

Further Use of Controlled Natural Language for Semantic Annotation of Wikis

Brian Davis and Siegfried Handschuh and Hamish Cunningham and Valentin Tablan

Digital Enterprise Research Institute, National University of Ireland, Galway

{bran.davis, siegfried.handschuh}@deri.org

Sheffield NLP Group, University of Sheffield

hamish@dcs.shef.ac.uk, V.Tablan@Sheffield.ac.uk

Abstract

Knowledge Acquisition through Semantic Annotation is vital to the evolution, growth and success of the Semantic Web. Both Semi-automatic and Manual Annotation are constricted by a knowledge acquisition bottleneck. Manual Semantic Annotation is a complex and arduous task both time-consuming and costly, often requiring specialist annotators. Therefore, automation of this process is essential to ease the constriction inherent to knowledge acquisition. Semi-automatic annotation tools detect instances of classes within text and relationships between classes; but their usage often requires knowledge of Natural Language Processing and/or formal ontological descriptions. However, one must offer an incentive for a user to annotate his/her respective documents in a user-friendly manner. We describe work in progress concerning the application of Controlled Language Information Extraction - CLIE to a Personal Semantic Wiki - SemperWiki, the goal being to permit users who have no specialist knowledge in ontology tools or languages to semi-automatically annotate their respective personal Wiki pages.

1 Introduction

Knowledge Acquisition through Semantic Annotation is vital to the evolution, growth and success of the Semantic Web. Both Semi-automatic and Manual Annotation are constricted by a knowledge acquisition bottleneck[5]. Manual Semantic Annotation is a complex and arduous task both time-consuming and costly often requiring specialist annotators. The automation of this process (via Information Extraction (IE)) is of crucial importance in order to break through the knowledge acquisition barrier. Semi-automatic Semantic annotation tools detect instances of classes within text and relationships between classes, however their usage often requires knowledge of Natural Language Processing(NLP) and/or formal ontological descriptions. The above requirements have an important impact on issues of HCI (Human

computer Interaction) with respect to the user's experience as an Annotator[12]. This challenges researchers to develop user-friendly authoring/annotation environments within the Knowledge Process[20]. In this paper, we will provide a brief overview of the relevant literature with respect to Controlled Natural Languages for the Semantic Annotation of Wikis. We will provide examples of a translation between a Controlled Language(based on CLIE[21]) and a formal ontology representational language - N3 and describe briefly our proposed implementation, which will combine both CLIE and SemperWiki[16], the purpose being to permit users to use simplified, unambiguous English to semantically annotate their personal Wiki pages.

2 Controlled Languages and Semantic Wikis

"Controlled Natural Languages are subsets of natural language whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity.¹"

Traditionally, controlled languages are split into two major categories: (1) CLs that improve human readability, mainly for non-native speakers, and (2) those that constrain the text for computational treatment. The original concept arose during the 1930s, when a number of influential linguists and scholars devoted considerable effort to establishing a 'minimal' variety of English; It's purpose being to make English accessible to and usable by as many individuals as possible world wide [18]. Early CLs include Caterpillar Fundamental English (CFE)[6] have since then evolved into many variations and flavors such as Smart's Plain English Program (PEP)[1], Whites International Language for Serving and Maintenance (ILSAM)[1], Attempto Controlled English (ACE)[9] and KANT[4]. Furthermore they have found particular favor in large multi-national corporations such as IBM, Rank, Xerox and Boeing amongst others usually within the context of user-documentation production and machine translation/machine-aided translation [1],[18].

¹<http://www.ics.mq.edu.au/~rolfs/controlled-natural-languages/>

Semantic Wiki prototypes have been available for some time. Whereas regular Wikis enable users to describe web resources in natural language, Semantic Wikis allow one to further identify information about Wiki pages i.e. metadata and their relations using a formal, machine-processable language. This allows one to query annotations directly and/or create views for queries [17]. Furthermore, the aforementioned formal language serves as the underlying basis for the knowledge model used to perform reasoning within the Semantic Web community. Semantic Wikis (in the context of annotation) include SemperWiki[16], Platypus² and Wiksar[2]. A subcategory of Semantic Wikis include the usage of Wikis as collaborative ontology editors. OntoWiki[13], WikiOnt[3] and DynamOnt[11] are examples of such editors. Their collective efforts focus however on easing the ontology engineering experience rather than augmenting Wikis with semantic annotations. It should be noted that there is a natural overlap in functionality between both ontology authoring and semantic annotation with respect to HCI and the user experience.

3 Controlled Languages for the Semantic Web

The use of CLs for ontology authoring and instance population is by no means a new concept and it has already evolved into quite an active research area. Additionally, as mentioned in section 2, and most importantly, we logically assume a natural overlap exists between enabling both ontology authoring and semantic annotation through the use of Controlled Natural Language. An example of previous work involving ontology authority via CLs include[19], who present and discuss translations of a CL - PENG-D to First Order Logic (FOL), the purpose being to target the CL to a knowledge representation language such as RDFS and/or OWL. The rationale behind this approach is that FOL has been proposed as the "semantic underpinning" of the semantic web [14]. A well known implementation of this approach (involving CL translation to FOL) is the use of the popular CL , Attempto Controlled English (ACE)[9] as an ontology authoring language. Interestingly, this process has been applied to a Semantic Wiki[15]. ACE can be translated unambiguously into First Order Logic, more specifically Discourse Representation Structure[10].

4 Controlled Language for Information Extraction - CLIE

CLIE - Controlled Language Information Extraction [21]uses CL technology to create and use knowledge repositories stored in KOAN and/or Sesame. It has been developed and is maintained by the Sheffield NLP group, University of Sheffield. CLIE maps the CL sentences directly into RDF/OWL triples using the Java Annotations Pattern Engine - JAPE[8]. JAPE is a processing resource within GATE - General Architecture for Text

²<http://platypuswiki.sourceforge.net>

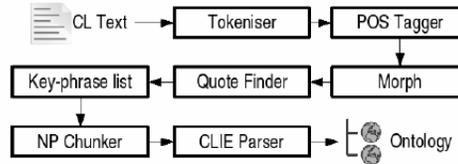


Figure 1: Overview of CLIE

Engineering. JAPE[7] and consists of a cascade of finite state transducers. CLIE[21]has been tested successfully using a Wiki as an ontology authoring environment, whereby the CL is embedded in a mark-up language YAM - (Yet another Markup Language), which is used to input text into the Wiki. As mentioned earlier, the components of the implementation are based on GATE's finite state transducer cascade[7]. The CLIE architecture contains the standard GATE pipeline consisting of the following language and processing resources: The GATE English tokeniser, the Hepple part-of-speech tagger, a morphological analyser, a finite state transducer for identifying quoted strings, a gazetteer list component for recognising useful key-phrases. Finally, two additional JAPE based finite-state transducers (FST) are applied (1) for chunking noun phrases and finally (2) the FST which parses the CL text and generates the ontology. After the initial preprocessing the CLIE parser must search for sentences that contain pre-determined types of key-phrases. Any remaining tokens from the sentence which are not recognised as key-phrases are used as names to generate ontological objects[21]. Figure 1 provides an overview of the CLIE architecture.

5 Proposed Implementation

A natural overlap exists between tools used for both ontology creation and semantic annotation, for instance CLIE permits ontology creation and population by mapping both concept definitions and instances of concepts to a ontological representation. However, there is a subtle difference between the process of ontology creation and population and the process of semantic annotation. We describe semantic annotation as "a process as well as the outcome of the process. Hence it describes i) the process of addition of semantic data or metadata to the content given an agreed ontology and ii) it describes the semantic data or metadata itself as a result of this process"[12]. Of particular importance is the notion of the addition of semantic data or metadata to *content*. In our scenario the content is defined as controlled/uncontrolled text within a Wiki Page and furthermore the metadata generated by Controlled Language is anchored to free/uncontrolled text.

Based upon the existing CLIE/Wiki integration[21],we aim to augment semantic annotation by i) supporting the annotation process and ii) providing a comprehensive model for the metadata itself. The annotation pro-

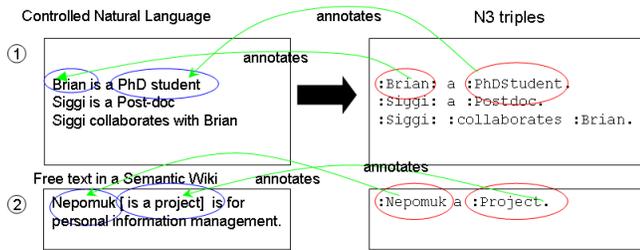


Figure 2: Annotation of Controlled English with N3 triples

cess will be supported by i) the integration of CLIE with a Semantic Wiki instead of a conventional Wiki system; A Semantic Wiki provides us already with a full-fledged RDF/S infrastructure and ii) through interactive user-feedback e.g. for word-sense disambiguation. The comprehensive model for annotation will be provided by a dedicated annotation ontology, which will allow one to glue the metadata to the CL text in the Wiki page.

We intend to use an existing Semantic Wiki - SemperWiki[16] as a test-bed for rapid prototyping. It is our intention to adapt the work accomplished with CLIE specifically for semantic annotation in SemperWiki, whereby the user will annotate her Wiki-pages using Controlled Natural Language. Ideally, the user will write their personal notes freely, however the CLIE enabled user interface will attempt to recognize snippets of controlled text, which will be annotated automatically. If the interface doesn't recognize the snippet, then the user is capable of manually annotating the free text by using a designated syntax to escape the free/uncontrolled text. Additionally, when in this mode, the user will be guided by the user interface with respect to expressivity of the sentences they will be allowed to input. Furthermore the subsequent translated instance metadata will be anchored to the corresponding controlled/uncontrolled text. The upper half of Figure 2 illustrates annotation of Controlled English with N3 triples, which would occur automatically upon recognition of controlled input. The lower part of Figure 2 demonstrates the second option of anchoring uncontrolled or free text in a Semantic Wiki to Controlled English which is in turn annotated with N3 triples. This process would require little intervention on the users part. All the user would need do is type the designated escape character(in Figure 2 it is represented as "[") and proceed to annotate in Controlled English, guided of course by the user interface. When the user is finished annotating, she can return to normal editing in free text by typing "]". In addition, an example of CL translation from a CL conjunction of subjects to N3 triples is provided in part one of Figure 3. Finally, an example of CL translation of a conjunction of objects to N3 triples is provided in part 2 of Figure 3.

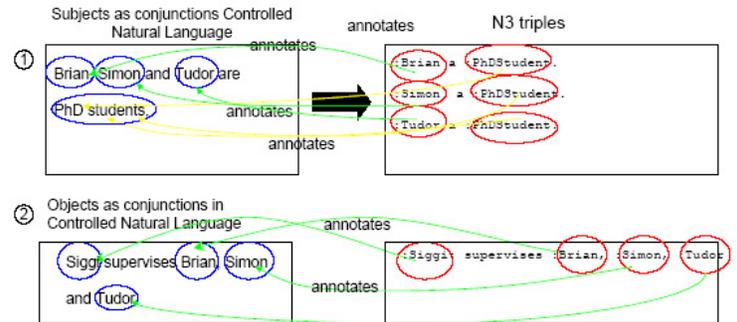


Figure 3: Annotation of noun phrases containing conjunctions in Controlled English with N3 triples

6 Conclusions and Future Work

Incentive for the user to annotate his/her respective documents plays an important role for the continued growth of the Semantic Web. We propose to enhance an existing Semantic Wiki- SemperWiki with Controlled Language-driven semi-automatic semantic annotation abilities using CLIE[21] and furthermore assess the HCI impact of our approach with regards to increasing the incentive for the user to annotate his/her respective documents. Finally, we intend to investigate the cost/benefit of enhancing the expressivity of the existing CL within the CLIE implementation with additional linguistic/structural features i.e. relative clauses, modifier phrases and bullet construction.

Acknowledgments

Special thanks to Diana Maynard, Eyal Oren and Tudor Groza for their comments, guidance and assistance.

References

- [1] Geert Adriaens and Dirk Schreors. From cogram to alcogram: toward a controlled english grammar checker. In *Proceedings of the 14th conference on Computational linguistics*, pages 595–601, Morristown, NJ, USA, 1992. Association for Computational Linguistics.
- [2] S. Aumueller, D. Auer. Towards a semantic wiki experience - desktop integration and interactivity in wiksar. 1st Workshop on The Semantic Desktop, Next Generation Personal Information Management and Collaboration Infrastructure, 2005.
- [3] Jie. Bao and Vasant. Honavar. Collaborative ontology building with wiki@nt - a multi-agent based ontology building environment. Technical report, TR-343, Computer Science, Iowa State University, 2004.
- [4] K. Christine, E. Adolphson, T. Mitamura, and E. Nyberg. Controlled language for multi-lingual document production: Experience with

- caterpillar technical english, 1998. See cite-seer.ist.psu.edu/kamprath98controlled.html.
- [5] Philipp Cimiano, Fabio Ciravegna, John Domingue, Siegfried Handschuh, Alberto Lavelli, Steffen Staab, and Mark Stevenson. Requirements for information extraction for knowledge management. Also available as <http://citeseer.ist.psu.edu/703515.html>.
- [6] Caterpillar Corporation. *Dictionary for Caterpillar Fundamental English*. 1974.
- [7] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 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*, 2002.
- [8] H. Cunningham, D. Maynard, and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS-00-10, Department of Computer Science, University of Sheffield, November 2000.
- [9] N. Fuchs and R. Schwitter. Attempto controlled english (ace), 1996. See cite-seer.ist.psu.edu/article/fuchs96attempto.html.
- [10] Norbert E. Fuchs, Stefan Hoefler, Kaarel Kaljurand, Gerold Schneider, and Uta Schwertel. Extended Discourse Representation Structures in Attempto Controlled English. Technical Report ifi-2005.08, Department of Informatics, University of Zurich, Zurich, Switzerland, 2005.
- [11] Eva Gahleitner, Wernher Behrendt, Juergen Palkoska, and Edgar Weippl. On cooperatively creating dynamic ontologies. In *HYPertext '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 208–210, New York, NY, USA, 2005. ACM Press.
- [12] Siegfried Handschuh. *Creating Ontology-based Metadata by Annotation for the Semantic Web*. PhD thesis, 2005.
- [13] M. Hepp, D. Bachlechner, and K. Siorpaes. Ontowiki: Community-driven ontology engineering and ontology usage based on wikis, 2005. See cite-seer.ist.psu.edu/hepp05ontowiki.html.
- [14] I. Horrocks and P. Patel-Schneider. Three theses of representation in the semantic web, 2003. See citeseer.ist.psu.edu/horrocks03three.html.
- [15] Tobias Kuhn. In *Attempto Controlled English as an Ontology Language*. REWERSE -Reasoning on the Web with Rules and Semantics, 2006.
- [16] Eyal Oren. SemperWiki: a semantic personal Wiki. In *Semantic Desktop (ISWC)*, November 2005.
- [17] Eyal Oren, John G. Breslin, and Stefan Decker. How semantics make better wikis. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 1071–1072, New York, NY, USA, 2006. ACM Press.
- [18] Rolf Schwitter. Controlled natural languages. Also available as <http://www.ics.mq.edu.au/~rolfs/controlled-natural-languages/>.
- [19] Rolf Schwitter and Marc Tilbrook. Controlled natural language meets the semantic web, 2004. See citeseer.ist.psu.edu/718300.html.
- [20] York Sure. *Methodology, Tools and Case Studies for Ontology based Knowledge Management*. PhD thesis, University of Karlsruhe, May 2003.
- [21] V. Tablan, T. Polajnar, H. Cunningham, and K. Bontcheva. User-friendly ontology authoring using a controlled language. In *5th Language Resources and Evaluation Conference*, 2006.