BOEMIE ontology-based text annotation tool

**Pavlina Fragkou, Georgios Petasis, Aris Theodorakos,
Vangelis Karkaletsis, Constantine D. Spyropoulos**

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Center for Scientific Research (N.C.S.R.) “Demokritos”
P. Grigoriou and Neapoleos Str., 15310 Aghia Paraskevi Attikis, Greece
E-mail: {fragou, petasis, artheo, vangelis, costass}@ iit.demokritos.gr

Abstract

The huge amount of the available information in the Web creates the need of effective information extraction systems that are able to produce metadata that satisfy user’s information needs. The development of such systems, in the majority of cases, depends on the availability of an appropriately annotated corpus in order to learn extraction models. The production of such corpora can be significantly facilitated by annotation tools that are able to annotate, according to a defined ontology, not only named entities but most importantly relations between them. This paper describes the BOEMIE ontology-based annotation tool which is able to locate blocks of text that correspond to specific types of named entities, fill tables corresponding to ontology concepts with those named entities and link the filled tables based on relations defined in the domain ontology. Additionally, it can perform annotation of blocks of text that refer to the same topic. The tool has a user-friendly interface, supports automatic pre-annotation, annotation comparison as well as customization to other annotation schemata. The annotation tool has been used in a large scale annotation task involving 3000 web pages regarding athletics. It has also been used in another annotation task involving 503 web pages with medical information, in different languages.

1. Introduction

Nowadays, the vast amount of information available in the Web remains in a considerable degree an unexploited thesaurus of knowledge resources. Among the approaches aiming to perform effective extraction, analysis and fast search is those that try to explore the actual content of the web pages by producing metadata. Metadata is usually defined as “data about data” aiming to express the “semantics” of information, thus improving information seeking, retrieval, understanding and use. Metadata can be applied both to a wide range of documents as well as to applications available in the web in the form of web services. The most important problems in extracting metadata from documents and applications firstly lie in the description of the “semantics” of the desired information and secondly in the requirements of information extraction systems to extract those metadata.

Regarding the first problem, semantics can be simply defined by the specification of a simple vocabulary however they are more effectively defined by ontologies. This is due to the fact that, ontologies not only specify the vocabulary of interest based on a set of agreed keywords each of which corresponding to a concept defined, but specify also the properties of those concepts, relations between concepts as well as formal axioms.

Regarding the second problem, information extraction systems in order to be able to extract metadata must be provided with an appropriately annotated corpus. The production of such corpora can be significantly facilitated by annotation tools.

Annotation tools ideally must support the annotation of the following tasks: (a) annotation of named entities i.e. of certain types of proper names (b) annotation of possible relations between the named entities, such as the relation between an employee and a company (c) annotation of a structure with extracted information involving several relations and events of interest, based on previous steps, such as annotation of the positions, companies and persons involved in high-level management succession events. In all steps, the annotation is guided by the semantics defined.

While a considerable number of annotation tools can be found in the literature, the majority of those are restricted to the annotation of named entities. Additionally, there are unable to cope with the annotation of large corpora. In order to overcome those problems and to perform annotation for the purposes of the BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction) project (www.boemie.org) for the text modality, a new text annotation tool was developed by extended an already existing one. This tool supports the annotation of named entities, the so-called Middle level concepts (MLCs) of the domain ontology, using the BOEMIE terminology. A part of the BOEMIE ontology concerning the text modality is depicted in Fig.1. The annotation tool enables also the annotation of relations between those named entities. These relations are grouped in tables of specific types. Tables correspond to High Level Concepts (HLCs) of the domain ontology, (according to BOEMIE terminology) having their fields filled with MLC instances. Furthermore, the tool enables the annotation of relations between HLC instances by creating linkages between tables in an effective and easy way. An additional property of the tool is that it is able to annotate blocks of texts in a web page that refer to the same topic defined by the presence of predefined types of MLC instances.

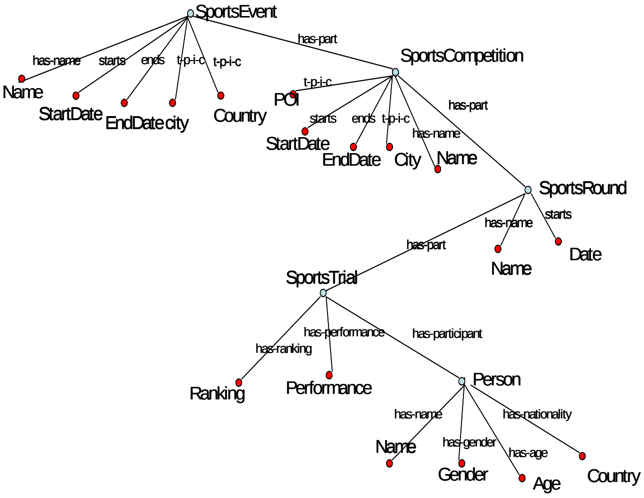


Figure 1: Part of the BOEMIE domain ontology

Finally, the BOEMIE text annotation tool supports a number of additional features that facilitate annotators' work. Among the most important ones is the handling of various file formats, the support of customized annotation schemata, the storage of annotations separately from the original documents and the rendering of html web pages. It is also capable of dealing with large amounts of documents, provides a comparison tool in order to compare annotations and performs automatic pre-annotation exploiting existing annotations.

Our semantic annotation tool was used in large scale annotations tasks involving the annotation of 3000 web pages regarding athletics and 503 web pages containing medical information. The annotation task proved to be significantly difficult due to the richness of information contained in them. However the annotation tool with the linkage property of tables corresponding to HLC instances as well as the automatic annotation of MLC instances proved to be extremely helpful to the annotators.

The structure of the paper is as follows: Section 2 presents relevant work in this research area, Section 3 describes in detail our annotation tool, Section 4 presents the annotation tasks performed using the BOEMIE annotation tool while Section 5 contains concluding remarks and future steps.

2. Related Work

An interest research study about existing annotation tools can be found in (Uren, 2005). However, even thought a variety of semantic annotation tools may be found in the literature, our study for existing annotation tools performed for the purposes of the BOEMIE project was focused on tools specifying the following criteria:

- use of standard formats,
- support user centered /collaborative design and user friendly interface,
- support of customized annotation schemata, rendering of html web pages,
- possibility of performing automatic annotation,
- annotation of relations between named entities,
- table filling with named entities where each table corresponds to a composite concept defined in the domain ontology (HLC),
- grouping of tables, and finally
- comparison facility.

Our study resulted in the comparison of the following annotation tools: Callisto [1], Wordfreak [2], GATE [3] (Cunningham, 2002), MMAX2 [4] (Müller and Strube, 2006), Knowtator [5] (Ogren, 2006), AeroSWARM [6] (Corcho, 2006), and Ellogon [7] (Petasis, 2003). The results of the study led to the following observations:

(a) all seven tools provide a user friendly interface (even with certain limitations) to perform named entity annotation,

(b) they all provide a tool for supporting the customization of annotation schemas; Knowtator and AeroSWARM provide an annotation schema defined by a corresponding ontology, while GATE, Callisto and MMAX2 support customized annotation schemas specified by XML file,

(c) Callisto, GATE, Ellogon and AeroSWARM provide a rendering of web pages,

(d) GATE, Ellogon, MMAX2 and Knowtator provide an annotation comparison functionality,

(e) Wordfreak, GATE and Ellogon support automatic annotation by the use of regular expressions – Wordfreak provides also active learning for human correction of automatically annotated data,

(f) MMAX2 and AeroSWARM enable the annotation of relations between entities. Knowtator has the ability to relate annotations to each other via slot definitions of the corresponding annotated classes. Annotation of relations is also performed by extensions of GATE i.e. by information extraction systems such as SEKT[8]. OBIE also sues GATE but extends it by providing the facility of automatic annotation.

Of special interest are systems that are extensions of Protégé. The Knowtator annotation tool (Fernández, 2005) belongs in this category. This can use ontologies both in RDF and OWL and perform semi-automatic

[1] http://callisto.mitre.org/
[2] http://wordfreak.sourceforge.net/
[3] http://gate.ac.uk/
[4] http://www.eml-research.de/english/research/nlp/download/mmax.php
[5] http://bionlp.sourceforge.net/Knowtator/index.shtml
[6] http://projects.semwebcentral.org/projects/aeroswarm/
[7] www.ellogon.org
[8] http://www.sekt-project.com/

annotation based on the text contained in labels that describe each concept. Another example is the iAnnotateTab [9] used in (Zhang, 2007) which is an OWL-based Protégé plug-in performing manual annotation of text with ontology concepts.

3. BOEMIE annotation tool

The majority of the aforementioned tools seem to work well for the case of the annotation of named entities. However, to the best of our knowledge, there is a lack of information regarding their ability to (a) annotate effectively and rapidly an important number of relations between named entities captured in a form of tables thus creating HLC instances and relations between them by appropriately relating their tables, (b) annotate blocks of texts which refer to a single topic based on predefined domain specific categories. These requirements were essential for the annotation of text corpus regarding athletics used by the BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction) information extraction system, (www.boemie.org) leading to the creation of a text annotation tool based on an already existing one satisfying the requirements listed below.

3.1 BOEMIE annotation requirements

In the case of text corpora the BOEMIE ontology-based information extraction system needs at a first step to locate those blocks of text that correspond to specific types of MLC instances (MLCIs) i.e. named entities as these are defined in the domain ontology (e.g. person name, person age, sports event name, sports event dates, etc. – see Fig. 1). At a second step it needs to fill tables corresponding to HLC's i.e. ontology concepts with the extracted information (e.g. fill an instance of the concept person by putting together the extracted person name, gender, age, nationality, as well as all of its synonyms for specific types of named entities). An HLC instance is composed by a specific set of MLC instances so the annotator should be able to create groups from the already annotated MLCIs. As HLCIs can be related with each other(s), the tool should have the functionality to create a relation between two or more HLCIs according to the defined HLC properties and rules of the domain ontology. In order to create those relations the annotation tool must be able to link the filled tables based on concept's relations defined in the domain ontology (e.g. link a person instance with a specific sports trial at a specific round of a sports competition in the context of a sports event). The extraction of such types of semantic information creates also the need for an annotation tool that will facilitate human annotators to annotate appropriately text corpora in order to to support the training and testing of BOEMIE text information extraction system. Furthermore, due to the fact that information extraction systems are more effective when applied to the relevant parts of a document (e.g. in the paragraph(s) talking for a specific athlete) instead of the whole document, the annotation tool must also support the annotation of text segments referring to a single piece of information.

Taking under consideration that the ontology may evolve over time, the need for support of a complex, ontology based annotation schema which should be easily customized by defining the annotation types in correspondence with a subset of ontological elements (concepts and relations) is high. Additionally, due to the fact that the amount of text may be significantly high, the annotation process must be assisted by additional functionalities such as automatic annotation of MLCIs matching either user-defined regular expression patterns or a previously annotated named entity. For example, if the user annotates "Tatyana Lebedeva" as a name, the automatic annotation system should be able to find and annotate every occurrence of that name in the document. Finally, the annotations must be saved in an accessible format for NLP tools so that they can be further processed and also be exported in an OWL Abox compliant with the ontology.

The above requirements led us to develop a new tool based on the text annotation tool component of the Ellogon text engineering platform (http://www.ellogon.org). Ellogon was chosen among the existing annotation tools due to its facility of developing components for the aforementioned requirements. In order to meet the BOEMIE requirements it was appropriately extended to support table filling, table linkage and automatic annotation by using a learning model. Detailed description of the basic and the extended functionalities of BOEMIE annotation tool is provided in the following sections. A screenshot of the BOEMIE annotation tool is depicted in Figure 2.

3.2 Basic functionalities

BOEMIE text annotation tool has been developed over the Ellogon text-engineering platform, extending the embedded annotation tool. It has been designed to run as a stand-alone application to avoid installation requirement and dependencies from other software even the whole interface of the Ellogon platform. The main functionalities that it offers are:

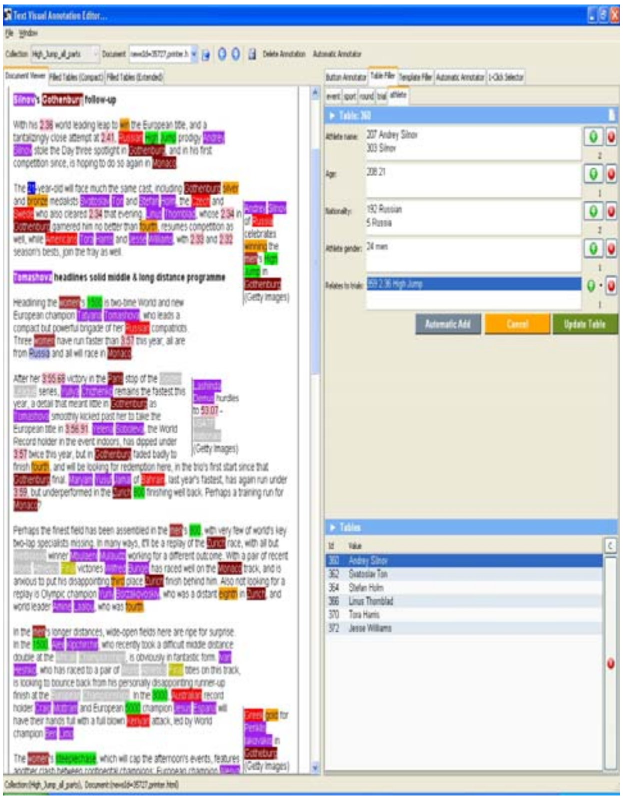
a) It creates and deletes corpus for annotation from html or text documents. Such a corpus is called an Ellogon collection. However, even though annotations may be performed to a collection, the original files are kept intact.

b) It displays html documents properly as it has a built-in html renderer.

c) The user can annotate both manually and automatically. Manual annotation is facilitated by a smart text-marking system by which the user selects with a mouse click words instead of single characters. In its initial version,

[9] www.dbmi.columbia.edu/~cop7001/iAnnotateTab/iannotate.htm

automatic annotation works by matching user-defined regular expression patterns.



Collection: High_Jump_all_parts
Document: newsId=35727,printer.h

Silnov's Gothenburg follow-up

With his 2.36 world leading leap to win the European title, and a tantalizingly close attempt at 2.41, Russian High Jump prodigy Andrei Silnov stole the Day three spotlight in Gothenburg, and in his first competition since, is hoping to do so again in Monaco.

The 21-year-old will face much the same cast, including Gothenburg silver and bronze medalists Svatoslav Ton and Stefan Holm, the Czech and Swede who also cleared 2.34 that evening. Linus Thornblad, whose 2.34 in Gothenburg garnered him no better than fourth, resumes competition as well, while Americans Tora Harris and Jesse Williams, with 2.33 and 2.32 season's bests, join the fray as well.

Andrei Silnov of Russia celebrates winning the men's High Jump in Gothenburg (Getty Images)

Tomashova headlines solid middle & long distance programme

Heading the women's 1500 is two-time World and new European champion Tatyana Tomashova, who leads a compact but powerful brigade of her Russian compatriots. Three women have run faster than 3:57 this year, all are from Russia and all will race in Monaco.

After her 3:55.68 victory in the Paris stop of the [illegible] series, Yulia Chizhenko remains the fastest this year, a detail that meant little in Gothenburg as Tomashova smoothly kicked past her to take the European title in 3:56.91. Yelena Soboleva, the World Record holder in the event indoors, has dipped under 3:57 twice this year, but in Gothenburg faded badly to finish fourth, and will be looking for redemption here, in the trio's first start since that Gothenburg final. Maryam Yusuf Jamal of Bahrain, last year's fastest, has again run under 3:59, but underperformed in the Zurich 800 finishing well back. Perhaps a training run for Monaco?

Lashinda Demus hurdles to 53.07 - [illegible] (Getty Images)

Perhaps the finest field has been assembled in the men's 800, with very few of world's key two-lap specialists missing. In many ways, it'll be a replay of the Zurich race, with all but [illegible] winner Mbulaeni Mulaudzi working for a different outcome. With a pair of recent [illegible] Fifth victories Wilfred Bungei has raced well on the Monaco track, and is anxious to put his disappointing third place Zurich finish behind him. Also not looking for a replay is Olympic champion Yuri Borzakovskiy, who was a distant eighth in Zurich, and world leader Amine Laalou, who was fourth.

In the men's longer distances, wide-open fields here are ripe for surprise. In the 1500, Alex Kipchirchir, who recently took a difficult middle distance double at the [illegible], is obviously in fantastic form. Ivan Heshko, who has raced to a pair of [illegible] titles on this track, is looking to bounce back from his personally disappointing runner-up finish at the [illegible]. In the 3000, Australian record holder Craig Mottram and European 5000 champion Jesus España will have their hands full with a full blown Kenyan attack, led by World champion [illegible]

Greek gold for Periklis Iakovakis in Gothenburg (Getty Images)

The women's steeplechase, which will cap the afternoon's events, features another clash between continental champions: European champion [illegible]

Table: 360

Athlete name: 207 Andrey Silnov
303 Silnov

Age: 208 21

Nationality: 192 Russian
5 Russia

Athlete gender: 24 men

Relates to trials: 359 2.36 High Jump

Tables

Id	Value
360	Andrey Silnov
362	Svatoslav Ton
364	Stefan Holm
366	Linus Thornblad
370	Tora Harris
372	Jesse Williams

Collection (High_Jump_all_parts), Document (newsId=35727,printer.html)

Figure 2: Screenshot of the BOEMIE text annotation tool

BOEMIE annotation tool has been extended to perform the process of annotation on html documents for the BOEMIE project. In short, the process of annotation consists of four basic steps: (a) Annotation of named entities (b) Table Filling (c) Table Linkage (d) Annotation of blocks of text corresponding to a specific topic. Each of those four steps is described with details in the sequel. It is worth mentioning that, no-textual media such as images are not represented in their original format by the annotation tool.

3.2.1 Annotation of Named Entities

For the purposes of the BOEMIE project, annotation of named entities involve the annotation of proper names as well as temporal expressions such as dates, numerical expressions denoting time and measures of distances such as meters. Additionally, alternative expressions, which are proven from the context of the web pages regarding athletics to be meaningful such as the gender of an athlete as well as other expressions that were proved to act as synonyms to possible values of specific types of named entities are also annotated. An example to this is that, the ranking position in the html pages regarding athletics is denoted not only by expressions such as “first”, “second”, “1rst”, “2nd”, but also by expression “winning” and “winner” referring explicitly to the first position.

The annotation of named entities of a specific, among the list of available collections, is performed by selecting a piece of text using the mouse and click on the appropriate button type. As a result, the color of the marked text changes according to the annotation type meaning that an annotation corresponds to an MCL instance was successfully created. Advanced annotations may be performed by: (a) the one-click selector that enables automatic mark-up of specific piece of text with one click of the mouse button (b) the creation of keyboard shortcuts that enable the association of a keyboard key with a specific type of concept property so that by pressing that key after some text is marked, the marked text will be annotated appropriately. Advanced annotation may also be performed automatically either by the use of appropriately defined regular expressions (which are defined manually by the user and preserved for future use) or the use of an appropriately learning annotation model which consists a novel functionality developed for the purposes of the BOEMIE project.

The later approach makes use of annotations created manually in previous steps. The original annotation tool was modified on order to be used in a type of a bootstrapping process during which the manual annotation performed in a part of the corpus can be used to learn a model that can identify named entity instances. The learned model is then applied to another part of the corpus in order to produce named entity annotations which can be corrected by human annotators. In this manner, the annotation time is further reduced leading to the production of named entities of high accuracy. For the purposes of the BOEMIE project, in order to learn a model for semantically segmenting a web page, the CRF++ algorithm (http://crfpp.sourceforge.net/) was chosen, due the fact that it has been proved to be a very effective framework for building probabilistic models to segment and label sequential data (Sha and Pereira, 2003). CRF++ is a simple, customizable, and open source implementation of Conditional Random Fields (CRFs) for segmenting/labeling sequential data. CRF++ has been designed for generic use, and has been applied on a variety of NLP tasks such as Named Entity Recognition, Relation Extraction and Text Chunking.

3.2.2 Table Filling

In its original version, Ellogon annotation tool supported only the annotation of named entities. In order to meet BOEMIE requirements, the Ellogon text annotation tool was easily extended to incorporate the information described into the BOEMIE ontology. More specifically, it was extended to support the annotation of relations between named entities instances (to form concept instances) in order to create HLC instances (HLCIs). HLC instances (HLCIs) are created through tables. Every HLC instance corresponds to a table having its fields filled with MLC instances.

The creation of an instance of a table, first of all, involves the selection of the desired type of table. After selecting the desired table type each of its fields are filled by selecting the desired named entity and by pressing a graphical button in order to add the annotation in the table. An alternative choice is the use of the functionality provided by the *Automatic Add* button which automatically adds into a field name the firstly occurring field name that matches its MCL type. For every table created an id number is assigned.

An important functionality provided is the visualization of the content of table instances either in compact or an extended form enabling their editing in an easy way by selecting a table from the list of created tables. Other useful functionalities regarding the manipulation of tables provided by the BOEMIE annotation tool are (a) cloning of a table which creates a duplicate of an existing table (with a different id) - in which the annotator can perform desired changes- and proves to be useful in the case of creating tables with partially common information, (b) deletion of more than one table which only requires their selection from the list of created tables using a combinations of hot keys.

3.2.3 Table Linkage

An additional extension for the purposes of the BOEMIE project is the annotation of relations between tables, i.e. HLCIs. It is worth mentioning that there is no restriction regarding the number of associations between tables but only regarding the types of tables that can be associated with, according to the HLC properties in the domain ontology, leading to a type of "hierarchical" linkage. This phase involves the linkage of the tables created in order to depict relations between ontology concepts. Tables that can be linked with others contain an additional field named "Relates to". The linkage between two tables is performed by selecting the id number (assigned to each table during its creation) to which the linkage must be performed and by pressing the corresponding graphical button. It must be stressed that the linkage between two types of tables can be performed in only one way (i.e. "Table Type 1 Relates to Table Type 2" but not also "Table Type 2 Relates to Table Type 1"). The impact of this is that the order of creating tables is very important. The types of tables that can be related are appropriately defined into the XML annotation schema.

3.2.4 Annotation of blocks of text corresponding to a specific topic

The annotation tool was also extended to meet the requirement of annotating text segments that refer to a single piece of information, such as a news item i.e. a specific topic. This additional annotation task is accomplished by exploiting both the visual and semantic information inside a web page. This is performed as a two steps process. The first step involves the exploitation of the nodes of the web page DOM tree[10] produced by the application of the VIPS (vision-based page segmentation) algorithm [11] (Cai, 2003). More specifically, at a first step, the basic units of the web page (nodes in the DOM tree) are annotated (e.g. captions and titles, columns, tables and page areas such as footers, headers, paragraphs and sentences) whereas at a second step, one or more of these units are annotated with a semantic category denoting a news item. It is worth mentioning that overlapping annotations may exist within a web page. Figure 3 depicts the use of the tool for the annotation of a web page with blocks.

3.3 Additional functionalities

The extension of the Ellogon text annotation tool involved also the addition and extension of functionalities to facilitate user's annotation task. These include the annotation comparison tool, the customization of the annotation schema and the export of annotations to OWL or XML format. Each of those functionalities is described in details below.

3.3.1 Automatic comparison tool

This functionality takes as input two identical collections and calculates the inter-annotation agreement of two separated tasks: (a) the agreement between the annotations of MLC instances i.e. named entities appearing in a collection (b) the agreement between the tables corresponding to HLC instances regarding their total number for a specific web page and their content i.e. MLC instances that each table contains. The comparison tools for both (a) and (b) calculates and displays the evaluation measures: precision, recall and f-measure. This functionality is performed via a graphical interface which allows the user to define the settings of the comparison in an easy way.

3.3.2 Customization of the annotation schema

The annotation schema is defined in a XML format and contains the definition of structures corresponding to MLCs i.e. names entities of the ontology in a form of tags. The definition of tables i.e. HLC's is performed by using the category names of named entities that each of it consists. Finally relations between tables are defined by another type of tags describing a relation to another table and it is contained into the definition of the involved table. Customization of the annotation schema is performed by simply editing the xml file containing the aforementioned definitions using a simple text editor.

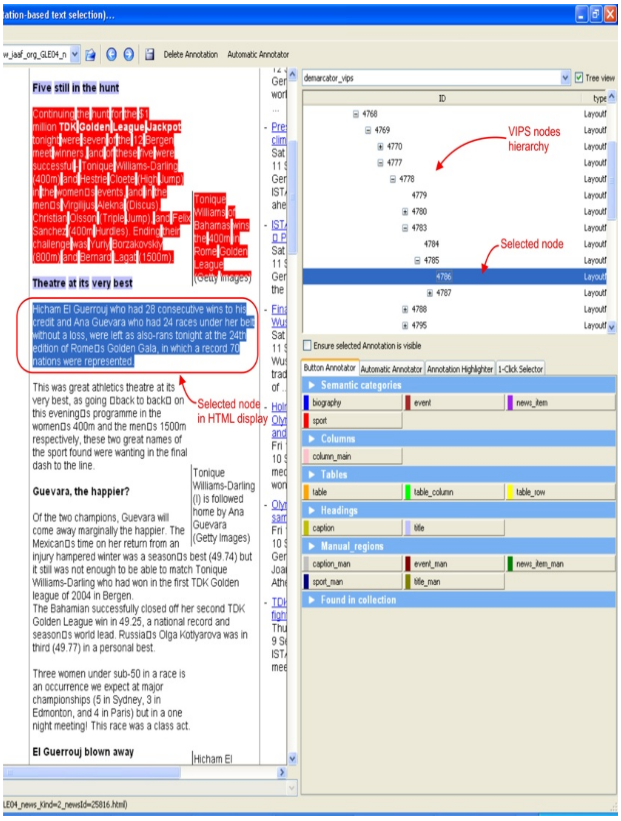
3.3.3 Export of annotations to OWL or XML format

This functionality was among those that were constructed

[10] DOM (Document Object Model) defines the logical structure of valid HTML and well formed XML documents and the way a document may be accessed and manipulated.

[11] http://www.ews.uiuc.edu/~dengcai2/VIPS/VIPS.hml

in order to meet the BOEMIE's requirements. However it has the advantage that it can be added to any system as it has no external dependencies and runs in a form of a component inside a system created in the Ellogon environment. The component takes as input an annotated document and produces an OWL ABox or XML file with respect to the concepts and relations given in the defined domain ontology.



Five still in the hunt

Continuing the hunt for the $1 million TDK Golden League Jackpot tonight were seven of the 12 Bergen meet winners, and of these five were successful – Tonique Williams-Darling (400m) and Hestrie Cloete (High Jump) in the women□s events, and in the men□s Virgilijus Alekna (Discus), Christian Olsson (Triple Jump), and Felix Sanchez (400m Hurdles). Ending their challenge was Yuriy Borzakovskiy (800m) and Bernard Lagat (1500m).

Tonique Williams of Bahamas wins the 400m in Rome Golden League (Getty Images)

Theatre at its very best

Hicham El Guerrouj who had 28 consecutive wins to his credit and Ana Guevara who had 24 races under her belt without a loss, were left as also-rans tonight at the 24th edition of Rome□s Golden Gala, in which a record 70 nations were represented.

This was great athletics theatre at its very best, as going □back to back□ on this evening□s programme in the women□s 400m and the men□s 1500m respectively, these two great names of the sport found were wanting in the final dash to the line.

Tonique Williams-Darling (l) is followed home by Ana Guevara (Getty Images)

Guevara, the happier?

Of the two champions, Guevara will come away marginally the happier. The Mexican□s time on her return from an injury hampered winter was a season□s best (49.74) but it still was not enough to be able to match Tonique Williams-Darling who had won in the first TDK Golden league of 2004 in Bergen.
The Bahamian successfully closed off her second TDK Golden League win in 49.25, a national record and season□s world lead. Russia□s Olga Kotlyarova was in third (49.77) in a personal best.

Three women under sub-50 in a race is an occurrence we expect at major championships (5 in Sydney, 3 in Edmonton, and 4 in Paris) but in a one night meeting! This race was a class act.

El Guerrouj blown away

Hicham El

Figure 3: Annotation of a single web page with blocks corresponding to a topic

4. Use of the tool in large scale annotation tasks

The BOEMIE text annotation tool was successfully used for the annotation of a corpus of a 3000 web pages collected from athletics web sites focusing on 10 different sports: Pole Vault, High Jump, Javelin, Hammer throw, Long Jump, Triple Jump, Marathon, Race Walking, 100m and 100/110m hurdles. For each of these sports 300 pages were collected. The majority of the web pages were collected from the following sites: http://www.iaaf.org, http://ww.european-athletics.org, http://ww.usatf.org/, http://www.ukathletics.net.

For every page belonging to a specific sport the annotators were asked to:

(a) annotate all the MLC i.e. instances of named entities that appear in the given pages regardless of the sport of interest;

(b) fill tables corresponding to HLC instances; and

(c) create the appropriate linkages between those tables in order to include all the information contained in the page under examination for the sport in question.

Tables were created in order to gather all the information related to the event(s) described in the pages, the sport in question, the various rounds of the sport performed and the athletes participating in those (i.e. the personal information provided as well as the information of their performance and ranking).

The average number of words per page was 1220, the average number of annotated named entities per page was 229, and the average number of created tables per page was 23,. The pre-annotation of web pages with named entities, resulted from the application of a machine learning based algorithm (CRF), proved to be beneficial for annotators. The time required for the annotation of a whole page, was calculated on average to be 30 minutes. The pre-annotation of web pages led approximately to one third reduction of the annotation time, although there were cases where it was less successful. This was observed when it was applied to a new domain i.e. a sport belonging to a different category (running, throwing, jumping) than the one(s) already being annotated. In such cases there were several errors which the annotators had to correct. But even in those cases, the annotation time was not higher compared to a page containing no annotations at all.

Regarding the second requirement of BOEMIE, involving the annotation of the blocks referring to a single topic, the dedicated Ellogon component constructed for this task was used in order to perform annotation on a subset of the 3000 web pages corpus. Most specifically it was performed on a corpus containing 100 web pages taken from eight different web sites, which are the following: IAAF (www.iaaf.org), USA Track & Field (http://www.usatf.org/), BBC (news.bbc.co.uk/), www.sportinglife.com/, www.scc-events.com/ and http://sportsofworld.com/. In all web pages manually annotation of the boundaries of the blocks of interest i.e. the "news items" was performed. In this case, the annotation of a single page varied from 20 to 30 minutes.

The BOEMIE annotation tool was also applied in a second annotation task for the purposes of the MedIEQ project (http://www.medieq.org) in a total of 503 web pages of medical content in English and Spanish, where the focus was the annotation of contact information of a person or organisation as well as the filling of the corresponding contact tables. Regarding the English part of the MEDIEQ corpus, this contained 147 texts, where the average number of words per page was 1234, the average number of annotated named entities per page was 43, and the average number of created tables per page was 7. Regarding the Spanish part of the MEDIEQ corpus, this contained 342 Spanish texts, where the average number of words per page was 225, and the average number of annotated named entities per page was 13 (this task did not involve the creation of tables). This annotation task will also be performed in corpora for five more languages in the context of the MedIEQ project.

5. Conclusion – Future Work

In this paper we presented the BOEMIE ontology based annotation tool which is able to annotate named entity instances that correspond to specific types of concepts (middle level concepts - MLCs) defined in the domain ontology, fill tables corresponding to ontology concepts (high level concepts – HLCs) with the annotated named entities and link the filled tables based on relations between HLCs defined in the domain ontology. Additionally, it can perform annotation of blocks of text that correspond to different news items. BOEMIE text annotation tool was proved extremely user friendly and robust in the two large scale annotation tasks it was recently used. Its functionalities have proved to be beneficial to the annotators especially in the case of the BOEMIE annotation task taking into account the complexity of the task as well as the size of the corpus. The automatic annotation of named entities proved to be very helpful.

In the future we aim to examine ways to perform (semi) automatic annotation of HLC instances i.e. table instances as well as to facilitate further the annotation of blocks corresponding to news items.

6. References

Cai, D., Yu, S., Wen, J-R., Ma, W-Y. (2003). VIPS: a Vision-based Page Segmentation Algorithm. Microsoft Technical Report (MSR-TR-2003-79).

Corcho O. (2006). Ontology based document annotation: trends and open research problems. *Int. J. Metadata, Semantics and Ontologies, 1(1),* pp.47--57.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Fernández, M., Vallet, D., Castells, P. (2005). Automatic Annotation and Semantic Search from Protégé. Demo at the 8th International Protégé Conference, Madrid, Spain.

Müller, C. and Strube, M. (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In *Sabine Braun, Kurt Kohn, and Joybrato Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. English Corpus Linguistics*, 3, pp: 197--214.

Ogren, P. V. (2006). Knowtator: A Protégé plug-in for annotated corpus construction. *Human Language Technology Conference Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp: 273—275.

Petasis, G., Karkaletsis, V., Paliouras, G., Spyropoulos, C. D. (2003). Using the Ellogon Natural Language Engineering Infrastructure. *In Proceedings of the Workshop on Balkan Language Resurces and Tools, 1st Balkan Conference in Informatics (BCI 2003)*, Thessaloniki, Greece.

Sha F., Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. In *Proceedings of HLT-NAACL 2003*, Edmonton, Canada, pp: 213--220.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna F. (2005). Semantic annotation for knowledge management: Requirements and a survey of the state of the art., *Journal of Web Semantics,* pp. 14--28.

Zhang, C., Du, J., Zhang, R., Fan, X., Yuan, Y., Ning, T. (2007). Extracting Information of Anti-AIDS Inhibitor from the Biological Literature Based on Ontology. *Frontiers in Algorithmics, Book Series Lecture Notes in Computer Science, Springer,* 4613, pp: 74—83