PhD topic

Dynamic semantic annotation: analysis, modeling and implementation

Mai 2014

Supervision: Adeline Nazarenko
Laboratory : LIPN

Key words

Semantic annotation, text mining, knowledge engineering, semantic web

Résumé

The semantic annotation of documents is playing a key role in the convergence of the textual web text and the web of data or semantic web.

Text annotation consists in attaching information or metadata whose semantics is given by a model (indexing language, thesaurus, ontology, for example) to textual documents or to some text fragments. The semantic annotation process associates to the text a formal semantic representation that can be exploited by semantic engines and software agents in the semantic web.

This thesis aims to develop tools and methods to dynamically build or update the semantic model being used for annotation during the annotation process.

Objectif

Text annotation consists in attaching to text fragments some metadata whose semantics is given by a model (indexing language, thesaurus, ontology, for example). It builds on the top of the text a formal semantic representation which granularity depends on the intended applications but which is formal. The content analysis operations that exploit the annotated corpus (e.g. document search, comparison, synthesis, navigation, segmentation) can thus rely on the plain source text, the added annotations and the underlying semantic model altogether. This annotation can be done automatically or manually as part of annotation campaigns. Manually annotated corpora are generally used as training data and evaluation.

Tools exist to annotate texts automatically or to guide the work of manual annotation wrt. a semantic model (usually a thesaurus or ontology). There are also methods and tools to build semantic models from texts, as texts are valuable sources of information for knowledge elicitation. These acquisition and annotation processes are usually considered as distinct. The semantic models are defined *a priori* and used as they are in semantic annotation.

However, this static vision of semantics is inadequate for most annotation tasks. It assumes that a suitable semantic model of sufficient quality already exists. In practice, the semantic model needs be built or updated dynamically in the course of the annotation process, which often shows the limitations of the initial model or the markup rules associated to it. The inability to annotate certain parts of the text and/or the poor quality of the resulting annotation often call for enriching or amending either the semantic model or the way it is used.

In contrast to the traditional sequential and static approach, this PhD thesis aims to develop a method of semantic annotation that allows for populating but also dynamically updating the semantic models in the course of the annotation process. The goal is to integrate the acquisition and annotation processes.

Approach

Modeling such a dynamic process coupling knowledge acquisition and semantic annotation involves several steps. It requires to

1. formally define the various types of target annotation and specify the tools to be used to implement them (new annotation tool(s) may need to be developed);
2. identify and model the conditions that call for triggering the model updates (e.g. a measure of coverage, the detection of an inconsistency);
3. define, formalize and implement mechanisms for updating the semantic model: the changes may concern the model itself (add / remove / edit a semantic entity or larger restructuration) but also the annotations rules used to project the model onto the text;
4. in some cases, revise the existing annotation to ensure its conformance with the updated model;
5. extend the approach to take into account several semantic resources, optionally partially aligned with each other.

This PhD work will benefit from the existing state of the art in ontology population [11, 5], semantic annotation [12, 6, 14, 1], evolution of semantic models (especially ontologies [7, 3, 13]) but will require to expand and articulate the existing solutions to take into account various or richer types of semantic annotation and tackle the dynamics problem.

The PhD student will rely on the RCLN skills: the tools developed for the acquisition of knowledge from texts (TERMINAE [2] and SemEx [8]), the experience in semantic annotation corpus, be it automatic [10, 9] or manual [4] and works on semantic search (to be published). She/He will initially work with traditional semantic models (thesauri, ontologies) for which there exists well-established formalisms (SKOS , OWL -DL) and technologies but other semantic models will eventually be considered.

As mentioned above, the problem of dynamic semantic annotation arises for both automatic and annual annotation approaches. The PhD work will either focus on the automatic annotation process or address both approaches in parallel.

Context

This PhD is part of the RCLN team research program in semantic annotation (collaborative projects such as Quaero, ONTORULE or Legilocal have all faced problems of semantic annotation).

These issues of semantic analysis and annotation of corpora are also an important issue for the labex "Empirical Foundations of Language", in which the RCLN team is involved, especially for the strand "computational semantic analysis" where the problems of corpus semantics and access to content are addressed.

Based on its past experience in the field of scientific and technical information management (*e.g.* collaborations with INRA and INIST), the RCLN team now tackles issues related to Humanities and Social Sciences (especially in the legal field), where the size of texts, the structure of document collections and the richness of the target interpretations show the limits of the static approach of semantic annotation and call for more dynamic and robust methods.

References

[1] Pierre Andrews, Ilya Zaihrayeu, and Juan Pane. A classification of semantic annotation systems. *Semant. web*, 3(3):223–248, August 2012.

[2] Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The TERMINAE Method and Platform for Ontology Engineering from texts. In Paul Buitelaar and Philipp Cimiano, editors, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pages 199–223. IOS Press, janvier 2008.

[3] Rim Djedidi and Marie-Aude Aufaure. Ontology change management. In A. Paschke, H. Weigand, W. Behrendt, K. Tochtermann, and T. Pellegrini, editors, *5th International Conference on Semantic Systems (I-Semantics 09), Proceedings of I-KNOW ?09 and I-SEMANTICS ?09*, pages 611–621, Graz, Austria, September 2009. Verlag der Technischen Universitt Graz.

[4] Karën Fort, Adeline Nazarenko, and Sophie Rosset. Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis. In *Proceedings of*

the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India, December 2012.

[5] C. Giuliano and A. Gliozzo. Instance-based ontology population exploiting named-entity substitution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 265–272, Manchester, August 2008.

[6] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[7] Pieter De Leenheer and Tom Mens. Ontology evolution : State of the art and future directions. In Martin Hepp, Pieter De Leenheer, Aldo de Moor, and York Sure, editors, *Ontology Management: Semantic Web, Semantic Web Services, and Business Applications*, pages 131–176. Springer, 2007.

[8] François Lévy, Adeline Nazarenko, and Abdoulaye Guissé. Annotation, indexation et parcours de documents numériques. *Revue des Sciences et Technologies de l'Information (Série Document Numérique)*, 13(3/2010):121–152, December 2010.

[9] Yue Ma, François Lévy, and Sudeep Ghimire. Reasoning with Annotations of Texts. In *The 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, pages 192–197, États-Unis, May 2011.

[10] Yue Ma, Adeline Nazarenko, and Laurent Audibert. Formal description of resources for ontology-based semantic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, Malta, May 2010. ELRA.

[11] Bernardo Magnini, Emanuele Pianta, Octavian Popescu, and Manuela Speranza. Ontology population from textual mentions: Task definition and benchmark. In *Proceedings of the OLP2 workshop on Ontology Population and Learning*, Sidney, Australia, 2006.

[12] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. Kim – a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4):375–392, 2004.

[13] Zied Sellami, Valérie Camps, and Nathalie Aussenac-Gilles. Dynamo-mas: a multi-agent system for ontology evolution from text. *J. Data Semantics*, 2(2-3):145–161, 2013.

[14] Victoria Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4, 2006.