Report on the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'11)

Omar Alonso[1] **Jaap Kamps**[2] **Jussi Karlgren**[3]

[1] Microsoft Corporation, USA
[2] University of Amsterdam, The Netherlands
[3] Gavagai & SICS Stockholm, Sweden

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 increasing the depth of analysis of today's systems. Currently, we have only started to explore the possibilities and only begun 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 remarkable commonalities in annotation tools, indexing methods, user interfaces, and general methodology. Second, we got insights in the "exploitation" aspects, leading to a clear separation between the low-level annotations giving context or meaning to small units of information (e.g., NLP, sentiments, entities), and annotations bringing out the structure inherent in the data (e.g., sources, data schema's, document genres). Third, the plan to enrich ClueWeb with various document level (e.g., pagerank and spam scores, but also reading level) and lower level (e.g., named entities or sentiments) annotations was embraced by the workshop as a concrete next step to promote research in semantic annotations.

1 Introduction

The goal of the fourth 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 micro-formats, 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.

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.

The first two ESAIR workshops, organized by Omar Alonso and Hugo Zaragoza, were held at ECIR 2008 [1] and WSDM 2009 [2]. The third ESAIR workshop, held at CIKM 2010 [10], 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. ESAIR'10 concluded with viewing semantic annotation as (i) a *linking* procedure, connecting (ii) an *analysis* of information objects with (iii) a *semantic model* of some sort. This linking is in some way intended to work towards an effective contribution to (iv) some gainful *task* of interest to end users.

This ESAIR'11 workshop further explored this view focusing on the "exploitation" aspects—how to leverage the "semantic model" that the annotation induces and what level of control does this give to the searcher and what does this enable—which led to a clear separation between the low-level annotations giving context or meaning to small units of information (e.g., NLP, sentiments, entities), and annotations bringing out the structure inherent in the data (e.g., sources, data schema's, document genres).

ESAIR'11 also embraced the plan to enrich ClueWeb with various document level (e.g., pagerank and spam scores, but also reading level) and lower level (e.g., named entities or sentiments) annotations, which will serve research in semantic annotations in two ways: first by creating optimal conditions for testing methods leveraging enriched data, and second to reduce variability and allow for reproducibility in the evaluation of research results.

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 first two workshops were exploratory workshops to discuss the research space around the topic. The third workshop took great strides in formulating a common framework for discussing family likeness, evaluation, and application of semantic technologies. This fourth workshop proposed 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 list of themes for the workshop included:

Application and use Case 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? How can the expressive power of semantic annotation best be put to use? What is keeping searchers from exploring these powerful search request?

Annotation and analysis 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. Are there other types of annotations that are within our grasp? What semantic theories do we need to formulate further annotation schemes?

Data Curation Annotations may live inside documents, or be stored externally (e.g., annotated by uncontrolled authors or tools) or both (e.g., annotation with linked data). How to keep data and metadata together? Does the annotation stop somewhere, or is all social or linked data of potential use for searching or navigating. How important is source of the annotations? Are there issues with credibility or trust that need to be taken into account?

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?

3 Keynote

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

3.1 What to do when one size does not fit all

Arjen de Vries (CWI, TU Delft, Spinque) talked about "What to do when one size does not fit all" [7], partly based on the paper "Searching by Strategy" [8] which won a best paper award at ESAIR'10. Current information access is complex: heterogeneous data sources (web, wikipedia, news, email, patents, twitter, personal information, ...), varying result types (traditional documents, tweets, courses, people/experts, gene expressions, temperatures, ...), multiple dimensions of relevance (topicality, recency, reading level, ...). How to represent this information? How to represent the information need and search request? And

how to represent the the objects shown in response? How to match information representations? Data retrieval (deductive) or information retrieval (inductive), or some mixture? The proposal is to have a parameterized search system, which effectively removes the IR engineer from the loop, and have end-users construct entire *search strategies* rather than individual queries. This is done by having end users construct comprehensive search strategies from basic building blocks—each of them representing data or a manipulation of data—in a visual interface. This concept resembles mashup builders like `http://pipes.yahoo.com/`, which also demonstrates the almost unlimited power of combining a small set of basic components. The proposed system is representing the data flow (resources and manipulations, including all IR specific handling such as indexing and stemming), which is translated into a probabilistic relational algebra, which runs in turn on an effective RDMS (MonetDB). This allows for exploratory search, and probabilistic faceted browsing over dynamic facets of semantically annotated data. The current system allows for drill-down search within a facet, but not for probabilistic joins combining information to create new facets on the fly—a hard and interesting research problem. The resulting search by strategy approach is a radical departure from the fast non-interactive one-shot search engines—and empowers the searcher to not only control the *search process*, but also to control the *search tools* at their disposal by allowing them to construct a tailored search engine on the fly.

4 Accepted papers

We requested the submission of short, 2 page papers to be presented as boaster and poster. We accepted a total of 13 papers out of 15 submissions. The papers were loosely grouped into three themes:

4.1 Users, usage and applications

There was a group of papers dealing with semantic annotations in action.

De Graaf [6] reported on the annotation and retrieval of knowledge in software documentation.

Kamps [9] constructed a model of interaction for complex tasks, and the different information flow and success criterion of each phase, framing the role of annotation throughout a search episode.

Marshall [12] studied the completeness and relative value of image tags and how this impacts image similarity evaluation.

Murakami and Ura [13] proposed a decimal classification system for people on the Web, leading to capture semantic labels and hierarchical relations.

Rój [17] discussed how the discovery and retrieval of application program interfaces (APIs) can benefit from rich semantic annotation.

4.2 Mining, extracting and enriching

There was a groups of papers focusing primarily on the annotation itself.

Narr et al. [14] applied NLP approach to annotate entities, persons, and events in tweets, improving access through normalization and taxonomic relations.

Ng [15] discussed the annotation of word senses and argues that renewed analysis will increase of understanding when it works and why.

Sapkota et al. [18] extracted models from regulatory texts (containing regulations, policies, mandates and guidelines for organizations) from different sources using semantic annotation.

Trandabat [19] proposed semantic role labeling as a means to encode context of and relations between entities occurring in texts.

4.3 Models and representations

The was a group of papers discussing the type of information modeled or represented by the semantic annotations.

Damljanović et al. [5] discuss virtual documents as a way to unify data driven approach in IR, and knowledge driven approaches in DB and KM.

Karlgren [11] addressed three distinct affective aspects of relevance to information access tasks: expressions of sentiment in texts, the searcher's own mood, and the emotive impact of the information access process.

Pareti [16] focused on identifying the source of a statement and the relation between the source and the message, and how this attribution helps retrieval and interpretation.

Tsatsaronis [20] studied sources of lexical ambiguity: syntactic ambiguity across syntactic categories and semantic ambiguity due to polysemy or homonymy, and their relative effect on information retrieval effectiveness.

5 Breakout Sessions

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

The original plan was to have three breakout groups according to the three themes of submitted papers (as shown in Section 4). However, Paul Bennet (Microsoft Research) staged a welcome "hostile takeover" of the breakout group on "mining, extracting and enriching" pushing his agenda to enrich ClueWeb with all sorts of annotation [4]. So eventually we settled for two breakout groups, one pushing the practical case of enriching a suitable corpus like ClueWeb, and one on the more theoretical aspects focusing on the "exploitation" of semantic annotations.

5.1 Models and representations

Jussi Karlgren (SICS Stockholm, Gavagai) chaired a breakout group on "Models and representations." The group started from the view of semantic annotation as a *linking* procedure, connecting an *analysis* of information objects with a *semantic model* of some sort, expressing relations that contribute to a *task* of interest to end users. The focus was on the "missing link" of the exploitation: where does it happen? are we missing a component?

There is an "exploitation bridge" in the connection of annotated documents with the semantic model—one could say the fruitful meaning of semantic annotations can be traced here. There is also a question of what is exploited exactly: is it the annotations themselves adding useful extra information (think of linguistic analysis of sentences or sentiments)? or is it the structure inherent in the data that is brought out by the annotations (think of document, link, or thematic structure)? In the first case we have to build specific tools

tailored to the annotation at hand. In the second case perhaps generic tools can be used. How do we know the exploitation is successful? For evaluation, there user should be placed central (where a user may also be a program or agent) and various overall measures, such as task success suggest themselves.

The discussion led to various new insights, even though the direct practical value will require further discussion. As the report of the breakout group concluded, "we got very far and nowhere."

5.2 Mining, extracting, and enriching

Paul Bennet (Microsoft Research) chaired a breakout group on "Mining, extracting, and enriching." The discussion centered on the creation of a semantically annotated corpus to serve research in semantic annotations in two ways: first by creating optimal conditions for testing methods leveraging enriched data, and second to reduce variability and allow for reproducibility in the evaluation of research results.

The most suitable candidate corpus is an enriched version of ClueWeb (either '09 or the new '12). In fact there is already a range of derivative data on ClueWeb'09 available, including pagerank and spam scores, duplicate URLs, redirects, web graph, anchor text. The wish list of other types of annotations was long, and included reading level, sentiments, named entities, geotagging, as well as advanced annotations such as a full parse, credibility, word sense disambiguation, twitter mentions. The initial focus should be on document-level annotations and on annotations generated using soft scores (probability or confidence), as to delay decisions on accuracy trade offs to the use case at hand. In particular the annotation of the queries should get priority. A simple distribution format, using document, attribute, value triples, was discussed.

The idea to enrich a central corpus like ClueWeb received great support, and the concrete ideas to accomplish this seem viable. There should be an open call to everyone with interesting tools, to annotate the ClueWeb corpus in whole or in part.

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, annotations and use cases come in many different shapes and forms depending on the domain at hand, but at a higher level there are remarkable commonalities in annotation tools, indexing methods, user interfaces, and general methodology. Second, we got insights in the "exploitation" aspects: how to leverage the "semantic model" the annotation induces? what level of control does this give to the searcher and what does this enable? This led to a clear separation between the low-level annotations giving context or meaning to small units of information (e.g., NLP, sentiments, entities), and annotations bringing out the structure inherent in the data (e.g., sources, data schema's, document genres). Third, the plan to enrich ClueWeb with various document level (e.g., pagerank and spam scores, but also reading level) and lower level (e.g., named entities or sentiments) annotations will serve research in semantic annotations in two ways: first by creating optimal conditions for testing methods leveraging enriched data, and second to reduce variability and

allow for reproducibility in the evaluation of research results.

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...

Last, but certainly not least, the workshop has gained a proud reputation with its earlier social events, leading to new papers, spinoff workshops, and new friendships. This tradition was continued with a informal program in the "*The Goat and Grill*" on Argyle Street, attended by workshop participants and other CIKM attendees interested in the workshop's topic, combining great discussion with great food and drinks. A special treat was the presentation of the most hyped paper of SIGIR in Beijing [3], whose presentation in China was prevented by an undisclosed red tape incident. Intense discussion about exploiting semantic annotations and (scientific) life in general continued far into the Glaswegian night.

Acknowledgments We would like to thank ACM and CIKM for hosting this workshop, in particular Craig Macdonald for his outstanding support in the organization. We would also like to thank the program committee: Hany Azzam, Pablo Castells, Shlomo Geva, Claudia Hauff, Vanja Josifovski, Noriko Kando, Aaron Kaplan, Ray Larson, Liz Liddy, Amelie Marian, Paul Ogilvie, Ralf Schenkel, Hinrich Schütze, Özlem Uzuner, Arjen de Vries, Andrew Trotman, Roman Yangarber, Hugo Zaragoza, and the three program chairs. Final thanks are due to the paper authors, the invited speaker Arjen de Vries, 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/esair2011/`. The contributed papers are available online at `http://dl.acm.org/citation.cfm?id=2064713`.

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] L. Azzopardi. Searching for unlawful carnal knowledge. In N. J. Belkin, C. L. A. Clarke, N. Gao, J. Kamps, and J. Karlgren, editors, *Proceedings of the SIGIR'11 Workshop on "entertain me" : Supporting Complex Search Tasks*, pages 17–18. ACM Press, 2011.

[4] P. N. Bennett, K. El-Arini, T. Joachims, and K. M. Svore. Enriching information retrieval. *SIGIR Forum*, 45(2):60–65, 2011.

[5] D. Damljanović, U. Kruschwitz, and M.-D. Albakour. Using virtual documents to move information retrieval and knowledge management closer together. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 3–4. ACM Press, 2011.

[6] K. A. De Graaf. Annotating software documentation in semantic wikis. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 5–6. ACM Press, 2011.

[7] A. P. de Vries. What to do when one size does not fit all? In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 1–2. ACM Press, 2011.

[8] A. P. de Vries, W. Alink, and R. Cornacchia. Search by strategy. In J. Kamps, J. Karlgren, and R. Schenkel, editors, *ESAIR'10: Proceedings of the CIKM'10 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 27–28, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0372-9.

[9] J. Kamps. Toward a model of interaction for complex search tasks. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 7–8. ACM Press, 2011.

[10] J. Kamps, J. Karlgren, and R. Schenkel. Report on third workshop on exploiting semantic annotations in information retrieval (ESAIR). *SIGIR Forum*, 45(1):33–41, 2011.

[11] J. Karlgren. The relation between author mood and affect to sentiment in text and text genre. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 9–10. ACM Press, 2011.

[12] B. Marshall. Context seeking with social tags. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 11–12. ACM Press, 2011.

[13] H. Murakami and Y. Ura. People search using ndc classification system. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 13–14. ACM Press, 2011.

[14] S. Narr, E. W. De Luca, and S. Albayrak. Extracting semantic annotations from twitter. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 15–16. ACM Press, 2011.

[15] H. T. Ng. Does word sense disambiguation improve information retrieval? In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 17–18. ACM Press, 2011.

[16] S. Pareti. Annotating attribution relations and their features. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 19–20. ACM Press, 2011.

[17] M. Rój. Exploiting user knowledge during retrieval of semantically annotated api operations. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 21–22. ACM Press, 2011.

[18] K. Sapkota, A. Aldea, D. Duce, M. Younas, and R. Bañares-Alcántara. Semantic-art: A framework for semantic annotation of regulatory text. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 23–24. ACM Press, 2011.

[19] D. Trandabat. Semantic role labeling for structured information extraction. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 25–26. ACM Press, 2011.

[20] G. Tsatsaronis. An experimental study on syntactic and semantic annotations in text retrieval. In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 27–28. ACM Press, 2011.