2000.