KIM – Semantic Annotation Platform

Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, Miroslav Goranov

Ontotext Lab, Sirma AI EOOD, 138 Tsarigradsko Shose, Sofia 1784, Bulgaria
{borislav, naso, angel, mitac, damyan, miro}@sirma.bg

Abstract. The KIM platform provides a novel Knowledge and Information Management infrastructure and services for automatic semantic annotation, indexing, and retrieval of documents. It provides mature infrastructure for scaleable and customizable information extraction (IE[1]) as well as annotation and document management, based on GATE[2]. In order to provide basic level of performance and allow easy bootstrapping of applications, KIM is equipped with an upper-level ontology and a knowledge base providing extensive coverage of entities of general importance. The ontologies and knowledge bases involved are handled using cutting edge Semantic Web technology and standards, including RDF(S) repositories, ontology middleware and reasoning.

From technical point of view, the platform allows KIM-based applications to use it for automatic semantic annotation, content retrieval based on semantic restrictions, and querying and modifying the underlying ontologies and knowledge bases. This paper presents the KIM platform, with emphasize on its architecture, interfaces, tools, and other technical issues.

1 Introduction

The acquisition of masses of metadata for the web content would allow various Semantic Web applications to emerge and gain wide acceptance. Such applications would provide and use new access methods based on the associated metadata. At present there are various IE technologies available that allow recognition of named entities within the text, and even the relations, events, and scenarios in which they take part. Thus, metadata could be assigned to the document, presenting part of its information content, suitable for further processing. Such metadata can range from formal reference to the author of the document, to annotations of all the companies and amounts of money referred in the text. It is an important question how to make this metadata machine readable for the purposes of effective structuring, discovery, automation, integration, and reuse.

The approach for automatic (versus manual) extraction of metadata is promising scalable, cheap, author-independent and (potentially) user-specific enrichment of the

[1] Information extraction, a relatively young discipline in the Natural Language Processing (NLP), which conducts partial analysis of text in order to extract specific information.

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

web content. However, at present there is no technology available to provide automatic semantic annotation in conceptually clear, intuitive, scalable, and accurate enough fashion. Even more, there is no clear vision regarding the approach and model for generation and representation of such annotations.

This paper presents first an innovative model for semantic content enrichment, named semantic annotation (section 2.) This model is implemented in the KIM platform which is presented in the third section with its architecture. KIM ontology and knowledge base are presented in sections 4 and 5. Section 6 offers a brief explanation of the semantic IE process. The indexing and retrieval by entities is outlined in section 7. In section 8 there is a brief overview of the KIM front-ends. The ninth section contains KIM performance and IE evaluation metrics. Short outline of the technologies used as basis for the KIM Platform is present in section 10. Discussion on related work is present in section 11. In the last section we present our conclusion and plans for future work.

2 Semantic Annotation

The semantic annotation offered here is a specific metadata generation and usage schema targeted to enable new information access methods and extend existing ones. It is based on the hypothesis that the named entities[3] mentioned in the documents constitute important part of their semantics (see [19], section 3.2 for discussion on this statement.) In a nutshell, we consider Semantic Annotation the idea of assigning to the entities in the text links to their semantic descriptions. The idea of this sort of metadata is to provide both class and instance information about the entities referred in the documents. It is a question of terminology whether these annotations should be called “semantic,” “entity” or some other way. To the best of our knowledge there is no well established term for this task; neither there is a well established meaning for the term “semantic annotation”[4].

The automatic semantic annotations enable new applications: highlighting, indexing and retrieval, categorization, generation of more advanced metadata, smooth traversal between unstructured text and available relevant knowledge. Semantic annotation is applicable to any sort of text – web pages, regular (non-web) documents, text fields in databases, etc. Further, knowledge acquisition can be performed based on extraction of more complex dependencies – analysis of relationships between entities, event and situation descriptions, etc. We believe that, defined this way, semantic annotation is clearly specified, easy to understand, and can serve as a basis for number of useful applications (some of those demonstrated in KIM).

The automatic semantic annotation can be seen in this case as a classical named-entity recognition (NER) and annotation process. The semantic annotation is specific for providing more precise type information than the systems based on flat NE type

[3] Named entities are people, organizations, locations, and others referred by name. The wide interpretation of the term includes any tokens referring something specific in the world: numbers, addresses, amounts of money, etc.

[4] The term is previously used in [29] in a bit more general sense compared to what we propose, but it didn’t got wide acceptance.

sets (not bound to a taxonomy or other definitions.) The NE type is specified by reference to an ontology, and more important, the semantic annotation requires identification of the entity. While in a classical NER task, guessing the type is everything to be achieved, a semantic annotation needs to recognize the entity (either out of a set of known ones either as a new, unknown, one) and refer to it.

Considerations about the structure and the representation of the semantic annotations, including the necessary knowledge and metadata are not presented in detail. These issues (and many others, not directly related to KIM) are discussed in [19]; here we will only sketch the conclusions promoted there:

– Semantic annotation system requires a light-weight upper-level ontology focused on named entity classes (unlike Apple, Worm, Chair, Key, etc.);
– RDF(S) with compliance and possible extensions to OWL Lite is the best choice for knowledge representation language for the ontology and the KB – more power will unnecessarily degrade the scale and performance;
– the documents and the metadata (annotations) should be kept decoupled from each other and separate from the ontology and the knowledge base.

3 The KIM platform

The KIM[5] platform provides services and infrastructure for semantic annotation, indexing, and retrieval. To do this in a consistent fashion, it performs information extraction based on an ontology and a massive knowledge base.

The traditional flat 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 the entities with same type are dividable in more specific classes from the average educated human (e.g. public companies, sport teams, and syndicates are all organizations.) We identified an inter-domain NE type hierarchy from a corpus of general news and integrated it in the KIM Ontology (KIMO). The ontology contains definitions of entity classes, attributes, and relations, as well as a branch of lexical resource types (e.g. Title, PersonFirstName, etc.). The semantic descriptions of entities and relations between them are kept in a knowledge base (KB) encoded in the KIM ontology and residing in the same semantic repository. Thus KIM provides for each entity reference in the text (i) a link (URI) to the most specific class in the ontology and (ii) a link to the specific instance in the KB. Each extracted NE is linked to its specific type information (thus Arabian Sea would be identified as **`Sea`**, instead of the traditional – **`Location`**). The KB has been pre-populated with entities of general importance, and is iteratively enriched with entity individuals and relations as a result of the IE process. Thus the extracted named entities could be further used for semantic indexing and retrieval of content with respect to entity instance and type, as well as name and attribute restrictions, and expected relations between these entities (e.g. look for a **`Sea`** that is a sub region of the Indian Ocean).

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

For the end-user, the KIM IE functionality is straightforward and simple to use – requesting annotation from a browser plug-in, which highlights the entities in the current web page and generates a hyperlink used for further exploring the available knowledge for the entity (as shown on Fig. 1). Various access methods are also available – entity pattern search, entity lookup, keyword and document attribute search. There is also an opportunity to create a composite query consisting of atomic searches of the above types.

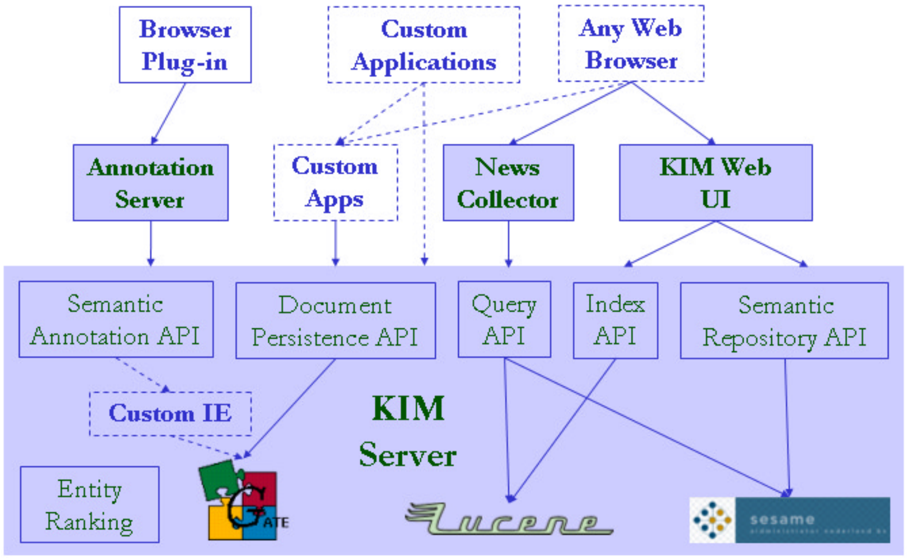


Fig. 1. KIM Architecture

3.1 KIM Architecture

The KIM platform consists of KIM Ontology (KIMO), knowledge base, KIM Server (with API for remote access, embedding, and integration), and front-ends (browser plug-in for Internet Explorer, KIM web user interface with various access methods, and Knowledge Explorer for KB navigation). The KIM API provides semantic annotation, indexing and retrieval services and infrastructure. KIM ontologies and knowledge bases are kept in semantic repositories based on cutting edge Semantic Web technology and standards, including RDF(S) repositories (SESAME[6]), and ontology middleware[7] [16]. KIM provides a mature infrastructure for scalable and customizable information extraction, as well as annotation and document management, based on GATE [7]. The Lucene[8] information retrieval engine has been

[6] http://sesame.aidministrator.nl/, RDF(S) repository by Aidministrator b.v.
[7] OMM, http://www.ontotext.com/omm.
[8] Lucene, http://jakarta.apache.org/lucene/, high performance full text search engine

adopted to index documents by entity types and measure relevance according to entities, along with tokens and stems. It is important to mention that KIM, as a software platform, is domain and task independent as are Gate, Sesame, and Lucene. The KIM Architecture diagram is depicted on Fig. 1.

3.2 KIM API

The KIM Server API for remote access, embedding, and integration provides functionality and infrastructure for semantic annotation, indexing, and retrieval, as well as document management, and KB navigation. The modules of the KIM API can be seen at the middle layer on Fig. 1. The Semantic Annotation API allows annotation of documents with respect to KIM ontology and KB. It also provides infrastructure for content and annotations management. Documents and the associated annotations could be stored to/loaded from a data store via the Document Persistence API. The Index API is based on the Lucene IR engine, which has been modified to allow indexing with respect to named entities.

The Query API could be thought as a semantic IR API, which allows the specification of traditional key word searches, and also other ontology-aware access methods. It provides the infrastructure for constructing composite searches, combining entity search, keyword search, and entity pattern search. Due to the indexing with respect to entities, one could request all documents that are referring an entity described via restrictions over its name, class, and attributes. Even more, it is possible to specify a pattern of entities with the corresponding relations between them and ask for the documents that are referring the entities that satisfy the query.

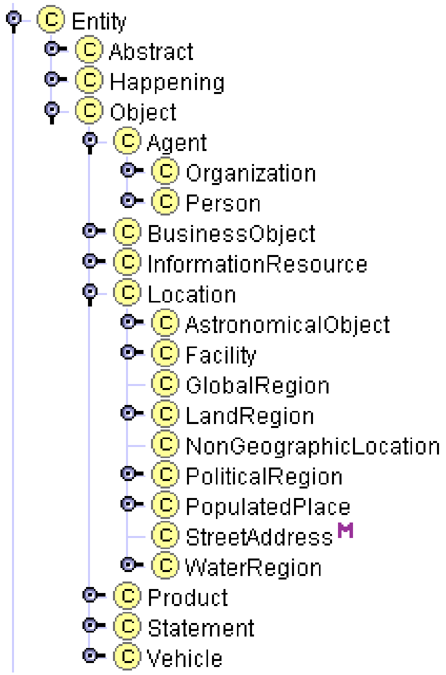
For the means of managing and accessing the KB there is a Semantic Repository API, allowing access to the underlying KB through an RDF(S) compliant infrastructure and method set.

4 KIM Ontology

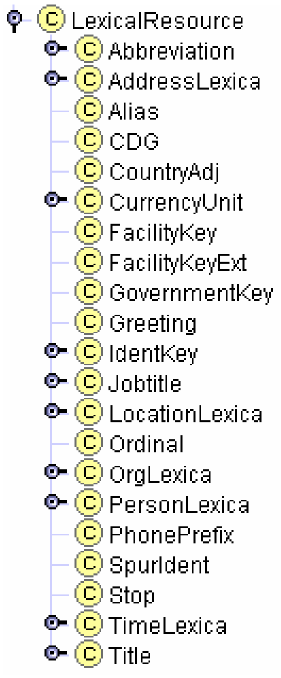
The rationale behind the KIM Ontology (KIMO) is to provide a minimal but sufficient ontology suitable for open-domain general purpose semantic annotations. It is designed from scratch for the purposes of KIM; inspirations are taken from number of upper-level resources: OpenCyc, WordNet 1.7, DOLCHE, EuroWordnet Top, and others. In order to keep it simple and easy to understand, it is kept small and naïve with respect to big number of philosophical, mathematical, and logical problems.

KIMO is a simplistic upper-level ontology starting with some basic distinctions between entity types (such as **`Object-`**s - existing entities such as locations and agents, **`Happening-`**s – defining events and situations, and **`Abstract-`**ions that are neither objects, neither happenings). Further on, the ontology goes in more details to such extent that real-world entity types of general importance are included (meetings, military conflicts, employment positions, commercial, government and other organizations, people, and various locations, etc.). The characteristic attributes and relations for the featured entity types, are defined (e.g. **`subRegionOf`** property for

Location-s, **hasPosition** for **Person**s, **locatedIn** for organizations, etc.) Having this ontology as basis, one could add domain-specific extensions to it easily, to profile the semantic annotation for concrete applications. The integration of multiple domain-specific ontologies for use within a single application is hardly possible without the intermediate role played by the upper-level ontology.



- Entity
  - Abstract
  - Happening
  - Object
    - Agent
      - Organization
      - Person
    - BusinessObject
    - InformationResource
    - Location
      - AstronomicalObject
      - Facility
      - GlobalRegion
      - LandRegion
      - NonGeographicLocation
      - PoliticalRegion
      - PopulatedPlace
      - StreetAddress
      - WaterRegion
    - Product
    - Statement
    - Vehicle



- LexicalResource
  - Abbreviation
  - AddressLexica
  - Alias
  - CDG
  - CountryAdj
  - CurrencyUnit
  - FacilityKey
  - FacilityKeyExt
  - GovernmentKey
  - Greeting
  - IdentKey
  - Jobtitle
  - LocationLexica
  - Ordinal
  - OrgLexica
  - PersonLexica
  - PhonePrefix
  - SpurIdent
  - Stop
  - TimeLexica
  - Title

Fig. 2. The top of KIMO class hierarchy with expanded Entity branch. (on the left)
Fig. 3. The Lexical Resources top class hierarchy.

The distribution of the most commonly referred entity types varies greatly from domain to domain. As researched in [23], despite the difference of type distributions, there are several general entity types that appear in all corpuses – Person, Location, Organization, Money (amount), Dates, etc. The proper representation and positioning of those basic types was one of the objectives behind the design of KIMO. Further the ontology defines more specific classes, e.g. **Mountain**, as a sub-class of **Location** (the location-related classes are extensively covered, see [21] for more details.)

The extent of specialization of the ontology is determined on the basis of research of the entity types in a corpus of general news (incl. political, sport, financial, etc.) The KIM ontology (KIMO)[9] consists of about 250 entity classes and 100 attributes and relations. The top classes are **Entity**, **EntitySource** and **LexicalResource**. The top entities classes could be seen in the type hierarchy of the KIM plug-in on Fig. 6, and separately on Fig. 2.

[9] http://www.ontotext.com/kim/kimo.rdfs always refers to the most recent version

The **`LexicalResource`** branch is dedicated to encoding various data aiding the IE process, such as company suffixes (AG, Ltd.), person first names, etc. (depicted on Fig. 3.) An important sub class of this branch is **`Alias`**, representing the alternative names for an **`Entity`** (see Fig. 4). The **`hasAlias`** relation is used to link an **`Entity`** to its alternative names. The official name of an entity is referred by the **`hasMainAlias`** property.

The instances of the **`EntitySource`** class are used to separate the trusted (pre-populated) information in the KB, from the automatically extracted one. This is indicated by the **`generatedBy`** property of the specific entity.

5 KIM Knowledge Base

No matter how sophisticated the automatic IE process is, one can still benefit from a starting KB to represent the entities that are important (popular, often cited, highly connected to others, etc) in the respective domain. Having a critical mass of trusted (in a sense handcrafted) knowledge can serve as a seed for further automatic population. Based on the determinism provided by this trusted knowledge, more advanced extraction methods could be applied to get it extended. In a sense, being able to grasp, learn, and infer is important, but knowing something in advance can be expected to improve the overall performance. Inferring or extracting with uncertain methods basic knowledge should be avoided in cases when it can be engineered in advance and maintained with reasonable efforts. This is the philosophy behind KIM, it reflects our intuition and can hardly be justified with something else than the overall performance of the KIM platform.

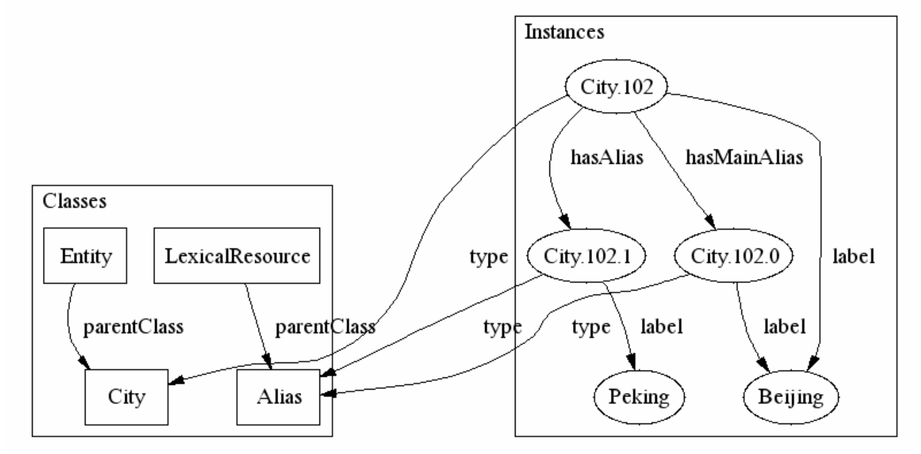


Fig. 4. RDF(S) Representation of an Entity description.

Combining the above philosophy with the aim of building a general-purpose, open-domain, semantic annotation platform leads us to genesis of the KIM KB. It is aiming to provide quasi exhaustive coverage of the most important entities in the world. In different words, the goal could be defined as: to cover what an undergraduate knows outside its country, specialization, and hobby. At the base line, this means the entities with their proper classes and aliases.

Obviously, building a domain-independent general knowledge base is too complex task and defined this way does not provide an obvious realization strategy. While still using various (including encyclopedic) sources and mental exercises, we substitute this task with an easier one which seems to serve as a good approximation: to provide good coverage of the entities mentioned in the international news. Here we mean those publications which use to cross the borders of the countries and feed the headlines of the global news wires. The specifics about this domain is that it covers (and also pre-determines) the most well known entities in the world.

KIM keeps the semantic descriptions of entities in the KIM KB, which is repeatedly enriched with newly recognized entities and relations. The entity descriptions are being stored in the same RDF(S) repository as the KIM ontology. Each entity has information about its specific type, aliases (incl. a main alias, expressing the most probable full or official name), attributes (e.g. latitude of a **`Location`**), and relations (e.g. a **`Location subRegionOf`** another **`Location`**). A schema of the entity representation in RDF(S) is depicted on Fig. 4.

5.1 Pre-population of KIM KB

KIM KB has been pre-populated with entities of general importance, that allow enough clues for the IE process to perform well on inter-domain web content. It consists of above 80,000 entities.

In its current state the KIM KB contains about 50,000 locations, including continents, global regions, 282 countries with their capitals, 4,700 cities (including all the cities with population over 100,000), mountains, big rivers, oceans, seas, and even oil fields. Each location has geographic coordinates and several aliases (usually including English, French, Spanish, and sometimes the local transcription of the location name) as well as co-positioning relations (e.g. **`subRegionOf`**.) Based on this spatial knowledge makes KIM a good basis for location-based services.

The organizations with highest general importance also have been pre-populated in the KB. Including the biggest world organizations (such as UN, NATO, OPEC), over 7,900 companies, and 140 stock exchanges for a total of 8,400 organization instances. For the public companies there are position relations of managing personnel. The organizations also have **`locatedIn`** relations to the corresponding **`Country`** instances. The additionally imported information about the companies consists of short description, URL, reference to an industry sector, reported sales, net income, and number of employees. A more extended report on the pre-population of the KIM KB can be found in [21].

Finally, in order to enable the IE process to recognize new entities and relations that are not a part of the KB, a collection of lexical resources is also presented in the KB. It covers organization suffixes, person names, time lexica, currency prefixes and others.

5.2 Controlling the Quality and Coverage of KIM KB

Ensuring the quality of the KB content, is not trivial and could not be performed manually (having more than 80,000 pre-populated entities, the manual approach simply not scale). The KIM KB is iteratively verified using an independently built KB of entities and relations collected manually from various web sources. An indirect verification is also performed during the evaluation of the performance of the KIM IE against a human annotated corpus.

The coverage of the KIM KB is being guaranteed through processing and analysis of the leading articles of the global news wires. The corpus of these articles is being constantly updated and bears on average 4000 documents per week – the top stories, as well as all other the economic and political news collected from about 15 sources. More details about the KIM KB are presented in [21]. On top of the corpus gathered this way, entity ranking is performed to detect the level of "popularity" of the specific entities. This allows proper manual handling at least of the most popular entities, as well, as early spotting of problems with the import strategy and sources.

6 Semantic Information Extraction

The essence of the KIM IE approach is the recognition of named entities with respect to a formal upper-level ontology (KIMO.) The NE annotations are typed with respect to the entity classes in the ontology. The entity instances all bear identifiers (URI,) that allow the annotations to be linked to the exact individual in the KB. The IE involved in KIM is currently concentrated mostly on the NER task, which is considered a step-stone for further attribute, relation, event, and scenario extraction. In order to identify the references of entity relations in the content, one should first have identified the related NE. Usually the entity references are associated with a NE type, such as Location, Person, etc. More and more hierarchical NE type sets appear, especially for domain-specific applications. This is due to the need for finer grained specification and identification of world concepts. For example, it would be natural for an IE application performing company intelligence to keep more specialized sub-classes of Organization (e.g. such as PublicCompany). A NE type taxonomy however brings in a new level of complexity and (as discussed in section 9) sets new challenges for the evaluation of the performance, since the traditional Precision/Recall metrics are not directly applicable.

The IE process presented here uses a light-weight ontology (KIMO) defining the entity classes. In addition to the hierarchical ordering, each class is coupled with its appropriate attributes. The relation types are also defined with their domain and range restrictions. Actually, the basic ontology language used (RDFS) considers both the relations and attributes as properties, which can also be organized in a hierarchy. Given the ontology, the entities in the text could be linked to their type, which is also feasible with just a type taxonomy. However we would like to go further, and identify not only the type of the NE but also keep its semantic description and extend it through the IE process. Thus, the NE references in the text are linked to an entity individual in the KB (section 5). The accessibility of the semantic descriptions of entities in the KB would allow the IE process to later base on attributes and relations as clues for recognition and disambiguation. For example, if a **`Person`** appears along with a **`Company`** in the content, and there are two companies that have the mentioned alias, we face an ambiguity. A possible approach would be to check whether the Person has some relations with one of the companies (e.g. working in it), and if so, the related **`Company`** to be chosen as a better candidate and associated with the NE reference in the content.

KIM IE is based on the GATE framework, which has proved its maturity, extensibility and task independency for IE and other NL applications. We have reused much of GATE's document management functionality, and generic NLP components as its Tokenizer, Part-of-Speech Tagger, and Sentence Splitter. These processing layers are provided by the GATE platform, along with pattern-matching grammars, NE coreference and others, as standardized building bricks for easy construction of sophisticated IE applications.

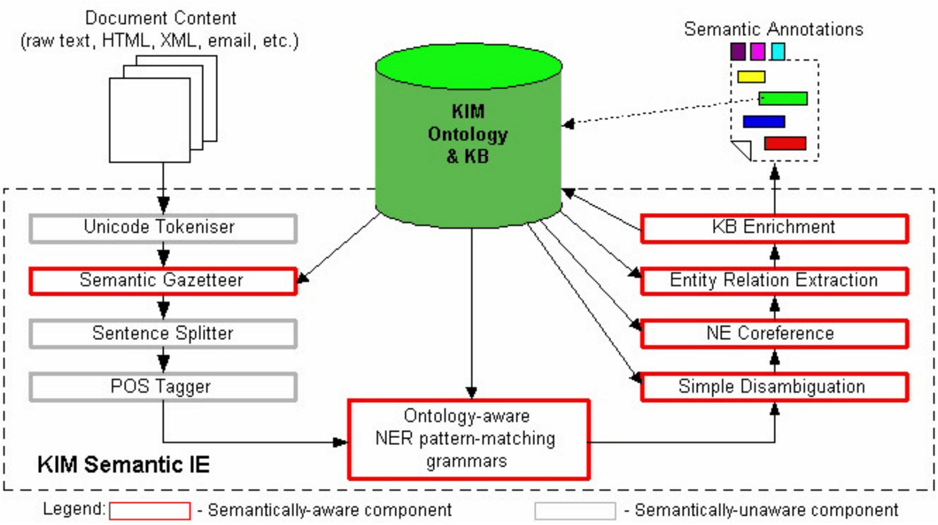


Fig. 5. KIM Semantic IE flow diagram.

For our purposes we changed the grammar components to handle entity class information and match rules according to it. The grammar rules are now based on the ontology classes, rather than on a flat set of NE types. This allows much more flexibility in the creation of NER rules at the most appropriate level of generality, giving both the opportunity to generalize and handle more specific NE types. A rule trying to extract relation between an organization and its point of presence can be specified at the level of the most general classes it applies to (Organization and Location) and still match a patterns with much more specific information (say, a radio station located in a county). On the other hand, instead of referring to all locations we could prefer to have rules that are especially applicable for Countries, Cities, or Seas.

The Semantic Gazetteer component is looking up in the text for entity aliases and lexemes loaded from the KB – its key role is to provide fast track for annotation of entities referred with known aliases. Along with it, all the reused components have been opened towards the semantic repository. For example, the NE co-reference module, in addition to the traditional ortho-matching techniques, handles the instance information of NE annotations and matches them according to it, as well as the traditional substring transformation matching. The Semantic Gazetteer, the simple disambiguation and annotation filtering components, as well as the final KB enrichment layer have been developed from scratch. These are not innate to a traditional NER and are inquired by the specifics of the Semantic IE, which takes care of the identification of NE references with respect to the ontology and KB.

The IE component flow diagram (Fig. 5) displays the sequential processing of content to the point where semantic annotations are produced over it. The semantic repository is also displayed and linked with the ontology and KB aware components.

7 Indexing and Retrieval

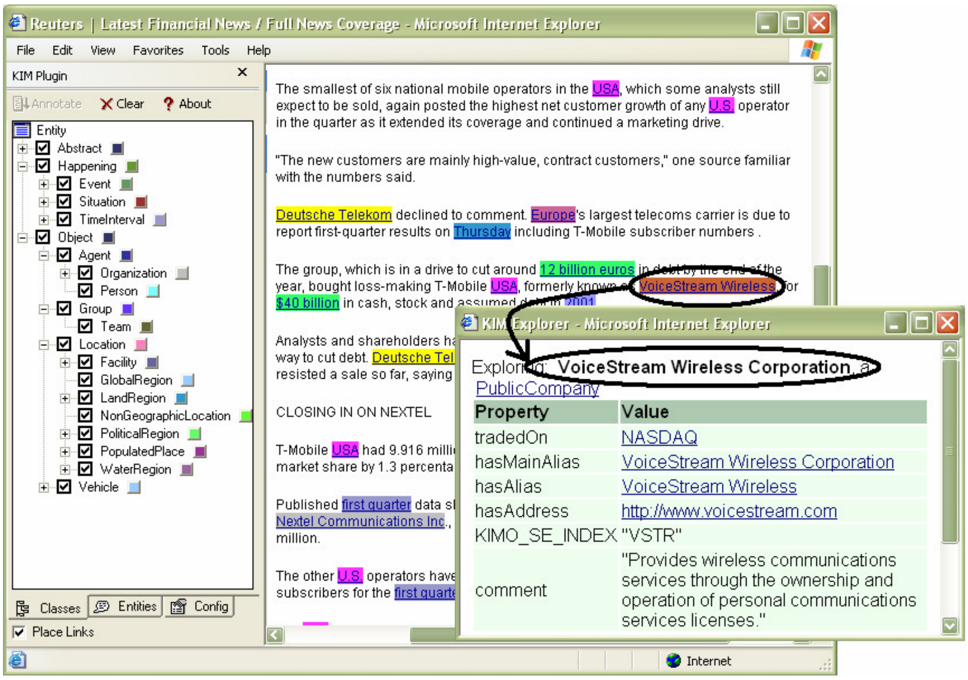
After semantic annotations of the documents, KIM allows those to be indexed with respect to the contained named entities. This allows later on document retrieval to be performed with respect to entities. One could specify the NEs that to be referred in the documents of interest, with name restrictions (e.g. a Person which name ends with ‘Alabama’). Further, pattern of entities, relations between them and attribute restrictions can be specified. To answer the query, KIM applies semantic restrictions over the entities in the KB. Then the documents referring the resulting entities are being retrieved with relevance ranking according to NEs. Technically, Lucene is adapted to perform full-text indexing, which is uniquely addressing each entity disregarding the alias used in the text.

The retrieval accuracy of KIM has not been evaluated against a traditional IR engine, and this is a topic that should be researched in the future. However, KIM has the potential to perform better not only on reducing the unrelated documents in the result set while still retrieving the relevant ones (as a NE indexing system with flat entity types as in [24] and [29]), but also to increase the number of relevant documents, with such that do not contain the alias that was used for the entity name restriction, but have the same entity mentioned with another of its aliases.

8 KIM Front-Ends

The KIM API allows the implementation of various front-end tools providing access to the KIM Server functionality and infrastructure. We have created a web user interface (KIM Web UI) allowing traditional (key word search), as well as semantic access methods (entity search, pattern search.) Along with these the KIM Web UI generic content management and IR functionality allows indexing and storing of semantically annotated content. The entity and pattern searches return either a set of entities that satisfy the query, either 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.) The resulting entities are highlighted in the content and linked to their semantic descriptions in the KB.

KIM is also equipped with a plug-in (Fig. 6) for the Internet Explorer browser. It provides light-weight semantic annotations delivery to the end user. On its first tab, the plug-in displays the entity type hierarchy (a branch of the KIM ontology). For each entity type there is an associated color used for highlighting the annotations of this type. Check boxes for each entity, allow the user to select the entity types of interest. Upon invoking annotation of the current browser content, the plug-in extracts the text of the currently displayed document and sends it to an Annotation Server which is in its turn using the KIM Server Semantic Annotation API. The servers return the annotations with their offsets, type and instance information. The annotations are highlighted in the content (in the color of the respective entity type), and are hyperlinked to the KIM KB Explorer (Fig. 6.) On the second tab of the plug-in, there is a list of all the recognized entities for the current document, sorted by appearance frequency. Upon choosing from the list of entities, or following a hyperlink over an annotated entity in the text 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 chosen entity (incl. type, aliases, relations, and attributes). This way, the user could easily navigate from the named entity annotations to the instances that they are linked to in the KB. Via this explorer, the KB could be further explored by choosing one of the related entities, or the entity type.



The smallest of six national mobile operators in the USA, which some analysts still expect to be sold, again posted the highest net customer growth of any U.S. operator in the quarter as it extended its coverage and continued a marketing drive.

"The new customers are mainly high-value, contract customers," one source familiar with the numbers said.

Deutsche Telekom declined to comment. Europe's largest telecoms carrier is due to report first-quarter results on Thursday including T-Mobile subscriber numbers .

The group, which is in a drive to cut around 12 billion euros in debt by the end of the year, bought loss-making T-Mobile USA, formerly known as VoiceStream Wireless, for $40 billion in cash, stock and assumed debt in 2001

Analysts and shareholders ha
way to cut debt. Deutsche Tel
resisted a sale so far, saying

CLOSING IN ON NEXTEL

T-Mobile USA had 9.916 milli
market share by 1.3 percenta

Published first quarter data s
Nextel Communications Inc.,
million.

The other U.S. operators have
subscribers for the first quarte

Exploring: VoiceStream Wireless Corporation, a PublicCompany

Property	Value
tradedOn	NASDAQ
hasMainAlias	VoiceStream Wireless Corporation
hasAlias	VoiceStream Wireless
hasAddress	http://www.voicestream.com
KIMO_SE_INDEX	"VSTR"
comment	"Provides wireless communications services through the ownership and operation of personal communications services licenses."

Fig. 6. The KIM plug-in with the top of the KIM Ontology, and KIM Explorer on top

The News Collector[10] is reusing some of the KIM Web UI and is an example of a KIM-based application which collects, annotates, and indexes news articles.

9 KIM Performance and IE Evaluation

Here we present brief performance metrics of the KIM Server and evaluation of KIM IE against a human annotated corpus. The evaluation corpus contains 100 documents of news articles from several UK media sources (Financial Times, Independent, Guardian) with total size close to 1 MB. We performed semantic annotation, indexing and storage of the metadata over the documents from the corpus on a dual-processor (1.13 GHz) PC server to acquire the following throughput metrics: annotation - 8kb/s; indexing 27 kb/s; storage 5kb/s; average 10 annotations/kb; 70 annotations/s. The speed of annotation depends on the document size, getting slower for bigger documents in a sort of logarithmic dependency.

The evaluation of the accuracy of the IE process has been performed over the same corpus human-annotated with the traditional flat NE types used by most of the NER systems. Despite the fact that KIM provides more specific type information, it is still possible to test it against the human annotated corpus (because something that is a **`Mountain`** is also a **`Location`**). In Table 1 we present the Precision, Recall and F-Measure of the automatically annotated corpus versus the human annotated one.

Table 1: KIM evaluation versus the human annotated corpus with respect to general NE types.

Annotation Type	**Precision**	**Recall**	**F-Measure**
Date	0.91	0.83	0.87
Person	0.90	0.85	0.87
Organization	0.77	0.56	0.64
Location	0.92	0.92	0.92
Percent	1	1	1
Money	1	1	1
Total	*0.86*	*0.82*	*0.84*

The above accuracy measures bear considerable distortion and only roughly approximate the accuracy of the semantic annotations. Sizeable semantically annotated corpus and new metrics are necessary for proper evaluation.

Given the complexity of the semantic IE process and the early development phase in which KIM is, these performance metrics encourage us to think that the automatic semantic annotation really could cope with the scale of the web content.

[10] A service which collects and semantically annotates the latest news from more than 10 of the most popular news sources. It is the one used to collect the corpus, further used for “entity ranking” and maintenance of the KIM KB. See http://news.ontotext.com

10 KIM Back-Ends

The technologies that KIM builds on have been carefully chosen, to be both mature enough, scalable and platform independent. The knowledge resources are kept in the Sesame RDF(S) repository, which provides storage and query functionality infrastructure. The Sesame repository is used with much over a million of RDF(S) statements. It is being queried by the semantic search methods to identify the entities according to the provided restrictions, and the result is further used for the retrieval of the referring documents. The IE process also relies on the semantic repository for its initialization and further processing.

The GATE platform has been used as a basis for the IE process and also for the management of content and annotations. It provided the basis text analysis technologies on top of which we have added the semantically aware extensions specific for KIM IE. The annotations and document management paradigms have been derived from the GATE infrastructure, but slightly simplified to avoid dependencies of the KIM clients on anything beyond the KIM API.

The Lucene IR engine has been adopted to perform indexing and retrieval with respect to named entities and evaluation of content relevance according to the entities which allows the semantic access methods described in section 7. The adoption of Lucene proves that it is easy to adjust a traditional IR engine to perform indexing with respect to metadata. Which means that there is no need of completely new technologies for the purposes of semantic indexing but rather the systems providing semantic IR could base on the existing quality IR engines.

11 Related Work

Semantic annotation of documents with respect to ontology and entity knowledge base is discussed in [5] and [14] – although presenting interesting and ambitious approaches, these do not discuss usage of information extraction for automatic annotation. The focus of [14] is manual semantic annotation for authoring web content, while [5] targets the creation of a web-based open hypermedia linking service, backed by a conceptual model of document terminology.

Semantic annotation is used also in the S-CREAM project presented in [13] – the approach there is interesting with the heavy involvement of machine learning techniques for extraction of relations between the entities being annotated. Similar approach is taken also within the MnM project [35], where the semantic annotations can be placed inline in the document content and refer to an ontology and KB server (WebOnto), accessible via standard API. Related to the previous is the project OFF, [9], which puts an emphasize on the collaborative ontology development.

Significant amount of research on IE has been performed in various projects related to GATE (see [22], [2], [6-8], [10], [23]). GATE provides tools such as tokenizers, part-of-speech taggers, gazetteer lookup components, pattern-matching grammars, coreference resolution tools and others that aid the construction of various NLP and especially IE applications. GATE is also a framework for content and annotation management. KIM's IE and content management is grounded in the GATE framework, and opens it towards Semantic Web knowledge representation and management technologies.

For some time now it has been obvious that the several general NE types used by the IE systems are not specific enough for many applications, that there are much more categories that matter. NE type hierarchies design has been discussed in [38].

All the semantic annotation techniques referred above lack the usage of upper-level ontologies and critical mass of world knowledge to serve as a trusted and reusable basis for the automatic recognition and annotation, as in the approach presented in [1] and discussed here. Also the IE processes involved in related work do not link the NE reference in the text with a NE individual in the KB. Because of this unique feature the semantic description of the entity instance reveals its attributes, aliases, type, origin source, and relations with other individuals.

12 Conclusion and Future Work

We had shortly introduced semantic annotations – an innovative notion for meta-data schema, generation and usage, discussed in more details in [19]. Next presented the KIM platform implementing it, based on GATE (which was substantially developed) and Sesame (which was substantially tuned to manage the size of the KB.)

Even linguistically simple, KIM platform provides a test bed and proofs number of hypothesis and design decisions: (i) its worth using almost-exhaustive entity knowledge (sort of super-gazetteers) for information extraction. The technology used (based on GATE) can manage the scale. Even without significant efforts on disambiguation, the precision drawbacks are acceptable for many applications; (ii) It is possible to adopt a traditional symbolic IE system to perform semantic annotations and thus provide its results in shape suitable for Semantic Web applications.

The implementation is currently under development, so, preliminary results are reported. In the future we would like to develop (or adapt) evaluation metric which properly measures the performance of a semantic annotation system; Experiment different approaches towards disambiguation of NE references; Make use of more advanced IE techniques for identification of relations, analysis of events and situations, etc. The KIM Ontology and KB as well as the methodology and procedure for their sustainable maintenance and improvement will be subject of future research.

References

1. Bontcheva K., Kiryakov A., Cunningham H., Popov B., Dimitrov M. *Semantic Web Enabled, Open Source Language Technology.* In proc. of EACL Workshop “Language Technology and the Semantic Web”, NLPXML-2003, 13 April, 2003
2. Bontcheva K., Dimitrov M., Maynard D., Tablan V., Cunningham H., *Shallow Methods for Named Entity Coreference Resolution.* Chaînes de références et résolveurs d'anaphores, workshop TALN 2002, Nancy, France, 2002.
5. Carr L., Bechhofer S., Goble C., Hall W. *Conceptual Linking: Ontology-based Open Hypermedia.* In The WWW10 Conference, Hong Kong, May, pp. 334-342.

6. Cunningham H., *Information Extraction: a User Guide* (revised version). Department of Computer Science, University of Sheffield, May, 1999
7. Cunningham H., Maynard D., Bontcheva K. and Tablan V., *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. In Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.
8. Cunningham, H. and Maynard D. and Tablan V., *JAPE: a Java Annotation Patterns Engine* (Second Edition). Technical report CS--00--10, Univ. of Sheffield, Department of Computer Science, 2000.
9. Collier N., Takeuchi K, Kawazoe A. *Open Ontology Forge: An Environment for Text Mining in a Semantic Web World.* In proc. of the International Workshop on Semantic Web Foundations and Application Technologies, Nara, Japan, 11th March.
10. Dimitrov M., Bontcheva K., Cunningham H., Maynard D., *A Light-weight Approach to Coreference Resolution for Named Entities in Text*. Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lisbon, September 2002.
13. Handschuh S., Staab St., Ciravegna F. *S-CREAM – Semi-automatic CREAtion of Metadata.*The 13th International Conference on Knowledge Engineering and Management (EKAW 2002), ed Gomez-Perez, A., Springer Verlag, 2002.
14. Kahan J., Koivunen M., Prud'Hommeaux E., Swick R.. *Annotea: An Open RDF Infrastructure for Shared Web Annotations*. In The WWW10 Conference, Hong Kong, May, pp. 623-632.
16. Kiryakov A., Simov K.Iv., Ognyanov D. *Ontology Middleware: Analysis and Design* Del. 38, On-To-Knowledge, March 2002. http://www.ontoknowledge.org/downl/del38.pdf
19. Kiryakov A., Popov B., Kirilov A., Manov D., Ognyanoff D., Goranov M. *Semantic Annotation, Indexing, and Retrieval.* In proc. of 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. To appear.
21. Manov D, Kiryakov A, Popov B, Bontcheva K, Maynard D, Cunningham H. *Experiments with geographic knowledge for information extraction*. NAACL-HLT 2003, Canada. Workshop on the Analysis of Geographic References, May 31 2003, Edmonton, Alberta.
22. Maynard D. ,Tablan D.,Ursu C., Cunningham H., Wilks Y., *Named Entity Recognition from Diverse Text Types*. RANLP 2001 Conference, Tzigov Chark, Bulgaria.
23. Maynard D., Tablan V., Bontcheva K., Cunningham H, and Wilks Y., *MUlti-Source Entity recognition – an Information Extraction System for Diverse Text Types*. Technical report CS--02--03, Univ. of Sheffield, Dep. of CS, 2003. http://gate.ac.uk/gate/doc/papers.html
24. Moldovan D., Mihalcea R.. *Document Indexing Using Named Entities*. In "Studies in Informatics and Control", Vol. 10, No. 1, March 2001.
29. Pustejovsky J., Boguraev B., Verhagen, M., Buitelaar P., and Johnston M., *Semantic Indexing and Typed hyperlinking*. In Proc. of the AAAI Conference, Spring Symposium, NLP for WWW, 120-128. Stanford University, CA, 1997.
35. Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A. and Ciravegna F, *MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup*, In Proc. Of EKAW 2002, ed Gomez-Perez, A., Springer Verlag, 2002.
38. Sekine S., Sudo K., Nobata Ch., *Extended Named Entity Hierarchy* (LREC 2002)