МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2007
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2007
Proceedings of the Thirty Sixth Spring Conference of
the Union of Bulgarian Mathematicians
St. Konstantin & Elena resort, Varna, April 2–6, 2007

THE SEMANTIC ANNOTATION TODAY AND SOME CHALLENGES FROM THE PAST

Anna Devreni-Koutsouki

The paper presents a survey of current semantic annotation platforms that can be used to perform semi-automatic annotation. The platforms vary in their architecture, information extraction tools and methods, initial ontology, amount of manual work required to perform annotation, performance and other features. The platforms for semi-automatic annotation of texts are considered and assessed according to a number of various scientific, technical and application/practical requirements. The current research and existing platforms and tools focus on the semantic annotation of already existing or currently-created texts in a definite format. The paper gives a proof of that it is necessary to develop a tool for annotation and intelligent search in the expanding repository of digital copies of materials of the State Archive Fund (SAF) of the Republic of Bulgaria.

1. Introduction. The Semantic Web community refers to semantic annotation as (i) a sort of meta-data and (ii) the process of generation of such meta-data.

Computing knowledge by using mark-up techniques and by supporting semantic annotation is a major technique for creating metadata. It is beneficial in a wide range of content-oriented intelligent applications. One important application of this type is the Semantic Web. The research about the WWW currently strives to augment syntactic information already present in the Web by semantic metadata.

Full implementation of the Semantic Web requires widespread availability of semantic annotations for existing and new documents on the Web. Manual annotation is more easily accomplished today, but it has lead to a knowledge acquisition bottleneck.

To overcome this bottleneck, semiautomatic annotation of documents has been proposed. Semiautomatic means, as opposed to completely automatic, are required because it is not yet possible automatically to identify and classify all entities in source documents with complete accuracy.

The platforms for semi-automatic annotation of texts are considered and assessed according to a number of various scientific, technical and application/practical requirements. There are multiple developed schemes, consistent with some of them, but there is not yet a completely integrated environment consistent with all these requirements.

There has been a standing issue coming from the past – the problem related to the storage and access provision to already created materials, which were not designed for computer processing.

We believe that from a practical and scientific point of view, it is necessary to develop a tool for annotation and intelligent search in the expanding repository of digital copies of materials of the State Archive Fund (SAF) of the Republic of Bulgaria.

The rest of the paper is organized as follows. Section 2 presents the classification of semantic annotation platforms. Section 3 describes platform overview. Section 4 presents some significant requirements towards semantic annotation platforms. Section 5 describes an evaluation of the platforms that are presented in section 3, according to the criteria in section 4. Section 6 focuses on the problem related to the storage and access provision to already created materials, which were not designed for computer processing. Section 7 concludes this paper.

2. Platform classification. Semantic annotation platforms (SAP) can be classified according to the type of annotation method used. There are two primary categories: Pattern-based and Machine Learning-based, as it is shown in Figure 1 [10]. In addition, platforms can use methods from both types of categories, called Multistrategy.

Pattern-based SAPs can perform pattern discovery or have patterns manually defined.

Machine learning-based SAPs utilize two methods: probabilistic and induction.

Multistrategy SAPs are able to combine methods from both pattern-based and machine learning-based systems when they are designed with extensible architectures.

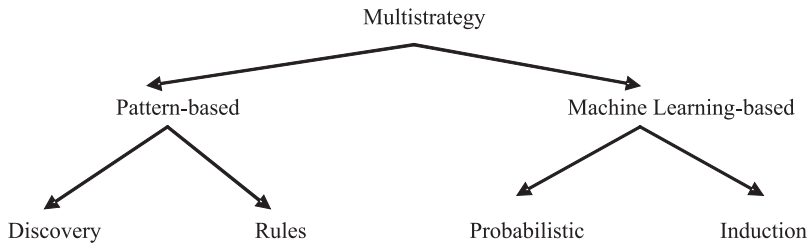

Fig. 1. Classification of SAPs

3. Platform overview. Semantic annotation platforms provide support for information extraction (IE) implementations, ontology and knowledge base management, access APIs, storage (e.g., RDF [12] repositories), and user interfaces for ontology and knowledge base editors [1]. Platforms may include only a subset of these features, and may include other features not generally included by all SAPs, such as annotation storage.

3.1. Armadillo. Armadillo [1] is a system for unsupervised creation of knowledge bases from large repositories (e.g. the Web) as well as document annotation. Armadillo uses the Amilcare IE system to perform wrapper induction on web pages to mine web sites that have a highly regular structure. Armadillo uses a pattern-based approach to find entities. Information redundancy, *via* queries to Web services such as Google and CiteSeer, is used to verify discovered entities by analyzing query results to confirm or deny the existence of an entity, similar to the way the PANKOW algorithm [7] operates.

3.2. KIM. The Knowledge and Information Management (KIM) platform [2] contains an ontology, a knowledgebase, a semantic annotation tool, an indexing and retrieval server, as well as tools for interfacing with the server. KIM uses information extraction techniques to build large knowledge base of annotations. The annotations in KIM are metadata in the form of named entities (people, places, etc.) which are defined in the KIMO ontology and identified mainly from reference to extremely large gazetteers.

The information extraction component of semantic annotation is performed using components of the GATE toolkit [4]. Some components of GATE have been modified to support the KIM server.

3.3. Ont-O-Mat/Amilcare. Ont-O-Mat [6] is an implementation of the S-CREAM (Semiautomatic CREAtion of Metadata) semantic annotation framework. The IE component is based on Amilcare.

Amilcare is a Supervised IE system. It learns how to recognize the objects that require annotation by learning from a collection of previously annotated documents.

Amilcare uses the ANNIE ("A Nearly-New IE system") part of the GATE toolkit to perform IE. The result of ANNIE processing is passed to Amilcare which then induces rules for IE.

Ont-O-Mat provides an extensible architecture to replace selected components.

3.4. SemTag. SemTag [8] is another example of a tool which focuses only on automatic mark-up. It is based on IBM's text analysis platform Seeker and uses similarity functions to recognize entities which occur in contexts similar to marked up examples. The key problem of large-scale automatic mark-up is identified as ambiguity. A Taxonomy Based Disambiguation (TBD) algorithm is proposed to tackle this problem. The annotations generated by SemTag are stored separate from the source document.

4. Requirements. The platforms for semi-automatic annotation of texts are considered and assessed according to a number of various scientific, technical and application/ practical requirements. Some of the most important include the following [9]:

4.1. Standard formats. Using standard formats is preferred, wherever possible. For annotation systems, in particular, standards can provide a bridging mechanism that allows heterogeneous resources to be accessed simultaneously and collaborating users and organizations to share annotations. Two types of standard are required: standards for describing ontologies such as the Web Ontology Language OWL [11] and standards for annotations such as the W3C's RDF annotation schema [12].

4.2. User centred/collaborative design. An ideal semantic annotation system would use a single point of entry approach in which annotation functionality, including access to maintain the underlying ontologies would be seamlessly integrated with other tools routinely used by knowledge workers to author and read documents. This does not yet exist although there are signs of a trend towards integrated authoring environments, such as WickOffice [16] and AktiveDoc [17].

4.3. Ontology support (multiple ontologies and evolution). Annotation tools have adapted rapidly to recent changes in ontology standards for the Web, with many of the more recent tools already supporting OWL. Ontology maintenance, which directly affects the maintenance of annotations, is poorly supported, or not supported at all, by the current generation of tools. However, there are signs that annotation systems are giving users more control of ontologies. Much more is still required. A genuinely

integrated semantic annotation environment should give the user automatic support for ontology maintenance.

4.4. Support of heterogeneous document formats. Satisfying this requirement is a prerequisite for producing integrated annotation environments and our survey suggests that the range of document types that can be handled is expanding, though few individual systems handle many different formats. Most of the annotation tools, we looked at supported only HTML and XML. WickOffice and OntoOffice [18], provide annotation for word processor files. Mangrove [19] and SMORE [20] provide facilities for handling emails. SMORE, Vannotea [21] and M-OntoMat-Annotizer provide means to annotate images and image regions.

4.5. Document evolution (document and annotation consistency). We believe that keeping annotations synchronized with changes to documents is challenging and this is one area in which the current annotation standards are inadequate.

5. Evaluation and comparative analysis of the platforms. The Table 1 describes an evaluation and comparative analysis of the platforms that are presented in section 3, according to the criteria that we saw in section 4.

Annotation tool	Standard formats	User-centred design	Ontology support	Document formats	Document evolution
Armadillo	RDF(S)	—	—	HTML	—
KIM	RDF(S), OWL	Various plug-in front ends	KIMO	HTML	—
OntoMat	DAML+OIL, OWL, SQL, XPionter	Drag & drop, create & annotate	OntoBroker annotation inference server	HTML, Deep Web	XPointer, pattern
SemTag	RDF(S)	—	—	HTML	—

Table 1. Comparison of annotation tools for requirements 1–5

6. Some challenges from the past. There has been a standing issue coming from the past – the problem related to the storage and access provision to already created materials, which were not designed for computer processing.

We envisage the vast amount of documents, forms, protocols, letters/correspondence, pictures, maps, images and other objects which can be found in private or public museum collections or in state, local or personal archives. They are a part of the cultural and historical heritage of humanity, in general, of a given country or region, in particular. Usually they are unique, more or less expensive, and they are stored under a special regime of protection. On the other hand, the interest in them is very big and comes from different directions – tourists, students, non-professionals, experts and research workers from the various fields of art and science, personally involved people etc. All this makes direct physical access inexpedient and in some occasions even impossible.

The national strategy of many countries, including private institutions, which possess such collections and archives, is making them widely-spread and accessible. The common

practice is the creation of repositories of images or digital copies which can already be accessed through the Web [13]. Each collection usually has its own (semi-) structured indexing scheme that typically supports a keyword-type search. However, finding the right image is often still problematic [14].

Over the past few years, various approaches have been proposed to effectively and manage digital image content on the Web. Traditionally, these have included techniques such as building keyword indices based on image content, embedding keyword-based labels into images, analyzing text immediately surrounding images on Web pages, etc. More recently, there is a research focus to develop techniques to annotate the content of images on the Semantic Web, using languages such as RDFS and OWL [15].

7. Conclusion. In this paper a short survey of semantic annotation platforms and a classification of them were presented. Semantic annotation platforms (SAPs) can be distinguished primarily by their annotation method.

In addition some requirements were developed. There are multiple developed systems consistent with some of them but there is not yet a completely integrated environment which can handle all of them.

The current research and existing platforms and tools focus on the semantic annotation of already existing or currently-created texts in a definite format, for example HTML, XML or word processor files.

We believe that it is necessary to develop a tool for annotation and intelligent search in the expanding repository of digital copies of materials of the State Archive Fund (SAF) of the Republic of Bulgaria.

The initial materials in CAD are stored in special repositories and take considerable amount of space. The documents of the State Archive Fund have been used in reading-rooms, classified in each archive and they are not allowed to leave this archive. The search for information in the documental data is performed with the help of a system of archive reference books. Part of this information is filed in computers [22].

From a practical point of view one tool for annotation and intelligent search in repositories of digital copies of these materials would solve a number of problems of different character such as difficult and slow access, storage of originals under suitable conditions and other items already mentioned in section 5.

From a scientific point of view such an environment would represent a new type of integrated means. It is expected to perform different classification, annotation and search in the base of the general description of the archive data and in accordance with the type of particular entities (letters, maps, images, forms etc.). Apart from that it should be stressed out that there are other non-trivial challenges. Some of the most significant include the following:

- Materials of different content (letters, protocols, pictures etc.) are presented as images or in PDF-format;
- Each material should be classified in accordance to its type so it may be properly processed;
- There are multiple heterogeneous materials whose individual components should be processed as required;

- Coordination and synchronization of the heterogeneous methods for annotation and search is needed;
- All available materials are in several different languages;
- In case of work with ready ontologies the problem related to the evolution of languages should be considered.

REFERENCES

[1] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, M. Goranov. KIM – Semantic Annotation Platform. In: 2nd International Semantic Web Conference (ISWC2003), Florida, USA, 2003, 834–849.

[2] A. Dingli, F. Ciravegna, Y. Wilks. Automatic Semantic Annotation using Unsupervised Information Extraction and Integration. In: K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation, 2003.

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

[4] M. Dowman, V. Tablan, H. Cunningham, B. Popov. Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News. In: Proc. of the 14th International World Wide Web Conference. Chiba, Japan, 2005.

[5] S. Handschuh, S. Staab, F. Ciravogna. S-CREAM – Semi-automatic CREAtion of Metadata. In: SAAKM 2002 – Semantic Authoring, Annotation & Knowledge Markup – Preliminary Workshop Programme, 2002.

[6] P. Cimiano, S. Handschuh, S. Staab. Towards the Self-Annotating Web. In: Thirteenth International Conference on World Wide Web, 2004, 462–471.

[7] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, **4**, No 1 (2006), 14–28.

[8] L. Reeve, H. Han. Survey of Semantic Annotation Platforms. In: SAC'05, March 13–17, 2005, Santa Fe, New Mexico, USA.

[9] M. Smith, C. Welty, D. McGuinness. OWL Web Ontology Language Guide. `http://www.w3.org/TR/owl-guide/`, last accessed May 31, 2006.

[10] W3C, Resource Description Framework (RDF). `http://www.w3.org/RDF/`, last accessed May 31, 2006.

[11] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, S. Kettula. MuseumFinland—Finnish museums on the semantic web. In: 3rd International Semantic Web Conference (ISWC2004), Hiroshima, Japan, 07-11 November 2004.

[12] L. Hollink, G. Schreiber, J. Wielemaker, B. Wielinga. Semantic Annotation of Image Collections. In: Knowledge Capture – Knowledge Markup & Semantic Annotation Workshop, 2003.

[13] C. Halaschek-Wiener, J. Golbeck, A. Schain, M. Grove, B. Parsia, J. Hendler. Annotation and Provenance Tracking in Semantic Web Photo Libraries.

[14] T. Miles-Board, A. Woukeu. Demo Abstract: Bringing the Semantic Web to the Office Desktop, `http://eprints.ecs.soton.ac.uk/12237/01/p229-miles-board_-_doceng05.pdf`, last accessed July 3, 2006.

[15] `http://nlp.shef.ac.uk/wig/aktivedoc.htm`, last accessed July 3, 2006.

[16] OntoOffice Tutorial. www.ontoprise.de/documents/tutorial_ontooffice.pdf, last accessed July 3, 2006.
[17] L. McDowell, O. Etzioni, S. Gribble, A. Halevy, H. Levy, W. Pentney, D. Verma, S. Vlasseva. Enticing ordinary people onto the Semantic Web *via* instant gratification. In: Proceedings of the 2nd International Semantic Web Conference (ISWC 2003), October, 2003.
[18] SMORE: Semantic Markup, Ontology and RDF Editor.
http://www.mindswap.org/2005/SMORE/, last accessed July 3, 2006.
[19] http://www.itee.uq.edu.au/∼eresearch/projects/vannotea/index.html, last accessed July 3, 2006.
[20] http://www.archives.government.bg/dostup.html, last accessed July 3, 2006.

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, J. Bouchier Str.
1164 Sofia, Bulgaria
e-mail: annadevreni@hotmail.com

СЕМАНТИЧНОТО АНОТИРАНЕ ДНЕС И НЯКОИ ПРЕДИЗВИКАТЕЛСТВА ОТ МИНАЛОТО

Анна Деврени-Куцуки

Материалът представя изследване на съвременните платформи за семантично анотиране, които могат да се използват за извършване на полуавтоматично анотиране. Тези платформи се различават по тяхната структура, методи и средства за извличане на информацията, начална онтология, обема на ръчния труд необходим за извършване на анотирането, действие и други. Платформите за полуавтоматично анотиране на текстове се разглеждат и оценяват в зависимост от редица научни, технически и приложни/ практически изисквания. Настоящите изследвания и съществуващите платформи и инструментални средства се съсредоточават върху семантичното анотиране на съществуващи или създавани в момента текстове в определен формат. Считаме, че е на лице необходимостта от разработването на инструментално средство за анотиране и интелигентно търсене в разрастващото се хранилище от дигитални копия на материали от Държавния архивен фонд (ДАФ) на република България.