SMC4LRT - Semantic Mapping Component for Language Resources and Technology

Durco Matej, 0005416

May 26, 2011

1 Main Goal

This work proposes a component that shall enhance search functionality over a *large heterogeneous collection of metadata descriptions* of Language Resources and Technology (LRT). By applying semantic web technology the user shall be given both better recall through *query expansion* based on related categories/concepts and new means of *exploring the dataset* via ontology-driven browsing.

Following two examples for better illustration. First a concept-based query expansion: Confronted with a user query: `Actor.Name = Sue` and knowing that `Actor` is synonym to `Person` and `Name` is synonym to `FullName` the expanded query could look like:

```
Actor.Name = Sue OR Actor.FullName = Sue OR
Person.Name = Sue OR Person.FullName = Sue
```

And second, an ontology-driven search: Starting from a list of topics the user can browse an ontology to find institutions concerned with those topics and retrieve a union of resources for the resulting cluster. Thus in general the user is enabled to work with the data based on information that is not present in the original dataset, but rather in external linked-in semantic resources.

Such **semantic search** functionality requires a preprocessing step, that produces the underlying linkage both between categories/concepts and on the instance level. We refer to this task as **semantic mapping**, that shall be realized by corresponding `Semantic Mapping Component`. In this work the focus lies on the method itself – expressed in the specification and operationalized in the (prototypical) implementation of the component – rather than trying to establish a final, accomplished alignment. Although a tentative, naïve mapping on a subset of the data will be proposed, this will be mainly used for evaluation and shall serve as basis for discussion with domain experts aimed at creating the actual sensible mappings usable for real tasks.

In fact, due to the great diversity of resources and research tasks, a "final" complete alignment does not seem achievable at all. Therefore also the focus shall be on "soft" dynamic mapping, i.e. to enable the users to adapt the mapping or apply different mappings depending on their current task or research question

essentially being able to actively manipulate the recall/precision ratio of the search results. This entails an examination of user interaction with and visualization of the relevant additional information in the user search interface. However this would open doors to a whole new (to this work) field of usability engineering and can be treated here only marginally.

2 Method

We start with examining the existing data and describing the evolving infrastructure in which the components are to be embedded. Then we formulate the function of **Semantic Search** distinguishing between the concept level – using semantic relations between concepts or categories for better retrieval – and the instances level – allowing the user to explore the primary data collection via semantic resources (ontologies, vocabularies).

Subsequently we introduce the underlying **Semantic Mapping Component** again distinguishing the two levels - concepts and instances. We describe the workflow and the central methods, building upon the existing pieces of the infrastructure (See *Infrastructure Components* in 4). A special focus will be put on the examination of the feasibility of employing ontology mapping and alignment techniques and tools for the creation of the mappings.

In the practical part - processing the data - a necessary prerequisite is the dataset being expressed in RDF. Independently, starting from a survey of existing semantic resources (ontologies, vocabularies), we identify an intial set of relevant ones. These will then be used in the exercise of mapping the literal values in the by then RDF-converted metadata descriptions onto externally defined entities, with the goal of interlinking the dataset with external resources (see *Linked Data* in 4).

Finally in a prototypical implementation of the two components we want to deliver a proof of the concept, supported by an evaluation in which we apply a set of test queries and compare a traditional search with a semantically expanded query in terms of recall/precision indicators. A separate evaluation of the usability of the Semantic Search component is indicated, however this issue can only be tackled marginally and will have to be outsourced into future work.

3 Expected Results

The primary concern of this work is the integrative effort, i.e. putting together existing pieces (resources, components and methods) especially the application of techniques from ontology mapping to the domain-specific data collection (the domain of LRT). Thus the main result of this work will be the *specification* of the two components `Semantic Search` and the underlying `Semantic Mapping`. This theoretical part will be accompanied by a proof-of-concept *implementation* of the components and the results and findings of the *evaluation*.

One promising by-product of the work will be the original dataset expressed as RDF with links into existing external resources (ontologies, knowledgebases, vocabularies), effectively laying a foundation for providing this dataset as *Linked Open Data*[1] in the *Web of Data*.

[1] `http://linkeddata.org/`

4 State of the Art

Infrastructure Components

There are multiple relevant activities being carried out in the context of research infrastructure initiatives for LRT. The most relevant ongoing effort is the `VLO - Virtual Language Observatory`[2][1], being developed within the CLARIN project. This application operates on roughly the same collection of data as is discussed in this work, however it employs a faceted search, mapping manually the appropriate metadata fields from the different schemas to 8 fixed facets. Although this is a very reductionist approach it is certainly a great starting point offering a core set of categories together with an initial set of category mappings.

`Component Registry` and `ISOcat`[3] are two integral components of the *CLARIN Metadata Infrastructure* maintaining the normative information. Especially `ISOcat` – the ISO-standardized Data Category Registry for registering and maintaining `Data Categories` as globally agreed upon incarnations of concepts in the domain of discourse – is the definitive primary reference vocabulary [2, 3]. A tightly related work is that on the so called `Relation Registry`, a separate component that allows to define arbitrary relations between data categories, however this activity is rather in an early prototypical phase.

And a last relevant intiative to mention is that of a `Vocabulary Alignment Service` being developed and run within the Dutch program CATCH[4], which serves as a neutral manager and provider of controlled vocabularies. There are plans to reuse or enhance this service for the needs of the CLARIN project.

All these components are running services, that this work shall directly build upon.

LRT Resources

The CLARIN project also delivers a valuable source of information on the normative resources in the domain in its current deliverable on *Interoperability and Standards* [4]. Next to covering ontologies as one type of resources this document offers an exhaustive collection of references to standards, vocabularies and other normative/standardization work in the field of Language Resources and Technology.

Regarding existing domain-specific semantic resources `LT-World`[5], the ontology-based portal covering primarily Language Technology being developed at DFKI[6], is a prominent resource providing information about the entities (Institutions, Persons, Projects, Tools, etc.) in this field of study. [5]

Ontology Mapping

As the main contribution shall be the application of *ontology mapping* techniques and technology, a comprehensive overview of this field and current developments is paramount. There seems to be a plethora of work on the topic and the difficult task

[2] `http://www.clarin.eu/vlo/`
[3] `http://www.isocat.org/`
[4] *Continuous Access To Cultural Heritage* - `http://www.catchplus.nl/en/`
[5] `http://www.lt-world.org/`
[6] *Deutsches Forschungszentrum für Künstliche Intelligenz* - `http://www.dfki.de`

will be to sort out the relevant contributions. The starting point for the investigation will be the overview of the field by Kalfoglou [6] and a more recent summary of the key challenges by Shvaiko and Euzenat [7].

In their rather theoretical work Ehrig and Sure [8] elaborate on the various similarity measures which are at the core of the mapping task. On the dedicated platform OAEI[7] an ongoing effort is being carried out and documented comparing various alignment methods applied on different domains.

One more specific recent inspirative work is that of Noah et. al [9] developing a semantic digital library for an academic institution. The scope is limited to document collections, but nevertheless many aspects seem very relevant for this work, like operating on document metadata, ontology population or sophisticated querying and searching.

Linked Open Data

As described previously one outcome of the work will be the dataset expressed in RDF interlinked with other semantic resources. This is very much in line with the broad *Linked Open Data* effort as proposed by Berners-Lee [10] and being pursuit across many discplines. (This topic is supported also by the EU Commission within the FP7.[8]) A very recent comprehensive overview of the principles of Linked Data and current applications is the book by Heath and Bizer [11], that shall serve as a practical guide for this specific task.

References

[1] D. V. Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardellini, "Virtual language observatory: The portal to the language resources and technology universe," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds.), (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[2] D. Broeder, M. Kemps-Snijders, D. V. Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn, "A data category registry- and component-based metadata framework," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds.), (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[3] ISO12620:2009, "Computer applications in terminology – data categories – specification of data categories and management of a data category registry for language resources," 2009.

[4] E. Hinrichs, P. Banski, K. Beck, G. Budin, T. Caselli, K. Eckart, K. Elenius, G. Faaß, M. Gavrilidou, V. Henrich, V. Quochi, L. Lemnitzer, W. Maier, M. Monachini, J. Odijk, M. Ogrodniczuk, P. Osenova, P. Pajas, M. Piasecki,

[7] Ontology Alignment Evalution Intiative - `http://oaei.ontologymatching.org/`

[8] `http://cordis.europa.eu/fetch?CALLER=PROJ\_ICT&ACTION=D&CAT=PROJ&RCN=95562`

A. Przepiórkowski, D. V. Uytvanck, T. Schmidt, I. Schuurman, K. Simov, C. Soria, I. Skadina, J. Stepanek, P. Stranak, P. Trilsbeek, T. Trippel, and I. Vogel, "Interoperability and standards," deliverable, CLARIN, March 2011.

[5] B. Jörg, H. Uszkoreit, and A. Burt, "Lt world: Ontology and reference information portal," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds.), (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[6] Y. Kalfoglou and M. Schorlemmer, "Ontology mapping: the state of the art," *The Knowledge Engineering Review*, vol. 18, pp. 1–31, Jan. 2003.

[7] P. Shvaiko and J. Euzenat, "Ten challenges for ontology matching," in *On the Move to Meaningful Internet Systems: OTM 2008* (R. Meersman and Z. Tari, eds.), vol. 5332 of *Lecture Notes in Computer Science*, pp. 1164–1182, Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-88873-4_18.

[8] M. Ehrig and Y. Sure, "Ontology mapping – an integrated approach," in *The Semantic Web: Research and Applications* (C. Bussler, J. Davies, D. Fensel, and R. Studer, eds.), vol. 3053 of *Lecture Notes in Computer Science*, pp. 76–91, Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-25956-5_6.

[9] S. Noah, N. Alias, N. Osman, Z. Abdullah, N. Omar, Y. Yahya, and M. Yusof, "Ontology-driven semantic digital library," in *Information Retrieval Technology* (P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov, eds.), vol. 6458 of *Lecture Notes in Computer Science*, pp. 141–150, Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-17187-1_13.

[10] T. Berners-Lee, "Linked data." online: http://www.w3.org/DesignIssues/LinkedData.html, 07 2006. Status: personal view only. Editing status: imperfect but published. Last visited: 2011-04-13.

[11] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 1, pp. 1–136, Feb 2011.