Report on the Third Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR)

Jaap Kamps[1] **Jussi Karlgren**[2] **Ralf Schenkel**[3]

[1] University of Amsterdam, The Netherlands
[2] SICS Stockholm, Sweden
[3] MPI/Saarland University, Germany

Abstract

There is an increasing amount of structure on the Web as a result of modern Web languages, user tagging and annotation, and emerging robust NLP tools. These meaningful, semantic, annotations hold the promise to significantly enhance information access, by enhancing the depth of analysis of today's systems. Currently, we have only started exploring the possibilities and only begin to understand how these valuable semantic cues can be put to fruitful use. The workshop had an interactive format consisting of keynotes, boasters and posters, breakout groups and reports, and a final discussion, which was prolonged into the evening. There was a strong feeling that we made substantial progress. Specifically, each of the breakout groups contributed to our understanding of the way forward. First, annotations and use cases come in many different shapes and forms depending on the domain at hand, but at a higher level there are commonalities in annotation tools, indexing methods, user interfaces, and general methodology. Second, there is a framework emerging to view annotation as (1) a *linking* procedure, connecting (2) an *analysis* of information objects with (3) a *semantic model* of some sort, expressing relations that contribute to (4) a *task* of interest to end users. Third, we should look at complex tasks that cannot be comprehensible articulated in a few keywords, and embrace interaction both to incrementally refine the search request and to explore the results at various stages, guided by the semantic structure.

1 Introduction

The goal of the third ESAIR workshop was to create a forum for researchers interested in the use of semantic annotations for information access tasks. By semantic annotations we refer to linguistic annotations (such as named entities, semantic classes or roles, etc.) as well as user annotations (such as microformats, RDF, tags, etc.). The aim of this workshop was not semantic annotation itself, but rather the *applications* of semantic annotation to information access tasks on various levels of abstraction such as ad-hoc retrieval, classification, browsing, textual mining, summarization, question answering, etc.

There are many forms of annotations and a growing array of techniques that identify or extract information automatically from texts: geo-positional markers; named entities; temporal information; semantic roles; opinion, sentiment, and attitude; certainty and hedging to name a few directions of more abstract information found in text. Furthermore, the number of collections which explicitly identify entities is growing fast with Web 2.0 and Semantic Web initiatives. In some cases semantic technologies are being deployed in active tasks, but there is no common direction to research initiatives nor in general technologies for exploitation of non-immediate textual information, in spite of a clear family resemblance both with respect to theoretical starting points and methodology.

The first two ESAIR workshops, organized by Omar Alonso and Hugo Zaragoza, were held at ECIR 2008 [1] and WSDM 2009 [2]. The previous ESAIR workshop at WSDM 2009 ended with the suggestion that semantic annotations might be the way to provide a path towards *making sense* of data on very various levels of abstraction, even non-textual data, by providing narratives and paths through an intractable information space. This is a first thought of how to conceptualise a framework to integrate the various analyses we have recourse to. A special issue of IPM on semantic annotations contains some of the results from the first two ESAIR workshops [2].

The ESAIR workshops, and in particular this third ESAIR at CIKM 2010, have made concrete progress in clarifying the exact role of semantic annotations in supporting complex search tasks as a means to construct more powerful queries. Such queries articulate far more than a typical Web-style, shallow, navigational information requests. The annotations provide the raw material for users to individually build more complex information structures that fit their information needs.

To move further beyond the current understanding of search as factoid queries or navigational requests, the various practices underlying semantic annotation need to be framed in a common intellectual structure. One of the pronouncements of the third ESAIR was to view semantic annotation as (1) a *linking* procedure, connecting (2) an *analysis* of information objects with (3) a *semantic model* of some sort. This linking is in some way intended to work towards an effective contribution to (4) some gainful *task* of interest to end users. All of these four facets of semantic annotation contribute towards the application of semantic annotation to information access tasks on various levels of abstraction such as ad-hoc retrieval, classification, browsing, textual mining, summarization, question answering, etc.

Unleashing the potential of semantic annotations requires us to combine the insights of NLP to go beyond bags of words, the insights of databases to use structure efficiently even when aggregating over millions of records, the insights of information retrieval in effective goal-directed search and evaluation, and the insights of knowledge management to get grips on the greater whole. CIKM provides the convergence of all these four strands of research, which was well proven at the conference itself by the keynote talks which to a large extent voiced the necessity of the information access field to move beyond string processing to semantic analysis and deeper understanding of the information being processed in order to be able to meet the next level of user needs and challenges [8, 10, 14].

2 Workshop

The workshop was structured around four groups of questions, and had a format that emphasized interaction—after all it was a *work*shop.

2.1 Many Open Questions

The previous two workshops were exploratory workshops to discuss the research space around the topic; this workshop intended to propose future directions for the benefit of the field as a whole. Specifically, we brought together a varied group of researchers covering NLP, IR, DB, and KM, and together identified the *barriers* to success and worked on ways of addressing them.

The workshop addressed a range of challenge questions, that can roughly be categorized into the four main themes of the third ESAIR workshop:

Applications and Use Cases What are *use cases* that make obvious the need for semantic annotation of information? What tasks cannot be solved by document retrieval using the traditional bag-of-words? What are the prerequisites of successful application?

Annotations What types of annotation are available? Are there crucial differences between author-, software-, user-, and machine-generated annotations? Named entities, temporal expressions on the one hand and sentiment and hedging on the other are examples of analyses beyond topic that have moved to profitable application. What is holding back the widespread use of these annotations? Are there other types of annotations that are within our grasp?

Result Aggregation Whereas IR focuses almost exclusively at finding individual chunks of information, DB naturally focuses on results that combine information and produce aggregated results (think of OLAP queries), and KM naturally deal with the whole information space. How can we fruitfully combine these strengths?

Searchers and Queries With shallow 2.4 word navigational queries, there may be little benefit in semantic annotations. What expressive power is hidden in the semantic annotation? What is keeping searchers from exploring these powerful search request?

2.2 Format

We started the day with a short introduction of the goals and schedule, and a "feature rally" in which each participant introduced her- or himself, and stated her or his particular interest in this area.

Next, we had two keynotes that helped frame the problem, and create a common understanding of the challenges. Liz Liddy (Syracuse University) looked at the problem from the long history of NLP and IR, and its bright future. Maarten Marx (University of Amsterdam) demonstrated the extraordinary power of querying annotated documents.

We continued with a boaster/poster session, where the papers from Section 4 were presented. The poster session continued over lunch.

After lunch, we had break-out sessions in parallel that focused on specific aspects or problems related to the four themes. After the afternoon coffee, we had reports of the breakout sessions, followed by a final discussion on what we achieved during the day and how to take it forward.

3 Keynotes

Two invited speakers helped frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees.

3.1 NLP's Role in Improving Semantic Annotation

Liz Liddy (Syracuse) started the day with her presentation entitled, "Questions to be Asked and Answered as to NLP's Role in Improving Semantic Annotation" [17]. In the realm of Information Retrieval, why is Semantic Annotation needed? What has changed? Is it the users, the sources, the genres, the technologies, the applications, the queries? If there are differences, why and how can Semantic Annotation help? And more specifically, how and what is Natural Language Processing (NLP) contributing? What is showing promise is the ability to understand how to utilize the higher levels of language processing to do Semantic Annotation. This is largely through the introduction of the Pragmatic level of language processing—the functional perspective which provides the extra understanding that comes from the study of language in actual use. Pragmatics is concerned with the aspects of language which require context to be understood. Basically, how situational context is lexicalized and grammaticalized. In Pragmatics, the goal is to recognize the extra meaning that humans read into utterances, which other levels of language processing have not recognized as being encoded in them. Semantic Annotation can then go the next step and amplify current annotations with this additional contextual and intentional knowledge. Examples were shown of what is being done with NLP today that couldn't be done, or simply wasn't being done in earlier days of IR. In applications of keenest interest today, there is an increased relative emphasis focus on dialogue, interaction, real-time, social, and exploratory search, where understanding the user's intent or plan in their query is key.

3.2 Surplus Value of Semantic Annotations

Maarten Marx (University of Amsterdam) talked about "The Surplus Value of Semantic Annotations" [20]. He compared the costs of semantic annotation of textual documents to its benefits for information processing tasks. Semantic annotation can improve the performance of retrieval tasks and facilitates an improved search experience through faceted search, focused retrieval, better document summaries, and result grouping. Applications which summarize large collections of text or explain real world phenomena based on textual evidence may receive even more benefit from semantic annotations. Semantic annotation creates surplus value if the annotated data can be used beyond any foreseen application. In particular by third parties linking your data by means of your semantic markup to other data with similar markup. He presented a list of properties of the annotated data which optimize this surplus value. They are derived from the principle which states that annotation should facilitate the reuse of data in a mashup without information being lost or distorted. For the Dutch House of Parliament the parliamentary proceedings are annotated based on this principle. Concrete examples from this data collection illustrated the surplus value enhancing properties.

4 Accepted papers

We requested the submission of short, 2 page papers to be presented as boaster and poster. We accepted a total of 16 papers out of 19 submissions. We loosely grouped the papers in four themes:

4.1 Applications

Gey et al. [13] discuss the use of semantic annotations in geo-temporal search, as done in geo-IR in general and at the NCTIR GeoTime Task in particular, and argues for date-stamped topics. Lagos et al. [16] discuss the role (semi) automatic annotations can play in solving the e-discovery problem. Velupillai [25] discusses "electronic patient records" which contain a combination of structured and unstructured content of subjective and objective variables encoded by multiple authors. Structuring the variables expressed in the unstructured text can be of help for further analysis, learning, categorisation and similarity computation, making content available for further research and clinical purposes.

4.2 Annotation

Anh and Takashi [3] investigate automatic annotation using a "concept base," and the use of the annotations for CLIR, to retrieve, to detect mistranslations, and to rerank results. Badia [5] asks whether formalizing events is necessary for their full exploitation, and studies the merits of different axiomatizations. Ferragina and Scaiella [11] re-presented a related CIKM poster at the workshop, introducing the TAGME system that enriches plain-text with links to Wikipedia pages. Marrero et al. [18] propose a specific formalization of rule-based patterns for semantic annotation and information extraction. Palacios et al. [21] describe an approach to the integration of semantic information that is associated with documents with heterogeneous semantic annotations. Tichy et al. [24] propose using semantic annotation as part of the software specification and life-cycle system, by using NLP to extract semantic tags from the specifications and following those tags through to the development process.

4.3 Aggregation

Azzam and Roelleke [4] propose the classification of queries in classes of varying semantic complexity. This classification can then be used for several purposes, one might be to call an appropriate search engine after a query is parsed and classified. Fortuna et al. [12] study predicting user demographics (such as age and gender) of a news-site from their visiting history. Performance is shown to improve when named entities and/or editorial annotations are taken into account. Shiells et al. [23] proposes the grouping of tweets by the URL they contain and then considering the textual content of these tweets as social annotations of the URL. de Vries et al. [9] advocates "search by strategy," a novel user-driven interactive search formalism that helps searchers construct complex queries exploiting the semantic annotations.

4.4 Searchers and Queries

Baskaya et al. [6] discuss "WebExplorer" a tool for constructing search ontologies containing synonyms and translations, and using this tool for cross-language information exploration. Bowers et al. [7] introduce a system for adding semantic annotations to observational datasets, with a use case from ecology, and discuss the use of the resulting annotation framework. Marshall [19] propose a graph representation of multimedia objects, including information in different media and from different sources such as user tags. Said et al. [22] investigate contextual recommendations based on hierarchical tags for various facets.

5 Breakout Sessions

The lively discussion of the poster session continued in three break out groups each discussing a particular aspect of exploiting semantic annotations in a forward looking way.

5.1 Applications and Use cases

Arjen de Vries (CWI and TU Delft) chaired a breakout group on "Applications and Use Cases." The group discussed a range of possible applications where semantic annotation could clearly contribute. Use cases ranged from a tax assistant automatically processing and classifying expenses and payments, to a CIKM 2011 attendee app that would predict what sessions you should go to based on your research interests and other preferences. Whilst a range of interesting applications is within our reach, it is difficult to identify the crucial commonalities between them, and it feels too early to provide a clear recipe for success—this is similar to the points made by Callan [8] in his CIKM keynote talk.

On the positive side, there is a range of semantic annotation tools available, including human annotation or tagging either by traditional means or through crowdsourcing, automatic named-entity recognition or other light-weight NLP tools for English and some other languages, comprehensive ontology management tools for the semantic Web, or simply by matching or linking to Wikipedia or DBpedia.

There is a number of generic types of annotations that seem fairly domain independent, and are frequently used in semantic search: geographical or spatial annotation, temporal annotation, types of events or processes, polarity or affective annotation, or even low-level named-entity recognition (persons, organizations, etc.). The key question remains if there is such a thing as a *domain independent* killer application. Many examples seem very specific to the case at hand. One way out is to view a framework for domain adaptation of semantic search tools as the killer application.

5.2 Annotation and Aggregation

Karen Shiells (Stanford University) chaired a breakout group on "Annotation and Aggregation." The group discussed semantic annotations of different types and attempted to establish some sort of family likeness between the various annotation efforts described and discussed in this workshop, suggested by the keynotes and presented in the main conference papers and posters. As a point of departure the group took the three diagnostics for semantic analysis given by Grefenstette [14] in his CIKM keynote talk: i) Is this *an example* of that? ii) Are these two *the same*? iii) What is *the relation* between these two?

After discussing various types of annotation a tentative general description of annotation as a *linking procedure* was settled on. This description is intended to provide a support for *evaluation* as well as establishing further commonalities between the various activities today vaguely grouped under the heading of semantic annotation. Firstly, semantic annotation typically relies on some sort of *analysis* or *extraction* technology. This can be a text analysis component, a sensor tool, a network analysis tool, an interaction widget—but something which identifies *features* in *information objects* under consideration. Secondly, semantic annotation typically bases its analyses on a *semantic model* of some sort: an ontology, a thesaurus, a conceptual model, a knowledge base, a typology, a symbol table, an alphabet. There are many conceivable models, ranging from simple flat lists, whether open-ended or

pre-coordinate, to complex graph structures of interrelated and non-disjoint categories or labels. Or something else. Thirdly, semantic annotation performs, on the basis of the feature analysis technology and the semantic model, a *linking* of items to the descriptive categories of the model. Fourthly, this resulting linking is somehow related to some gainful activity for the good of some category of end users. All of these four above are components in a description of semantic annotation procedures, methodologies, or technologies.

5.3 Searchers and their Queries

Maarten Marx (University of Amsterdam) chaired a breakout group on "Searchers and their Queries." The group discussed the searcher's role in exploiting semantic annotation. So assuming we have rich data with various types of annotation—and we are nicely progressing in that way—then what can we do with it! What potential added value is in the Semantic Annotation? What do users have to know or do in order to formulate an information need to a semantically rich system? Etc.

The breakout group came to the conclusion that, in fact, the searcher is the main bottleneck in exploiting semantic annotation: we need more than 2.5 keywords in order to use the annotation. This has a number of fundamental consequences. First, we should look at more complex tasks, rather than shallow navigational needs or even ad hoc informational requests, that cannot be comprehensible articulated in a few keywords. Currently complex tasks are solved by having a searcher combine the results of a series of sub-queries by hand. Second, formulating a fully explicit complex query may be impractical since it requires substantial effort, or even impossible since it requires intimate knowledge of the exact data and annotations. This suggests use cases where implicit information from profiles or localizations is available, think of searching in Facebook or in Maps. Third, interaction seems key. Searchers may interactively construct a complex query by incrementally refining their search request with constraints on both content and semantic structure as suggested by the result of a previous query. In addition, they may interactively explore the results at any given stage, again aided by the semantic structure of result space in ways similar to faceted search.

6 Conclusions

After the results of the breakout groups, as discussed in Section 5 above, were presented to the workshop in the final plenary session, there was a strong feeling that we made substantial progress. Specifically, each of the breakout groups contributed to our understanding of the way forward. First, while annotations and use cases come in many different shapes and forms depending on the domain at hand, at a higher level there are commonalities in annotation tools, indexing methods, user interfaces, and general methodology. Second, there is a framework emerging to view annotation as (1) a *linking* procedure, connecting (2) an *analysis* of information objects with (3) a *semantic model* of some sort, expressing relations that contribute to (4) a *task* of interest to end users. Third, we should look at complex tasks that cannot be comprehensible articulated in a few keywords, and embrace interaction both to incrementally refining the search request and to explore the results at various stages, guided by the semantic structure.

More generally, there was broad support for the workshop's interactive character and the group discussions, and how this perfectly complemented the more formal presentations

during the CIKM conference. Casting the gained insights into a clear statement or declaration turned out to be non-trivial: we could not come up with a statement that Jussi expected to convince his colleagues at the laboratory back in Stockholm of the crucial utility of semantic annotation for every future information access task of importance—admittedly a very hard success criterion...

The workshop offered two awards, a *best paper award* based on the reviews and a best paper committee consisting of the three organizers, and a *best presentation award* based on a popular vote amongst workshop participants. The best paper award went to Hany Azzam and Thomas Roelleke, for their paper "A Semantic Query Rating Scheme" [4]. The best presentation award was won by Arjen de Vries, Wouter Alink and Roberto Cornacchia for their paper "Search by Strategy" [9]. Congratulations to Hany and Arjen (and co-authors)!

Last, but certainly not least, the workshop continued with a more informal program in the "Loose Moose Tap and Grill" with continued, and even more intense, discussion about exploiting semantic annotations and (scientific) life in general. This then seamlessly continued into a spooky Halloween night in downtown Toronto...

Acknowledgments We would like to thank ACM and CIKM for hosting this workshop, in particular Mounia Lalmas, Aijun An and Jimmy Huang for their outstanding support in the organization. We would also like to thank the program committee: Omar Alonso, Pablo Castells, Shlomo Geva, Vanja Josifovski, Noriko Kando, Liz Liddy, Maarten Marx, Paul Ogilvie, Hinrich Schütze, Andrew Trotman, Ozlem Uzuner, Arjen de Vries, Roman Yangarber, Hugo Zaragoza, and the three program chairs. Final thanks are due to the paper authors, the invited speakers Liz Liddy and Maarten Marx, and the participants for a great and lively workshop. Details about the workshop including the presentations and slides are online at `http://www.sics.se/events/esair2010/`. The contributed papers are available online at `http://portal.acm.org/citation.cfm?id=1871962`.

References

[1] O. Alonso and H. Zaragoza. Exploiting semantic annotations in information retrieval: Esair '08. *SIGIR Forum*, 42:55–58, 2008.

[2] O. Alonso and H. Zaragoza. Editorial: Introduction. *Information Processing and Management*, 46:381–382, 2010. Special Issue on Semantic Annotations in Information Retrieval.

[3] P. H. Anh and Y. Takashi. Cross language information retrieval based on concept base and language grid. In Kamps et al. [15], pages 11–12.

[4] H. Azzam and T. Roelleke. A semantic query rating scheme. In Kamps et al. [15], pages 21–22.

[5] A. Badia. Is formalizing events necessary for full exploitation. In Kamps et al. [15], pages 13–14.

[6] F. Baskaya, J. Kekäläinen, and K. Järvelin. A tool for ontology-editing and ontology-based information exploration. In Kamps et al. [15], pages 29–30.

[7] S. Bowers, H. Cao, M. Schildhauer, M. Jones, B. Leinfelder, and M. O'Brien. A semantic annotation framework for retrieving and analyzing observational datasets. In Kamps et al. [15], pages 31–32.

[8] J. Callan. Search engine support for software applications. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*, pages 1–2. ACM, 2010.

[9] A. P. de Vries, W. Alink, and R. Cornacchia. Search by strategy. In Kamps et al. [15], pages 27–28.

[10] S. T. Dumais. Temporal dynamics and information retrieval. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*, pages 7–8. ACM, 2010.

[11] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international Conference on Information and Knowledge Management (CIKM '10)*, pages 1625–1628. ACM, 2010.

[12] B. Fortuna, D. Mladenić, and M. Grobelnik. Application of semantic annotations to predicting users' demographics. In Kamps et al. [15], pages 23–24.

[13] F. Gey, N. Kando, and R. R. Larson. The crucial role of semantic discovery and markup in geo-temporal search. In Kamps et al. [15], pages 5–6.

[14] G. Grefenstette. Use of semantics in real life applications. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*, pages 5–6. ACM, 2010.

[15] J. Kamps, J. Karlgren, and R. Schenkel, editors. *Proceedings of the Third Workshop on Exploiting Semantic Annotations for IR, ESAIR 2010, Toronto, Canada, October 30, 2010*, 2010. ACM.

[16] N. Lagos, S. Castellani, and A. Kaplan. Semantic annotations for digital investigations. In Kamps et al. [15], pages 7–8.

[17] E. D. Liddy. Questions to be asked & answered as to NLP's role in improving semantic annotation. In Kamps et al. [15], pages 1–2.

[18] M. Marrero, J. Urbano, J. Morato, and S. Sánchez-Cuadrado. On the definition of patterns for semantic annotation. In Kamps et al. [15], pages 15–16.

[19] B. Marshall. Modeling betweenness for question answering. In Kamps et al. [15], pages 33–34.

[20] M. Marx. The surplus value of semantic annotations. In Kamps et al. [15], pages 3–4.

[21] V. Palacios, J. Lloréns, S. Sánchez-Cuadrado, and M. Marrero. Tagging for improved semantic interpretation of xml documents. In Kamps et al. [15], pages 19–20.

[22] A. Said, J. Kunegis, E. W. D. Luca, and S. Albayrak. Exploiting hierarchical tags for context-awareness. In Kamps et al. [15], pages 35–36.

[23] K. Shiells, O. Alonso, and H. J. Lee. Generating document summaries from user annotations. In Kamps et al. [15], pages 25–26.

[24] W. Tichy, S. Körner, and M. Landhäußer. Creating software models with semantic annotation. In Kamps et al. [15], pages 17–18.

[25] S. Velupillai. Semantic annotations in clinical documentation: Exploring potentials for future information retrieval. In Kamps et al. [15], pages 9–10.