KEYSTONE WG3: Activities and Results Overview on User Interaction

Omar Boucelma(✉)

Aix-Marseille University, CNRS, ENSAM, Toulon University,
LSIS UMR 7296, 13397 Marseille, France
omar.boucelma@univ-amu.fr

1 WG3 Objectives

User Interaction WG investigates issues related to the semantic disambiguation of the queries based on the context and on the keyword annotations with respect to some reference ontologies, the development of languages for keyword searching and the use of users' feedbacks for improving results. Moreover, the WG studies techniques for identifying the "scope" of a keyword query, i.e. determining what are the data source elements to be returned to the user and in which form (e.g., in a graphical way).

WG3 is composed of 50 members who conduct research in various fields. At the Keystone winter 2017 meeting[1], WG3 participants enumerate three main research topics categorized as follows:

1. Information Retrieval/Natural Language Processing
   - Natural language disambiguation
   - Named entity recognition
   - Keyword query cleaning
   - Semantic relatedness
   - Exploratory Search
2. Databases
   - Keyword search over Relational Data/Linked Data
   - Examplar query
   - Query augmentation
3. Machine Learning/Information Extraction
   - Query disambiguation
   - Document annotation

In the sequel we highlight some of WG3 publications.

[1] http://www.keystone-cost.eu/keystone/6th-mc-meeting-and-winter-2017-wg-meeting/.

J. Szymański and Y. Velegrakis (Eds.): IKC 2017, LNCS 10546, pp. 224–229, 2018.
https://doi.org/10.1007/978-3-319-74497-1_22

2 Review of Selected Papers

2.1 Improving Document Retrieval in Large Domain Specific Textual Databases Using Lexical Resources

2.2 Selectivity-Based Keyword Extraction Method

In [4] authors propose a novel Selectivity-Based Keyword Extraction (SBKE) method, which extracts keywords from the source text represented as a network. The node selectivity value is calculated from a weighted network as the average weight distributed on the links of a single node and is used in the procedure of keyword candidate ranking and extraction. Authors show that selectivity-based keyword extraction slightly outperforms an extraction based on the standard centrality measures: in/out-degree, betweenness and closeness. Therefore, they include selectivity and its modification - generalized selectivity as node centrality measures in the SBKE method. Selectivity-based extraction does not require linguistic knowledge as it is derived purely from statistical and structural information of the network. The experimental results point out that selectivity-based keyword extraction has a great potential for the collection-oriented keyword extraction task.

In [9], large collections of textual documents represent an example of big data that requires the solution of three basic problems: the representation of documents, the representation of information needs and the matching of the two representations. This paper outlines the introduction of document indexing as a possible solution to document representation. Documents within a large textual database developed for geological projects in the Republic of Serbia for many years were indexed using methods developed within digital humanities: bag-of-words and named entity recognition. Documents in this geological database are described by a summary report, and other data, such as title, domain, keywords, abstract, and geographical location. These metadata were used for generating a bag of words for each document with the aid of morphological dictionaries and transducers. Named entities within metadata were also recognized with the help of a rule-based system. Both the bag of words and the metadata were then used for pre-indexing each document. A combination of several tf_idf based measures was applied for selecting and ranking of retrieval results of indexed documents for a specific query and the results were compared with the initial retrieval system that was already in place. In general, a significant improvement has been achieved according to the standard information retrieval performance measures, where the InQuery method performed the best.

2.3 Uncertainty Detection in Natural Language

Designing approaches able to automatically detect uncertain expressions within natural language is central to design efficient models based on text analysis, in particular in domains such as question-answering, approximate reasoning, knowledge-based population. In [6], authors, first, review several contributions and classifications defining the concept of uncertainty expressions in natural language, and the related detection methods that have been proposed so far. Then, they introduce a new supervised and generic approach for detecting uncertainty. The approach is based on the statistical analysis of multiple lexical and syntactic features used to characterize sentences through vector-based representations that can be analyzed by proven classification methods. The global performance of the approach is demonstrated and discussed with regard to various dimensions of uncertainty and text specificities.

2.4 Disambiguation of User Sentiment

In [1], usage of an opinion process mining method ABSA (Aspect Based Sentiment Analysis) is described. In ABSA, texts are analyzed to extract the sentiments that their authors express towards certain features and characteristics of particular entities, such as products or persons. Key role in the effectiveness of this process plays the accurate and complete identification of the entities' discussed aspects within the text, as well as of the evaluation expressions that accompany these aspects. Nevertheless, what entities may be considered as aspects and what evaluation expressions may characterize them, depends largely on the domain at hand. With that in mind, in this paper we propose an approach for representing and populating semantic lexicons that contain domain-specific aspect-evaluation-polarity relations and, as such, can be (re-)used towards more effective ABSA in concrete domains and scenarios.

2.5 Collective Intelligence for Exploratory Keyword Search

In [10], authors address an exploratory search challenge by presenting a new (structure-driven) collaborative filtering technique. The aim is to increase search effectiveness by predicting implicit seeker's intents at an early stage of the search process. This is achieved by uncovering behavioral patterns within large datasets of preserved collective search experience. Authors apply a specific tree-based data structure called a TB (There-and-Back) structure for compact storage of search history in the form of merged query trails' sequences of queries approaching iteratively a seeker's goal. The organization of TB-structures allows inferring new implicit trails for the prediction of a seeker's intents. Experiments that have been conducted demonstrate both: the storage compactness and inference potential of the proposed structure.

2.6 Exploiting Linguistic Analysis on URLs for Recommending Web Pages: A Comparative Study

In this paper, [5], authors analyze and compare three different approaches to leverage information embedded in the structure of web sites and the logs of their web servers to improve the effectiveness of web page recommendation. They propose to exploit the context of the users' navigations. These approaches do not require either information about the personal preferences of the users to be stored and processed, or complex structures to be created and maintained. The paper also reports some comparative experiments using a real-world website to analyze the performance of the proposed approaches.

2.7 Semantic Description of Liver Computerized Tomography Images

Semantic representations and querying are critical in interpretation of data, which is otherwise very difficult (if not impossible) to develop 'natural' and subjective (i.e. tailored to the users' needs/foci) user interfaces and to have an intuition about the results of any analysis and/or processing. Medicine is a special domain where such a semantic representation is a keystone in developing computerized methods due to the critical (and required) human component in all decision making processes.

The paper [7] is focused on demonstrating how an ontology can indeed be beneficial for such semantic processing. Authors developed ONLIRA (Ontology of the Liver for Radiology) and used it (as a sample application) to search for similar radiology reports. Then they studied the performance of searching the ontology-based radiology reports in comparison to searching free text reports using NLP techniques.

Experiments have been conducted on the basis of 30 radiology reports of different patients written in natural language and converted into ONLIRA instances. To highlight differences between two search/retrieval approaches (Keyword based vs. Semantic), five queries expressed in both (Description Logic) DL query and keywords, have been tested. To establish a gold standard, two board certified radiologists manually evaluated each query to decide which reports should be retrieved. Both approaches have been evaluated against the gold standard by comparing their precision and recall.

2.8 Keyword-Based Search of Workflow Fragments and Their Composition

Workflow specification, in science as in business, can be a difficult task, since it requires a deep knowledge of the domain to be able to model the chaining of the steps that compose the process of interest, as well as awareness of the computational tools, e.g., services, that can be utilized to enact such steps. To assist designers in this task, authors in [3], propose a methodology and an associated mechanisms for the specifications of scientific workflows by reusing workflow fragments. In particular, they show how semantic keyword search can be utilized to effectively identify the workflow fragments that are relevant for the composition of new workflow. They describe a methodology that consists in exploiting existing workflow specifications that are stored and shared in repositories, to identify workflow fragments that can be re-utilized and re-purposed by designers when specifying new workflows. Specifically, they present a method for identifying fragments that are frequently used across workflows in existing repositories, and therefore are likely to incarnate patterns that can be reused in new workflows. They describe a keyword-based search method for identifying the fragments that are relevant for the needs of a given workflow designer. They present an algorithm for composing the retrieved fragments with the initial (incomplete) workflow that the user designed, based on compatibility rules that have been identified, and showcase how the algorithm operates using an example from eScience.

2.9 Discovery and Recommendation of Web Services

A web services recommendation system is described in [8]. The salient idea of the paper is to go beyond traditional matchmaking techniques in taking leverage of the available information on objects (services and users) for structuring the ecosystem of web services as a heterogeneous multigraph where nodes (services and users) are connected by labeled edges having different semantics, i.e., similarity, popularity, follow and track relations, etc. A service may be recommended to a given user either as a response to his/her request or based on his/her profile. The contribution of this work is twofold: (i) the design of a multigraph model where intra-services, intra-users and inter services/users links are exhibited; (ii) a novel recommendation approach based multigraph search is proposed. A prototype has been implemented on top of a Neo4j graph database, enabling both keyword-based queries and graph analytics.

3 Conclusion

The first conclusion to draw is related to the COST EU instrument. Thanks to the flexibility and the bottom-up research approach promoted by this instrument, researchers from different countries and different disciplines had the opportunity to share their know-how.

From the research perspective, while Human-Computer Speech (HCI) is gaining momentum as a technique of computer interaction, we still need to get access to large unstructured datasets in order extract relevant information from them. As illustrated in [2], combining HCI with keyword-search is probably the next challenge to achieve.

References

1. Alexopoulos, P., Wallace, M.: Creating domain-specific semantic lexicons for aspect-based sentiment analysis. In: 10th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2015, Trento, Italy, 5–6 November 2015, pp. 1–6 (2015). https://doi.org/10.1109/SMAP.2015.7370083
2. Audhkhasi, K., Rosenberg, A., Sethy, A., Ramabhadran, B., Kingsbury, B.: End-to-end ASR-free keyword search from speech. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 March 2017, pp. 4840–4844 (2017). https://doi.org/10.1109/ICASSP.2017.7953076

3. Belhajjame, K., Grigori, D., Harmassi, M., Yahia, M.B.: Keyword-based search of workflow fragments and their composition. Trans. Comput. Collect. Intell. **26**, 67–90 (2017). https://doi.org/10.1007/978-3-319-59268-8_4
4. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Selectivity-based keyword extraction method. Int. J. Semantic Web Inf. Syst. (IJSWIS) **12**(3), 1–26 (2016)
5. Cadegnani, S., Guerra, F., Ilarri, S., del Carmen Rodríguez-Hernández, M., Lado, R.T., Velegrakis, Y., Amaro, R.: Exploiting linguistic analysis on urls for recommending web pages: a comparative study. Trans. Comput. Collect. Intell. **26**, 26–45 (2017). https://doi.org/10.1007/978-3-319-59268-8_2
6. Jean, P., Harispe, S., Ranwez, S., Bellot, P., Montmain, J.: Uncertainty detection in natural language: a probabilistic model. In: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, 13–15 June 2016, pp. 10:1–10:10 (2016). https://doi.org/10.1145/2912845.2912873
7. Kökciyan, N., Türkay, R., Üsküdarli, S., Yolum, P., Bakir, B., Acar, B.: Semantic description of liver CT images: an ontological approach. IEEE J. Biomed. Health Inform. **18**(4), 1363–1369 (2014). https://doi.org/10.1109/JBHI.2014.2298880
8. Slaimi, F., Sellami, S., Boucelma, O., Ben, H.A.: A multigraph approach for web services recommendation. In: Debruyne, C., Panetto, H., Meersman, R., Dillon, T., Kühn, E., O'Sullivan, D., Ardagna, C.A. (eds.) OTM 2016. LNCS, vol. 10033, pp. 282–299. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48472-3_16
9. Stanković, R., Krstev, C., Obradović, I., Kitanović, O.: Improving document retrieval in large domain specific textual databases using lexical resources. Trans. Comput. Collect. Intell. **26**, 162–185 (2017). https://doi.org/10.1007/978-3-319-59268-8_8
10. Terziyan, V., Golovianko, M., Cochez, M.: TB-structure: collective intelligence for exploratory keyword search. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) KEYSTONE 2016. LNCS, vol. 10151, pp. 171–178. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_15