Advanced Features of Collaborative Semantic Annotators – the 4A System

Pavel Smrz and Jaroslav Dytrych
Brno University of Technology, Faculty of Information Technology, IT4Innovations Centre of Excellence
Bozetechova 2, 612 66 Brno, Czech Republic
Email: {smrz,idytrych}@fit.vutbr.cz

Abstract

This paper deals with collaborative knowledge engineering, particularly focusing on collective editing and semantic annotation of hypertext. It discusses state-of-the-art functions of the 4A (Annotations Anywhere, Annotations Anytime) system that has been recently extended to be applicable in a broad range of annotation contexts. We introduce advanced features and recent improvements that make the tool unique in many aspects. A special attention is paid to the social way of semantic tagging – complex annotations can be created by a single click and immediately shared with other interested users or reused by external systems. We also compare the 4A system to similar software solutions and show their similarities and differences.

Introduction

Despite many efforts in formalising knowledge, a vast majority of textual content on the current Web takes form of natural language sentences with no explicit semantics. It is not easy to make this content understandable by machines and thus to realize the original vision of the Semantic Web to its full extent. Moreover, unless there is a very simple, generally accepted mechanism that would allow laymen to better express meaning and that would bring benefits immediately, the situation will not change in near future.

Information extraction from text offers a solution for this problem. However, cutting-edge text mining systems are limited to a narrow set of cases they were trained for. In other cases, there is still a need for manual semantic metadata creation. This paper discusses how the process of manual text annotation can be assisted by an intuitive user interface that presents annotation suggestions generated by an external semantic enrichment system.

To motivate particular functions, we use examples from the cultural heritage domain. It corresponds to the application area of the DECIPHER project[1] in which the tool was employed. DECIPHER was an EC FP7 project aiming to support discovery and exploration of cultural heritage through story and narrative. Semantic enrichment of underlying texts brought new quality to the whole range of narrative construction, knowledge visualisation and display for museum professionals involved in the project. Yet, the 4A framework is generally applicable in other knowledge engineering contexts, e.g., for biomedical text annotation.

There are three particular cases in which the preferable text mining scenario cannot be (fully) applied. First, a variability of natural language constructs to express a semantic relation can be high and there can be insufficient data to train a machine learning model. For example, relations of artistic influences (among artists, artworks, themes, styles, techniques, places, etc.) have been studied within the DECIPHER project and it showed up that despite the effort, expressions such as *pays tribute/homage to* are not well covered by the resulting system. Although various bootstrapping approaches on web-scale data provide a help (Zhu et al. 2009), at least an initial seed of examples needs to be provided by users. Manual annotation serves then clearly as a source of more training data and generally improves performance of automatic annotation procedures.

Second, the structure of knowledge (a template to be filled in by an automatic method) can be complex and natural language processing and machine learning techniques can be unable to deal with it. In the cultural heritage domain, a typical example is a knowledge scheme analysing different attitudes to an artist and his or her work. Many books can be written about the topic, people can have opposite meanings and it is very difficult for automatic methods to generalize in such situations. As complex structures often consist of sub-components that can be recognized automatically, annotation suggestions can significantly speed up the process of the semantic enrichment of text in this case.

Finally, the knowledge structure itself can be unclear, not well understood, or fuzzy. When annotating particular pieces of relevant texts, users often become aware of a general semantic pattern of knowledge represented by the texts, they better realize what attributes are crucial for a task in hand and can easily draft a knowledge scheme that reflects their specific needs. This aspect showed many times in DECIPHER. Although we maximally re-used existing knowledge resources such as well-established ontologies and conceptual hierarchies (CIDOC CRM, Getty Thesaurus, etc.), many tasks required specific knowledge structures that had to be created "on the fly". One cannot expect that ordinary end-users will adopt sophisticated ontology tools such as Protégé to suggest specific additions to knowledge specifi-

[1] http://www.decipher-research.eu/

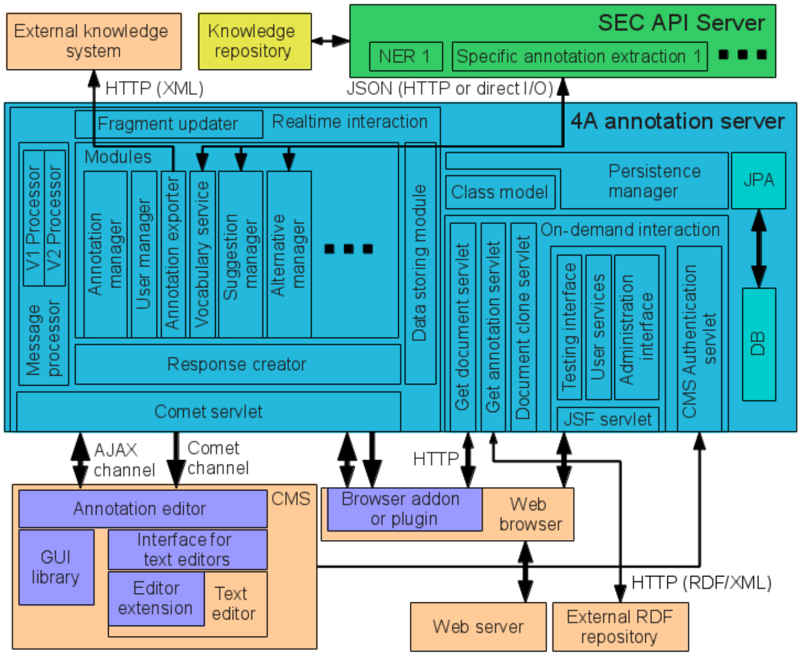


Figure 1: Architecture of the 4A system

cation schemata. It is thus beneficial if an annotation tool is able to play this role too.

The 4A tool reflects the needs discussed above. It was introduced in a workshop paper in 2011 (Smrz and Dytrych 2011). Since then, the system has been significantly improved and extended. Its new modular design can cope with diverse requirements defined by various user groups and particular annotating contexts. The paper presents the current version 2.0 of the system.

Advanced Features of the 4A System

The 4A system consists of a back-end server generating annotation suggestions, an annotation server managing semantic annotations and clients presenting the suggestions and annotations to users, transferring their feedbacks and visualising results of the annotation. A general architectural schema is presented in Figure 1.

The stateless SEC API Server invokes semantic enrichment components – named entity recognizers, co-reference resolution packages, and specific relation extractors – that pre-annotate an input text and mark potential occurrences of relevant entities and relations in it. The SEC API Server can run as a remote service or it can be started and managed by the 4A Annotation server. Variable deployment models are supported. For example, the system can be configured so that a local SEC API Server is accessed by many remote annotation servers.

Two kinds of clients access the system. The first group of clients is intended for annotating any web page or an existing document viewed in a browser. A client realized as an add-on for Mozilla Firefox is currently available, MS Internet Explorer, Opera and Google Chrome versions are being developed. To edit and annotate a new text at the same time, the Annotation Editor for JavaScript WYSIWYG editors can be employed. An extension for TinyMCE[2] is currently available. Other editors can be easily supported through a defined abstraction level. This form of the client provides also a way to integrate the tool into popular content management systems (such as Drupal in the case of DECIPHER).

An essential feature of the 4A system lies in anchoring annotations in text. Although resulting knowledge structures (e.g., in RDF) can be attached to the whole document, it is useful to associate semantic interpretation directly with particular words, phrases, sentences, paragraphs or any other piece of text (referred to as textual fragments in this paper). The fine-grained annotation is critical for further processing of semantically enriched data and accountability of results. Moreover, pinpointing a source of information helps machine learning methods to infer better models from annotated examples.

A text can come into existence and be immediately annotated. Texts and annotations are processed separately to support interleaved editing and annotating. A sophisticated annotation management guarantees that most of annotations remain valid after each text editing step and only disqualified annotations are thrown away. The 4A server takes care of annotation updates. To find the best match between a stored and an edited version of an annotated text, a cascade of methods with variable sensitivity is applied. A current node in a hierarchical representation of the text is searched forward and backward first. If no match is found, the content is searched from the current node to other nodes. A fragment in

[2] http://www.tinymce.com/

an edited text is taken as matching if specific criteria are met (a threshold on the Levenshtein distance, correspondence of first and last letters, etc.)

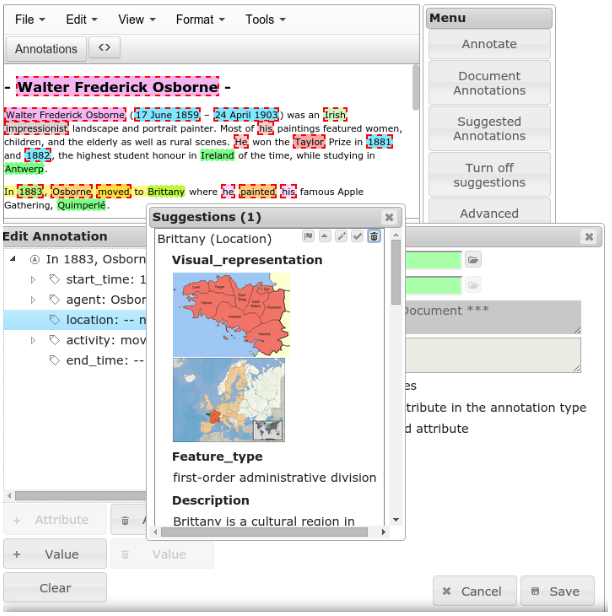
Although annotation structures can be complex, accepting or rejecting suggestions need to be very easy – it should correspond to a single click in most cases. As automatic methods never recognize all potential entities, there also needs to be a simple mechanism to add new entries into an underlying knowledge base and to maximally reuse relevant existing content. To realize this, the 4A system employs semantic templates that lead the user through the annotation process. For example, when annotating a text corresponding to creating an artwork, the system displays common attributes from the domain ontology model CIDOC CRM. When clicking on attribute Author which is known to be typically filled by a URI corresponding to a person, the system suggests primarily fragments that fit this semantic preference. Semantic templates, derived from ontologies in the initialization phase, effectively get users over complexities of existing knowledge structures – concepts are suggested as semantic types of attributes, constraints are transformed into template structures and existing annotations are used to improve the results. Any use of an attribute is also linked back to the original ontology. Thus, suggested changes in knowledge structures can be immediately supported by real world examples.

Annotations can link to other annotations that exist either independently or as a part of a superordinate annotation. The 4A system supports unlimited nesting of annotations. For example, an annotation referring to an event can include annotations of complex attributes and, at the same time, form a part of another annotation expressing a cause relation between two events, which is further attributed to a belief state of a person.

There is a non-trivial support for overlapping and interleaving annotations in the 4A system. To preserve well-formedness rules of XML documents and other hierarchical formats, the 4A system automatically splits fragments on "seams" and joins the parts again when representing annotation results. For example, the advanced mechanism enables annotating a sentence *Renoir and Manet made copies of Delacroix' paintings* by two separate events with different actors (*Renoir* and *Manet*) and shared textual fragments representing the verb phrase and the object of the relations.

Annotations are displayed in popups on moving the mouse over a relevant fragment. Links to other annotations can be clicked on and any content of linked and nested annotations can be displayed directly in a particular annotation popup too. Document-level annotations are displayed in a separate window.

Suggested annotations can be accepted one by one (either directly in the text – see Figure 2 – or in a separate window with a list of all suggestions for a quick review). The system can be also set to confirm all suggestions with a confidence value higher than a specified threshold. Users can further manually reject annotations confirmed in a previous step. Low confidence suggestions can be filtered out according to user's preference. The user feedback is stored and used for improving automatic suggestions.



- Walter Frederick Osborne -

Walter Frederick Osborne (17 June 1859 – 24 April 1903) was an Irish impressionist landscape and portrait painter. Most of his paintings featured women, children, and the elderly as well as rural scenes. He won the Taylor Prize in 1881 and 1882, the highest student honour in Ireland of the time, while studying in Antwerp.

In 1883, Osborne moved to Brittany where he painted his famous Apple Gathering, Quimperlé.

Brittany (Location)

Visual_representation

Feature_type

first-order administrative division

Description

Brittany is a cultural region in

In 1883, Osborn
start_time: 1
agent: Osbor
location: -- n
activity: mov
end_time: --

ocument ***

es
tribute in the annotation type
d attribute

Figure 2: An annotation suggestion

To prevent problems related to concurrent work of users on the same document, 4A Annotation editors and the server employ a real-time protocol and send changes immediately to all involved parties. If a new user opens a document being edited and/or annotated by other users, the system offers applying changes of others or start with a separate version of the text.

To cope with new requirements that were not envisaged in the beginning of the tool development, various enhancements of the 4A system have been recently realized. The key change is the move from a proprietary annotation format to the W3C Open annotation format[3]. This guarantees interoperability with other annotation systems and enables using external RDF reasoners to gain additional information.

Accuracy of generated annotation suggestions varies, it can be very high for some semantic types and very low for others. If the suggested annotation is not correct, the 4A framework enables creating a correct annotation manually. However, this presents a tedious task involving the search for a correct entity link and editing of additional attributes. To simplify this process, we developed a new module of the annotation server which manages alternative annotations for a given textual fragment. If the user rejects a current (best) suggestion, the system shows all other known alternatives for a particular textual fragment and the user can choose one by a single click again.

The annotation subscription mechanism has been also extended to allow fine grained setting of filters defining which annotations shared by other users should be shown. It is possible to create sets of filters with annotation sources and semantic types and then apply a given set in a particular in-

[3] http://www.openannotation.org/spec/core/

stance of the editor. This mechanism enables opening a document in more tabs and showing different kinds of annotations (different perspectives) in individual tabs.

An advanced scheme generating modification sequence numbers and server side conflict checking and approving of modifications have been also added. These mechanisms minimize delays between modifications in multi-user settings. It is not necessary to apply a change on all clients before a next modification request is processed (as, for example, Google Documents need to do). The system guarantees that conflicting changes will be performed in a right order but non-interfering changes can proceed immediately. More instances of the 4A editor can also newly appear at the same web page.

Related Work

The RDFaCE content editor[4] (Khalili, Auer, and Hladky 2012) is the most similar tool from the user interface perspective. It is implemented as a TinyMCE plugin and enables annotating textual fragments and linking them to ontology concepts. Both the systems allow users to create simple tags as well as complex annotations with attributes. However, RDFaCE does not support real-time collaboration among users. While the 4A system synchronizes both the textual content of a document and its annotations, RDFaCE needs to store the annotated text before others can annotate it.

Pundit[5] (Grassi et al. 2013) is an annotation tool that stores annotations on the server side in a similar way as the 4A framework. The system also works with textual fragments. It employs ontologies and controlled vocabularies. On the other hand, the frontend tool is implemented as a bookmarklet so that the text being annotated cannot be modified during the annotation process. Simple attributes can be of a plain text type only and links to ontology concepts need to be entered directly as RDF triples. This slows down the annotation process and makes it less comfortable than in the case of 4A or RDFaCE. Pundit does not allow adding new attributes and nesting annotations. As opposed to RDFaCE, however, it is possible to create links between annotations in Pundit.

Storing annotation in the Open Annotation format on the server is a feature that the 4A system shares with Domeo[6] (Ciccarese, Ocana, and Clark 2012). The frontend takes form of a browser plugin which is also one of the 4A client types. Domeo excels in the support for work with images – a quality no other semantic annotation tool currently matches. It enables creating simple textual annotations as well as linking fragments to ontology concepts. Adding attributes is complicated as the tool dedicates attribute manipulation to external plugins. Compared to the 4A direct search for a term in a controlled vocabulary by autocomplete functions with previews, Domeo supports the functionality by a simple search field in a separate tab. It also displays annotations differently – in a sidebar rather than together with annotated fragments as other tools do.

Conclusions and Future Directions

The 4A system was successfully deployed in the DECIPHER project. It is currently used for various annotation tasks in the cultural heritage domain. Museum professionals appreciate advanced functionality of the tool and introduce it to new environments of their interest.

Although the 4A systems overcomes other annotation tools in various aspects, there are still many enhancements that wait for integration to next versions. We are currently working on advanced export functions for knowledge structure extensions proposed by users and on improving interoperability with other tools. Annotation of images will be also introduced. Finally, the 4A system will newly support a broad range of WYSIWYG editors including Aloha[7] and CKEditor[8] which will help to apply the tool in new settings.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under the IT4Innovations Centre of Excellence project, Registration number CZ.1.05/1.1.00/02. 0070, supported by Operational Programme – Research and Development for Innovations – funded by Structural Funds of the European Union and the state budget of the Czech Republic.

References

Ciccarese, P.; Ocana, M.; and Clark, T. 2012. Open semantic annotation of scientific publications using DOMEO. *Journal of Biomedical Semantics* 3(Suppl 1). `http://www.jbiomedsem.com/content/3/S1/S1`.

Grassi, M.; Morbidoni, C.; Nucci, M.; Fonda, S.; and Donato, F. D. 2013. Pundit: Creating, exploring and consuming semantic annotations. In *Proceedings of the 3nd International Workshop on Semantic Digital Archives, Valletta, Malta, September 26, 2013*.

Khalili, A.; Auer, S.; and Hladky, D. 2012. The RDFa Content Editor – From WYSIWYG to WYSIWYM. In *Proceedings of COMPSAC 2012 – Trustworthy Software Systems for the Digital Society*. `http://svn.aksw.org/papers/2012/COMPSAC_RDFaCE/public.pdf`.

Smrz, P., and Dytrych, J. 2011. Towards new scholarly communication: A case study of the 4a framework. In *SePublica*, volume 721 of *CEUR Workshop Proceedings*.

Zhu, J.; Nie, Z.; Liu, X.; Zhang, B.; and Wen, J.-R. 2009. StatSnowball: A statistical approach to extracting entity relationships. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 101–110. New York, NY, USA: ACM.

[4] `http://rdface.aksw.org/`

[5] `http://www.thepund.it/`

[6] `http://swan.mindinformatics.org/`

[7] `http://aloha-editor.org`

[8] `http://ckeditor.com/`