Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR)

CIKM 2011 Workshop

Omar Alonso
Microsoft

Jaap Kamps
University of Amsterdam

Jussi Karlgren
SICS Stockholm

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. Unleashing the potential of semantic annotations requires us to think outside the box, by combining the insights of natural language processing (NLP) to go beyond bags of words, the insights of databases (DB) to use structure efficiently even when aggregating over millions of records, the insights of information retrieval (IR) in effective goal-directed search and evaluation, and the insights of knowledge management (KM) to get grips on the greater whole.

The Workshop aimed to bring together researchers from these different disciplines and work together on one of the greatest challenges in the years to come. The desired result of the workshop was concrete insight into the potential of semantic annotations, and in concrete steps to take this research forward; to synchronize related research happening in NLP, DB, IR, and KM, in ways that combine the strengths of each discipline; and to have a lively, interactive workshop where every participant contributed actively and which inspired attendees to think freely and creatively, working towards a common goal.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process, Selection process*
General Terms: Algorithms, Experimentation, Theory
Keywords: Semantic Annotation

1. THEME AND TOPICS

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. We believe further research is needed before we can unleash the potential of annotations!

The previous ESAIR workshops, and in particular the third ESAIR at CIKM 2010 [5], made concrete progress in clarifying the exact role of semantic annotations in support complex search tasks: both as a means to construct more powerful queries that articulate far more than a typical Web-style, shallow, navigational information need, and in terms of *making sense* of the retrieved results on very various levels of abstraction, even non-textual data, providing narratives and paths through an intractable information space.

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 were of interest to the fourth workshop—the aim of this workshop was not the technologies for semantic annotation itself, but rather the *applications* and *contributions* 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.

2. CHALLENGE 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/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 an-

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
ACM 978-1-4503-0717-8/11/10.

notations 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?

The Workshop concluded with a final session addressing the best way forward to unleash the potential of semantic annotation.

3. 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.

Damljanović et al. [2] discuss virtual documents as a way to unify data driven approach in IR, and knowledge driven approaches in DB and KM.

De Graaf [3] reported on the annotation and retrieval of knowledge in software documentation.

Kamps [4] 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.

Karlgren [6] 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.

Marshall [7] studied the completeness and relative value of image tags and how this impacts image similarity evaluation.

Murakami and Ura [8] proposed a decimal classification system for people on the Web, leading to capture semantic labels and hierarchical relations.

Narr et al. [9] applied NLP approach to annotate entities, persons, and events in tweets, improving access through normalization and taxonomic relations.

Ng [10] discussed the annotation of word senses and argues that renewed analysis will increase of understanding when it works and why.

Pareti [11] 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.

Rój [12] discussed how the discovery and retrieval of application program interfaces (APIs) can benefit from rich semantic annotation.

Sapkota et al. [13] extracted models from regulatory texts (containing regulations, policies, mandates and guidelines for organizations) from different sources using semantic annotation.

Trandabat [14] proposed semantic role labeling as a means to encode context of and relations between entities occurring in texts.

Tsatsaronis [15] 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.

4. 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. Arjen de Vries (CWI and TU Delft) looked at the problem from the intersection of databases, information retrieval, and web technology. XXX (YYY) demonstrated the extraordinary power of querying annotated documents.

We continued with a boaster/poster session, where the papers from Section 3 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.

The workshop continued with a more informal part, over drinks and dinner with all attendees of the workshop.

REFERENCES

[1] O. Alonso, J. Kamps, and J. Karlgren, editors. *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2011. ACM Press.

[2] D. Damljanović, U. Kruschwitz, and M.-D. Albakour. Using virtual documents to move information retrieval and knowledge management closer together. In Alonso et al. [1], pages 2–3.

[3] K. A. De Graaf. Annotating software documentation in semantic wikis. In Alonso et al. [1], pages 4–5.

[4] J. Kamps. Toward a model of interaction for complex search tasks. In Alonso et al. [1], pages 6–7.

[5] J. Kamps, J. Karlgren, and R. Schenkel, editors. *ESAIR'10: Proceedings of the CIKM'10 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2010. ACM Press.

[6] J. Karlgren. The relation between author mood and affect to sentiment in text and text genre. In Alonso et al. [1], pages 8–9.

[7] B. Marshall. Context seeking with social tags. In Alonso et al. [1], pages 10–11.

[8] H. Murakami and Y. Ura. People search using ndc classification system. In Alonso et al. [1], pages 12–13.

[9] S. Narr, E. W. De Luca, and S. Albayrak. Extracting semantic annotations from twitter. In Alonso et al. [1], pages 14–15.

[10] H. T. Ng. Does word sense disambiguation improve information retrieval? In Alonso et al. [1], pages 16–17.

[11] S. Pareti. Annotating attribution relations and their features. In Alonso et al. [1], pages 18–19.

[12] M. Rój. Exploiting user knowledge during retrieval of semantically annotated api operations. In Alonso et al. [1], pages 20–21.

[13] K. Sapkota, A. Aldea, D. Duce, M. Younas, and R. Bañares-Alcántara. Semantic-art: A framework for semantic annotation of regulatory text. In Alonso et al. [1], pages 22–23.

[14] D. Trandabat. Semantic role labeling for structured information extraction. In Alonso et al. [1], pages 24–25.

[15] G. Tsatsaronis. An experimental study on syntactic and semantic annotations in text retrieval. In Alonso et al. [1], pages 26–27.