

Semi-Automatic Annotation of Contested Knowledge on the World Wide Web

Bertrand Sereno
b.sereno@open.ac.uk

Simon Buckingham
Shum
s.buckingham.shum
@open.ac.uk

Enrico Motta
e.motta@open.ac.uk

Knowledge Media Institute
The Open University
Milton Keynes MK7 6AA
United Kingdom

ABSTRACT

We describe a strategy to support the semantic annotation of contested knowledge, in the context of the Scholarly Ontologies project, which aims at building a network of interpretations enriching a corpus of scholarly papers. To model such knowledge, which does not have 'right' and 'wrong' values, we are building on the notion of active recommendations as a means to sparkle annotators' interest. We finally argue for a different approach to the evaluation of its impact.

Categories and Subject Descriptors

H.1.2 [Information Systems]: User/Machine Systems—*Human information processing*; H.3.7 [Information Systems]: Digital Libraries; H.5.2 [Information Systems]: User Interfaces

General Terms

Human Factors, Experimentation

Keywords

Sense-making, Annotation, Contesting Interpretations, Interface

1. INTRODUCTION

The Semantic Web relies on a precise and as exact as possible annotation of the multiple resources it connects. Annotating a document with the information it contains is being addressed through a number of projects (see [1] for instance). Such knowledge is to be accepted 'as it is' by the agent to whom it is intended, and does not allow for multiple interpretations.

The Scholarly Ontologies project [6], on the other hand, proposes an approach to structurally represent knowledge which by essence is open to debate and interpretations, which can be found for instance in academic publications. Scholarly documents are annotated (or enriched) with the (possibly contradicting) interpretations made by their readers, who then become annotators adding signification to their contents and eventually enabling sense-making. Annotations are formalized as triples (or claims) [node, relation, node], where nodes can be chunks of text or (typed) concepts, and the relation is an instance of a class defined in a formal ontology

Copyright is held by the author/owner(s).
WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

of discourse, which organizes the way interpretations can be articulated. Therefore, the formality is moved from the nodes (in 'typical' applications) to the relations, which we expect to be more stable. In turn, the collection of these semi-formalized utterances in a repository let us envisage a number of *intelligent* services, like the tracing of the use of a particular idea by a community, or the discovery of the documents which take an unsupportive stance on it.

Unlike traditional Semantic Web approaches, the knowledge we are interested in, deriving from a sense-making process, might not appear *per se* in the actual document, and may (or *should*) furthermore be different for different annotators, raising some questions about the level of support one can provide to assist this formalization. In this paper, we describe an approach, complementing the work we described last year [6], to support the formalization of such interpretative knowledge.

2. DEFINITION OF THE TASK

Annotators will have to translate their opinions in a claim compatible form, and we expect the transition from a set of utterances expressed in natural language to a set of Scholarly Ontologies (or ScholOnto) claims not to be straightforward. Interpretations are obviously personal. They contain what will have been understood from a document; and they will also be (and should be) influenced by a number of factors upon which we will have no control, the most obvious being the annotator's personal research interests. To rephrase it, interpreting a document implies to take a perspective on its contents, to view it through a prism which reflects one's own interests.

Because of the underlying formalization, yet another difficulty resides in the elicitation of actually *what* to use as nodes and relations, how long (or detailed) should they be and so on, a problem which is likely to be faced by newcomers to any application requiring formalization, as noted by Shipman and McCall: "*Users are hesitant about formalization because of a fear of prematurely committing to a specific perspective on their tasks; this may be especially true in a collaborative setting, where people must agree on an appropriate formalism.*" [2].

In the end, one could easily argue that no actual support can be provided, as there is no way to infer the points an annotator will consider relevant. These problems might appear insurmountable at first; we do not solve them, but instead aim at providing ways to maybe lighten them, by helping users feeling more confident with

the system and by helping them as much as possible in their formalization task. In other words, we are hoping to help them bridge the gap between their interpretation of a scholarly document expressing the position defended by an author, and the canvas imposed by the ontology of discourse. One desirable side-effect would be that this support could raise some questions in an annotator's mind, possibly enticing her to formulate additional claims.

3. ACTIVE RECOMMENDERS

We define the identification of particular components from the text as the first step of a dual-annotation process, composed of an annotation with 'simple' claims, for which machine tools can help by spotting potentially relevant claim elements or valuable areas of the document; and in a second step, an annotation with 'complex' claims, which result from a human sense-making process.

An early experiment gave us some clues about the particular components one might use when faced with the task of approaching and retaining the salient points of a document. We obtained a set of signals to look for in a document, and the confirmation that annotators would have very different needs according to their ways of approaching a document. Therefore, we have decided to go for a recommending approach, by proposing different components grabbed from the text, leaving it to the annotator to decide whether or not to use them. Some of the components we have retained include:

- Instances of the discourse relations (e.g. *addresses* or *uses/applies*) identified in the ontology. These provide the ready-made claims of the author.
- Candidate concepts (most frequent noun groups, previously encoded (in any document) concepts found in the current document, ...) and claims. We believe that the provision of the annotations made by fellow annotators would help newcomers by showing them what is *feasible*.
- Finally, we would like to state that interpreting a document implies taking a perspective on its contents, and viewing it through a prism which bends it to one's own interests. We are also assuming that authors have to defend their position and their contributions, and relate them (through praise or criticism) to the positions of their colleagues [3]. We would therefore like to get as much insight as possible from the document authors' intentions, and from what they wanted to express. The ability to identify areas (or sets of sentences) describing for instance the research work being carried out by the author or the work being attributed to external (to the document) authors is therefore very useful [4].

These elements can be imported in one of the interfaces for claim formulation, like for instance ClaiMapper [5] or ClaimSpotter (cf. figure 1).

4. EVALUATION CRITERIA

The open-endedness of the task means that the evaluation of any supporting tool is not going to be straightforward. Indeed, annotators being constrained 'only' by the discourse relations, and not in what the nodes at the extremities of these are, there is no notion of 'correct' and 'incorrect' interpretation. Again, any element of the considered document might be of interest to at least one annotator. Therefore, typical measures for assessing the quality of an annotation do not hold here, whether it is reproducibility or consistency. We have to take a broader view of the process of claim elicitation, and maybe look at the impact the recommendations identified

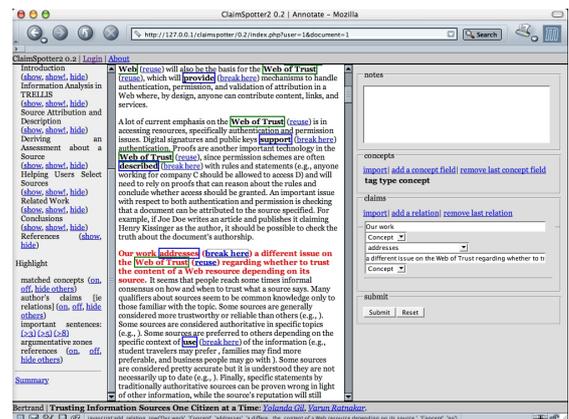


Figure 1: Using ClaimSpotter to provide active recommendations (left) on the document (middle) as a basis for annotation with knowledge claims (right).

above have on it. Or in other words, to assess "in which ways do the recommendations assist the annotators in their formalization task?".

As we mentioned earlier, the main rationale behind them is to stimulate an annotator's interpretation, by activating and highlighting some elements of the described work which she might have overseen. This aspect suggests the creation of a dialogue between the system, extracting as many (potentially) interesting elements (candidate concepts, ...) from the document, and the user, retaining the ones of interest and contextualising them, making sense out of them.

The next phase of work will include the recording of interactions between a user and the system, and the design of a questionnaire assessing the impact (whether positive or negative) of the recommendations and of their presentation.

5. CONCLUSION

This document has briefly presented the Scholarly Ontologies project and has proposed our approach to support semi-automatically the insertion of knowledge constructs (or claims) representing the interpretation one makes of a scholarly document. Although we will have a limited control on the very contents of a claim, we have suggested that the ability to get some insight in the author's argument and to propose these elements for consideration might be a successful approach. We have also presented some elements of evaluation which we will now start to perform.

6. REFERENCES

- [1] S. Handschuh and S. Staab. Authoring and Annotation of Web Pages in CREAM. In *Proceedings of the 11th International World Wide Web Conference (WWW2002)*, 2002.
- [2] F. M. Shipman and R. McCall. Supporting Knowledge Base Evolution with Incremental Formalization. In *Proceedings of Human Factors in Computing Systems conference*, pages 285–291, April 1994.
- [3] J. M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.
- [4] S. Teufel and M. Moens. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409–445, December 2002.
- [5] V. Uren, B. Sereno, S. Buckingham Shum, and G. Li. Interfaces for Capturing Interpretations of Research Literature. In *Proceedings of the Distributed and Collective Knowledge Capture Workshop, held with the Knowledge Capture Conference (KCAP), FL, USA, October 2003*.
- [6] V. Uren, S. Buckingham Shum, G. Li, J. Domingue, and E. Motta. Scholarly Publishing and Argument in Hyperspace. In *Proc. of the 12th International World Wide Web Conference (WWW2003), Budapest, Hungary, May 2003*.