Ontology-based information extraction for market monitoring and technology watch*

Diana Maynard[1], Milena Yankova[1], Alexandros Kourakis[2], Antonis Kokossis[2]

[1]Department of Computer Science, University of Sheffield, UK
[2]University of Surrey, UK
diana,milena@dcs.shef.ac.uk
a.kourakis,a.kokossis@surrey.ac.uk

Abstract. The h-TechSight Knowledge Management Portal (KMP) enables support for knowledge-intensive industries in monitoring information resources on the Web, as an important factor in business competitiveness. The portal contains tools for identification of concepts and terms from an ontology relevant to the user's interests, and enables the user to monitor them over time. It also contains tools for ontology management and modification, based on the results of targeted knowledge extraction from the web. The platform provides a means for businesses to keep track of trends and topics of interest in their field, and alert them to changes. In this paper we focus on the tools for targeted search and ontology management, driven by an ontology-based information extraction system, which has been evaluated over a test set of 38 documents and achieves 97% Precision and 92% Recall.

1 Introduction

The growing pervasiveness of Knowledge Management (KM) in industry marks an important new watershed. KM has become embedded in the strategy, policy and implementation processes of institutions and organisations worldwide. The global KM market has doubled in size since 1991 and is projected to exceed US$8.8 billion in 2005. KM applications are expected to save Fortune 500 companies around $31 billion, and the broader application cost has similar projected forecasts. Although the tools and resources developed in h-TechSight are targeted towards SMEs, there are important implications for the growth and dispersion of such new technologies to industry as a whole. h-TechSight aims to pave the way for such development by providing a variety of knowledge management tools in its portal. In this paper, we focus particularly on the underlying Information Extraction (IE) technology, and show how enhancing traditional IE with ontological information can lead to more interesting and useful acquisition of knowledge and benefit real users in industry.

The h-TechSight KMP is a knowledge management platform with intelligence and insight capabilities for technology intensive industries. It integrates a variety of next generation knowledge management (NGKM) technologies in order to observe information resources automatically on the internet, and notify users about changes occurring in their domain of interest. There are various new technologies developed in this research:

- a tool/model for the development of ontologies, which can be used to describe concepts and trends in the user's domain of interest;
- a tool/model for the development of generic and targeted search agents which can use these ontologies to search for business intelligence from diverse web-based sources;
- a platform for integrating information from various sources and consolidating, analysing and publishing this information.

There are also new competences in the form of knowledge about porting the tools and methodologies into any industry/technology, and about localising support services throughout Europe.

* This work is partially supported by the EU-funded Knowledge Web network of excellence ((IST-2004-507482) and SEKT project (IST-2004-506826)

1.1 Technology watch in the employment domain

Employment is a generic domain into which a great deal of effort in terms of knowledge management has been placed, because every company, organization and business unit must encounter it. Human Resources departments often have an eye open for knowledge management in order to monitor their environment in the best way, and many recruitment consultant companies have watchdogs to monitor and alert them to changes. There exist a variety of job search engines (portals) which use knowledge management extensively to link employees and employers, e.g. JobSearch[1] and Job Portals[2].

The employment domain is also chosen for h-TechSight because it contains many generic kinds of concepts. First this means that an existing IE system can more easily be ported to this domain (because it does not require too much adaptation), and second, it does not require a domain expert to understand the terms and concepts involved, so the system can easily be created by a developer without special domain skills. These two considerations are very important in the fast development of a system to be used as an example application [9].

The employment application in the KMP aims to alert users to technological changes, since job advertisements are a very good indicator of moving trends in the field. By monitoring these advertisements over a period of months or even years, we can examine, for example, changes in the requirements for particular skills and kinds of expertise required, how salaries fluctuate, what kinds of qualifications are being demanded, and benefits awarded to employees.

1.2 Monitoring of the news domain

The news domain is another clear area where it is important for companies to keep a close eye on technological developments in their field. Primary market players for this are the pharmaceutical industry and the oil and gas industry. Pharmaceutical companies need to extract knowledge from diverse sources in order to predict pharmacological and toxicological effects, for example integrating knowledge from newly acquired organisations and keeping a close watch on news of and reports from their competitors. The oil and gas industry is currently faced with increasing pressures to create higher quality and more environmentally friendly products, and therefore such companies need up-to-the-minute access to news, reports, and experiences of colleagues around the world in order to leverage such information and respond to critical information requests from government agencies. Our application for the news domain is aimed at helping companies to access and monitor such information quickly and accurately, bringing new products, processes and technologies to their attention, as well as tracking the progress of rival companies in the field.

1.3 The h-Techsight Knowledge Management Platform

In this paper we shall focus on the application mode of the KMP, which is used for analysing and enhancing previously discovered information. The Targeted Search Module (Application Mode) can either be used as standalone, if the user already has access to the information sources required, or combined with the other tools in the platform such as the Generic Search module in order to discover such sources. In the following sections, we shall describe the tools for the data-driven analysis of terminology in the portal. These aim at creating semantic metadata automatically from web-mined documents, and monitoring concepts and instances (domain-specific terms) extracted over time. We have developed sample applications in the employment and news domains in the field of chemical engineering.

[1] http://www.job-search.com/
[2] http://www.aspanet.org/solutionstemp/jobport.html

2 Ontology-based Information Extraction

The advent of tools and resources for the semantic web brings new challenges to the field of Information Extraction (IE), and in particular with respect to Ontology-Based IE (OBIE). Such tools are being developed within the context of projects such as SEKT[3] and others (see Section 5). One of the important differences between traditional IE and OBIE is the use of a formal ontology rather than a flat lexicon or gazetteer structure. This may also involve reasoning. Another difference is that OBIE not only finds the (most specific) type of the extracted entity, but it also identifies it, by linking it to its semantic description in the ontology. This allows entities to be traced across documents and their descriptions to be enriched through the IE process.

If the ontology is already populated with appropriate instances, the task of an OBIE system may be simply to identify instances from the ontology in the text. Similar methodologies can be used for this as for traditional IE systems, but using an ontology rather than a flat gazetteer. For rule-based systems, this is relatively straightforward, other than in the case of ambiguity. For learning-based systems, however, this is more problematic because training data is required and collecting such training data is likely to be a large bottleneck. Unlike traditional IE systems for which training data exists in domains like news texts in plentiful form, there is a dearth of material currently available for semantic web applications. New training data needs to be created manually or semi-automatically, which is a time-consuming and onerous task, although systems to aid such metadata creation are currently being developed (see Section 5).

The advantage of OBIE over traditional IE is that the output (semantic metadata about the text) is linked to an ontology, so this enables us to extract much more meaningful information about the text, for example making use of relational information or performing reasoning. We therefore can get a much better "snapshot" of the text and draw more meaningful and useful conclusions from it. For example, in the employment domain, identifying the locations where there are job vacancies is handy (as can be done with traditional IE), but linking towns and cities to areas and countries provides us with much more useful information, because we can then perform analyses about specific areas (for example, that the computer industry is growing in the North of England, or that London-based jobs are providing better benefits packages than those in the rest of the UK).

3 GATE

GATE is an architecture for language engineering developed at the University of Sheffield [1], containing a suite of tools for language processing, and in particular, a vanilla IE system ANNIE. In traditional IE applications, GATE is run over a corpus of texts to produce a set of annotated texts. In h-TechSight, the input to GATE takes the form of a set of URLs of target webpages, and an ontology of the domain. Its output comprises annotated instances of the concepts from the ontology. The ontology sets the domain structure and priorities with respect to relevant concepts with which the application is concerned.

GATE's IE system is rule-based, which means that unlike machine-learning based approaches, it requires no training data [8]. On the other hand, it requires a developer to create rules manually, so it is not totally dynamic. The architecture consists of a pipeline of processing resources which run in series. Many of these processing resources are language and domain-independent, so that they do not need to be adapted to new applications [6]. Pre-processing stages include word tokenisation, sentence splitting, and part-of-speech tagging, while the main processing is carried out by a gazetteer and a set of grammar rules. These generally need to be modified for each domain and application, though the extent to which this is necessary depends on the complexity and generality of the domain. The gazetteer contains a set of lists which help identify instances in the text. Traditionally, this is a flat structure, but in an OBIE application, these lists can be linked directly to an ontology, such that instances found in the text are then related back to the ontology.

[3] http://www.sekt.semanticweb.org

3.1 GATE in h-TechSight

The GATE application performs targeted information extraction relative to a domain and ontology, enabling statistical information to be gathered about the data collected. Inferences drawn from this information pave the way for the monitoring of trends of new and existing concepts and instances. For example, companies can track information about their rivals over time, and check for the emergence of new companies, products and technologies.

The GATE application consists of 5 basic stages:

1. web mining application to find relevant documents (or manual input of relevant documents);
2. selection of concepts in which the user is interested;
3. information extraction;
4. visual presentation of results (annotation of instances) and statistical analysis
5. ontology modification (an ontology editor is used to enrich the existing ontology from the results of the analysis)

The application uses two main inputs: a web mining application which feeds relevant URLs to GATE based on the user's query, and a domain ontology. Alternatively, the user can input their own relevant documents to GATE. The texts are automatically annotated with semantic information based on the concepts in the ontology. Instances in the text can not only be visualised (through colour-coding) but can also be output in two forms: into a database for further processing, and in the form of a new ontology (DAML+OIL or RDF).

h-TechSight proceeds a stage further than traditional IE systems and other systems performing OBIE (see Section 5), by not only performing metadata generation and ontology population (by adding new instances to the ontology), but also by enabling the process of ontology evolution. By this we mean that the IE application serves not only to populate the ontology with instances, but also to modify and improve the ontology itself on the conceptual level. Statistical analysis of the data generated can be used to determine how and where this should take place. For example, a set of instances will be linked to a concept in the ontology, but this concept may be too general. A clustering algorithm can be used to group such instances into more fine-grained sets, and thereby lead to the addition of new subconcepts in the hierarchy. h-TechSight is unique in performing monitoring of the data over time, which can also lead to suggested changes in the ontology.

3.2 Application for the employment domain

For the employment domain in h-TechSight, a domain-specific application has been created, which searches for instances of concepts present in a sample employment ontology. The ontology has 9 main concepts: Location, Organisation, Sectors, JobTitle, Salary, Expertise, Person and Skill. Each concept in the ontology has a set of gazetteer lists associated with it. Some of these (generic lists) are reused from previous applications, while others (domain-specific lists) need to be created from scratch. In total there are around 60 domain-specific lists, and 50 generic lists. The generic lists are quite large (around 29,000 entries) and contain common entities such as first names of persons, locations, abbreviations etc. Collection of lists is done through corpus analysis (examining the texts manually and/or performing statistical analysis to spot important instances and concepts), unless a set of texts has been manually annotated by a user, in which case, the list collection process can be automatic [5]. For the employment domain, we used a combination of methods. We annotated around 20 documents manually and used this to collect lists automatically. This enabled us to bootstrap the development of the system and then complete the lists through further text analysis methods.

Grammar rules for recognition of new types of entities mainly use the gazetteer lists. However, not all entities can be recognised just from gazetteer lists. Some entities require more complex rules based on contextual

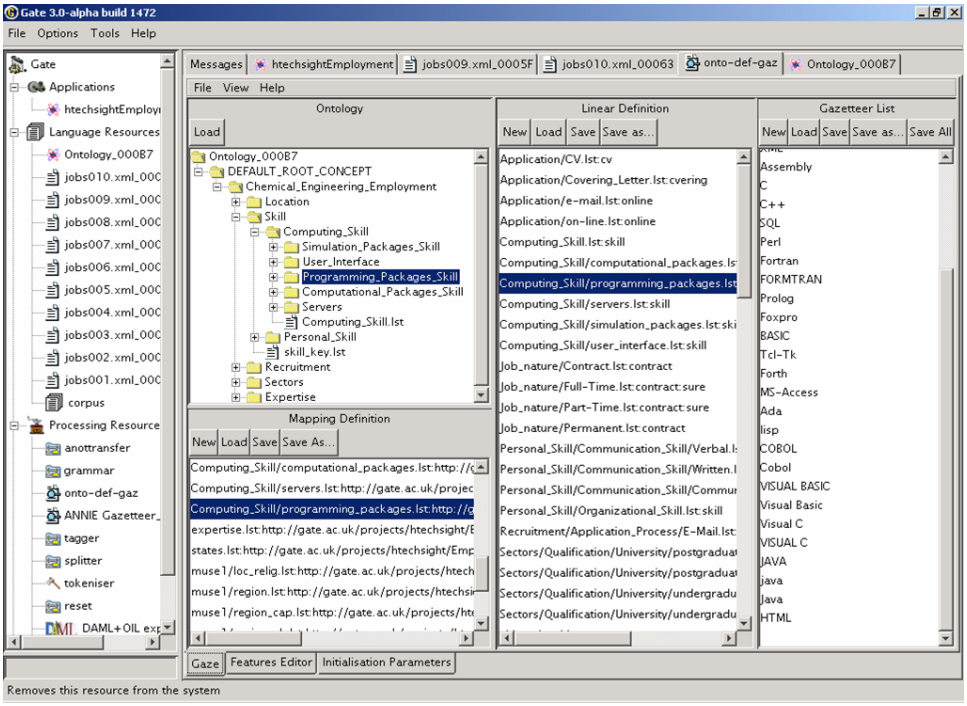

Fig. 1. Screenshot of Populated Employment Ontology in GATE

information. These may also use special lists that contain keywords and are used to assist such contextually-based rules. Some of the keyword lists are also attached to the ontology, because they clearly show the class to which the identified entity belongs. All lists that correspond to the ontology are ordered in a hierarchy similar to the class hierarchy in the ontology. A section of the ontology, the mappings from the lists to the ontology, and the contents of a list is shown in Figure 1.

The concepts in which we are interested can be separated into 3 groups. The first consists of classic named entities which are general kinds of concepts such as Person, Location, Organisation. The second is more specific to the chosen domain of employment, and consists of the following types:

- JobId - shows the ID of posted job advertisements;
- Reference - shows the reference code of the job position;
- Status - shows the employment/position type;
- Application - shows the documents necessary and the method of job application (e.g. by email, letter, whether a CV should be sent, etc.);
- Salary - shows the information available in the text about salary rates, bonus packages, compensations, benefits etc.;
- Qualification - shows the qualifications required for the advertised position, mainly a University degree;
- Citizenship - shows restrictions about the applicant's citizenship, eligibility, etc.;
- Expertise - shows the required expertise / skills for the job.

For both groups, the grammar rules check if instances found in the text belong to a class in the ontology and if so, they link the recognised instance to that same class and add the following features:

```
EntityType.ontology = ontology url,
EntityType.class = class name
```

The third group presents instances already annotated with HTML or XML tags (if such exist), and consists of the following:

- Company - contains the name of the organisation advertising the job;
- Date_Posted - shows the date when the job advertisement was posted;
- Title - shows the job title;
- Sector - shows the sector of the job that is advertised.

If these are not already annotated in the texts, they are identified using further rules.

The grammar rules for creating annotations are written in a language called JAPE [2]. The rules are implemented in a set of finite-state transducers, each transducer usually containing rules of a different type, and are based on pattern-matching. In traditional IE applications, the rules find a pattern on the LHS, in the form of annotations, and on the RHS an action such as creating a new annotation for the pattern. In OBIE applications such as this, the rules also add information about the class and ontology on the RHS of the rule. So for example the string "PhD" found in the text might be annotated with the features:

```
{class = Postgraduate}
{ontology = http://gate.ac.uk/projects/htechsight/Employment}
```

This information is taken from the gazetteer, which is mapped to an ontology, as described earlier. In total the application contains 33 grammars, which run sequentially over the text. Each grammar contains anything from 1 to about 20 rules, depending on the complexity of the annotation type.

3.3 Adaptation to the news domain

The GATE application for the news domain is focused on the area of chemical technologies. For this, a new domain-specific ontology and set of texts is required as input to the system. We have constructed a sample ontology consisting of 13 concepts related to the technologies domain, such as Corrosion, Thermodynamics, Optimization, Reaction, Equipment, etc. Some gazetteer lists were reused from the employment domain, while others needed to be created from scratch and mapped to the ontology in the correct place. In total there are 181 lists.

Some of the grammar rules used for the employment domain were directly reused for the news domain, while others had to be created from scratch. The aim was to minimise the amount of adaptation necessary; however, the nature of technical terminology makes generic kinds of rules very difficult to implement, because of its specialised nature and the fact that not only are the terms different, but the syntax and structure of more technical texts can be very different. A discussion of the problems in adapting an IE system to different genres and domains can be found in [10,7]. For this domain, the help of a chemical engineering expert was required, since it was impossible for a non-domain expert to understand correctly which instances should be linked with which concepts, and therefore to construct appropriate rules.

Unlike the employment domain, the news application also finds relations between entities in the text. This is accomplished by using JAPE grammar rules to search for instances belonging to two different concepts in the hierarchy, and analysing the syntax of the text between the two instances using a Noun Phrase and Verb Phrase chunker (also developed using JAPE grammars), to extract relevant relations based on verbal groups. So for example, we use patterns such as $< instance >< Verb >< Instance >$ to extract triples like $< TIPunit >< upgrades >< octane >$ (where "upgrades" is a relation between the two terms TIP unit and octane) from the sentence "The TIP unit upgrades the octane of the feed to achieve research octanes of close to 90 in the product". We can then form clusters between related concepts, using information collected about such relations and also infer other important knowledge. This is another example of how ontologies

can help us to extract more useful information, because instead of simply linking instances from certain annotation types (e.g. finding relations between Person and Organisation), we can progress up or down the hierarchy in order to obtain more or less fine-grained information.

4 Presentation and analysis of results

The GATE application for the employment domain has been implemented in the h-TechSight portal as a web service. The user may select a URL and choose the concepts for the ontology. Then by invoking the service, a new web page is created with highlighted colour-coded annotations of the web page selected. The results are collected by dynamically populating a Microsoft Access database, and their statistical analysis is presented inside the KMP. The database has the following structure:

- Concepts: the concept which the record set of the database is about;
- Annotations: the instance of the record set annotated inside a document;
- Document _ID: a unique ID for the document;
- Time_Stamp: a time stamp found inside the document.

4.1 Monitoring instance-based dynamics

One of the most primitive dimensions of ontologies is the display of data as concrete representations of abstract concepts, i.e. as instances. GATE leads the data-driven analysis in h-TechSight, as it is responsible for extracting from the text instances represented in the ontology. Statistical analysis is then invoked to present instance-based dynamics.

In the h-TechSight platform, we try to monitor the dynamics of ontologies using two approaches: dynamics of concepts and dynamics of instances. Users may not only annotate their own websites according to their ontology, but may also see the results of a dynamic analysis of the respective domain. They may see tabular results of statistical data about how many annotations each concept had in the previous months, as well as seeing the progress of each instance in previous time intervals (months). Following this analysis, end users may also see the dynamics of instances by means of an elasticity metric that indicates the trend of each individual instance. Developments in the GATE results analysis have eliminated human intervention, as the results are created automatically in a dynamic way. The two approaches to the monitoring of dynamics are described in more detail below.

Dynamic metrics of concepts are calculated by counting the total occurrences of annotated instances over time intervals (per month). By clicking on the concepts, a user may see the instances related to a concept. Instances are presented in a time series where the total occurrences per month and a calculation of an elasticity metric of instances are presented in tabular form. The elasticity metric (Dynamic Factor) counts the differences between the total occurrences of every instance over time intervals (per month) taking into consideration the volume of data of each time period. The mathematical type that calculates the DF takes into consideration the differences of volume of data (documents annotated by GATE) of each time period (months).

4.2 Analysis of results

From Table 1 we can examine how particular kinds of expertise are being sought over a period of time. Clearly, looking at just 3 months of data is not sufficient to make an informed analysis about trends, but looking at data over a longer period of time will be a useful indicator. Instances with a negative Dynamic Factor (DF) show an overall downward trend. The higher the dynamic factor, the greater the upward trend.

Instances	Dynamic Factor	Jan	Feb	Mar
1 year as a J2EE designer	-1	0	1	0
1 year JSP experience	48	0	0	2
2 years banking	23	0	0	1
2EE	145	0	15	6

Table 1. Dynamics of Instances for the Concept "Expertise"

Instances	Dynamic Factor	Jan	Feb	Mar
ARC	145	0	12	6
Archimedia SA	-1	0	1	0
Army	23	0	2	1
AT&T	-1	0	2	0
AT&T Wireless	-1	0	3	0
BA	23	0	3	1
BMI British Midland	-335	1	3	0
British Airways	-163	1	11	7

Table 2. Dynamics of Instances for the Concept "Organisation"

From Table 2 we can see how frequently different companies are placing job advertisements on the portals under scrutiny. One important fact to notice is that at the moment, if the same company is referred to in two (or more) different ways, the results will be stored individually, thus skewing the figures. For example, the counts for BA and British Airways are stored separately, because the system does not recognise that these refer to the same company We are currently implementing a coreference mechanism to cluster such term variants together, so that we only calculate one overall score rather than two separate ones. This is also extended to cluster more loosely connected variants, so for example the term C and C++ might be grouped together. In this way we can show two separate views of such clusters an overall count and DF for the cluster, and a table showing details of the individual instances that form the cluster.

4.3 Evaluation of the IE technology

We conducted an initial evaluation of the IE application to see how well the system found relevant instances of the concepts. We tested the system on a small set of 38 documents containing job advertisements in the Chemical Engineering domain, mined from the website http://www.jobserve.com. The web portal is mined dynamically using a web content agent written in WebQL, a commercial web crawling software[4]. We manually annotated these documents with the concepts used in the application, and used the evaluation tools provided in GATE to compare the system results with the gold standard. Overall, the system achieved 97% Precision and 91.5% Recall, with an F-Measure of 94.2%.

4.4 User Feedback

The KMP has been tested by real users in industry, such as Bayer Technology Services and IChemE. Users found that it was very helpful in increasing the efficiency of acquiring knowledge and supporting project work in industry, by helping to automatically scan, filter, structure and store the wealth of information available on the web related to their needs. For Bayer, the potential areas of application spanned from research and development, engineering and production, to marketing and management.

[4] http://www.webql.com

Users at IChemE, a leading international body which provides services for chemical engineers world-wide, claimed that the employment application was a very sound idea, and that it "would be a very valuable means of graduates gaining a fresh insight into their jobs and related training which may be narrower than ideally it should be due to company constraints (i.e. time and money for development!)".

One important fact to note is that due to the complexity of the underlying system, it is not really feasible for non-IE experts to adapt the system to new domains. However, since the system runs as a web service, the end user need have no knowledge of the underlying technology in order to use the system, so this is not necessarily a problem.

5 Related Work

There currently exist several other systems for automatic semantic metadata creation of web-based documents.

Magpie [4] is a suite of tools which supports the interpretation of webpages and "collaborative sense-making", by annotating a text with instances from a known ontology. These instances can be used as a confidence measure for carrying out some services. The principle behind it is that it uses an ontology to provide a very specific and personalised viewpoint of the webpages the user wishes to browse. This is important because different users often have different degrees of knowledge and/or familiarity with the information presented, and have different browsing needs and objectives.

KIM [11] is an architecture for automatic semantic annotation developed within a platform for semantic-based indexing and retrieval from large document collections. KIM contains an instance base which has been pre-populated with 200,000 entities (mostly locations), and performs information extraction based on GATE. Essentially, KIM recognises entities in the text with respect to the KIM ontology, and adds new instances where they do not already exist.

The SemTag system [3] performs large-scale semantic annotation with respect to the TAP ontology. It first performs a lookup phase annotating all possible mentions of instances from the TAP ontology, and then performs disambiguation, using a vector-space model to assign the correct ontological class or determine that this mention does not correspond to a class in the ontology.

h-TechSight and KIM both use the same core IE system, although KIM uses a general IE application while h-TechSight uses one tuned to the specific domain and ontology being used. KIM supports ontology modification in that it identifies new instances and adds them to the ontology. h-TechSight also supports ontology evolution, whereby the actual structure of the ontology can be modified as a result of the instances discovered, in a semi-automatic way (making suggestions to the user).

h-TechSight also has a slightly different goal from systems such as SemTag and KIM, in that these are domain-independent, large-scale approaches, while in h-TechSight the IE algorithms have been specifically created for particular domains and therefore can offer the extended functionality. Finding the balance between sophisticated functionality and good IE performance and domain independence is always difficult. The approaches used in KIM and SemTag are more appropriate for large-scale automatic annotation systems, while user involvement in the process of adding new instances is more beneficial for domain-specific applications which can afford to be semi-automatic and which, by their nature, are more suitable for user involvement.

There are also many other research efforts in the area of NGKM. Two major projects in this area, which are frequently referred to as grounding initiatives, are On-To-Knowledge[5] and Vision[6]. Related research on

[5] Content-driven Knowledge Management Tools through Evolving Ontologies IST-1999-10132
[6] http://km-aifb.uni-karlsruhe.de/fzi/vision/

Knowledge Management System development is discussed in detail in [12], but is not so relevant to this work, where we focus on the application-specific tools in the KMP.

6 Conclusions

In this paper we have presented an application for automatic knowledge extraction, management and monitoring in the Chemical Engineering domain, integrated in a dynamic knowledge management portal. Combined with the other tools and applications for knowledge engineering found within the portal, it forms the basis of a system for information retrieval, terminology acquisition and technology watch. GATE makes use of terminological processing and domain-specific IE to evolve existing ontologies automatically and to enable thte monitoring of domain-specific information relevant to the user. The application has been tested in the Employment sector with excellent results, and has been successfully ported to other genres of text such as news items and company reports.

References

1. 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.
2. 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.
3. S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of WWW'03*, 2003.
4. J. Domingue, M. Dzbor, and E. Motta. Magpie: Supporting Browsing and Navigation on the Semantic Web. In N. Nunes and C. Rich, editors, *Proceedings ACM Conference on Intelligent User Interfaces (IUI)*, pages 191–197, 2004.
5. D. Maynard, K. Bontcheva, and H. Cunningham. Automatic Language-Independent Induction of Gazetteer Lists. In *Proceedings of 4th Language Resources and Evaluation Conference (LREC'04)*, 2004. `http://gate.ac.uk/sale/lrec2004/gazcollector.pdf`.
6. D. Maynard and H. Cunningham. Multilingual Adaptations of a Reusable Information Extraction Tool. In *Proceedings of the Demo Sessions of EACL'03*, Budapest, Hungary, 2003. `http://gate.ac.uk/sale/eacl03/demo.pdf`.
7. D. Maynard, V. Tablan, K. Bontcheva, H. Cunningham, and Y.Wilks. Multi-source entity recognition – an information extraction system for diverse text types. Research Memorandum CS–03–02, Department of Computer Science, University of Sheffield, April 2003.
8. D. Maynard, V. Tablan, and H. Cunningham. NE recognition without training data on a language you don't speak. In *ACL Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, Sapporo, Japan, 2003. `http://gate.ac.uk/sale/acl03/surprise.pdf`.
9. D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274, 2002. `http://gate.ac.uk/sale/robust/robust.pdf`.
10. 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`.
11. B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM ?Semantic Annotation Platform. In *2nd International Semantic Web Conference (ISWC2003)*, pages 484–499, Berlin, 2003. Springer.
12. M. Stollberg, A. Zhdanova, and D. Fensel. h-TechSight – A Next Generation Knowledge Management Platform. *Journal of Information and Knowledge Management*, 3 (1):1–22, 2004.