From Provenance-awareness to Explanation-awareness—When Linked Data Is Used for Case Acquisition from Texts

Thomas Roth-Berghofer[1,2] and Benjamin Adrian[1,2]

[1] Knowledge Management Department,
German Research Center for Artificial Intelligence (DFKI) GmbH
Trippstadter Straße 122, 67663 Kaiserslautern, Germany

[2] Knowledge-Based Systems Group, Department of Computer Science,
University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern

{firstname.lastname}@dfki.de

Abstract. Explanation-awareness in Case-Based Reasoning system development aims at making such systems smarter in interactions with their users. When using Linked Data for case acquisition from text one aspect of being smarter is the provision of evidence for the trustworthiness of the acquired cases. The Trustworthiness of such cases relies not only on the provenance of the text but also on the provenance of the used ontological knowledge. Users can only assess the quality of the case-based reasoner's results, i.e., the cases, if the system provides provenance information and if such a system can justify its results. In fact, explanation capabilities very much rely on provenance information.

Key words: Textual Case-Based Reasoning, Ontology-based Information Extraction, Linked Data, Web of Data, Explanation, Provenance

Explanation-awareness in Case-Based Reasoning system development aims at making such systems much smarter in interactions with their users. It looks at new ways to guide software designers and engineers to a purposeful explanation-aware software system by making their designers and engineers explanation-aware [14]. Systems that intend to exhibit explanation-awareness must be more than simple reactive systems. When the word "awareness" is used in conjunction with the word "explanation" it implies some consciousness about explanation. When a system exhibits explanation awareness, it is capable of reasoning about explanations. When we use the word aware make a strong statement about the capabilities of the entity described. As being knowledgeable is central to being aware, some kind of reasoning capabilities, or intelligence, is implied. Thus, whether we aim at creating computer systems that are explanation-aware—or provenance-aware for that matter—such systems must regard explanations and provenance information from the knowledge level [11].

It has been established that there is a certain explanation quality in the use of cases as explanations [5]. The general argument being that a similar past case

will contain additional information, which (together with the target information asked for) provides a context around the target information that should be viewed as a situation-oriented explanation. However, there are also the other knowledge containers [12] related to CBR, i.e., vocabulary, similarity measures, and adaptation knowledge, that can be tapped into for providing additional explanation capabilities [17]. As soon as the vocabulary container is not a closed container anymore, but linked to the Web of Data, provenance issues such as source of information, ownership and authorship arise and need to be dealt with.

Taking the term "provenance-awareness" one of the questions coming to mind probably is: "Who needs to be aware?" Surely designers and users of such software systems, e.g., CBR system developers and knowledge engineers, and the users of the respective CBR system. CBR system developers need to provide means for integrating provenance information and knowledge engineers need to model provenance information or make it available to the CBR system user.

The Linked Open Data project is a Semantic Web effort using the Web (i.e., Web standards such as HTTP, URIs and RDF) to connect related data that was not previously linked together [3]. One of the most visible examples for Linked Open Data is DBpedia, "a community effort to extract structured information from Wikipedia and to make this information available on the Web". Governments and organisations have begun to provide Linked Data in vast amounts.

In [15], Roth-Berghofer et al. describe that case models published as Linked Open Data can be interconnected to information from various data sources. Thus, they can be enriched with additional knowledge from other ontologies. Combining SCOOBIE, an ontology-based information extraction system [1], and *myCBR*, a tool for rapid prototyping of similarity-based retrieval applications [18], allows leveraging available knowledge of the Web of Data for acquiring cases from texts. As standard ontology representation formalism for expressing the case model the Resource Description Framework language (RDF) [19] was chosen, where meaning is expressed by facts encoded in sets of triples [2].

Representing a case model in RDF results in URIs for each attribute and its possible values. If an HTTP request of these URIs delivers RDF triples with additional information about the resource identified with this URI and if several URIs in multiple web domains are interlinked with special types of predicates (e.g., `owl:sameAs`) whether they represent a similar resource, we talk about Linked (Open) Data .

Generating cases from texts using Linked Data from the Web of Data provides new challenges. Textual case-based reasoning (CBR) systems that combine Linked Data from different sources are faced with quality and trust issues. These issues have been recognised and are already addressed by the Semantic Web community (e.g., [10, 9] and by the Web of Data community (e.g., [7]). Both communities provide terminologies for explicitly describing provenance information such as source name, date and author(s) of original information and any updates. Typically, based on the provenance information the quality is assessed. But the vocabularies have lots of weaknesses still and tools are missing.

As soon as CBR system developers and knowledge engineers rely on Linked Data they need to have a vocabulary and tools to express provenance information. They as well as need to provide access to that information for the final recipient of provenance information, the CBR system user who should not only be aware of that kind of information but should demand it.

myCBR provides some general means for extending the vocabulary knowledge container and for accessing it in order to support CBR system users with knowledge about the used terminology [16], namely *concept explanations*. A concept explanation links concepts unknown to a user to already known concepts. It is a comprehensive description in the form of a definition, an example, or a reference to further characterisations, for which any kind of medium such as text, images, audio, or video can be used.

CBR system users can only assess the quality of the case-based reasoner's results if the system provides access to provenance information and if such a system can justify its results. The users might have questions about the history of the case model, which part of the case model was generated by the knowledge engineer and which part comes from other sources. They might ask about why, where, and how provenance information came to pass.

In database research, Buneman et al. [4] distinguished exactly this: why- and where provenance. Why-provenance provides information on the origins that were involved in computing a result, and where-provenance describes the location of the source. Green et al. [6], finally, added how-provenance, which describes how the origins were involved in the computation. These questions are directly feeding into why- and action explanations [13, 17].

Making CBR system developers and knowledge engineers provenance-aware is one step in the direction of making a CBR system provenance-aware. As soon as such CBR systems have provenance information they can act upon it, giving the CBR system user a feeling of how trustworthy a case is or if the data on which the case relies has lost its reliability, as Linked Data today may be different from Linked Data tomorrow [8]. Last but not least, provenance information is also be a good opportunity for triggering maintenance operations.

Acknowledgments

This work was funded by the BMBF project Perspecting (Grant 01IW08002).

References

1. Adrian, B., Hees, J., van Elst, L., Dengel, A.: iDocument: Using ontologies for extracting and annotating information from unstructured text. In: Mertsching, B., Hund, M., Aziz, Z. (eds.) KI 2009: Advances in Artificial Intelligence. Künstliche Intelligenz (KI-2009), Sept. 15–18, Paderborn, Germany. Lecture Notes in Artificial Intelligence, LNAI, vol. 5803, pp. 249–256. Springer, Heidelberg (2009)
2. Bergmann, R., Schaaf, M.: Structural Case-Based Reasoning and Ontology-Based Knowledge Management: A Perfect Match? Journal of Universal Computer Science

9(7), 608–626 (2003), http://www.jucs.org/jucs_9_7/structural_case_based_reasoning [Last access: 2010-02-26]
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009), http://dblp.uni-trier.de/db/journals/ijswis/ijswis5.html#BizerHB09 [Last access: 2010-02-26]
4. Buneman, P., Khanna, S., Tan, W.C.: Why and where: A characterization of data provenance. In: ICDT '01: Proceedings of the 8th International Conference on Database Theory. pp. 316–330. Springer, London, UK (2001)
5. Cunningham, P., Doyle, D., Loughrey, J.: An evaluation of the usefulness of case-based explanation. Tech. Rep. TCD-CS-2003-04., Trinity College Dublin (2003)
6. Green, T.J., Karvounarakis, G., Tannen, V.: Provenance semirings. In: PODS '07: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 31–40. ACM, New York, NY, USA (2007)
7. Hartig, O.: Provenance Information in the Web of Data. In: Proceedings of the Linked Data on the Web LDOW Workshop at WWW (2009)
8. Hellmann, S., Stadler, C., Lehmann, J., Auer, S.: Dbpedia live extraction. In: Debbabi, M., Caragea, D., Yang, X.C. (eds.) Proceedings of the 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE). Lecture Notes in Computer Science, vol. 5871, pp. 1209—1223. Springer (2009)
9. McGuinness, D.L., Ding, L., da Silva, P.P., Chang, C.: Pml 2: A modular explanation interlingua. In: Roth-Berghofer, T.R., Schulz, S., Leake, D.B. (eds.) ExaCt. pp. 49–55. AAAI Press (July 2007), technical Report WS-07-06
10. McGuinness, D.L., Pinheiro da Silva, P.: Infrastructure for web explanations. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) The Semantic Web — ISWC 2003. pp. 113–129 (2003)
11. Newell, A.: The knowledge level. AI Magazine 2(2), 1–20 (1981)
12. Richter, M.M.: The knowledge contained in similarity measures. Invited Talk at the First International Conference on Case-Based Reasoning, ICCBR'95, Sesimbra, Portugal (1995)
13. Roth-Berghofer, T. (ed.): Künstliche Intelligenz: Anwendungen im Semantischen Web (Integriertes Seminar). Technische Universität Kaiserslautern (Juli 2004)
14. Roth-Berghofer, T.: ExaCt manifesto: Explanation-aware computing (July 2009), http://on-explanation.net/content/ExaCt_Manifesto.html
15. Roth-Berghofer, T., Adrian, B., Dengel, A.: Case acquisition from text: Ontology-based information extraction with SCOOBIE for myCBR. In: Montani, S., Bichindaritz, I. (eds.) Advances in Case-Based Reasoning. Springer, Heidelberg (2010)
16. Roth-Berghofer, T.R., Bahls, D.: Explanation capabilities of the open source case-based reasoning tool mycbr. In: Petridis, M., Wiratunga, N. (eds.) Proceedings of the thirteenth UK workshop on Case-Based Reasoning UKCBR 2008. pp. 23–34. University of Greenwich, London, UK (2008)
17. Roth-Berghofer, T.R., Cassens, J.: Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In: Muñoz-Avila, H., Ricci, F. (eds.) Case-Based Reasoning Research and Development, ICCBR 2005, Chicago, IL, USA, August 2005, Proceedings. pp. 451–464. No. 3620 in Lecture Notes in Artificial Intelligence LNAI, Springer, Heidelberg (2005)
18. Stahl, A., Roth-Berghofer, T.R.: Rapid prototyping of CBR applications with the open source tool myCBR. In: Bergmann, R., Althoff, K.D. (eds.) Advances in Case-Based Reasoning. Springer (2008)
19. W3C: Rdf primer (February 2004), http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ [Last access: 2010-02-26]