Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Semantic Annotation of Text

Gurinder Pal Singh Gosal
Department of Computer Science,
Punjabi University, Patiala, Punjab, India

Abstract -The semantic annotation with manual means is an expensive process and often does not consider the multiple perspectives of a data source. The automation of the annotation process is essential to provide the scalability needed to annotate existing documents and reduce the burden of annotating new documents considering we have to deal with large collections of data. The automatic annotations bring with them the benefits of improved information retrieval and enhanced interoperability. In this paper the issues related to automatic representation and uses of the semantic annotation have been looked at and some important semantic works and platforms are investigated.

Keywords— Semantic Annotation, Information Extraction, IE, Ontology, Knowledge Acquisition, Semantic Markup

I. INTRODUCTION

Annotation in simple terms can be described as a markup of a text span in a specific format that indicates a feature or features of the text within the span. Therefore, semantic annotation is going beyond the simple textual annotations of the text of the documents with associating the entities in the text to their semantic descriptions. This kind of assignment provides both class and instance information about the entities and may be known by multiple names, however, 'semantic' is the term widely acceptable in the literature.

A. Representation and Usage of Semantic Annotation

The word semantic is derived from the Greek "semasia" which means "signification" or "meaning". Cunningham et al. (2011) defined **semantic annotation** as the process of assigning metadata tags and/or ontology classes to text segments, as a further basis for knowledge access and retrieval mechanisms [1]. The sphere of semantic annotation at the very simple level starts with the premise that the named entities stated in the text themselves constitute important part of their semantics. The entities in the text assigned links to their semantic descriptions are annotations that are perceived as semantic annotations [2]. Although it has been argued also by many researchers that there is no well-established term for this task and neither there is a well-established meaning for the term "semantic annotation".

The semantic annotation with manual means is an expensive process and often does not consider the multiple perspectives of a data source. As we usually deal with large collections of text, the automation of the annotation process is essential to provide the scalability needed to annotate existing documents and reduce the burden of annotating new documents [1]. This automation is typically used with information extraction techniques, among which named entity recognition (NER) is used to identify concepts to annotate. Afterwards knowledge sources about a particular domain, such as, ontologies can be used for semantic annotation of the identified concepts.

The ontologies can integrate knowledge from different information sources having heterogeneous of database schemas [3]. The ontologies provide syntax and semantic to define complex domain vocabularies, leverage interoperability and offer powerful way of semantic annotation [4]. The knowledge from the ontologies can be harvested for semantic annotation. *For example*, a semantic annotation might relate the term *AKT1* to an ontology identifying it as an instance of the abstract concept *Protein Kinase*, and linking it to the instance *carcinoma*of the abstract concept *Cancer*. The automatic annotations bring with them the benefits of improved information retrieval and enhanced interoperability [5].

II. SEMANTIC ANNOTATION

The transformation of texts into formal representations of the contained knowledge, in terms of annotation, paves the way for the application of sophisticated computational methods and hence helping the researchers and advance science [6]. There are many efforts in the past to build and use annotation system and some important works are reviewed here.

A. Annotea

Annotea [7] was a Web-based shared annotation system in which annotations, viewed as statements made by an author about a Web document, were modeled as a class of metadata. The Annotea system was based on a general-purpose open RDF infrastructure and therefore allows different applications to reuse the information that is present in an Annotea system. Notably the authors considered annotations external to the documents being stored externally.

The users can utilize Annotea objects for their own requirements to build SW metadata which is easily reusable for the other applications. As an example, the trusted users can annotate the spam messages from discussion lists and hence filtering away the spams without losing information. The salient feature of Annotea metaphors is that it hides the underlying SW technologies and the users can use SW without knowing the inner complex details.

The advantage of using SW technologies and metadata is directly leveraged by combining and reusing metadata generated by users. This can be easily reused in different other applications like data mining, search engine applications and user profiles for services.

B. COHSE (Conceptual Open Hypermedia Service)

COHSE (Conceptual Open Hypermedia Service) project [8], another effort in the domain of semantic annotation, was based on ontological reasoning service used to represent a sophisticated conceptual model of document terms and their relationships. Further this service was integrated with a Web-based open hypermedia link service that can offer a range of various link-providing facilities in a scalable manner. Both these services formed a *conceptual hypermedia* system that enables documents to be linked with description of the contents via metadata and therefore helps to improve the reliability and scope of linking of WWW documents at retrieval and authoring times. COHSE project was aimed at researching into methods for quality improvement, consistency and breadth of linking of documents of WWW at the time of retrieval and at the time of authoring.

Both Annotea and COHSE systems, however, were not elaborate on the procedure of information extraction for automatic annotation.

C. S-CREAM (Semi-automatic CREAtion of Metadata)

S-CREAM (Semi-automatic CREAtion of Metadata) [9] was another project of semi-automatic annotation of webpages which heavily depended upon machine learning techniques for extraction of relations between the entities. The new version of S-CREAM, in addition to all the other good features of CREAM, supported metadata creation with the help of information extraction. The utilities provided are inference services, document management system, crawler, ontology guidance, document editors and viewers, and a meta ontology.

Ont-O-Mat [9] is an implementation of the S-CREAM framework. Ont-O-Mat is Java-based and also provides a plugin interface for having further extensions and advancements, e.g. collaborative metadata creation or integrated ontology editing and evolution. The IE component in Ont-O-Mat is based on Amilcare which is based on machine learning technique and needs a training corpus of documents which are manually annotated. Amilcare uses the ANNIE ("A Nearly-New IE system") component of the GATE resources to perform IE.

S-CREAME has four distinctive attributes or dimensions when we compare it with other semantic annotation frameworks:

1. It can be regarded as a framework for mark-up in the Semantic Web.
2. It is an annotation framework for sure although may be different focus than the other ones.
3. It also can be considered to be a knowledge acquisition framework.
4. It utilizes information extraction as a support for generating semantic mark-up.

D. MnM

MnM is a Web-based annotation tool [10] which is oriented to semantic markup and provides mechanisms for large-scale automatic markup of Web pages with semantic contents. MnM offers an open API for connecting to ontology servers and for combining tools of information extraction and integrates a Web browser with an ontology editor. In another work Sætre et al. reported a novel approach by using Google for semantic annotation of the biomedical words [11].

To perform the semantic annotation, dedicated framework or platforms have been devised in the recent past by the researchers. Semantic annotation platforms (SAPs) provide support for IE applications, ontology and knowledgebase management utilities, Application Program Interfaces (APIs) for access, storage mechanisms (e.g. RDF repositories), and interfaces or editors for viewing and browsing for ontology and knowledge base [3]. The platforms may not be having all these utilities and may be including a subset of these utilities.

III. SEMANTIC ANNOTATION WORKS USING GATE

The General Architecture for Text Engineering (GATE) [12] is an architecture, framework and development environment for language processing R&D that can be used to build applications and resources in multiple languages. The GATE architecture provides a platform to develop a number of successful applications for various natural language processing tasks, such as, information extraction and at the same time also helps to develop and annotate corpora and perform evaluations on the applications developed [13-14]. GATE framework, over the time has proved its maturity and extensibility for information extraction and other NLP applications in multiple languages.

Significant amount of research on information extraction and semantic annotation has been performed in various projects related to GATE architecture. Maynard et al. in their work described a system, MUSE [15] for NER from texts of widely differing domain, format and genre by using GATE platform. The MUSE system executes IE components, called *processing resources* (PRs) in GATE that form a processing pipeline, conditionally based on text attributes and demonstrates that a rule-based system can perform as well as a machine learning system.

In another work, Bontcheva et al. discussed the shallow methods for resolving named entity co-reference and constructing of the co-reference chains that was developed as modules in the ANNIE Information Extraction system [16] in GATE environment. Dimitrovet al. provided an implementation for resolution of pronoun anaphora in the case where the antecedent is a named entity by using GATE [17].

We know that GATE is also a framework for content and annotation management. Semantically annotating documents with the usage of ontologies and knowledge-bases has been used in many works accomplished in the GATE environment.

A. The Knowledge and Information Management (KIM)

The Knowledge and Information Management (KIM) [3] is a prominent work done in semantic annotation area using GATE. In 2007 Cunningham et al. in their work on adapting an IE and semantic annotation system to patent data by developing ANNIE (a Nearly-New IE pipeline) and ANNIC (ANNotations In Context) using GATE. These are now part of the diverse set of development tools for language processing used in GATE [1]. ANNIC was designed with the objective of providing support to the development of finite state transduction patterns in JAPE language in GATE.

The Knowledge and Information Management (KIM) platform supported a vision that for realization of Semantic Web, the most of metadata needed will be generated by a massive automatic semantic annotation. KIM for each reference to the designated entity in the text generates a link (URI) to the most pertinent class in the ontology, and also a link to the definite instance in the knowledge base. The outcome of the process of automatic semantic annotation is the generation of metadata which is not embedded in the document being processed, thus enabling different semantic annotation tasks to occur.

IV. CONCLUSION

In this paper, a short survey of semantic annotation frameworks was undertaken. The multiple efforts to provide the semantic annotation of text underline the importance of the automation of the annotation process to deal with large scale collection of documents. However, it is observed that although there has been a lot of progress made in the research area of automatic Information Extraction (IE) tasks, the semantic annotation has been lagging behind. This can be due to fewer efforts in the direction of linking IE tasks with formal knowledge sources and ontology systems.

REFERENCES

[1] Cunningham, H., Tablan, V., Roberts, I., Greenwood, M. A., & Aswani, N. (2011). Information extraction and semantic annotation for multi-paradigm information management. In Current Challenges in Patent Information Retrieval (pp. 307-327). Springer Berlin Heidelberg.

[2] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). KIM–semantic annotation platform. In *The Semantic Web-ISWC 2003* (pp. 834-849). Springer Berlin Heidelberg.

[3] Noy, N. F., & Klein, M. (2004). Ontology evolution: Not the same as schema evolution. *Knowledge and information systems*, *6*(4), 428-440.

[4] Handschuh, S., & Staab, S. (Eds.). (2003). *Annotation for the semantic web* (Vol. 96). IOS Press.

[5] Wang, B., Huang, H., Wang, X., & Chen, W. (2009, December). An Ontology-Based NLP Approach to Semantic Annotation of Annual Report. In *Computational Intelligence and Security, 2009. CIS'09. International Conference on* (Vol. 1, pp. 180-183). IEEE.

[6] Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., ... & Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC bioinformatics*, *13*(1), 161.

[7] Kahan, J., Koivunen, M. R., Prud'Hommeaux, E., & Swick, R. R. (2002). Annotea: an open RDF infrastructure for shared Web annotations. *Computer Networks*, *39*(5), 589-608.

[8] Carr, L., Hall, W., Bechhofer, S., & Goble, C. (2001, April). Conceptual linking: ontology-based open hypermedia. In *Proceedings of the 10th international conference on World Wide Web* (pp. 334-342). ACM.

[9] Handschuh, S., Staab, S., & Ciravegna, F. (2002). S-CREAM—semi-automatic creation of metadata. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web* (pp. 358-372). Springer Berlin Heidelberg.

[10] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., & Ciravegna, F. (2002). MnM: Ontology driven semi-automatic and automatic support for semantic markup. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web* (pp. 379-391). Springer Berlin Heidelberg.

[11] Sætre, R., Tveit, A., Steigedal, T. S., & Lægreid, A. (2005). Semantic annotation of biomedical literature using google. In *Computational Science and Its Applications–ICCSA 2005* (pp. 327-337). Springer Berlin Heidelberg.

[12] Cunningham H, Maynard D, Bontcheva K, Tablan V, Dimitrov M, Dowman M, Aswani N, Roberts I, Li Y, Funk A (2000) Developing language processing components with GATE Version 6.0 (a user guide). http://gate.ac.uk/

[13] Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002, July). GATE: an architecture for development of robust HLT applications. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 168-175). Association for Computational Linguistics.

[14] Cunningham H,Maynard D, Bontcheva K, Tablan V (2002) GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th anniversary meeting of the association for computational linguistics (ACL'02)

[15] Maynard, D., Tablan, V., Ursu, C., Cunningham, H., & Wilks, Y. (2001, September). Named entity recognition from diverse text types. In *Recent Advances in Natural Language Processing 2001 Conference* (pp. 257-274).

[16] Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., & Cunningham, H. (2002, June). Shallow methods for named entity coreference resolution. In *Chaınes de références et résolveurs d'anaphores, workshop TALN*.

[17] Dimitrov, M. (2005). A Lightweight Approach to Coreference Resolution for Named Entities in Text Marin Dimitrov, Kalina Bontcheva, Hamish Cunningham and Diana Maynard. *Anaphora Processing: Linguistic, cognitive and computational modelling*, *263*, 97.

