WORKSHOP REPORT

Exploiting Semantic Annotations in Information Retrieval – ESAIR'08

Omar Alonso
A9.com
Palo Alto, CA
oalonso@a9.com

Hugo Zaragoza
Yahoo! Research
Barcelona, Spain
hugoz@yahoo-inc.com

Abstract

The following report summarizes the highlights of the first workshop on exploiting semantic annotations in information retrieval (ESAIR'08). The workshop format included paper and demo presentations as well as breakout sessions and a panel discussion.

1 Introduction

The goal of this workshop was to create a forum for researchers interested in the use of semantic annotations for information retrieval. By semantic annotations we refer to linguistic annotations (such as named entities, semantic classes, etc.) as well as user annotations such as microformats, RDF, tags, etc. We were not interested in the annotations themselves, but on their application to information retrieval tasks such as ad-hoc retrieval, classification, browsing, textual mining, summarization, question answering, etc.

In the recent years there has been a lot of discussion about semantic annotation of documents. There are many forms of annotations and many techniques that identify or extract them. As NLP tagging techniques mature, more and more annotations can be automatically extracted from free text. In particular, techniques have been developed to ground named entities in terms of geo-codes, ISO time codes, Gene Ontology ids, etc. Furthermore, the number of collections, which explicitly identify entities, is growing fast with Web 2.0 and Semantic Web initiatives.

Despite the growing number and complexity of annotations, and despite the potential impact that these may have in information retrieval tasks, annotations have not yet made a significant impact in Information Retrieval research or applications. Further research is needed before annotations become truly useful for retrieval. This research is necessarily multi-disciplinary, encompassing areas such as information retrieval, natural language processing, web mining, semantic web, etc. Furthermore, for this research to be effective, researchers from the different areas need to share and agree on goals, tools and evaluation methodology.

We asked a number of leading researchers interested in search with annotated data[1] if they thought a workshop with such as title would be interesting to them, and their feedback was unanimously positive. They all agreed that it made sense to group researchers around this topic, crossing the disciplinary borders that often separate us. Furthermore they provided great feedback which allowed us to fine-tune the workshop focus and call for papers.

The call for papers was designed to encourage the submission of early and novel work, without any restrictions on format or length, and de-emphasizing evaluation and reproducibility. Our objective was to get a sample as large as possible of the different lines of work in which researchers on this topic are involved. The workshop format was designed to foster audience participation. The morning session had most of the paper presentations and the rest of the day was dedicated to demos, a break out session and a final panel. We also created a Wiki page for the workshop, which was opened to workshop participants to comment on other papers, as there were presented, suggest topics for discussion in the afternoon sessions, and add general comments and feedback[2].

2 Presentations

We had a wide range of topics presented as papers, from traditional use of NLP technologies to enhance retrieval to novel uses of social metadata for navigation and browsing. Due to lack of space we cannot review all the papers here, but we list the titles and authors bellow to give you a flavor of the presentations (papers are available online[1]):

- *Training-less Ontology-based Text Categorization* by Maciej Janik and Krys Kochut.
- *Optimizing single term queries using a personalized Markov random walk over the social graph* by Maarten Clements, Arjen P. de Vries, Marcel J.T. Reinders.
- *Collaborative Annotation for Pseudo Relevance Feedback* by Christina Lioma, Marie-Francine Moens and Leif Azzopardi.
- *Web Search Disambiguation by Collaborative Tagging* by Ching-man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt.
- *Introducing Triple Play for Improved Resource Retrieval in Collaborative Tagging Systems* by Rabeeh Ayaz Abbasi and Steffen Staab.
- *Keyword Suggestion Using Concept Graph Construction from Wikipedia Rich Documents* by Hadi Amiri, Abolfazl AleAhmad, Masoud Rahgozar, Farhad Oroumchian.
- *Annotation of Scientific Summaries for Information Retrieval* by Fidelia Ibekwe-SanJuan, Silvia Fernandez, Eric SanJuan, Eric Charton.
- *A Combined Method of Frequency & Markup Analysis for Terminological Ontologies* (demo) by Roman Schneider.

3 Interactive Session

For the interactive session we asked each participant to suggest i) three application areas which may benefit from semantic annotations, and ii) three theoretical problems that constitute a bottleneck to

[1] ESAIR'2008 PC committee: John Atkinson, Jamie Callan, Arjen de Vries, Michael Gertz, Marko Grobelnik, Peter Jackson, Aaron Kaplan, Mounia Lalmas, Hang Li, Peter Mika, Inderjeet Mani, Mark Stevenson, Anne-Marie Vercoustre.

[2] `http://www.yr-bcn.es/esair08`

further progress in this area. This lead to a number of topics, out of which the following were chosen as most interesting further discussion in smaller groups:

- Industrial: main interests in building products and exploration of new verticals (4 participants.)
- Search: main interests in search interaction, models, and discovery (8 participants.)
- Social search: main interests in how to use the social aspect and folksonomies (5 participants.)
- Natural Language Processing: main interest in the NLP side of semantics (5 participants.)

These break-out sessions in small, coherent groups were greatly appreciated by the workshop participants. After the small parallel sessions, a member of each group gave an overview of their discussion to the rest of the workshop participants. We provide here an overview of their comments:

Industrial: As expected this group was heavy into tools (and lack of), integration, and benefits of using named-entity extraction and semantics for different search vertical needs. The main themes identified:

- The group identified the need for a good tool or platform that allows good quality of extracted annotations with rich extensibility features and training sets.
- Semantics or a WordNet equivalent for product and services.
- How annotations impact advertisement and other revenue models?

Search: According to the group leader, there were more questions than answers. That said, here is an outline of the main themes:

- The most challenging issue for semantic search and discovery is to find a general model or solutions, which can be applied in a majority of cases.
- It is important to be realistic and thorough when evaluating the "success" of work in this area, since too much has been claimed in the past with few results.
- People may benefit from exposure of semantics in the results, but these cannot be strict or formally defined. We associate facts, but out associations are fuzzy.

Social search: The group concentrated in trying to identify scenarios where social search could be beneficial. They described the following three:

- Searching relevant multimedia content from a search engine.
- Getting to know other people and using other people to find interesting information.
- Find interesting events or pictures at a certain location at a certain time.

NLP: The group discussed the ability to use NLP in tasks like content analysis, fact-finding, question answering, inference, translation, and sense making. The main topics discussed were:

- Applications: a) Text-based applications such as in IR, TDT, medical domains, language translations and text summarization and b) dialogue based applications such as in question answering, education and teaching systems and speech recognition.
- Techniques: semantic grammars, syntactically driven parsing and pattern matching
- Tools: GATE, Stanford parser, CGParser (a linear parser of conceptual graphs).

4 Results and Future Directions

It was very interesting to see different views on how annotations can be used for search. Furthermore, we found that despite the wide range of disciplines in which researchers are involved, there was a high degree of agreement on what was needed to advance. We summarize these views bellow:

1. Identifying key tasks and their evaluation. It is unclear if there are tasks that are general enough to be of interest to a large part of the community, while remaining specific enough to make them realistic and interesting. Furthermore, it is unclear if a proper evaluation framework could be set up for such tasks to foster competition and reproducibility of results. This seems key for the development of scientific theories with some generality.

2. Data availability. Having access to large annotated data sets is problematic for several reasons. First, some of this data is proprietary and not easily accessible (for example, social annotations). Even when it is available, it is important to agree on a particular version of the data in order to compare results. This is hard because online data changes fast, it needs to be preprocessed in many ways, etc. Finally, due to the interdisciplinary nature of this area, it is often hard for small groups to use all the different tools necessary to preprocess, annotate and index the data. For this reason, it is seen as very favorable to share collections and to do so in several pre-processing stages.

3. Tool availability. Similarly, any project that involves semantic annotations will require using a number of technologies from different disciplines (NLP, web mining, semantic web, etc). It is important to share information about the basic building blocks, their advantages and disadvantages, and whatever necessary to speed-up the learning and set-up phase.

These are difficult issues that cannot be solved in an afternoon. However, it was encouraging to see such a high level of agreement on the main difficulties that we face. Many researchers in the workshop expressed an interest to work together on these issues and start sharing tools, data and problems. The exact vehicle for this cooperation has not yet been decided, but we plan to continue the discussion through the ESAIR wiki page.

5 Acknowledgements

We would like to thank the participants for their active participation during the breakout sessions and for placing excellent comments on the Wiki as the workshop was going on. Last, but not least, a big thank to the program committee who were able to provided detailed feedback in a very short period of time.