Automatic Annotation and Ontology Population for Business Intelligence

Horacio Saggion, Adam Funk, Kalina Bontcheva, Diana Maynard
Department of Computer Science
University of Sheffield
211 Portobello Street - Sheffield, England, UK
saggion@dcs.shef.ac.uk
Tel: +44-114-222-1947
Fax: +44-114-222-1810

1. Introduction

Business intelligence (BI) can be defined as the process of finding, gathering, aggregating, and analysing information for decision making (See (Chung et al., 2003) for example). Current business intelligence systems (see (Marshall et al., 2004) for example) are portals which allow users access to full documents using technologies such as information retrieval, summarization, and classification. The MUSING project is developing a set of tools for Business Intelligence based on semantic technology. Is in this context that we are developing extraction components which automatically annotate documents from multiple multilingual sources with respect to a domain ontology in the business domain and create instances of concepts and relations for ontology population. Once populated, instances and theirs property values can be retrieved in order to perform a number of BI analytical tasks such as assessing credit worthiness; measuring likelihood of success in international ventures; assessing company reputation; etc. The ontology which is been manually built with the assistance of domain experts and relying on a number of international standards (e.g. Extensible Business Reporting Language - XBRL `www.xbrl.org`) and classification systems (e.g., NACE codes) gives both domain knowledge and a standard mechanism for representation and storage of extracted information. For this project, we have adapted for different BI applications existent extraction technology to provide functionalities for ontology population with emphasis on instance identification and relation extraction.

2. Mining Business Information and Ontology-based Annotation

One of the target applications we have developed deals with the extraction and consolidation of company information from multiple sources. Relevant concepts expressed in an ontology for company information are key executives; number of employees; web site; industry sector; address information; products and services; etc. This information is extracted as values of particular properties of company instances (e.g. *company* `hasWebSite` *web site*).

In order to develop and evaluate the extraction system, we have mined a number of available Web sources to create a corpus which is used for development and evaluation of the company information extraction system.

One technique we have used to create such a corpus consisted on targeting sites which list company names and company tickers; once the ticker of a particular company is known, additional pages can be searched using specific requests to obtain company profiles from well known and trusted sites (Yahoo! Finance, MarketWatch, etc.).

Another technique used is to target the company web site and from there, pages such as “contact us” and “about us” using a number of strategies to be described in the paper.

Another target application we have developed deals with the extraction of relevant information about countries and regions. In this case we have mined information from trusted Web sites such as Wikipedia and the CIA World Fact Book. The target information to extract is a number of social, political, geographical and economic indicators which are used by statistical models which can predict the success of investment in particular countries or regions.

An ontology-based annotation tool has been implemented which allows domain experts in our project to identify concepts and relations of interest in text. The tool which is been used to annotate texts in English, Italian, and German allows the user to activate an ontology, select a text span and annotate it with an ontology class or instance (existent or new). Additionally, instances can be related together by the use of property values. The annotated information is being used for evaluation of a rule based system. In the future, the annotated corpus will be used in machine learning experiments.

2.1. Multimedia Information

Valuable information about companies, organisations, and countries can also be found in sources other than text (e.g. tv/radio programmes, presentations, images), here we concentrate on the problem of extracting information from business images. We have collected a corpus of images from a variety of Web sources. The technique applied used an initial set of Web pages containing business graphics (e.g., pie, bar, line chars, tables). From these texts queries were created using appropriate keywords to find additional images. Many images have been collected following this methodology, however we have kept for the purpose of this project only pages which contain images and text.

While in-depth analysis of the image is beyond the scope of our project, interesting information (names of companies, presence of particular information types) can be extracted from the texts appearing in the images for the purpose semantic indexing or metadata production (e.g., a pie chart which contains information about the organisation OPEC - the Organisation for Petroleum Exporting Countries; image contains percents; data is from year 1999). These images are linked to concepts in our domain ontology, so they can later be retrieved for human analysis.

3. Natural Language Processing Tools

We have used as tool for developing information extraction applications GATE components such as tokenisers, sentence splitters, parts of speech taggers (Cunningham et al., 2002). Although GATE comes with a number of default resources, most of them require adaptation to a new domain. In our case we have created not only new recognisers that target domain concepts and different types of sources, but also have created resources for property-value identification and ontology population.

3.1. Rule Based System

Our initial system developed from manual corpus analysis is a rule based system which uses lexical information and syntactic information (POS and NP chunking) to assist a pattern matching process which identifies target concepts in text. A number of lexical resources are in place for the different applications; as an example in order to identify company information a number of gazetteer lists have been created which contain verbs and other linguistic constructs which mark ways in which products and services are expressed in unstructured company profiles. Using these gazetteer lists and syntactic information, rules have been created to identify company activities (“company produces X, Y, and Z” or “services provided include A, B, and C”). Other gazetteer lists are in place to identify for example structured parts of a Web page describing a company profile. Labels such as “Website:”, “Phone:”, and “Fax:” provide useful context for disambiguation of specific types of information (e.g. a web site mention in a page may not be the web site of the company described in the profile.)

3.2. OCR Processing System

Our objective is to analyse images to extract useful semantic information for ontology-based image indexing. Because of the many errors one usually finds in found in Optical Character Recognition (OCR), we are applying a methodology which first attempts to correct the OCR and then applies an extraction system to the corrected text. The methodology consists of the following steps: (i) OCR using off-the shelf software (ABBYY system `http://www.abbyy.com`); (ii) correction of OCR based on a set of candidates from collateral sources and the edit distance function; (iii) recognition of entities on corrected OCR. Using corrected texts before applying the extraction system shows a positive improvement in performance.

3.3. Consolidation

One of the problems we face with extraction from multiple sources is to decide whether an instance or fact is already present in the ontology. We are working with an identify resolution framework which uses a rule-based mechanism for deciding upon the similarity between two extracted entities. Rules to decide whether two instances are similar are defined on per class basis. So, rules for deciding on the identity of companies may differ from rules deciding on the identity of people. The rules use predicates which compute values, they either reason on the structure of the ontology (a city is sub class of location) or compute similarity values between attribute values of compared instances (e.g. the name “Metaware” is similar to the name “MetaWare Co. Ltd.”). Computed values are weighted and combined to produce a final score which is then used to rank candidates for resolution.

4. Experiments and Results

So far we have completed applications for information extraction on company profiles and extraction of factual information on countries and regions. Our extraction components have F-measure performance (on the concept level) superior to 80%, details will be given in the full paper. In the full paper we will also report results on consolidation experiments on extraction of company information from different sources.

5. Related Work

Information extraction in the business domain has a long tradition: one typical scenario for information extraction in the business domain is the case of insurance companies tracking information about ship sinkings around the globe (Wilks and Catizone, 1999). (Yangarber et al., 2000) 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.Another typical IE scenario is the extraction of information about joint ventures or other types of commercial company agreements from unstructured documents (Appelt et al., 1993; Jacobs and Rau, 1990). 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

6. Outline of the Paper

The paper will describe the adaptation to the business domain of traditional information extraction tools to create an ontology-based information extraction and ontology population system. We will describe the language resources and algorithms used to implement different business applications including the analysis of multi-media information, present evaluation results on extraction and information merging, and discuss current trends and future work.

7. References

D.E. Appelt, J.R. Hobbs, J. Bear, D. Israel, M. Kameyama, and M. Tyson. 1993. 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.

W Chung, H. Chen, and Nunamaker Jr. J.F. 2003. 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. IEEE Computer Society.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. 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)*.

Marty Ellingsworth and Dan Sullivan. 2003. Text mining improves business intelligence and predictive modeling in insurance. *DM Review Magazine*.

P.S. Jacobs and L.F. Rau. 1990. Scisor: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97.

A. Marshall, D. McDonald, H. Chen, and W. Chung. 2004. EBizPort: Collecting and Analysing Business Intelligence Iformation. *Journal of the American Society for Information Science and Technology*, 55(10):873–891.

Yorick Wilks and Roberta Catizone. 1999. Can We Make Information Extraction More Adaptive? In *M. Pazienza (ed.) Proceedings of the SCIE99 Workshop*, pages 1–16, Rome, Italy.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. 2000. Unsupervised Discovery of Scenario-level Patterns for Information Extraction. In *Proceedings of ANLP-NAACL'00*, Seattle, WA.