Report on the Fifth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'12)

Jaap Kamps[1] **Jussi Karlgren**[2] **Peter Mika**[3] **Vanessa Murdock**[4]

[1] University of Amsterdam, The Netherlands
[2] Gavagai, Sweden
[3] Yahoo! Research Barcelona, Spain
[4] Microsoft Bing, USA

Abstract

There is an increasing amount of structure on the web as a result of modern web languages, user tagging and annotation, emerging robust NLP tools, and an ever growing volume of linked data. 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. To complicate matters, standard text search excels at shallow information needs expressed by short keyword queries, and here semantic annotation contributes very little, if anything. The main questions for the workshop are how to leverage the rich context currently available, especially in a mobile search scenario, giving powerful new handles to exploit semantic annotations. And how can we fruitfully combine information retrieval and knowledge intensive approaches, and for the first time work actively toward a unified view on exploiting semantic annotations.

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, there is a need for further integration of symbolic and statistical methods with each adopting parts of the other's strengths, by focusing on types of annotations that are informed by and meaningful for the task at hand, and relying on automatic information extraction and annotation based on web scale observations. Second, the discussion contributed to the creation of a concrete shared corpus with state of the art semantic annotation—in particular a web crawl annotated with Freebase concepts—that will benefit research in this area for years to come.

1 Introduction

The goal of the ESAIR workshop series is to create a forum for researchers interested in the use of application 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.).

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 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 various levels of abstraction, even non-textual data, providing narratives and paths through an intractable information space.

First, one of the main outcomes of ESAIR'10 was to recognize that semantic annotations are no panacea, and have clearly more potential in areas characterized by the need for i) rich context, ii) for interaction, and iii) for combining different types of data. The potential of semantic annotations in this setting is huge, but this may result in our searcher needing to articulate a complex information request in a complex query language, requiring full awareness of the used annotation schemes. It is crucial to prevent that the onus for exploiting semantic annotation is put on the searcher. The mobile search scenario, which is particularly context-rich, is an ideal scenario to push the ESAIR agenda. Processing data from mobile users allows a wide range of contextual information not available in many other usage situations. Besides personalization and geo-positional information, mobiles have a wide and growing range of locational, mechanical and even biometrical sensor data available to them. In an information retrieval situation this allows the system to infer task and situational context to flesh out the topical content of the query itself.

Second, one of the main outcomes of ESAIR'10 and '11 of was a clearer "theoretical" view on the role of semantic annotations. ESAIR'10 concluded with viewing semantic annotation as a *linking* procedure, connecting an *analysis* of information objects with a *semantic model* of some sort. ESAIR'11 further explored this view focusing on the "exploitation" aspects—how this can be leveraged to some gainful *task* of interest to end users. Interestingly, the resulting view still allows for a wide range of views on semantic annotations—including radically opposing views as held in *information retrieval* (relying on statistical methods modeling uncertainty) and *semantic web* (relying on knowledge-intensive methods based on certainty). These opposing views did surface during the breakout groups at earlier ESAIRs, highlighting different underlying assumptions, and different modes of information access assumed. Both views respond differently to the trade-off between the desire to enforce a messy world into clean data structures, and the need to do justice to every unique searcher and search request, in a world of partial and uncertain information. ESAIR'12 delved deeper into the underlying assumptions, tried to clear up under which conditions each approach has benefits, and worked toward an integrated view on semantic annotations for information access tasks.

The rest of this report will follow the program of the workshop. The workshop started with a round of introductions where each attendee introduced him- or herself, and explained their own interest in the area. Next, featured two keynotes (discussed in §2) who helped

frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees. Ron Kaplan of Nuance talked about conversation user interfaces, and Evgeniy Gabrilovich of Google discussed how we can start to "understand" the web using large-scale knowledge repositories. This was followed by a boaster and poster session in which ten papers (discussed in §3) were presented. The lively discussion extended over lunch. In the next session, participants divided over two discussion groups (discussed in §4). One group discussed the plans to annotate research corpus of crawled web data with Freebase entities, and the other group discussed the ways in which statistical and symbolic methods can be fruitfully combined. In the final session report of the break out groups were presented, the results and progress of the workshop was discussed and preliminary conclusions were drawn (discussed in §5).

2 Keynotes

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

2.1 The conversational user interface

Ronald M. Kaplan (Nuance) talked about "The Conversational User Interface". Ron's talk started from the long history of building intelligent dialog systems since the 1970s, and how the many obstacles from the past are currently within our grasp to be resolved. Speech recognition, one of the key ingredients of natural dialog, now reached a satisfactory level of performance due to massive training data and computing power, rapidly approaching the level of accuracy in human to human communication. This allows for reducing traditional user interface interactions like: 1) activate device, 2) find app, 3) open app, 4) interact with app, to a straightforward conversational user interface asking a natural language question like "Computer, what's my schedule for tomorrow?" as current employed in systems like Siri on mobile devices. [Whether Apple's personal assistant Siri is driven by Nuance speech recognition technology was neither explicitly confirmed nor denied.]

However, a lot more is needed than speech recognition. There is the whole linguistic pipeline, starting from morphology, to syntax, semantics, pragmatics, and discourse and dialog management, as well as AI and reasoning in a modern form, such as inferring intent and preferences, representing knowledge, bridging language and logic, and modeling collaboration. This requires close integration of symbolic and knowledge intensive methods with massive machine learning. Semantic annotation is crucial to derive simple patterns that predict the behavior of future users, with an emphasis on annotations that support context and intent rather than (only) domain taxonomy.

The areas of speech, NLP, reasoning and dialog, are rapidly evolving due to the available data and computational resources—with many exciting options for PhD students to work on now (as well as open positions at the natural language lab of Nuance).

2.2 Understanding the web using large-scale knowledge repositories

Evgeniy Gabrilovich (Google) talked about "Understanding the Web using Large-Scale Knowledge Repositories". Evgeniy's talk start from the need for semantic annotations and how this is just a stepping stone toward the goal of understanding of textual content. A silent revolution is taking place right now: with the massive resources and computing power available, a crucial threshold is reached that will enable us to realize much of the old dreams of Artificial Intelligence.

There are massive knowledge repositories, such as ODP, Wikipedia, Freebase, YAGO2, and Knowledge Graph (reported to contain 500 million entities and 3.5 billion relations relations in May 2012) that can function as a *lingua franca* for annotation, by linking text to concepts, and linking concepts to other concepts in the same or other knowledge bases. The state of the art is the use of web scale annotation tools identifying the main entities occurring on a page, as well tools that populate the knowledge repository with new instances of existing entities and relations, or new entities, relations, or types/schemas. There is rapid development of these tools, now working on long-tail concepts extracted from web pages using patterns derived from occurrences of known concepts in existing data. The resulting more informed representations of information is rapidly being embedded in all aspects of information access: showing structured results about entities in web search, supporting exploratory search over the whole result space, and by triggering actions or proactively showing facts based on the predicted intent of the overall session.

The areas of web scale information extraction, inference, and text understanding, are rapidly evolving due to the available data and computational resources—with many exciting options for PhD students to work on now (as well as open positions at the knowledge graph team at Google).

3 Accepted papers

We requested the submission of short, 2 page papers to be presented as boaster and poster. We accepted a total of 10 papers out of 13 submissions.

Balog and Nørvåg [1] suggest to extend existing work on entity search with the temporal dimension, i.e. searching over knowledge bases where the temporal validity of facts is well defined and the information needs may have temporal constraints.

Das and Gambäck [2] investigates the 5W annotation (Who, What, When, Where, Why) of a sentiment/opinion corpus that pilots this kind of annotation in Bengali.

Eklund [3] investigates mapping "end-user" search terms to the appropriate medical terminology using the UMLS, addressing the problem of dealing with natural language searches in systems that use controlled vocabularies.

Fujita et al. [4] presents several ways to identify query rewrites based on the click behavior of users, and a topic hierarchy of the Yahoo directory.

Mishra et al. [6] describes the creation of an important new benchmark corpus, integrating Wikipedia with the knowledge bases DBpedia and Yago. In addition it comes with 90 SParQL queries based on Jeopardy questions that are conjunctive queries on the structure part plus free text queries on the textual part of the corpus.

Nomoto and Kando [7] address the problem of labeling unstructured documents with labels generated from combinations of Wikipedia article titles and section headers.

Sellami and Rodríguez [8] address the task of measuring the quality of annotations for Semantic Web services, in terms mappings between schema elements and ontological concepts in a reference ontology.

Sojka [9] discusses the semantic annotation of mathematics in large scale digital libraries, by augmenting surface texts (including math formulae) with additional linked representations providing semantic information (expanded formulas as text, canonicalized text and subformulas).

Yoko Kristianto et al. [10] propose a framework for annotating scientific papers for mathematical formulae search,which in essence extracts surrounding text and classifies that text.

Yoshioka and Kando [11] presents a system that supports news searches where the user can specify hybrid structured queries involving explicit named entities, news metadata (source, date), and text keywords.

For further details we gladly refer to the proceedings available online at the ACM digital library at `http://dl.acm.org/citation.cfm?id=2390148`.

4 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. One group addressed shared data, corpora, tasks, and models, and the other group discussed leveraging semantic search and IR.

4.1 Shared Resources

Jussi Karlgren (Gavagai) chaired a breakout group on "Shared data, corpora, tasks, and models" which focused exclusively on the annotation of two web crawls (ClueWeb09 and ClueWeb12) with Freebase entities. This discussion was in a sense a continuation of the discussion at ESAIR'12 on a shared corpus, and the concrete plan to annotate the ClueWeb corpus as prepared by CMU was warmly welcomed by the participants.

The breakout group brainstormed on the ongoing efforts, and generated a wish list for the annotation to be added to ClueWeb. There was a preference for stand-off annotation, where the annotation layer is stored in a separate file with byte offsets to the exact piece of text annotated—to allow for maximal flexibility also when adding other layers of annotation to the collection. Given that the automatic annotation is noisy, the inclusion of some probability/confidence level information seemed essential. Perhaps the main request was to not only annotate the corpus of web data, but also relevant topic sets as used in the TREC Web track 2009–2012, as well as a sample of queries resembling those encountered on the web. All these requests have been realized at the time of writing—with the volunteer efforts of many individuals.

There were further wishes, such as a mapping between the Freebase entities and DBpedia/LOD (which is available from Freebase for the parts based on Wikipedia); some top-level ontology for grouping facets (Wikipedia seems again a useful pivot to YAGO2, DBpedia, and Freebase); the special treatment of geographic (place names, or longitude/latitude points or regions) and temporal references (both explicit dates as well as relative statements); as well

as a second round of annotation in 2014 to allow for experiments with the growth of Freebase coverage and its impact on information access tasks.

One strand of discussion was of the granularity of the annotation. Will we be annotating single entities observed in text or the character of the entire document or the discourse it participates in. One of the thoughts picked up from the earliest ESAIRs was that semantic analysis involves the seamless aggregation of information on many levels of abstraction simultaneously, from the level of sensory data to conceptually complex intellectual structures, processing them into an actionable and single analysis. This approach, by necessity, will involve annotation on several levels and from several aspects of sensemaking, from entities, constructions, linguistic items on the one hand to documents, data sources, and entire discourses.

The discussion contributed to the creation of a concrete shared corpus with state of the art semantic annotation that will benefit research in this area for years to come.

4.2 Leveraging semantic search and IR

Vanessa Murdock (Microsoft Bing) chaired a breakout group on "Leveraging semantic search and IR." The discussion focused on the ways in which statistical and symbolic methods can be fruitfully combined.

The group addressed the question: do systems need to understand meaning? This largely depends on what is understood by "meaning" in this context—which is a philosophical question. It could be just a purely symbolic string processing: if a system just observes language, with context missing, it will have to act based on whatever input it reads. Is semantic annotation useful for improving the search task? Mobile is a likely application area because of its limited real estate, in fact any application demanding high precision is of potential interest. Arguably, search is far from a solved problem: people are satisfied with current search technology because there is not anything better. People has been gotten used to what a search engine can produce for them (and what not), and current systems may even limit the types of tasks people engage in.

So far annotation has been used primarily as (additional) features in the machine learning system. What if we find ways to unleash the meaning and inferences behind the annotation? This may help narrow down the search to the most promising results. Statistical approaches might be able to pick up any feature given that there is enough training data, but where do these features come from? If we start with a too small set we may end up missing some salient aspects. If we increase the number of features, we may need more data and better algorithms to compensate for that. There is a need to better understand the types of features in the application space: some having to do with the searcher and her intent, some having to do with the task, some having to do with the interface and resources available, some having to do with the result space and success criteria etc. Such an understanding of the feature space feels closely related to the knowledge structures with symbolic methods.

As an aside it was mentioned that artificial intelligence fell out of favor in the IR community, and that IR has lost the connection to other types of learning and inference that were pervasive in the 1980s. The time has come to build new connections in ways that merge insights from both areas. For example, knowledge is statistically derived from observations rather than from expert generated rules, and annotations are automatically generated with confidence values rather than manually assigned by experts.

The main conclusion was further integration of symbolic and statistical methods is needed

with each adopting parts of the other's strengths.

5 Conclusions

After the results of the breakout groups, as discussed in Section 4 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, there is a need for further integration of symbolic and statistical methods with each adopting parts of the other's strengths, by focusing on types of annotations that are informed by and meaningful for the task at hand, and relying on automatic information extraction and annotation based on web scale observations. Second, the discussion contributed to the creation of a concrete shared corpus with state of the art semantic annotation—in particular a web crawl annotated with Freebase concepts—that will benefit research in this area for years to come.

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 social events in earlier years, leading to new papers, spinoff workshops, and new friendships. This tradition was continued with a informal program in the "*Castaway Cafe*" on the other side of the rock in Lahaina, Maui, Hawaii, attended by workshop participants and other CIKM attendees interested in the workshop's topic, combining great discussion with an almost endless supply of food and drinks. Intense discussion about exploiting semantic annotations and (scientific) life in general continued far into the Hawaiian night...

Acknowledgments We are very thankful to the EU FP7 Parlance project (`https://sites.google.com/site/parlanceprojectofficial/`) for sponsoring the workshop.

We would like to thank ACM and CIKM for hosting this workshop, the CIKM workshop chairs Dimitrios Gunopulos and Alin Dobra, and in particular Lipyeow Lim for his outstanding support in the organization.

We would also like to thank the program committee: Omar Alonso, Hany Azzam, Krisztian Balog, Pablo Castells, Henriette Cramer, Lars Erik Holmquist, Vanja Josifovski, Noriko Kando, Ray Larson, Liz Liddy, Maarten Marx, Livia Polanyi, Ralf Schenkel, Arash Termehchy, Andrew Trotman, Özlem Uzuner, Arjen de Vries, and the four program chairs.

Final thanks are due to the paper authors, the invited speakers Ron Kaplan and Evgeniy Gabrilovich, and the participants for a great and lively workshop.

Details about the workshop including the presentations and slides are online at `http://staff.science.uva.nl/~kamps/esair12/`. The contributed papers are available online at `http://dl.acm.org/citation.cfm?id=2390148`.

References

[1] K. Balog and K. Nørvåg. On the use of semantic knowledge bases for temporally-aware entity retrieval. In Kamps et al. [5], pages 1–2. http://dx.doi.org/10.1145/2390148.2390150.

[2] A. Das and B. Gambäck. Exploiting 5w annotations for opinion tracking. In Kamps et al. [5], pages 3–4. http://dx.doi.org/10.1145/2390148.2390151.

[3] A.-M. Eklund. Why query annotations may help in providing accurate public health information. In Kamps et al. [5], pages 5–6. http://dx.doi.org/10.1145/2390148.2390152.

[4] S. Fujita, G. Dupret, and R. Baeza-Yates. Semantics of query rewriting patterns in search logs. In Kamps et al. [5], pages 7–8. http://dx.doi.org/10.1145/2390148.2390153.

[5] J. Kamps, J. Karlgren, P. Mika, and V. Murdock, editors. *ESAIR'12: Proceedings of the CIKM'12 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2012. ACM Press.

[6] A. Mishra, S. Gurajada, and M. Theobald. Design and evaluation of an ir-benchmark for sparql queries with full-text conditions. In Kamps et al. [5], pages 9–10. http://dx.doi.org/10.1145/2390148.2390154.

[7] T. Nomoto and N. Kando. Conceptualizing documents with wikipedia. In Kamps et al. [5], pages 11–12. http://dx.doi.org/10.1145/2390148.2390155.

[8] S. Sellami and C. C. G. Rodríguez. Semantic annotation: What about quality? In Kamps et al. [5], pages 13–14. http://dx.doi.org/10.1145/2390148.2390156.

[9] P. Sojka. Exploiting semantic annotations in math information retrieval. In Kamps et al. [5], pages 15–16. http://dx.doi.org/10.1145/2390148.2390157.

[10] G. Yoko Kristianto, G. Topic, M.-Q. Nghiem, and A. Aizawa. Annotating scientific papers for mathematical formulae search. In Kamps et al. [5], pages 17–18. http://dx.doi.org/10.1145/2390148.2390158.

[11] M. Yoshioka and N. Kando. Multifaceted analysis of news articles by using semantic annotated information. In Kamps et al. [5], pages 19–20. http://dx.doi.org/10.1145/2390148.2390159.