Report on the Sixth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'13)

Paul N. Bennett[1] **Evgeniy Gabrilovich**[2] **Jaap Kamps**[3] **Jussi Karlgren**[4,5]

[1] Microsoft Research, USA
[2] Google, USA
[3] University of Amsterdam, The Netherlands
[4] Gavagai, Sweden
[5] KTH Royal Institute of Technology, Sweden

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.

ESAIR'13 focuses on two of the most challenging aspects to address in the coming years. First, there is a need to include the currently emerging knowledge resources (such as DBpedia, Freebase) as underlying semantic model giving access to an unprecedented scope and detail of factual information. Second, there is a need to include annotations beyond the topical dimension (think of sentiment, reading level, prerequisite level, etc) that contain vital cues for matching the specific needs and profile of the searcher at hand.

There was a strong feeling that we made substantial progress. Specifically, the discussion contributed to our understanding of the way forward. First, emerging large scale knowledge bases form a crucial component for semantic search, providing a unified framework with zillions of entities and relations. Second, in addition to low level factual annotation, non-topical annotation of larger chunks of text can provide powerful cues on the expertise of the search and (un)suitability of information. Third, novel user interfaces are key to unleash powerful structured querying enabled by semantic annotation—the potential of rich document annotations can only be realized if matched by more articulate queries exploiting these powerful retrieval cues—and a more dynamic approach is emerging by exploiting new forms of query autosuggest.

1 Introduction

The goal of the sixth ESAIR workshop is to create a forum for researchers interested in the use of application of semantic annotations for information access tasks. 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. Yet 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. Previous ESAIRs 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.

The general aim of ESAIR'13 is not the technologies for semantic annotation itself, but rather the *applications* and *contributions* of semantic annotation to information access tasks. While the goal remains to advance the general research agenda on this core problem, there is an explicit focus on two of the most challenging aspects to address in the coming years.

First, one of the main outcomes of the previous ESAIRs is a view of semantic annotation as a *linking procedure*, connecting a *content analysis* of information objects with a *semantic model* of some sort. All three are objects of study in their own right; the point of the ESAIR series is linking those three activities into a coherent and practical whole. The obvious next step in the discussion is how to leverage known semantic resources (such as knowledge bases, ontologies, folksonomies, lexical resources, hand-annotated or not) to streaming realistic-scale data ("big data"), to be processed in real time, with incrementally evolving knowledge models. The challenge is to use an existing resource as a semantic model, provide an effective and practicable content analysis, and a scalable linking procedure which can handle the data flows of real life data.

Second, whilst the exact scope and reach of the emerging knowledge resources (such as DBpedia, Freebase) is not yet clear, there is a clear focus on enumerating factual content that can fruitfully be complemented by non-topical aspects. There is a massive interest in annotations on non-topical dimensions, such as opinions, sentiment or attitude, reading level, prerequisite level, authoritativeness, credibility, etc. These annotations contain vital cues for matching information to the specific needs and profile of the searcher at hand, yet there is no consensus on how to exploit them, either as additional criteria on the "relevance" of results in traditional search tasks, or in specific use cases where non-topical cues are key, or in contextual or personalized search factoring in the searcher's state.

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, it featured three keynotes (discussed in §2) who helped frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees. Dan Roth talked about computational frameworks for semantic analysis and wikification, Kevyn Collins Thompson talked about enriching the web by modeling reading difficulty, and Marti Hearst talked about search interfaces to enhance the value of semantic annotations. This was followed by a boaster and poster session in which fourteen 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 focused on what is "semantics" and extending the model in the context of large knowledge bases, and the other group discussed the validity of the model: how do we know it is any good?

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

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

2.1 Computational Frameworks for Semantic Analysis and Wikification

The opening keynote was given by Dan Roth (University of Illinois at Urbana-Champaign) on "computational frameworks for semantic analysis and wikification."

Computational approaches to problems in natural language understanding and information extraction are often modeled as structured predictions—predictions that involve assigning values to sets of interdependent variables. Over the last few years, one of the most successful approaches to studying these problems involves constrained conditional models, an integer linear programming formulation that augments probabilistic models with declarative constraints as a way to support such decisions. Dan focused on two examples of this framework: extended semantic role labeling and wikification in the context of developing better semantic analysis of sentences. The first example was extended semantic role labeling, a typical information extraction problem of concept identification and typing, event identification, etc. The keynote focused on preposition relations in addition to verb predicates, and the joint estimation of both. The second example was wikification, the identification of concepts and entities in text and disambiguating them into Wikipedia or other knowledge bases. In particular relational analysis and the use of relations between concepts to generate and disambiguate candidate concepts was discussed.

Ron's keynote gave great examples of learning and inferencing for high level NLP tasks, using statistically learned models with declarative constraints, allow for addressing different layers of semantic annotations.

2.2 Enriching the Web by Modeling Reading Difficulty

The second keynote in the morning was given by Kevyn Collins-Thompson (University of Michigan), and he talked about "enriching the web by modeling reading difficulty."

The ability to read and understand a text would seem to be a basic aspect of interacting with a rich information source like the Web, yet little is currently known about the nature of the Web, its users, and how users interact with content when seen through the lens of reading difficulty. For example, a document isn't relevant to a person's information need—at least, not immediately—if they can't understand it, yet Web search engines have traditionally ignored the problem of finding or providing content at the right level of difficulty as an aspect of relevance. Kevyn focused not so much on reading difficulty prediction itself, but on its use in a web search environment. He showed how computing and applying metadata based on text readability at Web scale opens up new and sometimes surprising possibilities for enriching our interactions with the Web: from personalizing Web search results, to predicting user and site expertise, to estimating searcher motivation. He also highlighted future challenges and opportunities in improving text readability analysis, particularly in light of the rapidly growing interest in large-scale applications for online education.

Kevyn's keynote gave a great example of non-topical annotation of larger chunks of text, and on the creative use of a text difficulty classifier, to both characterize text complexity as well as the searcher's reading level and topical expertise.

2.3 How Can Search Interfaces Enhance the Value of Semantic Annotations (and Vice Versa)

After lunch, Marti A. Hearst (UC Berkeley) gave a keynote lecture on "how can search interfaces enhance the value of semantic annotations (and vice versa)?"

Marti's keynote was structured around four interlocking points: First, faceted search solves (or solved) a search UI problem. Faceted navigation is used in thousands of sites and tools, and is highly successful because it helps exploration and navigation beyond keyword search alone, and it conforms to user expectations even though they don't fully grasp the underlying model. Second, auto-suggest is a good search UI paradigm. Current faceted search has limitations in terms of inflexible hierarchies, its inability to capture main themes, and lacks ways to express explicit relations between concepts. A natural next step is to allow for expressing relations between facets or concepts, and autosuggest is good model to allow user to express complex relations between concepts. Third, behavior log analysis has made great strides. Query autocompletion can suggest ways to organize the possible results over various dimensions. The trick is knowing which categories are relevant. Extensive search and browse log analysis can suggest what parts of the knowledge base are actuated for this queries, which concept combination co-occur, and use this information to encourage searchers to explore longer relational queries. Fourth, current knowledge bases use focuses on head queries. Current usage of knowledge bases in web search engines, by surfacing entity results in Google, Yahoo! and Bing, is focusing entirely on head queries. While this is useful in itself, the entity results show rather generic biographical information that is helpful to identify the person, but is unlikely to answer the deeper information need of the searcher.

Marti's keynote gave a great outline of how the UI can better support querying semantically annotated data, with the particular suggestion that query autocompletion can help encourage the use of longer queries based on concepts and relations would give far more powerful handles to searchers.

3 Accepted papers

We requested the submission of short, 3 page papers to be presented as boaster and poster. We accepted a total of 14 papers out of 21 submissions after peer review (a 67% acceptance rate).

Almasri et al. [1] propose to enrich short queries by adding terms taken from Wikipedia article titles, where the Wikipedia link graph is used to include conceptually related articles that do not match the initial query. The experiments use CLEF/CHIC's Europeana data.

Alonso et al. [2] propose to annotate entities in tweets and exploit these annotations for improving the web search experience. The paper uses clickstream analysis to identify entities, exploiting queries and clicks on canonical pages.

Buscaldi and Zargayouna [4] present an extension of Lucene providing concept-based information retrieval, by using SKOS/OWL terminologies, by annotating documents and queries, and by combining textual and conceptual matching scores in the ranking.

Ceccarelli et al. [5] propose a general framework for entity linking systems, allowing researchers to compare entity linking methods under the exact same conditions. Three state-of-the-art entity linking algorithms are available within the framework.

De Ribaupierre and Falquet [6] propose a user-centric annotation model based on discourse elements (defined as an OWL ontology) and annotate a corpus of scientific articles in gender studies. The paper shows how complex queries, proposed by scientists, can be expressed in this model and solved by a description logic reasoner.

Friberg Heppin [7] investigates "semantic frames", essentially templates based on the lexical units in FrameNet, as a way to improve search results. Experiments on a Swedish corpus shows that the majority of matches conforms to the FrameNet meaning of the pattern, suggesting their potential for conceptual search.

Garkavijs [8] discusses exploratory image search by building a textual representation of a search trail based on viewed images. The paper proposes a simple algorithm for system training, that uses dwell-time data as input parameters for relevance recalculation, which is implemented a the prototype image search system.

Guha [9] investigates the problem of customizing web search results to suit a particular context derived from a user profile or use case, focusing on the context of a 'high school US history course'. The approach compares web content to Wikipedia pages of relevant entities (anchored by comparing the websites to a textbook).

Habib and Keulen [10] argue that named entity disambiguation and extraction are intimately linked and as such should be implemented together. One approach is to use the extraction confidence to maximize recall, and use this extra information to filter down to the best extracted entities and to disambiguate results.

Janowicz and Hitzler [11] is a position paper on how linked data and semantic annotation changes the interaction from the user's point of view, and tries to disentangle some of the complexities focusing on geo-search. There is a persuasive argument for the implications for building systems consistent with these views.

Kaptein et al. [12] discusse a a number of possible approaches for reusing multiple existing web search engines to create a recall-oriented search engine. Specifically, three abstract techniques to re-order the retrieved results are discussed: clustering, reranking, or aggregation ("analysis").

Kim et al. [13] propose a method that mines subtopics based on the clusters of relevant documents. The approach uses simple patterns to mine candidate subtopics that partly match the original topic, and use an hierarchical sub topic ranker.

Leber et al. [14] investigate annotating legal documents with semantic elements extracted from the text by off-the-shelf NLP techniques. The approach deals with partly changed or updated documents, in particular by parsing contract amendments to understand how the original contract is altered.

Yan [15] studies the use of Systemic Functional Analysis (a branch of linguistics) to capture the communicative context. A small corpus is manually annotated, and an initial classifier performs reasonably, opening up the possibility to deploy SFA in information access-related tasks.

For further details we gladly refer to the proceedings available online at the ACM digital library at `http://dl.acm.org/citation.cfm?id=2513204`.

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.

Jussi Karlgren started the discussion by raising two challenge questions in light of the findings of earlier workshops. The first challenge question was on extending the model: what is "semantics?" Discussion at earlier workshops regarded annotation as a linking procedure from data to a model. The discussion raised many fundamental and philosophical points. One interesting line of discussion was taking the availability of large scale knowledge bases into account, and investigated strengths and weaknesses of using such knowledge bases as semantic model.

The second challenge question was on exploring validity: is the model any good? The discussion started from a broad perspective, trying to get a grasp on what is important for (future) real life application. This raised consider discussion on future applications, such as those explored by the participants, and similarities and differences between the various classes of applications. There was general discussion on the coverage and validity of the model, as well as its scalability and robustness, and appropriate measures. One interesting line of discussion was on the requirements that various applications pose on the model.

The breakout group discussion blended seamlessly into the earlier points raised by the keynote speakers. In particular the suggestion of Marti Hearst to focus on the query suggest stage rather than static navigational hierarchies was received with much support. Highly structured data allows for powerful forms of auto suggest or query completion as a model to build powerful structured queries in natural language. A promising example is Facebook's graph search, where the search results are ultimately personalized to the requester, and the query autocompletion is actively suggesting users to explore different slices of the data.

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, the discussion contributed to our understanding of the way forward. First, emerging large scale knowledge bases form a crucial component for semantic search, providing a unified framework with zillions of entities and relations. Second, in addition to low level factual annotation, non-topical annotation of larger chunks of text can provide powerful cues on the expertise of the search and (un)suitability of information. Third, novel user interfaces are key to unleash powerful structured querying enabled by semantic annotation—the potential of rich document annotations can only be realized if matched by more articulate queries exploiting these powerful retrieval cues—and a more dynamic approach is emerging by exploiting new forms of query autosuggest.

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 at 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. In recent years, we visited the "Loose Moose Tap and Grill" in Toronto in 2010, the "The Goat and Grill" in Glasgow in 2011, and the "Castaway Cafe" in Lahaina, Maui in 2012. This tradition was continued with a informal program in the "*Elephant Bar*," one of the few highlights of Burlingame, attended by workshop participants and other CIKM attendees interested in the workshop's topic, combining great discussion with a sheer endless supply of food and drinks. Intense discussion about exploiting semantic annotations and (scientific) life in general continued far into the Californian night...

Acknowledgments We are very thankful to Google for sponsoring the workshop.

We would like to thank ACM and CIKM for hosting this workshop, the CIKM workshop chairs Jianshu Weng and Esteban Zimanyi, and in particular Yi Fang for his outstanding support in the organization.

We would also like to thank the program committee: Omar Alonso, Krisztian Balog, Pablo Castells, Arjen de Vries, Noriko Kando, Ronald Kaplan, Emre Kiciman, Liz Liddy, Maarten Marx, Peter Mika, Paul Ogilvie, Livia Polanyi, Ralf Schenkel, Andrew Trotman, and the four program chairs.

Final thanks are due to the paper authors, the invited speakers Kevyn Collins-Thompson, Marti Hearst, and Dan Roth, 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/esair13/`. The contributed papers are available online at `http://dl.acm.org/citation.cfm?id=2513204`.

References

[1] M. Almasri, J.-P. Chevallet, and C. Berrut. Wikipedia-based semantic query enrichment. In Bennett et al. [3], pages 4–5.

[2] O. Alonso, Q. Ke, K. Khandelwal, and S. Vadrevu. Exploiting entities in social media. In Bennett et al. [3], pages 6–7.

[3] P. N. Bennett, E. Gabrilovich, J. Kamps, and J. Karlgren, editors. *ESAIR'13: Proceedings of the CIKM'13 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2013. ACM Press.

[4] D. Buscaldi and H. Zargayouna. Yasemir: Yet another semantic information retrieval system. In Bennett et al. [3], pages 8–9.

[5] D. Ceccarelli, C. Lucchese, R. Perego, S. Orlando, and S. Trani. Dexter: an open source framework for entity linking. In Bennett et al. [3], pages 10–11.

[6] H. De Ribaupierre and G. Falquet. A user-centric model to semantically annotate and retrieve scientific documents. In Bennett et al. [3], pages 12–13.

[7] K. Friberg Heppin. Search using semantic framenet frames as variables. In Bennett et al. [3], pages 14–15.

[8] V. Garkavijs. Learning user's intent using user tags - intelligent interactive image search system. In Bennett et al. [3], pages 16–17.

[9] N. Guha. Course specific search engines: A study in incorporating context into search. In Bennett et al. [3], pages 18–19.

[10] M. Habib and M. V. Keulen. Named entity extraction and disambiguation: The missing link. In Bennett et al. [3], pages 20–21.

[11] K. Janowicz and P. Hitzler. Thoughts on the complex relation between linked data, semantic annotations, and ontologies. In Bennett et al. [3], pages 22–23.

[12] R. Kaptein, E. L. Van Den Broek, G. Koot, and M. Huis In 'T Veld. Recall oriented search on the web using semantic annotation. In Bennett et al. [3], pages 24–25.

[13] S.-J. Kim, K.-Y. Shin, and J.-H. Lee. Hierarchical subtopic mining for topic annotation. In Bennett et al. [3], pages 28–29.

[14] C. Leber, D. Yang, L. Tari, A. Chandramouli, and A. Crapo. Using semantics to process legal document updates. In Bennett et al. [3], pages 26–27.

[15] H. Yan. Annotation of clausal functional information for semantic retrieval. In Bennett et al. [3], pages 30–31.