Report on the Seventh Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'14)

Omar Alonso[1] **Jaap Kamps**[2] **Jussi Karlgren**[3,4]

[1] Microsoft, Mountain View CA, USA
[2] University of Amsterdam, The Netherlands
[3] Gavagai, Stockholm, Sweden
[4] 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. The goal of the ESAIR'14 workshop remained to advance the general research agenda on this core problem, with an explicit focus on one of the most challenging aspects to address in the coming years. The main remaining challenge is on the user's side—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. How can the query suggestion paradigm be used to encourage searcher to articulate longer queries, with concepts and relations linking their statement of request to existing semantic models? How do entity results and social network data in "graph search" change the classic division between searchers and information and lead to extreme personalization—are you the query? How to leverage transaction logs and recommendation, and how adaptive should we make the system? What are the privacy ramifications and the UX aspects—how to not creep out users?

There was a strong feeling that we made substantial progress. Specifically, the discussion contributed to our understanding of the way forward. First, for notable (head, shoulder, but not tail) entities in semantic search we have reached the level of quality at minimal costs allowing for deployment in major web search engines—the dream has become a reality. Second, entity detection is moving fast into domain specific, personal, and business domains, and has become a vital component for a range of applications. Third, semantic web has exchanged logic for machine learning approaches, and machine learning is the natural unification of semantic web and information retrieval approaches.

1 Introduction

The goal of the seventh ESAIR workshop was 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 semantic 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 very various levels of abstraction, even non-textual data, providing narratives and paths through an intractable information space. The ESAIR'14 workshop had far more focus than the earlier ESAIRs. While the goal remained to advance the general research agenda on this core problem, there was an explicit focus on the main remaining challenge of exploiting semantic annotations in the coming years. One of the main outcomes of the previous ESAIRs has been not only an overview of various domains of application and experiments on real life data, but also a clearer "theoretical" view on the role of semantic annotations. The starting point, based on discussions at 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 we can expect in real life data. 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. Over the last years there has been a massive interest in annotations on non-topical dimensions, such as opinions, sentiment or attitude, reading level, prerequisite level, authoritativeness, credibility, etc, both at the level of individual sentences or utterances as well as at more aggregative levels. It is clear that such annotations contain vital cues for matching information to the specific needs and profile of the searcher at hand, yet it is an open question how such annotations can be fruitfully exploited in information retrieval, 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 that takes

the searcher's state into account.

Both in terms of knowledge bases and in terms of non-topical annotation significant progress has been made in recent years. The main remaining challenge is on the user's side—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. How can the query suggestion paradigm be used to encourage searcher to articulate longer queries, with concepts and relations linking their statement of request to existing semantic models? How do entity results and social network data in "graph search" change the classic division between searchers and information and lead to extreme personalization—are you the query? How to leverage transaction logs and recommendation, and how adaptive should we make the system? What are the privacy ramifications and the UX aspects—how to not creep out users?

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 two keynotes (discussed in §2) who helped frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees. Peter Mika (Yahoo Labs) talked about understanding queries through entities, and Silviu Cucerzan (Microsoft Research) talked about entity extraction and linking with applications to web search. 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 preparing arguments against or in favor of a strict approach to semantic annotation, and fought this out as in an academic debate (discussed in §4). In the final session 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 Understanding Queries through Entities

The opening keynote was given by Peter Mika (Yahoo Labs) on "understanding queries through entities" [10].

Peter gave a comprehensive overview of semantic search motivated by the observation that improvements in IR are increasing hard to make, and the observation that the main bottleneck is not computational, but cognitive, in the need to understand the query, the content, and the relation between the two. The main part of his presentation was about explicit semantics, where queries and content is linked to items in a knowledge graph. The knowledge graph is build from many different sources, that are normalized and unified into a massive knowledge base. Entity linking is the task to find entity mentions in queries and documents, and to identify the correct entity in the knowledge graph it corresponds to. The potential of semantic search to improve an ever growing fraction of queries was illustrated by a number of examples of entity retrieval and related entity recommendations as implemented in Yahoo Search. An advanced entity linking approach was outlined, that combined solid performance with efficiency in CPU time and memory footprint, both essential

for deployment at scale. A literal quote "slow parsing is killing people."

Peter's keynote gave an excellent overview of the state of the art in semantic search and what it takes for an algorithm to be ready for deployment in a major web search engine, plus he managed to include Roi Blanco in most of his examples—also a non-trivial achievement.

2.2 Entity Extraction and Linking with Applications to Web Search

The second keynote in the morning was given by Silviu Cucerzan (Microsoft Research), and he talked about "entity extraction and linking with applications to web search" [2].

Silviu gave an overview of how Wikipedia was a game changer for entity detection and what next steps are in progress, in particular the unification of the entities ocurring in many thousands of domain specific, personal, and business collections. He detailed highly advanced entity extraction and linking algorithms, relying on context to disambiguate entity mentions, and the impact of knowledge base coverage for entity linking. The combination of deep linguistic processing and advanced machine learning addresses semantics beyond the sentence and paragraph level (still largely ignored in NLP). Silviu also covered a range of applications of entity repositories in conjunction with query logs of commercial Web search engines. These included context-aware search, query suggestion, question answering, retrieval of support for factual statements, and the automatic aggregation of topic pages as an alternative to the ten blue links.

Silviu's keynote gave great insight in how entity detection has reached a level of maturity and scale that it has become a vital component for a range of applications.

3 Accepted papers

We requested the submission of short, 3 page papers to be presented as boaster and poster. We accepted a total of 11 papers out of 15 submissions after peer review (a 73% acceptance rate).

Cotelo et al. [1] investigate semantic cues to articulate more expressive queries by reviving various query operators and explore their value in a preliminary evaluation.

De Nies et al. [3] give a broad overview of the challenges in the context of entity tagged corpora, focusing on the annotation quality, appropriate similarity measures, data quality, and access problems.

Deolalikar [4] investigates within corpus text mining to cluster documents and combine cluster and document scores, demonstrating that coarse grained clusters are unable to capture specific intent of topically focused queries.

Ibrahim et al. [5] address the problem of entity linking in social streaming data, looking into the normalization of mentions due to cryptic abbreviations, the contextualization of short postings by shared hashtags, persons, and links, and the temporal trends of attention to time-sensitive entities.

Jan et al. [6] study the specific domain of searching IT service desk tickets, based on topic modeling, concept analysis, and clustering, leading to increased performance on a corpus of noisy statements of IT related problems.

Jiang et al. [7] investigate some heuristics to improve "explicit semantic annotation" by labeling documents with Wikipedia concepts.

Li et al. [8] revisit the answer type prediction problem of question answering systems, using dependency parsing and semantic role labeling rather than ad hoc heuristics.

Mao and Lu [9] focus medical literature search and return to the old problem of using controlled subject headings with a mixture language model and show that this promotes retrieval effectiveness.

Verma and Ceccarelli [11] study the problem of entity detection in non-head queries, observing similarities and differences in the types of entities occurring in slices of queries.

Yang [12] studies concept similarity measures comparing tree edit distance with textual similarity of subtrees or fragments over the open directory project's concept hierarchy.

Zuccon et al. [13] investigates reasoning with rigorous semantic concept hierarchies in medical literature search, and discusses the potential benefits of semantic-based retrieval as well as the risks of unconditionally embracing such inferences.

For further details we gladly refer to the proceedings available online at the ACM digital library at `http://dl.acm.org/citation.cfm?id=2663712`.

4 Debate

The lively discussion of the poster session continued with an academic debate on "information retrieval" approaches (based on vagueness and uncertainty) versus "semantic web" approaches (based on logic and certainty). The participants were split into two teams—one advocating information retrieval approaches and one advocating semantic web approaches—and prepared the arguments in favor of, or against a proposition. The proposition was:

> *The world simply isn't structured enough to expect users to put their information in a neat little package, hence enforcing such a structure will eliminate crucial information.*

Each team had one hour to prepare the debate, and nominate four members with an active role in the debate in the next hour. The affirmative team (advocating vague IR approaches) consisted of Bo-Wen Zhang (captain), Yusra Ibrahim (first affirmative speaker), Zhuoren Jiang (second affirmative speaker), and Alexander Hinneburg (third affirmative speaker). The negative team (advocating strict SW approaches) consisted of: Jussi Karlgren (captain), Roi Blanco (first negative speaker), Robert Meusel (second negative speaker), and Peter Mika (third negative speaker).

The debate started with a short introduction on the proposition and the fundamentally different views on the way forward, and then the speakers of each team took turns and made there arguments, both by reacting to the previous speaker and by new arguments. The affirmative team argued for statistical models capable of dealing with uncertainty and making graceful estimates even in lieu of complete information. The negative team stressed the need for meaningful categories and clean data. The discussion was both strong and entertaining, with agreement on the potential of meaningful, semantic, annotations to significantly enhance information access, and fierce disagreement on the approaches privileging *control*, *authority*, and *certainty* versus approaches privileging *uncertainty*, *diversity*, and *lack of control*.

Both teams managed to convince the audience for their position at different times, making it a difficult call to determine which side won the debate. In the eventual audience vote the negative team was declared the winner—plausibly due to their quite liberal interpretation of semantic web approaches already accepting many information retrieval aspects into their position. However, the best speaker of the debate was Yusra Ibrahim of the affirmative team.

The debate led to further insight in the trade-off between the desire to have clean and well organized information, and the ability to do justice to a unique searcher with a unique tasks at a given time and place.

5 Conclusions

After the results of the debate, as discussed in Section 4 above, were discussed 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, for notable (head, shoulder, but not tail) entities in semantic search we have reached the level of quality at minimal costs allowing for deployment in major web search engines—the dream has become a reality. Second, entity detection is moving fast into domain specific, personal, and business domains, and has become a vital component for a range of applications. Third, semantic web has exchanged logic for machine learning approaches, and machine learning is the natural unification of semantic web and information retrieval approaches.

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; the "Castaway Cafe" in Lahaina, Maui in 2012; and the "Elephant Bar" in Burlingame, CA in 2013. This tradition was continued with an informal program in the "*Shining Hot Pot*" on Heng Shan Road in the French concession of Shanghai, 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 Shanghai night...

Acknowledgments We would like to thank ACM and CIKM for hosting this workshop, the CIKM workshop chairs Xiaofeng Meng and Huan Liu, and in particular general chair X. Sean Wang, for their outstanding support in the organization.

We would also like to thank the program committee: Krisztian Balog, Shlomo Geva, Djoerd Hiemstra, Noriko Kando, Ray Larson, Liz Liddy, Maarten Marx, Peter Mika, Patrick Pantel, Livia Polanyi, Ralf Schenkel, Andrew Trotman, Roman Yangarber, and the three program chairs.

Final thanks are due to the paper authors, the invited speakers Peter Mika and Silviu Cucerzan, 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/esair14/`. The contributed papers are available online at `http://dl.acm.org/citation.cfm?id=2663712`.

References

[1] S. Cotelo, A. Makowski, L. Chiruzzo, and D. Wonsever. Documents search using semantics criteria. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 5–7, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666187`.

[2] S. P. Cucerzan. Linking to web knowledge bases and applications to web search. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 3–3, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666199`.

[3] T. De Nies, C. Beecks, W. De Neve, T. Seidl, E. Mannens, and R. Van de Walle. Towards named-entity-based similarity measures: Challenges and opportunities. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 9–11, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666194`.

[4] V. Deolalikar. Can corpus similarity-based self-annotation assist information retrieval? In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 13–15, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666191`.

[5] Y. Ibrahim, M. Amir Yosef, and G. Weikum. Aida-social: Entity linking on the social stream. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 17–19, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666185`.

[6] E.-E. Jan, K.-Y. Chen, and T. Ide. A probabilistic concept annotation for it service desk tickets. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 21–23, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666193`.

[7] Z. Jiang, M. Chen, and X. Liu. Semantic annotation with rescoredesa: Rescoring concept features generated from explicit semantic analysis. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 25–27, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666192`.

[8] Z. Li, P. Exner, and P. Nugues. Using semantic role labeling to predict answer types. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 29–31, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666186`.

[9] J. Mao and K. Lu. Leverage the associations between documents, subject headings and terms to enhance retrieval. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 33–35, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666195`.

[10] P. Mika. Semantic search at yahoo! In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 1–1, New York, NY, USA, 2014. ACM. URL `http://doi.acm.org/10.1145/2663712.2666198`.

[11] M. Verma and D. Ceccarelli. Bringing head closer to the tail with entity linking. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in*

Information Retrieval, ESAIR '14, pages 37–39, New York, NY, USA, 2014. ACM. URL http://doi.acm.org/10.1145/2663712.2666196.

[12] H. Yang. A fragment-based similarity measure for concept hierarchies and ontologies. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 41–42, New York, NY, USA, 2014. ACM. URL http://doi.acm.org/10.1145/2663712.2666188.

[13] G. Zuccon, B. Koopman, and P. Bruza. Exploiting inference from semantic annotations for information retrieval: Reflections from medical ir. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, pages 43–45, New York, NY, USA, 2014. ACM. URL http://doi.acm.org/10.1145/2663712.2666197.