KNOWLEDGE MEDIA

KMi

INSTITUTE

Semantic Annotation Support in the Absence of Consensus

Tech Report kmi-04-1
March 2004

Bertrand Sereno, Victoria Uren, Simon Buckingham Shum and Enrico Motta

The Open University

Semantic Annotation Support in the Absence of Consensus

Bertrand Sereno, Victoria Uren, Simon Buckingham Shum, and Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes MK7 6AA, UK
b.sereno, v.s.uren, s.buckingham.shum, e.motta@open.ac.uk

Abstract. We are interested in the annotation of knowledge which does not necessarily require a consensus. Scholarly debate is an example of such a category of knowledge where disagreement and contest are widespread and desirable, and unlike many Semantic Web approaches, we are interested in the capture and the compilation of these conflicting viewpoints and perspectives. The Scholarly Ontologies project provides the underlying formalism to represent this meta-knowledge, and we will look at ways to lighten the burden of its creation. After having described some particularities of this kind of knowledge, we introduce ClaimSpotter, our approach to support its 'capture', based on the elicitation of a number of recommendations which are presented for consideration to our annotators (or analysts), and give some elements of evaluation.

1 Introduction

While the Semantic Web starts to soar, it is nevertheless relying on a precise and exact annotation of the multiple resources it connects. Annotating a document with the actual information it contains is being addressed through a number of projects ([1] for instance), but all of them have in common the desire to translate, in a more formal way, information which is already present in the actual page (*e.g.* the price of an item in an online store or the affiliation of a researcher), and more importantly, which is not going to be contested.

Indeed, such knowledge is to be accepted 'as it is' by the application, the knowledge expert, or the end user. We are on the other hand interested in the annotation of knowledge which does not necessarily fit this description. Consider scholarly discourse: there can be many interpretations about a particular piece of research, and disagreement is an inherent part of it. Unlike many Semantic Web approaches, disagreement and contest are highly desirable here, as we want all the conflicting viewpoints and perspectives to be captured.

Arguments are not necessarily constrained to the field of scholarly debate though, and we can witness their emergence in many domains nowadays, as analysts express their viewpoints about the direction their company should take piece or publish their review of the latest movies, and allow their readers to comment on them, by providing their own arguments for or against.

We are introducing in this article an approach, ClaimSpotter, to assist the formalisation of such knowledge, and we will focus on scholarly debate. We describe a strategy to provide our annotators (or analysts) with relevant information extracted from the document under consideration. We describe firstly the inherent formalism of this approach and introduce in more detail the characteristics of the knowledge we are interested in, and the difficulties associated with its capture. We will then present the architecture and the components of ClaimSpotter, and report on some preliminary elements of its evaluation. Finally, we will conclude by a discussion and the presentation of some related work.

2 The Scholarly Ontologies project

The Scholarly Ontologies (or ScholOnto) project [2] aims at implementing a Semantic Web of scholarly documents, enriched with the (possibly contradicting) interpretations made by their readers, who become analysts. These interpretations summarize the core contributions of an article and its connections to related work, which are deemed relevant in the eyes of an analyst. They are formalized as triples (or claims) <node, relation, node>, where the nodes can be chunks of text or (typed) concepts (like a theory, a methodology or an approach), and the relation is an instance of a class defined in a formal ontology of discourse, which organizes the way interpretations can be articulated; figure 1 gives some examples of relations which can be drawn between nodes. Utterances like *In my opinion, the document [3] describes a mechanism to enhance documents with machine understandable information, which supports the notion of Semantic Web, as introduced in [4]*, can be encoded as <enhancing documents with machine understandable information, supports, Semantic Web>, where [enhancing documents with machine understandable information] and [Semantic Web] are two concepts defined by the current analyst and associated to their respective document ([3] and [4]), and connected with an instance of the relation class [example of] (cf. figure 1).

For their annotation, users are encouraged to make links to concepts backed by other documents (*e.g.* [Semantic Web] in the previous example) and to reuse concepts. They may extend the models built by other contributors, adding further claims, or take issue with them if they feel the original interpretation is flawed. Thus, a claim space emerges collaboratively and cumulatively as a complex web of interrelated claims (cf. figure 1), which represents the perspective adopted upon a particular problem by a community of researchers. Representing annotation as claims allows a number of intelligent services, like for instance the tracking of a particular idea and the elicitation of its subsequent reuses [2].

However, we expect that moving from utterances expressed in natural language to a set of ScholOnto claims is not going to be straightforward, as analysts will have to translate their opinions in a claim-compatible form. In the following paragraphs, we will look at the characteristics of this problem.

2.1 Difficulties

We start by emphasising that interpretations are necessarily personal. They contain what has been understood from a document; and they will also be (and

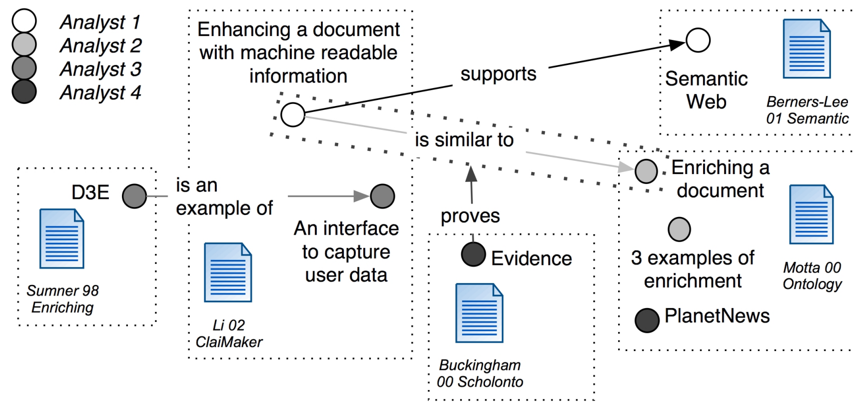

Fig. 1. Multiple interpretations encoded as sets of ScholOnto claims, expressing contributions and connections to and from documents. Analysts 1 and 3 have created two concepts defined in two different documents and linked them with a relation; analyst 2 reuses a concept created by analyst 1 and creates a link to one of his own concept defined in a new document; analyst 4, finally, creates a new concept in his document and connects it to another claim.

should be) influenced by an number of factors upon which we have no control, the most obvious being the analyst's personal research interests. To rephrase it, interpreting a document implies taking a perspective on its contents, and viewing it through a prism which bends it to one's own interests.

Because of the underlying formalisation, yet another difficulty resides in the elicitation of actually *what* to use as nodes and relations, how long (or detailed) they should 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 [5]:

> "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."

Formalising means translating, and potentially losing, a part of the original opinion held by an analyst. These opinions will have to fit in the schema of relations, which means leaving aside all the nuances that could not be represented by it. Opinions will appear more clear-cut, and added to their increased visibility, we might witness the rise of a legitimate fear of commitment.

2.2 Our approach

We are not pretending to solve these problems, but instead seek to provide ways to lighten them, by helping users feeling more confident with the overall process and helping them as much as possible in their formalisation task. In other words, we seek to help them bridge the gap between their interpretation of a scholarly document expressing the position defended by an author, and the schema imposed by the ontology of discourse relations.

Our answer to this daunting problem lies in two steps. We have firstly developed two generations of interfaces to make the process of inputting interpretations as easy as possible. These interfaces have been described in [6]. Secondly, we have conducted an observation of analysts' needs to identify potentially interesting components from the text. Our goal is *to get sense out of documents* and help analysts *put their own sense* on the knowledge structures they build.

3 A recommendation-based approach

To *get sense out of documents*, we are going first of all to look at the task in more detail and get some insight on the underlying formalization process. Although we have to bear in mind that the kind of knowledge we are interested in would be found only implicitly in scholarly documents [7], as it results from a sense-making process, we are investigating the following research question:

> what are the limits of text-based approaches to assist the translation of one's interpretation into a set of formalized knowledge constructs ?

We believe that by combining the Scholarly Ontologies (ScholOnto) repository of claims and some carefully selected components of the document, we can assist the task of claim formulation. We also argue that providing such resources within an interface will substantially improve the overall experience. Our global vision is therefore based on the following aspects: (1) assisting the formulation of claims by proposing an alternative, ScholOnto-aware, view of the document; and (2), wrapping these resources in an environment to actively support the formulation of claims. This enhanced environment would support 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.

3.1 Observation studies

Our previous use of the expression 'relevant claim elements and areas' was maybe an improper one. We should stress that we are not interested in actually summarising a scholarly document and 'reducing' it to a number of claim-worthy components. Indeed, any particular aspect of it might be of interest to at least one analyst. However, by lifting up some of its components and proposing them to the analyst for further analysis, we hope to reduce the cognitive overload, while still providing her full access to the whole document.

So, how do analysts approach a document, when faced with the task of expressing their interpretation? There has been some literature about how people

	a1	a2	a3	a4	a5	a6	a7
Title							Q1
Abstract							
The paper introduces an approach...	Q2	Q2	Q1	Q2	Q2 Q3	Q1	
Latent Semantic Analysis...	Q2	Q3	Q2		Q2 Q3		Q1
A modified Boltzman...	Q2	Q3	Q3		Q2 Q3		
The approach was implemented...	Q3						
Introduction							
The wealth of digitally stored...	Q1		Q1	Q1	Q1		
Keyword searches over...		Q1		Q1			
...							
Data Analysis						Q3	
Latent Semantic Analysis (LSA) [4]...			Q3	Q3			Q2 Q4
It overcomes...							
We apply LSA to extract...		Q3	Q3	Q2	Q3		
...							
Data Visualisation						Q3	
Rather than being a static...	Q4			Q2			
Data is displayed in an initially...		Q3	Q2 Q3	Q3	Q3	Q2	Q2 Q4
...							
Prototype Systems							
Conclusions							
Initial tests show that the...							
Detailed user studies are in preparation							
First resutls on using an... [3].							
An extended version of this paper...							

Table 1. Partial results of the experimental process, displaying, for each component (section and sentences) of our test document [9] and for each analyst (a1...a7), the question(s) it helped to answer.

approach a document (see [8] for instance), and how some components are more likely to be retained for attention than some others. However, we wanted to see if the claim-formulation process had some characteristics of its own. Therefore, we devised an initial experiment. Seven persons (all researchers) were given a short paper (2 pages) and a marker pen. They were asked to answer a number of questions, and to highlight, for each question, the parts (or components) of the document that they were going to use to formulate their answer. The questionnaire was designed to allow an easy mapping into ScholOnto claims, with questions about both the contributions made in the document and its connections to the rest of the literature:

- Q1: what is the problem tackled in this document?
- Q2: how does the work presented try to address this problem?
- Q3: what previous work does it build on?
- Q4: what previous work does it critique?

Our initial hope was to identify a set of components which would be widely used and therefore which could be *recommended* to novice analysts. However, we have witnessed a number of different approaches to answer these questions, as highlighted in the results table (cf. table 1).

Some persons highlighted the keywords and made use of them in their answers. Another person used the title. One person marked an entire section (*Data analysis*) without giving any more detail about which parts of it were going to be used. Some sections were also much more used than some others: most of the participants found their answers to questions Q1 and Q2 in the abstract and the introduction and dismissed nearly completely the remainder of the document. It suggests that the ability to access these components directly would be useful.

A majority of subjects used sentences spread in multiple sections however. Some of these sentences were used consistently to answer the same question: for instance, "*The wealth of digitally stored data available today increases the demand to provide effective tools to retrieve and manage relevant data.*" is used 4 times out of 4 to answer question 1. On the other hand, some sentences, like for instance "*Rather than being a static visualization of data, the interface is self organizing and highly interactive.*" is used three times, and to answer three different questions. It might mean that the questions were not well defined enough, that they were maybe overlapping, resulting in some confusion in our subjects' minds.

In addition to the section they belong to, other aspects or features of a sentence that we could identify include the presence of a citation (especially, of course, to answer questions Q3 and Q4), like "*Latent Semantic Analysis (LSA) [4] has demonstrated improved performance over the traditional vector space techniques.*". Such sentences are considered as describing related work and therefore make valuable elements to consider when interpreting a document's connections to the literature.

Finally, we also noticed that sentences containing an instance of what we assessed to be a highly-subjective verb (like 'to apply', 'to demonstrate', 'to overcome'), which showed a strong level of commitment by the author, were picked. Sentences like "*The paper introduces an approach that organises retrieval results semantically and displays them spatially for browsing.*" clearly describe the authors' intention and, therefore, also provide valuable information about the document's contributions and/or the authors' position, which in turn provide valuable material to write a claim.

We understand from this observation that our analysts would have very different needs according to their ways of approaching a document and that no one approach would be suitable for everyone. Among the different elements that we could propose for consideration, the most important to us seems to be the ability to identify areas in the document where an author defends her position and relates it (through praise or criticism) to the literature [10]. What is asked of analysts here is eventually to interpret a document. And interpreting a document also means positioning oneself (by agreeing or disagreeing) with respect to the research carried out and presented in the document, positioning oneself with respect to the arguments being proposed by the authors to defend themselves,

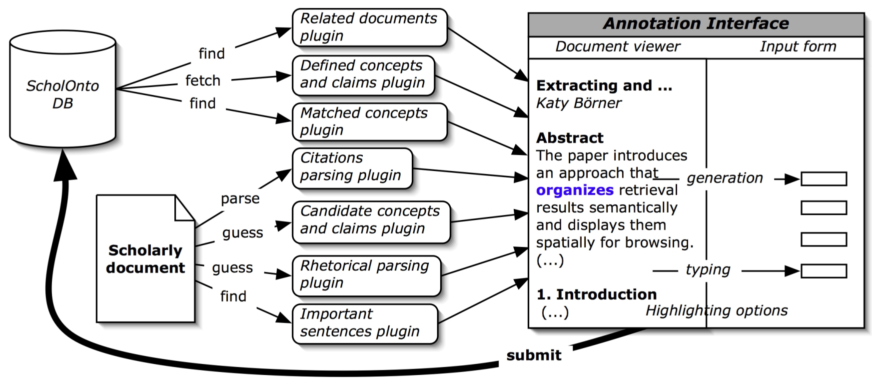

Fig. 2. ClaimSpotter architecture

and positioning oneself with the citations being made and their underlying motivation [11].

We are proposing a *recommending* approach (similar to the one described in [12]), named ClaimSpotter, based on the recommendation of different components grabbed from the original text and from the ScholOnto repository, leaving it to the analyst to decide whether to use them or not. It is our belief (a belief which has been supported by the observation study) that the ability to identify and recommend such elements from the text would help analysts *get sense out* of the document.

3.2 Recommendors

The first element of the ClaimSpotter architecture is a set of recommendors. We have developed preliminary versions of components to extract the previously mentioned elements from the documents, like references signals or sentences describing the work carried out by the authors. We should stress here that we were not initially interested in the development of up-to-date components; rather, we have developed an architecture for these elements based on a shared XML formalism, which allows them to communicate in a standardized way. More importantly, we should be able to plug into the system more robust and up-to-date components. Figure 2 presents the way these components are organized.

In the following paragraphs, we will describe the initial recommendation components (or plug-ins) we have developed so far.

Defined concepts and claims. Once an annotation is stored in a repository, it becomes an additional source of information for a new analyst. Therefore, we start by extracting as much relevant information as possible from the repository of annotations, including: (i) previously encoded (in any document) concepts matched in the contents of the current document; (ii) concepts defined by fellow analysts for the current document, and (iii) claims defined over the current document (i.e., the claims for which the current document was used as backing). We also fetch concepts and claims made by each of the document's authors, over the whole repository.

Candidate concepts and claims. Going back to the document itself, we identify elements from its contents, which, once again, *might* be used as an object in a claim. We look at elements like acronyms, proper nouns and frequent noun groups. They are presented by order of frequency.

Because our discourse ontology has natural language labels (*uses*, *applies*...), which can be changed to fit the dialect of the domain, we also implement a complementary approach, based on the identifications of areas where these labels appear, augmented with some selected synonyms from the WordNet resource [13]. The ability to recommend a sentence like "*Latent Semantic Analysis as well as (...)* ***are applied*** *for semantic data analysis (...).*" (taken from our test document [9]) is potentially very interesting. This should give us a first indication of the particular locations in the document where the author defends her argument. If the analyst shares the point of view of the author, further processing is made on such sentences in order to generate candidate claims like <semantic data analysis, uses/applies, LSA>.

Cited documents. We are also providing a first look at citation contextual parsing, to try to get some insight on the reason motivating each citation. Citation signals (identified manually) are extracted from the text with their context (a couple of sentences before and after the occurrence of the citation). We perform a basic parse of this contextual information (by looking for typical expressions) to guess whether the citation was being supportive or unsupportive. We present these cited documents with their concepts and claims.

Related documents. Additionally, we look at related documents: we extend the notion of relationship between two documents [14], by looking at documents which are related through a claim (in other words, the ones for which at least one analyst has seen a connection). For instance, in figure 1, although documents [4] and [3] are not in a cite/cited relationship, one analyst has related them through a claim. Once such documents are identified, we also fetch their concepts and claims and include them for consideration.

Important areas. We have implemented a simple approach to look at candidate important areas (sentences or sets of sentences) in the document, importance being merely defined here as a combination of the presence of title, abstract and header words in a passage.

Rhetorical parsing The experiment showed the importance of identifying areas containing a description of the author's work. A particularly efficient approach to identify areas where an author defends her position is found in Teufel and

	Category	Confidence
The paper introduces an approach...	OWN	*0.97*
Latent Semantic Analysis...	BACKGROUND	0.67
A modified Boltzman...	BACKGROUND	0.58
The approach was implemented...	OWN	*0.61*

Table 2. Output from the rhetorical parser with the abstract section of our test document [9]. Each sentence is associated with its most likely category. The column on the far right displays the classifier's confidence in the prediction (from 0 (uncertainty) to 1 (certainty)).

Moens' summarizing system [15]. The role played by each sentence (*e.g.* introducing the authors' work, providing background information, or supporting a cited work) is guessed from a number of annotated examples described in terms of an exhaustive range of features including, among many others, the contents of the sentence (presence of meta-discourse constructs [16] or linguistic cues like adverbs and citation signals [10]) and its position in the document.

We experimented with this idea by developing a rhetorical parsing approach, once again to see how it would work. We have focused on a three-category scheme, dealing only with the notion of scientific attribution [17]. We are interested in the sentences describing the research work being carried out by the author (OWN), the work being attributed to external (to the document) authors (OTHER), and the work (or ideas, assumptions, hypotheses, ...) attributed to a research community in general (i.e., where no explicit mention of a person's name is given) (BACKGROUND).

We trained a naïve Bayes classifier with a limited corpus of 230 sentences for the OWN category, 135 sentences for the OTHER category, and 244 sentences for the BACKGROUND category, and a features set composed of the words of each sentence. We nevertheless got some interesting results from the classifier. Table 2 gives an example of rhetorical filtering output. The first sentence, which was heavily used as a basis for annotation in our experiment (cf. table 1) could be lifted up (because of its guessed rhetorical role) and proposed for further consideration.

We also got additional insight in the classifier by looking at its most significant terms (*id est*, the terms which contribute the most to the decision to put a sentence in one category rather than another). We selected the 10 most relevant terms of the classifier, using a χ^2-based filtering computation [18]. Although there was some noise (provided mostly by domain-dependent words, resulting from our annotation corpus), some of the results were rather interesting: the most significant terms for the OWN category included *we*, *section*, *paper*, *our*, *this* and *describe*; while the terms for the OTHER category were including *[citation]* (a generic expression for each citation in the training documents) and *his*. We can then see that many terms can be used to infer the role played by a sentence, some of them being already captured by the relations labels of the ontology, and some of them having been guessed from our annotated training instances.

3.3 Interface

The second element of our approach is to help analysts put their own sense in documents, which we achieved through the realisation of several input interfaces. We have for instance generated an output filter for the recommendors which allow their integration in ClaiMapper [6], a graphical interface that partially realises the notion of claim space presented earlier (cf. figure 1) and allowing the creation of nodes and their connection into claims.

The ClaimSpotter annotation interface (cf. figure 3) basically acts as a document reader [19] [20], but also allow analysts to access the recommendations '*in situ*', through different highlighting options in the original text. It is also possible to drag and drop elements from the text into the 'notes panel' on the right side of the interface, allowing one to keep a trace of the elements of interest. Concepts defined by fellow analysts over the current document can also be accessed and reused; claims made by fellow analysts can be duplicated or debated through the creation of additional claims too. Cited and related documents' ScholOnto information (concepts and claims) can be accessed too and imported into the current document.

4 Evaluation

As we have just started our experimental validation, we will simply introduce the course of actions we have taken so far, and present the aspects we want to focus on. To summarise, we are interested in two main aspects: (i) the quality and usefulness of the recommendations, and (ii) their presentation in an interface.

Assessing the recommendations' relevance is going to be a highly subjective matter. Were we interested in capturing a fact in a document, we could check if the correct instance has been recognised, which is impossible here. Because we cannot measure two seminal aspects of annotations, their stability and their reproducibility [21], we have to fall back on some other aspects, like for instance the usefulness of the tool. We could look at the number of claims submitted by the analysts, and try to answer the following questions: *do analysts make more claims when presented with such recommendations?* or *does the presentation of fellows' claims give something for analysts to react against ?* More interestingly, we can also evaluate the intrinsic quality of the claims. If we assume that a claim using a relation 'is evidence for' is stronger than another one using 'is about' (because it bears more commitment for its author), we might want to see how these relations are used over time, and if the presence of the recommended information inspires confidence for the analyst and encourages her to believe that she can make such claims herself. Table 3 lists a number of these aspects that we have started to study.

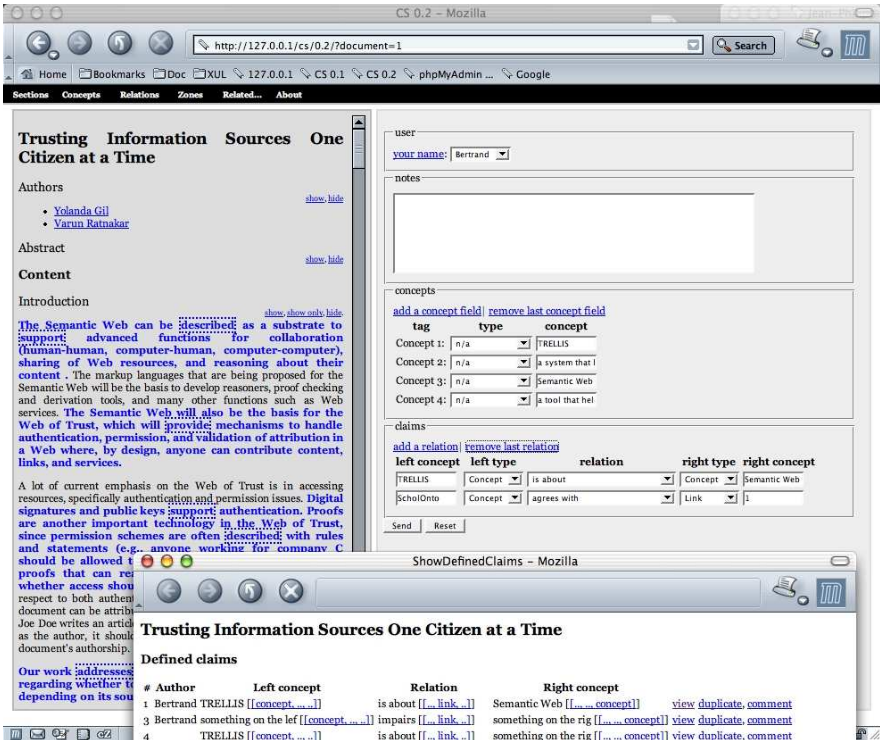

Fig. 3. ClaimSpotter's annotation interface is used to interact with the output of the different recommendors. The main screen separates the workspace into two main areas: a document frame where the view can be customised through highlighting and/or hiding some specific components; and a form frame on the right side where knowledge triples can be created and finally submitted to the ScholOnto database.

We started with four analysts, and we should stress immediately that we will not be able to derive any strong conclusions from such a limited sample, but the goal of this section is rather to introduce the experimental process and state a number of observations and comments we have made during the process. These four people were given instructions about the task and the Scholarly Ontologies project (two of them were novices, and the remaining two experienced users). They were then asked, for a given paper, to spend a reasonable amount of time annotating it, depending on how much they wanted to say about it. After this time, they were provided with the output from the recommendors and left to decide whether they would like to make changes and add any claims to their interpretation.

To summarise these preliminary results, we could say that there were as many stories as analysts, which is not surprising in itself. For instance, one of our expert users made a lot of use of the recommendations provided, by doubling the number of concepts (from 9 to 16) and claims (from 9 to 18) created. Also of interest was

impact of the recommendors	
a)	Do they help her understand what the document is about ?
b)	Do they help her hold an internal model of the document and of its connections to the literature ?
c)	Do they help her break her interpretation into 'acceptable' chunks of text and relations and model it as a structured network of nodes and relations ? [5]
d)	Does the presentation of fellow analysts' claims help her overcome her concerns: commitment, 'what to use as a node ?', 'what did they say ?', ...
e)	Do the recommendations give her additional ideas about the document (which she might not have thought about), additional claims to express, or to counter-argue about a particular point (claim) made by someone else ?
impact of the interface	
f)	Is the ability to browse the text, hide some of its components (sections), and to show the recommendations in situation (through highlighting) helping ?
g)	Are the recommendations easily available and accessible ?
h)	Does their presentation make sense ?

Table 3. Evaluation dimensions for our ClaimSpotter interface and the recommendors

the nature of these newly created claims, 5 out of these 9 being claims about the problem addressed in the document, highlighting maybe the potential usefulness of the tool at spotting OWN statements during the rhetorical parsing step. Our second analyst, who was also well acquainted with the approach, also added concepts as a result of the recommendations being provided, and expressed how some of the claims she had made had been found by the tool itself, helping her to (partly) check her results. In that case, the ability to 'understand', as much as possible, the author's argument was useful, by helping the analyst position herself with respect to that argument. Finally, our two novices also had different experiences with the recommendations. One highlighted how it was helping him to go beyond 'simple' claims like <this document, is about, X> and to be able to formulate more in-depth claims. Our last analyst made only limited use of the recommendations, as she was feeling more confident with the document itself, and did not want to be influenced by external factors. Having been able to provide some support (although not easily quantifiable) to three of our four analysts gives us the motivation to perform a more thorough study, including more detailed aspects such as the ones presented in table 3.

5 Related work

We are expecting these recommendations to support our ontology-supported sense-making annotation process, which is at the boundary of two areas: (1) sense-making approaches typically use free text to capture user stances (*e.g.* [22]); whereas (2) ontology-supported annotation tools aim at capturing knowledge which does not require interpretation and is not likely to be contested.

On the right side of this spectrum, ontology-based annotation tools like CREAM [1] or MnM [23] rely on the presence of the knowledge to capture directly in the document. We have seen that our task, on the contrary, would be more complex, particularly because what we are trying to capture, what we want to "remember, think, clarify and share" [24], will appear only implicitly. For that reason, our approach has tried to build a bridge to assist the formalisation and the population of the ontology as much as possible.

On the left side, the D3E environment, for instance, provides a publishing framework for research papers that allows readers to comment on and to discuss their contents [22]. These comments (and their author) are shown as a threaded discussion, with the ability to initiate several discussions about particular points. An optional categorisation of the nature of the comment can be provided, either 'agreement' or 'disagreement'. Because ScholOnto is based on a formal structure, it is believed that more uses (including intelligent services) could be made of the annotations, admittedly at the cost of formalising this information. TRELLIS is another system which adds formal structure to the semantic annotation by linking statements drawn from web documents using a set of discourse, logical and temporal connectives [25]. TRELLIS is designed to assist analysis of multiple documents in a Web context, implying collaboration between multiple users. However, Tim Chklovski *et alii.* have not reported the use of semantic relations in automatic analysis of data.

We have proposed a bridge between these two worlds by developing and proposing a set of recommendations. They are wrapped up in an approach to assist the formalisation of one's interpretation, which is inspired by the work of Leake *et alii* [12]. This approach proposes methods to assist experts and beginners alike in their task of building and extending a knowledge map by adding concepts and connections (or propositions). Our work shares the same goal, which is to support the construction of a knowledge map (an aspect which is made more obvious when we are using the ClaiMapper interface [6]).

Turning to the recommendors we have implemented, we have already introduced some of the work in rhetorically-directed parsing [10] [26] [15]. Another of our recommendors is based on the parsing of citations. Much work has tried to make use of citation context, for instance, to retrieve documents or to index the contents of the cited document [27] [28]. Additional work has also tried to *understand* the motivation underlying a citation, based on provided set of keywords [26] [29], or on learned lists of contextual words [15]. We expect to bring in some of these works into our architecture and build more and more interesting recommendors.

6 Conclusion

We have presented in this article an approach to support the annotation of a particular class of knowledge which does not necessarily imply a consensus on its interpretation. We have focused on the field of scholarly debate and more precisely on the Scholarly Ontologies project, where multiple and possibly contradicting interpretations can be expressed as a set of knowledge constructs, or claims. Although a claim is, by definition, a statement, and although it does require some elements over which we will not have any control, we have made the hypothesis that the ability to get as much insight as possible into the author's argument would help. Thus we defined an experiment where the subjects were explicitly asked to specify which parts of a document they were more likely to use to answer four questions about the contributions and the connections of a document. Initial conclusions on this experiment have allowed us to characterize the range of elements in the text that these analysts were more likely to consider as a basis to formulate their interpretation as a set of claims. Based on these observations, we hypothesized that a number of components (from claim elements to relevant document areas) would help and we provided mechanisms to extract them from the document, in an attempt to provide support to analysts.

References

1. Handschuh, S., Staab, S.: Authoring and Annotation of Web Pages in CREAM. In: Proceedings of the 11^{th} International World Wide Web Conference (WWW2002). (2002)
2. Buckingham Shum, S., Motta, E., Domingue, J.: ScholOnto : an Ontology-based Digital Library Server for Research Documents and Discourse. International Journal on Digital Libraries **3** (2000) 237–248
3. Li, G., Uren, V., Motta, E., Buckingham Shum, S., Domingue, J.: ClaiMaker: Weaving a Semantic Web of Research Papers. [30]
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. The Scientific American (2001) 34–43
5. Shipman, F.M., McCall, R.: Supporting Knowledge Base Evolution with Incremental Formalization. In: Proceedings of Human Factors in Computing Systems conference. (1994) 285–291
6. Uren, V., Sereno, B., Buckingham Shum, S., Li, G.: Interfaces for Capturing Interpretations of Research Literature. In: Proceedings of the Distributed and Collective Knowledge Capture Workshop, a part of the Knowledge Capture Conference (KCAP), FL, USA. (2003)
7. Motta, E., Buckingham Shum, S., Domingue, J.: Ontology-Driven Document Enrichment: Principles, Tools and Applications. International Journal on Human Computer Studies **50** (2000) 1071–1109
8. Bishop, A.P.: Digital Libraries and Knowledge Disaggregation: the Use of Journal Article Components. In: Proceedings of the 3^{rd} ACM International Conference on Digital Libraries. (1998)
9. Börner, K.: Extracting and Visualizing Semantic Structures in Retrieval Results for Browsing. In: Proceedings of the 5^{th} ACM International Conference on Digital Libraries. (2000)
10. Swales, J.M.: Genre Analysis: English in Academic and Research Settings. Cambridge University Press (1990)
11. Weinstock, M.: Citation Indexes. In: Encyclopedia of Library and Information Science. Volume 5. (1971) 16–40
12. Leake, D.B., Maguitman, A., Reichherzer, T., Cañas, A., Carvalho, M., Arguedas, M., Brenes, S., Eskridge, T.: Aiding Knowledge Capture by Searching for Extensions of Knowledge Models. In: Proceedings of the International Conference On Knowledge Capture (KCAP), FL, USA. (2003)

13. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet : an Online Lexical Database. Technical Report CSL 43, Cognitive Science Laboratory, Princeton University (1993)
14. Small, H.: Co-citation in the Scientific Literature: a New Measure of the Relationship Between Two Documents. Journal of the American Society for Information Science **24** (1973) 265–269
15. Teufel, S., Moens, M.: Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. Computational Linguistics **28** (2002) 409–445
16. Hyland, K.: Persuasion and Context: the Pragmatics of Academic Metadiscourse. Journal of Pragmatics **30** (1998) 437–455
17. Teufel, S., Moens, M.: What's Yours and What's Mine: Determining Intellectual Attribution in Scientific Text. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong. (2000)
18. Yang, Y., Pedersen, J.O.: A Comparative Study of Feature Selection in Text Categorization. In: Proceedings of the 14$^{\text{th}}$ International Conference on Machine Learning, Morgan Kaufmann (1997) 412–420
19. Graham, J.: The Reader's Helper: a Personalized Document Reading Environment. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '99). (1999)
20. Boguraev, B., Kennedy, C., Bellamy, R., Brawer, S., Wong, Y.Y., Swartz, J.: Dynamic Presentation of Document Content for Rapid On-line Skimming. In: Proceedings of the AAAI Spring 1998 Symposium on Intelligent Text Summarization. (1998)
21. Carletta, J.: Assessing Agreement on Classification Tasks: the Kappa Statistic. Computational Linguistics **22** (1996) 249–254
22. Sumner, T., Buckingham Shum, S.: From Documents to Discourse: Shifting Conceptions of Scholarly Publishing. In: Proceedings of the ACM SIGCHI 1998 Conference on Human Factors in Computing Systems, Association for Computing Machinery (1998)
23. Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F.: MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup. In: Proceedings of the 13th International Conference on Knowledge Engineering and Management (EKAW 2002). (2002)
24. Ovsiannikov, I., Arbib, M., McNeill, T.: Annotation Technology. International Journal on Human Computer Studies **50** (1999) 329–362
25. Gil, Y., Ratnakar, V.: Trusting Information Sources One Citizen at a Time. [30]
26. Miike, S., Itoh, E., Ono, K., Sumita, K.: A Full-text Retrieval System with a Dynamic Abstract Generation Function. In: Proceedings of the 17th Annual International ACM-SIGIR Conference. (1994) 152–161
27. Lawrence, S., Bollacker, K., Giles, C.L.: Indexing and Retrieval of Scientific Literature. In: Proceedings of the 8$^{\text{th}}$ Conference on Information and Knowledge Management (CIKM'99). (1999) 139–146
28. Bradshaw, S., Hammond, K.: Automatically Indexing Documents: Content vs. Reference. In: Proceedings of the 6th International Conference on Intelligent User Interfaces, IUI'02, Association for Computing Machinery (2002)
29. Nanba, H., Okumura, M.: Towards Multi-paper Summarization using Reference Information. In: Proceedings of the IJCAI'99 Conference. (1999) 926–931
30. Proceedings of the 1st International Semantic Web Conference (ISWC 2002), Sardinia, Italy. (2002)