Automatic Semantic Annotation with KIM

Atanas Kiryakov, Borislav Popov, Dimitar Manov, Damyan Ognyanoff,
Rosen Marinov, Ivan Terziev

Ontotext Lab, Sirma AI EAD, 135 Tsarigradsko Shose, Sofia 1784, Bulgaria

{naso, borislav, mitac, damyan, rosenm, ivanterziev}@sirma.bg
Tel: +359 2 9768 303; Fax: +359 2 9768 311

1 Executive Summary

The KIM platform[1] provides infrastructure and services for automatic semantic annotation, indexing, and retrieval of documents. It allows scaleable and customizable ontology-based information extraction (IE) as well as annotation and document management, based on GATE[2]. In order to provide a basic level of performance and to allow the easy bootstrapping of applications, KIM is equipped with an upper-level ontology and a massive knowledge base, providing extensive coverage of entities of general importance. The latter are accessed and managed through the use of a high-performance semantic repository (a tuning of Sesame, handling about 30M statements.) Semantically-enhanced information retrieval is provided on the basis of Lucene[3]. The vision and the modelling assumptions behind KIM are introduced in [KiryakovEtAl2004]; the platform is presented in [PopovEtAl2003].

The demonstration will go to show the recent developments of the KIM platform. It will consist of a set of scenarios, including:

- Instant semantic annotation, including the annotating of documents set forth by the audience. To be performed through the KIM Internet Explorer plug-in for highlighting and hyperlinking, based on semantic annotations (Fig. 2);
- Transition from a (text) document towards navigation over the knowledge in the semantic repository. To be demonstrated both hypertext and hyperbolic graph views.
- Exploration of the coverage and richness of the KIM Knowledge Base, again taking request from the audience. To be performed through a web UI;
- “Intelligence” research, based on a corpus of news articles. To include one pre-defined exercise and another one, based on a request from the audience (Fig. 3);

---

[1] http://www.ontotext.com/kim

[2] General Architecture for Text Engineering, http://gate.ac.uk, leading NLP and IE platform from the University of Sheffield.

[3] An open source full-text indexing engine from Jakarta, http://jakarta.apache.org/lucene/

- An overview of the top-100 most popular entities in the world, based on news articles.

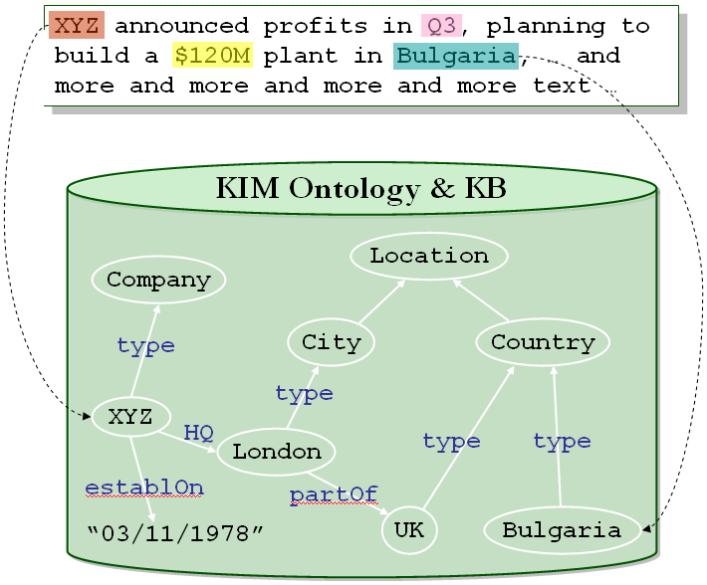

Fig. 1. Semantic Annotation

2 The KIM Platform

The Knowledge and Information Management (KIM) platform has been implemented in order to embody our vision of semantic annotation, indexing, and retrieval services and infrastructure. An essential notion in KIM is the semantic annotation. Here we bet on our vision that massive *automatic* semantic annotation is the prerequisite for the build-up of most of the metadata, needed for the Semantic Web to happen. For each entity, mentioned in the text, KIM provides references (URI) (i) to the most relevant class in the ontology, and (ii) to the specific instance in the knowledge base (see Fig. 1). As a result of the automatic semantic annotation, metadata is generated and associated with the resource processed. This metadata is not embedded in the processed document, thus allowing different semantic annotation tasks to take place, accordingly resulting in diverse sets of metadata.

To do this in a consistent fashion, KIM performs information extraction based on an ontology and a massive knowledge base. The traditional, flat named entity (NE) type sets consist of several general types (such as Organization, Person, Date, Location, Percent, Money). Although these represent the most important domain-independent NE types, still an extension is feasible. We identified an inter-domain NE type hierarchy from a corpus of general news and we integrated it within the KIM Ontology (KIMO). It contains definitions of about 300 entity classes and 100 attributes and relations. The semantic descriptions of entities and relations between them are kept in a knowledge base (KB), encoded in the KIMO. The semantic repository is initially pre-populated with the KIM

World KB, consisting of about 200 000 descriptions of entities of general importance, compiled from a number of reliable ("trusted") sources. The KB is being constantly enriched with new entities and relations, found during the annotation process.

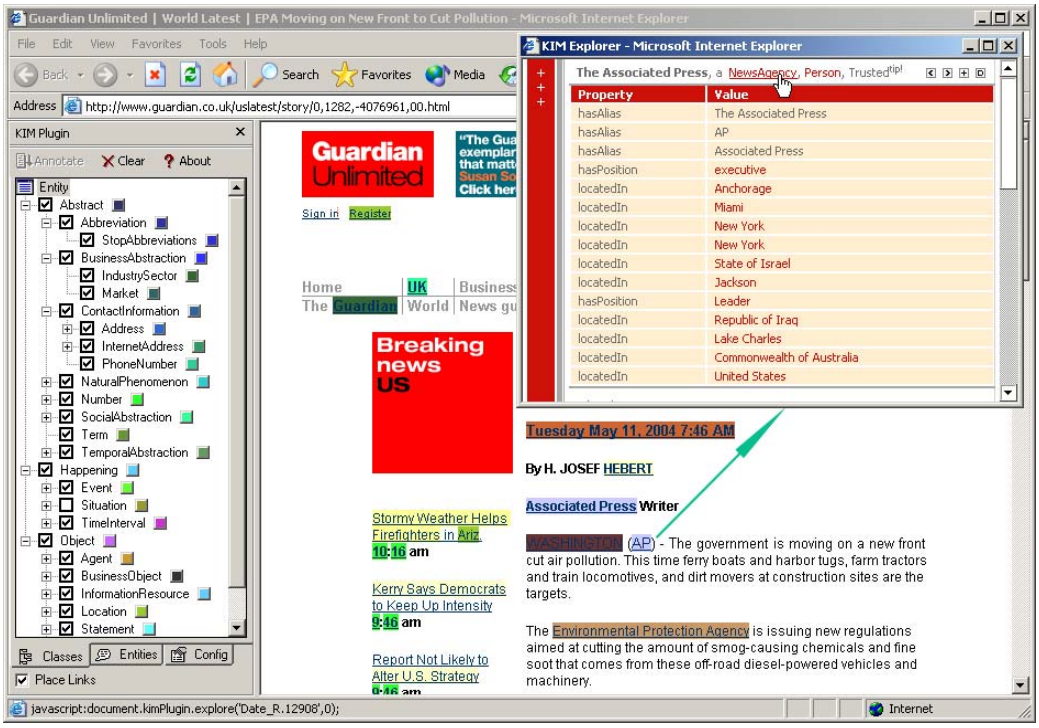

Fig. 2. The plug-in with the KIM Ontology shown in the left pane and the KIM KB Explorer on top

3 KIM Front-Ends

The KIM API allows the implementation of various front-ends, providing access to the KIM Server functionality and infrastructure. We have developed a web user interface (KIM Web UI, Fig. 3) that allows traditional access methods (key word search) and semantic ones (entity search, pattern search), too. The latter return either a set of entities that satisfy the query or a set of documents that refer to these entities. One could see the content with the associated metadata on the document level (title, author, etc.)

KIM is also equipped with a plug-in (Fig. 2) for Internet Explorer. It provides the delivery of light-weight semantic annotations to the end user. On its first tab, the plug-in displays the entity type hierarchy. For each entity class there is an associated colour, used for highlighting of the annotations of this class. Upon the user's request, the current page is processed and the annotations are highlighted and hyperlinked. On the second tab there is a list of all the recognized entities on the page. Upon selecting from the list of entities, or following a hyperlink over an entity, the user invokes the KIM KB Explorer – a web-form, which provides a view of the part of the KB and the ontology that are directly related to the respective (selected) entity (incl. type, aliases, relations, attributes). In this

way, the user could navigate from the annotations to the instances in the KB and further explore the KB by choosing one of the related entities, or the entity type.

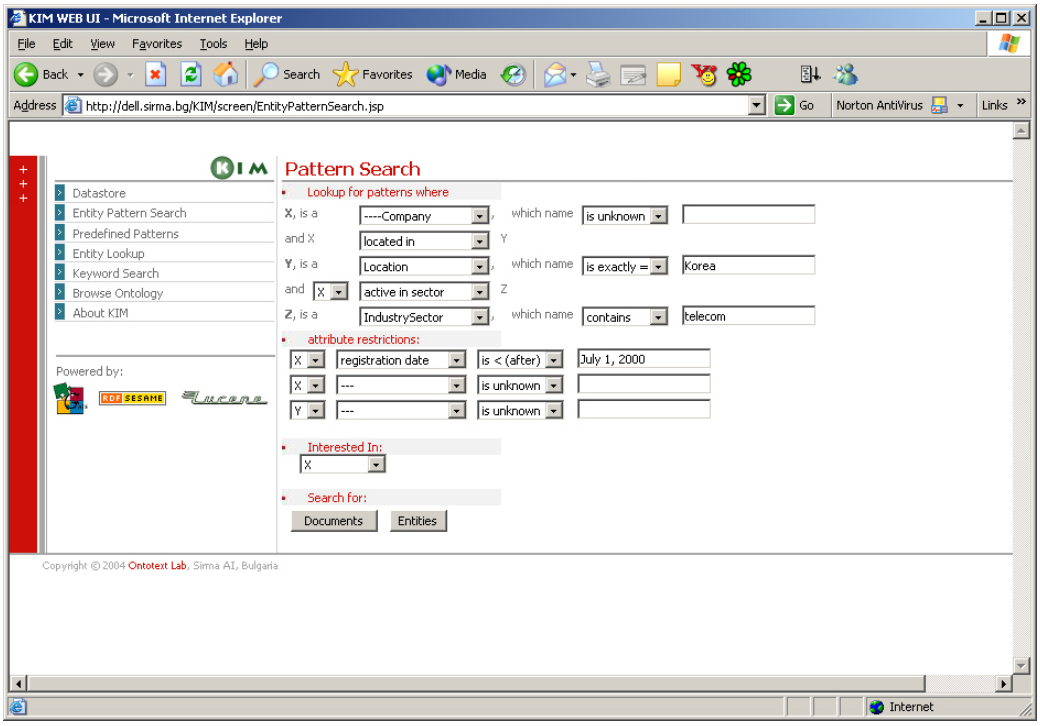

Fig. 3. Entity Pattern Search from the Web UI - looking for a telecom company in Korea

References

[KiryakovEtAl2004] Atanas Kiryakov, Borislav Popov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, Miroslav Goranov.
Semantic Annotation, Indexing, and Retrieval. To appear in Elsevier's Journal of Web Semantics, Vol. 1, ISWC2003 special issue (2), 2004. http://www.websemanticsjournal.org/

[PopovEtAl2003] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, Miroslav Goranov.
KIM – Semantic Annotation Platform. 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. LNAI Vol. 2870, pp. 484-499, Springer-Verlag Berlin Heidelberg 2003.