VISA - Corpus Annotation with OWL [*]

Stephanie Becker and **Thomas Kleinbauer** and **Stephan Lesch**
DFKI Gmbh
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
<firstname.lastname>@dfki.de

Abstract

We present VISA, a graphical annotation tool for OWL-based annotation schemes with a focus on generality and usability.

1 Introduction

The W3C standard OWL was originally designed as an ontology language for the semantic web, but it is progressively finding its way into various other fields of application. Annotated (linguistic) corpora, on the other hand, still often rely on their own specific data storage formats, although newer developments show a trend towards the use of XML (Carletta et al., 2005).

We believe that OWL is a suitable format for future corpora and annotations thereof, as it provides a semantically potent language based on a simple and open format. The main advantage is that further processing of corpus data can make use of automatic inference mechanisms, working only on one underlying formalism for all annotations. Existing annotation schemes can easily be expressed in OWL; annotation then becomes a process of assigning instances of ontology classes to corpus segments.

A number of tools specialized for different kind of annotations exist, as well as programs for working with OWL data. However, the number of tools for annotating OWL ontologies is rather small. One way to build such tools is to combine existing software for annotation and for OWL – a procedure taken for instance by (Bontcheva et al., 2004) or (Lauer et al., 2005) which both integrate the Protégé [1] editor for OWL into their own annotation framework.

But this approach suffers from the fact that Protégé was not originally designed for annotation work. Ontology instances, for example, are displayed as a flat list which makes it difficult for the annotator to discern which corpus segment was annotated with which instances. Relations between instances are displayed in a similar fashion. Furthermore, we found that Protégé reactivity decreases notably with increasing ontology size.

Hence, although a tool that combines existing programs is commendable in principal, practical application may prove very difficult under certain circumstances in which the user might prefer a tool tailored specifically to annotation with OWL. Furthermore, these observations illustrate the importance of good usability for annotation tools.

2 The VISA Annotation Tool

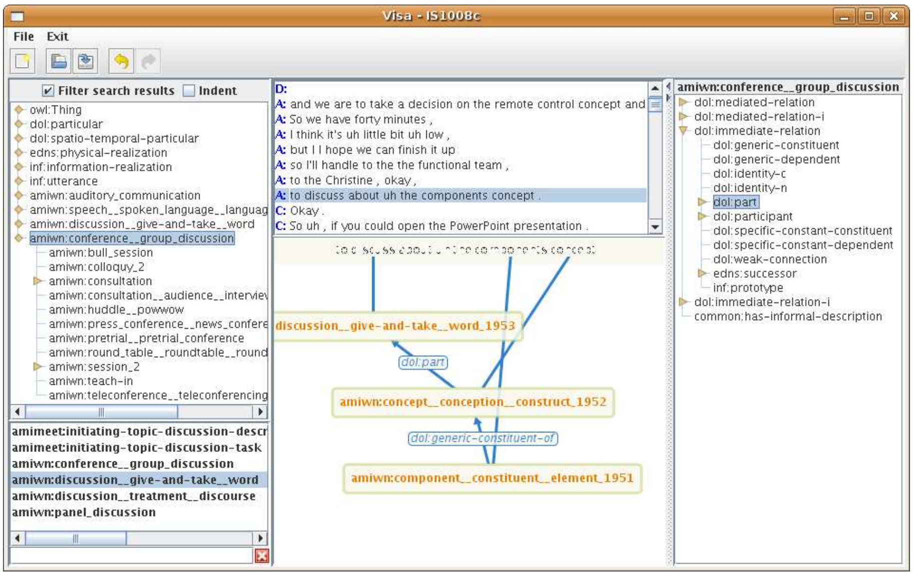
Based on the analysis of deficiencies of existing annotation tools we derived a first requirements specification for a new tool which was followed by the development of a prototype. The further development process has been accompanied by further theoretical considerations with respect to the possible extension of the requirements specification. Moreover we have conducted practical evaluations in form of repeated testing and the prototype has continuously been adapted according to the extended requirements specification.

The following screenshot displays the VISA tool. On the left hand side the classes of the ontology are displayed with their hierarchical relationships, on the right hand side the relation hierarchy of the ontology is shown. In the middle of the window an annotation panel and the text segments that are to be annotated are displayed.

To create a class instance during the annotation process, the corresponding class is selected in the

[*] The research presented here is funded by the EU under the grant FP6-506811 (AMI).

[1] http://protege.stanford.edu


Visa - IS1008c
File
Exit
☑ Filter search results
□ Indent
owl:Thing
dol:particular
dol:spatio-temporal-particular
edns:physical-realization
inf:information-realization
inf:utterance
amiwn:auditory_communication
amiwn:speech__spoken_language__languag
amiwn:discussion__give-and-take__word
amiwn:conference__group_discussion
amiwn:bull_session
amiwn:colloquy_2
amiwn:consultation
amiwn:consultation__audience__intervie
amiwn:huddle__powwow
amiwn:press_conference__news_confere
amiwn:pretrial__pretrial_conference
amiwn:round_table__roundtable__round
amiwn:session_2
amiwn:teach-in
amiwn:teleconference__teleconferencing
amimeet:initiating-topic-discussion-descr
amimeet:initiating-topic-discussion-task
amiwn:conference__group_discussion
amiwn:discussion__give-and-take__word
amiwn:discussion__treatment__discourse
amiwn:panel_discussion
discussion__give-and-take__word_1953
dol:part
amiwn:concept__conception__construct_1952
dol:generic-constituent-of
amiwn:component__constituent__element_1951
amiwn:conference__group_discussion
dol:mediated-relation
dol:mediated-relation-i
dol:immediate-relation
dol:generic-constituent
dol:generic-dependent
dol:identity-c
dol:identity-n
dol:part
dol:participant
dol:specific-constant-constituent
dol:specific-constant-dependent
dol:weak-connection
edns:successor
inf:prototype
dol:immediate-relation-i
common:has-informal-description

D:
A: and we are to take a decision on the remote control concept and
A: So we have forty minutes ,
A: I think it's uh little bit uh low ,
A: but I I hope we can finish it up
A: so I'll handle to the the functional team ,
A: to the Christine , okay ,
A: to discuss about uh the components concept .
C: Okay .
C: So uh , if you could open the PowerPoint presentation .

to discuss about uh the components concept

hierarchy. An instance of the selected class is then created on the annotation panel by drag and drop.

Class instances can be connected with one or several words of the current text segment by dragging from the instances to the words. Relations between instances can be annotated by selecting a relation from the relation hierarchy and dragging from the instance of the corresponding domain class to the instance of the range class.

The graphical instances are arranged automatically on the annotation panel, thus the annotator does not need to take care of the graphical layout of the annotation. To facilitate navigation in the ontology, keyword search functions are available.

VISA is capable of dealing with large-sized ontologies without slowing down the annotation process. One of the ontologies we tested VISA with , e. g., contains more than 60.000 concepts.

VISA is based on NXT (Carletta et al., 2003) which supports the development of corpus tools through the provision of an open source Java API. However, through its modular architecture, VISA allows the integration of other data formats as well.

3 Conclusion and Future Work

We developed a tool for the annotation of text segments with OWL-based ontologies, focussing on a rich feature set an good usability. VISA can deal with large-sized ontologies without slowing down the annotation process.

VISA requires that the text to be annotated is pre-segmented. Furthermore an already existing ontology is required. As our primary concern is to provide an appropriate tool for annotation, VISA does not provide functions for creating or editing ontologies, nor for segmenting or editing of the corpus.

Currently, VISA should still be considered as a prototype. Several features are planned to be added, particularly with regard to the further facilitation of the annotation process, but also features like a reasoning function in order to prohibit inconsistent annotations.

References

K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. 2004. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10(3/4):349–373.

Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holger Voormann. 2003. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior*, 35(3):353–363.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: A pre-annoncement. In *Proceedings of MLMI 2005*.

Christoph Lauer, Jochen Frey, Benjamin Lang, Jan Alexandersson, Tilman Becker, Thomas Kleinbauer, and Harald Lochert. 2005. Amigram–a general-purpose tool for multimodal corpus annotation. In *Proceedings of MLMI 2005*, Royal College of Physicians, Edinburgh, UK, 11-13 July.