Report on the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '15)

Krisztian Balog
University of Stavanger
Norway
krisztian.balog@uis.no

Jeffrey Dalton
Google Research
USA
jeffdalton@google.com

Antoine Doucet
University of La Rochelle
France
antoine.doucet@univ-lr.fr

Yusra Ibrahim
Max Planck Institute for Informatics
Saarbrücken, Germany
yibrahim@mpi-inf.mpg.de

Abstract

The amount of structured content published on the Web has been growing rapidly, making it possible to address increasingly complex information access tasks. Recent years have witnessed the emergence of large scale human-curated knowledge bases as well as a growing array of techniques that identify or extract information automatically from unstructured and semi-structured sources. The ESAIR workshop series aims to advance the general research agenda on the problem of creating and exploiting semantic annotations. The eighth edition of ESAIR took place at CIKM 2015 in Melbourne, Australia, on the 23rd of October. Having a special focus on applications, we dedicated an "annotations in action" track to demonstrations that showcase innovative prototype systems, in addition to the regular research and position paper contributions. The workshop also featured invited talks from leaders in the field. This report presents an overview of the event and its major outcomes.

1 Introduction

Over the recent years, major search engines have redesigned their search results to accommodate semantic information, including rich search results and direct question answering. They have also developed large-scale entity repositories, such as Google's Knowledge Graph, Yahoo!'s Web of Objects, and Bing's Satori. At the heart of these efforts is large-scale semantic annotation of both queries and documents.

The goal of the ESAIR workshop series is to create a forum for researchers interested in the 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.). ESAIR'15 continued, with

a renewed group of organizers, on the path set by the previous edition(s) of the workshop: clarifying the exact role of semantic annotations in supporting complex search tasks.

Annotations come in a variety of flavors (named entities, temporal information, geo-positional markers, semantic roles, sentiment, etc.) and there is a growing repertoire of tools and techniques available for extracting these annotations automatically from text. The question then presents itself: what, if anything, is missing? We seek to answer this question by focusing on applications that are rooted in specific, real-world use cases. To further sharpen this application-oriented focus, the workshop featured a dedicated "annotations in action" track for presenting innovative prototype systems. A Best Demonstration Award, sponsored by Google, was presented to the participant(s) with the most outstanding demo at the workshop.

The workshop called for regular papers (4+ pages), position papers (2+1 pages), and demo papers (4 papers), which were each reviewed by at least 3 members of the program committee. The committee was keen on selecting papers that introduce novel ideas and will stimulate discussion in the workshop. The accepted contributions cover a wide spectrum of topics related to semantic annotations and information retrieval, including: healthcare, automatic translation, semantic search, question answering, knowledge profiling, topic modeling, contextual enrichment, named entity disambiguation, temporal tagging, and semantic annotation of video content. Out of the 19 submissions, the committee accepted a total of 10 papers (5 regular, 2 position, and 3 demo papers) to be presented at the workshop. In addition to the regular paper presentations and demonstration session, the programme included invited talks by Maarten de Rijke (University of Amsterdam) and Matthew Kelcey (Google, Inc.). It was our aim to inspire attendees to think "outside the box" and to invite everyone to contribute. Thus, the workshop included a focused discussion sessions at which researchers with similar interests were able to showcase their ideas and concerns. Participants were divided into smaller groups based on their areas of interest; then, each group was asked to come up with a breakthrough application that utilizes semantic annotations. A final wrap-up session concluded the event at which each discussion group presented their application in a five minutes sales-pitch. The discussions were effective in a way that helped participants to formulate critical insights into the potential of semantic annotations, the barriers to success, and specific steps to take this research forward with useful applications. We continued with the tradition of earlier ESAIR editions and organized a social event for further, more informal discussions among workshop participants and other CIKM attendees.

The list of accepted contributions, corresponding resources (slides, links, etc.), and all other outcomes of the workshop are available at `http://esair.org`.

2 Keynotes

2.1 Semantic Entities (Maarten de Rijke)

The first keynote talk of the workshop was given by Maarten de Rijke (University of Amsterdam) on "semantic entities" [16]. Maarten set the stage by emphasizing the importance of entities in terms of query volume, but also in terms of "attention volume" (large number of entity-oriented papers and industry talks at CIKM this year).

First, he presented an unsupervised, discriminative model with efficient inference capabilities for expertise retrieval [17]. The proposed log-linear model is similar to neural language

models, with the difference that it predicts a candidate expert instead of the next word. This model improves retrieval performance compared to vector space-based and generative language models, mainly due to its ability to perform semantic matching.

Next, Maarten addressed the problem of organizing information around entities through *entity aspects*: attributes, actions, or topics [13]. These aspects may be displayed on a result page or "entity card," may provide direct action or support for complex search tasks, or could be used to inform knowledge base design/construction. This general problem can be subdivided into three specific tasks: mining ("What do people search for in the context of an entity?"), ranking ("How should we rank aspects?"), and recommending ("Given an aspect, which related aspects should we recommend?"). Entity aspects are mined from query logs through entity linking and clustering context segments; ranking is done in a query-independent fashion; related aspects are recommended based on semantic and behavioral relatedness (query-flow graph).

The next part of the presentation focused on *entity relations*. Entity relationships are represented in knowledge graphs using formal descriptions, which are not suitable for presenting to the end user. The task ahead, then, is to explain the relationships using human-readable descriptions. This is approached as a sentence retrieval problem. Given an entity pair and a relationship, candidate sentences are first extracted and enriched (by applying coreference resolution and entity linking), then ranked by how well they describe the relationship of interest (using a learning-to-rank approach, where entity and relationship features are found to be the most important) [18].

Maarten then turned to the issue of *entity representations*, in particular, how to update the description of long tail entities with information from a range of sources [8]. The proposed method represents entities as fielded documents, using information from knowledge bases, the Web, Twitter, as well as user queries. The method adapts to changing entity representations by adding new descriptions in real time and learning the best representation (weights for combining content from different sources), based on implicit feedback, as time evolves.

Now that we have all that information around an entity, what is next? Maarten discussed some early work on *narrative search*: teaching the search engine to "tell a story" (by learning templates for SERP generation and learning to re-order elements on the SERP). He concluded his talk by highlighting that there is still a lot to do at the "usual" level of semantics (entities, aspects, relations). However, one should not stop there, but should also look into issues at a higher level: entity SERPs and narratives.

2.2 Open and Closed Schema for Aligning Knowledge and Text Collections (Matthew Kelcey)

The second keynote talk of the workshop was given by Matthew Kelcey [10], (Google Research). In the first part of the talk Matthew compared and contrasted "closed schema" versus "open schema" knowledge collections. In the second part of his talk, he outlined reasons and methods for reconciling the collections. Matthew's presentation concluded with applications of these knowledge collections to the task of question answering.

The main portion of Matthew's presentation contrasted two different approaches to representing and extracting knowledge. He first described differences in knowledge representation. The first approach he described as closed schema. A closed schema is similar to a knowledge base like Freebase, where knowledge is represented as (subject, predicate, object) triples. It is symbolic and the schema, predicates, and entity types are typically manually curated. A

closed schema typically represents objective facts. In contrast, an open schema representation is based on (subject, relation, object) triples represented as text chunks. In an open schema, the representation is drawn from sentences that generate triples. In this part of the presentation Matthew was clearly showing the limitations of closed schema approach as somewhat brittle and requiring significant effort to maintain. He continued to compare them across several other dimensions.

Beyond representation Matthew contrasted the closed versus open schema across dimensions including: tools, subjectivity, popularity, trustworthiness, and freshness. First, he contrasted the typical tools used (structured query systems vs. unstructured with NLP). The issue of subjectivity and popularity—the closed schema typically represents an objective fact once, an open schema represents statements (that are inherently subjective) observed across all mentions. He contrasted reliability—typically clean and human curated knowledge in a closed schema versus noisy facts from a variety of sources of varying trustworthiness in an open schema. Lastly, he touched on freshness—closed schema knowledge bases are typically staler than open schema approaches.

In the second part of the presentation Matthew addressed the problem of aligning closed and open schema knowledge representations. He argued that it was important to align them because they are closely related to each other and hold complementary information. First, the information from open schema can be used to identify and fill in knowledge gaps in a closed knowledge base, especially for tail entities. A closed schema can also be used to improve open schema knowledge bases by providing confidence in the open schema and adding trust to facts that can be aligned.

To align relationships between open and closed schema Matthew presented two approaches, statistical and distributed neural representations. The first way to align a predication and relation in the schemas is one based on set overlap statistics. A simple and effective one is the Jaccard overlap. The challenge here is determining equality between (subject, object) pairs; the entities need to be matched between symbolic and text representations. The second approach Matthew presented is to used a deep neural network architecture. In this architecture the network is passed in both closed $< s, p, o >$ and open $< s, r, o >$ examples at the same time. This allows for learning similarity between predicates and relations in an embedding space, and supports similarity search in both directions.

Matthew concluded with applications of open schema knowledge bases to retrieval tasks, in particular open domain question answering. He provided an overview of the paper by Bordes et al. [2] on open question answering. In particular, he emphasized the needs for lots of data—one of the main things he highlighted was the approach to generate large quantities of weakly supervised data with pseudo question-answer pairs using open schema patterns from a large web text collection, ClueWeb09. Another aspect he highlighted was finding cheap paraphrase data in the wild from WikiAnswers. The result is that open schema knowledge bases can be used to effectively answer complex questions.

3 Presented Papers

3.1 Full Papers

The first paper, entitled *Applying Semantic Web Technologies for Improving the Visibility of Tourism Data* [15], was presented by Fayrouz Soualah-Alila and discusses the problem of

representing touristic information in an ontology. It describes a solution that integrates the existing TourInFrance (TIF) ontology with schema.org. With the goal to generate tourism dashboards, the authors were confronted to the fact that the numerous tourism actors each adapted the TIF standard to their own needs. This effectively made it impossible to render their systems interoperable, creating the need to provide methods and concepts facilitating the effective integration of such touristic information originating from various sources. The paper describes such a technique and provides the corresponding downloadable ontology.

The second paper, *Contextualizing Data on a Content Management System* [12], was presented by Lara Santos. It addresses the problem of bringing semantic context to the contents of a content management system (CMS), with a focus application on information retrieval. There, context information is an especially important addition to improve performance, therefore user satisfaction. The authors built contextualized data in three steps, namely, language identification, keyphrase extraction, and domain expansion. The last two steps were conducted using Wikipedia, so as construct sets of expanded terms, to be updated periodically using a scheduler. Early experiments over a Portuguese data set showed that adding context triggered an improvement in terms of recall performance.

The third paper, *Harnessing Semantics for Answer Sentence Retrieval* [4], presented by Ruey-Cheng Chen, deals with the problem of anwser passage retrieval, a specialized question answering task that looks for non-factoid, multiple-sentence answers from the Web, such as "what was the role of Portugal in word war II?". Using a learning-to-rank retrieval setting, the authors experiment with two semantic approaches—Explicit Semantic Analysis (ESA) and Word2Vec—for finding non-factoid answers. The results showed that combining both ESA and Word2Vec led to slight improvements and indicated that ESA and Word2Vec tended to have different effects to retrieve different relevant sentences, suggesting that more adequate combination might provide stronger improvement.

Finally, the fourth paper, *Named Entity Disambiguation for Resource-Poor Languages* [7], presented by Mohamed Amir Yosef, proposes an approach for named entity disambiguation (NED) in languages for which linguistic resources are scarce. This can mean, for instance, that some of the required language-specific resources are non-existent or insufficient: named-entity dictionaries, named-entity descriptions, annotated training corpora, etc. The authors propose to compensate for this lack of resources by leveraging resources available in English, and exploiting them using Statistical Machine Translation (SMT) to translate English resources into the target language. With the running case of Arabic and further experiments in Italian and Spanish, they built and tested AIDA-ML++, their language-agnostic NED system, which improved the state-of-art in terms of disambiguation precision.

3.2 Position Papers

The first position paper, *Knowledge-Driven Video Information Retrieval with LOD: From Semi-Structured to Structured Video Metadata* [14], was presented by Leslie Sikos. The paper presents a comparison between different state-of-the-art video annotations (namely, Dublin Core, MPEG-7, MPEG-21, NewsML, TTML, and TV-Anytime), shows the limitations of each of them, and suggests possible solutions to overcome these limitations. The study further identifies the downsides of using text-based manually added video annotations, such as user added tags in Youtube videos; this suffers from tag variations, polysemy, misspelling and ambiguity. Thus, machine interpretable metadata is desired, but it has structural complexity; this can be targeted using a well-documented standard ontology. The paper concluded that

there is a need to adopt RDF-based multimedia annotations in Content Management Systems and video editing tools.

The paper titled *Temporal Reconciliation for Dating Photographs Using Entity Information* [11] was presented by Antoine Doucet. This work aims to find temporal information for photographs. Finding the date at which a photograph was taken is challenging, as there might be different dates mentioned in the surrounding context. Moreover, the date of the photograph might not be explicitly stated but rather to be inferred from the text describing the photograph. In such cases, finding the date of the photograph is difficult. The paper presents a novel approach which relies on the entity mentions in the context to deduce the date of the photo. The proposed technique starts by extracting photos from web documents, their relative positions, and the surrounding text. Then, it extracts all the temporal information that can be found on the web page. After that, it extracts all the entity mentions from the surrounding text, links them to YAGO knowledge base, and extracts all the temporal relations associated with them. Eventually, it creates a range of possible dates out of the extracted data, from which the photograph's date is deduced. The proposed technique is evaluated on a corpus of photographs extracted over the period from 1820 to 2010. The initial results show 61% accuracy in detecting the year of the photographs.

4 Presented Demos

The first demo, *An Interface Sketch for Queripidia: Query-driven Knowledge Portfolios from the Web* [5], was presented by Michael Schuhmacher. Queripidia is a web-based application that creates a knowledge portfolio for a given query. A knowledge portfolio is defined as a collection of entities associated with snippets of text, to show the relation between the entity and the given query. The application aims to enhance entity search by providing explanations for each retrieved entity. Queripidia uses ClueWeb12[1] with the FACC1 [6] annotations to extract snippets of text to display with each entity. It starts by retrieving a ranked list of entities for the given query. Then, for each entity, it extracts support passages from ClueWeb12 using the FACC1 entity annotations. This way Queripidia can augment the description of entities found on Wikipedia which, according to the authors, is insufficient in 50% of the cases.

The second demo, titled *CADEminer: A System for Mining Consumer Reports on Adverse Drug Side Effects* [9], was presented by Sarvnaz Karimi. CADEminer aims to discover unreported drug side effects by mining online forums. First, it extracts the mentions of relevant concepts and links them to an ontology (SNOMED CT or AMT). Then, it extracts the relations between those concepts and a possible drug side effect. After that, it filters out the known drug side effects. Further, an interface is provided to visually display the potential side effects along with some statistics. CADEminer acts as a surveillance tool and may be used by regulatory agencies, pharmacies, or drug producers to uncover adverse drug reactions reported on social media.

Finally, Ben Hachey presented *Hugo: Entity-based News Search and Summarisation* [3]. Hugo is a mobile application (available on iOS) that allows users to track entities that are of interest to them in the news. It acts as a personal assistant that aims to provide the user with a digest of recent relevant news related to the selected entities of interest (defined by name, organization, or location). Entity linking and summarization are performed offline for

[1] http://lemurproject.org/clueweb12

the news articles. In addition, Hugo makes sure to retrieve a diverse set of news articles. Hugo is primarily designed for business users who want to get quick, reliable, and up-to-date information about a person before a meeting, or news about other competitors.

4.1 Best Demo Award

ESIAR'15 presented a Best Demo Award, which was given to *Hugo: Entity-based News Search and Summarisation* [3]. The award came with a cash prize of 500$, generously sponsored by Google.

5 Discussion Session

In the afternoon, the workshop included a breakout session on the future of semantic annotations. After an initial round of collecting ideas, we identified four main application areas:

- Question answering and dialog systems
- Health and medicine
- Personal knowledge graph
- eGovernement applications

Participants were split into four groups and each group was asked to propose a new startup idea or research proposal that would represent a significant multi-year effort. In particular, one prompt question focused on coming up with new classes of semantic annotations.

Some of the motivations the groups developed were quite interesting. For example, the dialog focus group came up with the idea to create topical and personalized chat-bots. Examples included national chat-bots to dialogue about large-scale political issues, as well as a personal movie chat-bot that could be invited to recommend and describe the buzz on the latest movies. The health group was motivated by poor-doctor patient experience and proposed annotations of one's vitals, diet, and symptoms that could be summarized and shared with doctors. One surprising application was the local eGovernment focus group, driven by a delegation from the Kenyan government. Their goal was to increase transparency and empower citizens by making government data more freely available. One of the themes across multiple groups was an emphasis on multimodal annotations of personal data, such as images and sensors (fitness and activity) available from mobile devices.

6 Conclusions

ESAIR'15 succeeded in bringing researchers together that are interested in semantic annotations across different domains and in directing their focus to application-oriented use-cases. The papers presented in the workshop covered a broad range of domains, including multimedia, e-health, e-government, social media, question answering, and entity disambiguation. The two excellent keynotes by Maarten de Rijke and Matthew Kelcey gave insights into research trends, both in academia and in industry; these talks were stimulating for the discussions we had towards the end of the workshop. Overall, the workshop offered a highly

interactive environment with lively discussions throughout the day, which then also continued over the by now traditional ESAIR social event. We feel that there is still a lot of ground to cover in terms of applications, and we aim to continue this investigation at ESAIR'16.

7 Acknowledgements

We are grateful for the generous support we have received from Google, Inc. Google sponsored the Best Demonstration Award and the social event. We would also like to thank CIKM for hosting the workshop. Final thanks are due to the program committee,[2] paper authors and workshop attendees, without whom the workshop would not have been the success it was.

References

[1] K. Balog, J. Dalton, A. Doucet, and Y. Ibrahim, editors. *ESAIR '15: Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval*, New York, NY, USA, 2015. ACM.

[2] A. Bordes, J. Weston, and N. Usunier. Open question answering with weakly supervised embedding models. *CoRR*, abs/1404.4326, 2014. URL `http://arxiv.org/abs/1404.4326`.

[3] A. Cadilhac, A. Chisholm, B. Hachey, and S. Kharazmi. Hugo: Entity-based news search and summarisation. In Balog et al. [1], pages 51–54.

[4] R.-C. Chen, D. Spina, W. B. Croft, M. Sanderson, and F. Scholer. Harnessing semantics for answer sentence retrieval. In Balog et al. [1], pages 21–27.

[5] L. Dietz and M. Schuhmacher. An interface sketch for queripidia: Query-driven knowledge portfolios from the web. In Balog et al. [1], pages 43–46.

[6] E. Gabrilovich, M. Ringgaard, and A. Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0), June 2013.

[7] M. H. Gad-Elrab, M. A. Yosef, and G. Weikum. Named entity disambiguation for resource-poor languages. In Balog et al. [1], pages 29–34.

[8] D. Graus, M. Tsagkias, W. Weerkamp, E. Meij, and M. de Rijke. Dynamic collective entity representations for entity ranking. In *WSDM 2016: The 9th International Conference on Web Search and Data Mining*, 2016.

[9] S. Karimi, A. Metke-Jimenez, and A. Nguyen. Cademiner: A system for mining consumer reports on adverse drug side effects. In Balog et al. [1], pages 47–50.

[10] M. Kelcey. Open and closed schema for aligning knowledge and text collections. In Balog et al. [1], pages 3–3.

[2] http://esair.org/esair15-cfp/

[11] P. Martin, M. Spaniol, and A. Doucet. Temporal reconciliation for dating photographs using entity information. In Balog et al. [1], pages 39–41.

[12] C. Moreira, J. a. Taborda, R. Del Gaudio, L. dos Santos, and P. Pereira. Contextualizing data on a content management system. In Balog et al. [1], pages 11–19.

[13] R. Reinanda, E. Meij, and M. de Rijke. Mining, ranking and recommending entity aspects. In *SIGIR 2015: 38th international ACM SIGIR conference on Research and development in information retrieval*, 2015.

[14] L. F. Sikos and D. M. Powers. Knowledge-driven video information retrieval with lod: From semi-structured to structured video metadata. In Balog et al. [1], pages 35–37.

[15] F. Soualah-Alila, C. Faucher, F. Bertrand, M. Coustaty, and A. Doucet. Applying semantic web technologies for improving the visibility of tourism data. In Balog et al. [1], pages 5–10.

[16] C. Van Gysel, M. de Rijke, and M. Worring. Semantic entities. In Balog et al. [1], pages 1–2.

[17] C. Van Gysel, M. de Rijke, and M. Worring. Unsupervised, efficient and semantic expertise retrieval. In *WWW 2016: 25th International World Wide Web Conference*, 2016.

[18] N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke, and W. Weerkamp. Learning to explain entity relationships in knowledge graphs. In *ACL-IJCNLP 2015*, 2015.