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

CIKM 2013 Workshop

Paul N. Bennett
Microsoft Research

Evgeniy Gabrilovich
Google

Jaap Kamps
University of Amsterdam

Jussi Karlgren
Gavagai Stockholm

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.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords: Knowledge resources, Non-topicality, Semantic annotation

1. THEME AND TOPICS

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.

2. OBJECTIVES, GOALS, AND OUTCOME

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.

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

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
ACM 978-1-4503-2263-8/13/10.
http:/dx.doi.org/10.1145/2505515.2505808.

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.

4. FORMAT

We start 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 have three keynotes that help frame the problem, and create a common understanding of the challenges: Kevyn Collins-Thompson (University of Michigan); Marti A. Hearst (University of California, Berkeley); and Dan Roth (University of Illinois at Urbana-Champaign). We continue with a boaster/poster session, where the papers from Section 3 are presented. The poster session continues over lunch. After lunch, we have break-out sessions in parallel that focused on specific aspects or problems related to the four themes. After the afternoon coffee, we have 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 will continue with a more informal part, over drinks and dinner with all attendees of the workshop.

Acknowledgments We thank Google for sponsoring the workshop.

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.