The Semantic Web: A New Opportunity and Challenge for Human Language Technology

Kalina Bontcheva and Hamish Cunningham

Department of Computer Science, University of Sheffield
211 Portobello St, Sheffield, UK S1 4DP
{kalina,hamish}@dcs.shef.ac.uk
http://gate.ac.uk

Abstract. This position paper motivates the need for Semantic Web enabled Human Language Technology (HLT) tools and discusses the major outstanding challenges in this area. It introduces the idea of a "language loop" and shows how HLT can be used to bridge the gap between the current web of language and the Semantic Web. We also argue for a closer integration between HLT and Semantic Web tools and infrastructures. These challenges are at the core of the research agenda of the upcoming EU-funded SEKT project[1].

1 Introduction

The Semantic Web aims to add a machine tractable, re-purposeable layer to compliment the existing web of natural language hypertext. In order to realise this vision, the creation of semantic annotation, the linking of web pages to ontologies, and the creation, evolution and interrelation of ontologies must become automatic or semi-automatic processes.

In the context of new work on distributed computation, Semantic Web Services (SWSs) go beyond current services by adding ontologies and formal knowledge to support description, discovery, negotiation, mediation and composition. This formal knowledge is often strongly related to informal materials. For example, a service for multi-media content delivery over broadband networks might incorporate conceptual indices of the content, so that a smart VCR (such as next generation TiVO) can reason about programmes to suggest to its owner. Alternatively, a service for B2B catalogue publication has to translate between existing semi-structured catalogues and the more formal catalogues required for SWS purposes. To make these types of services cost-effective we need automatic knowledge harvesting from all forms of content that contain natural language text or spoken data.

[1] http://sekt.semanticweb.org. The SEKT partners are: British Telecommunications Plc.; Empolis GmbH; University of Sheffield; University of Karlsruhe; Jozef Stefan Institute; Institut fur Informatik der Universitat Innsbruck; Intelligent Software Components S. A.; Kea-pro GmbH; Ontoprise GmbH; Sirma AI Ltd; Vrije Universiteit Amsterdam; Autonomous University of Barcelona.

Other services do not have this close connection with informal content, or will be created from scratch using Semantic Web authoring tools. For example, printing or compute cycle or storage services. In these cases the opposite need is present: to document services for the human reader using natural language generation.

Finally, tools and infrastructures for the Semantic Web on the one hand and language technology on the other have so far remained largely independent from each other, despite the fact that they share a number of components, namely ontologies and reasoning mechanisms. HLT systems can benefit from new developments like the Ontology Middleware Module (OMM – an extension of the SESAME RDF(S) repository, see [9]) which will enable HLT tools to index and retrieve language data like annotations and gazetteers in RDF(S). It will also enable the use of Semantic Web reasoning tools within HLT components.

To summarise, recent developments in the Semantic Web field have created new opportunities and challenges for Human Language Technology.

In this position paper we discuss the role of HLT in closing the language loop, provide brief overviews of state-of-the-art approaches to tackling some aspects of the problem, and discuss a number of open issues that remain to be solved. The paper is organised as follows. Section 2 provides an overview of relevant HLT technologies. Section 3 focuses on automatic metadata extraction and document annotation for the Semantic Web. Section 4 discusses language generation from formal knowledge. Finally, Section 5 argues for the closer integration between infrastructures for HLT and the Semantic Web.

2 The Role of HLT

The web revolution has been based largely on human language materials, and in making the shift to the next generation knowledge-based web, human language will remain key. Human Language Technology involves the analysis, mining and production of natural language. HLT has matured over the last decade to a point at which robust and scaleable applications are possible in a variety of areas, and new projects in the Semantic Web area (e.g. SEKT – `http://sekt.semanticweb.org`) are now poised to exploit this development.

Figure 1 illustrates the way in which Human Language Technology can be used to bring together the natural language upon which the current web is mainly based and the formal knowledge at the basis of next generation Semantic Web.

Information Extraction (IE) is a process which takes unseen texts as input and produces fixed-format, unambiguous data as output. This data may be used directly for display to users, or may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in Information Retrieval (IR) applications. It is instructive to compare IE and IR: whereas IR simply finds texts and presents them to the user, the typical IE application analyses texts and presents only the specific information from them that the user is interested in. For example, a user of an IR system wanting information on the share price movements of companies with holdings in Bolivian raw materials would typically

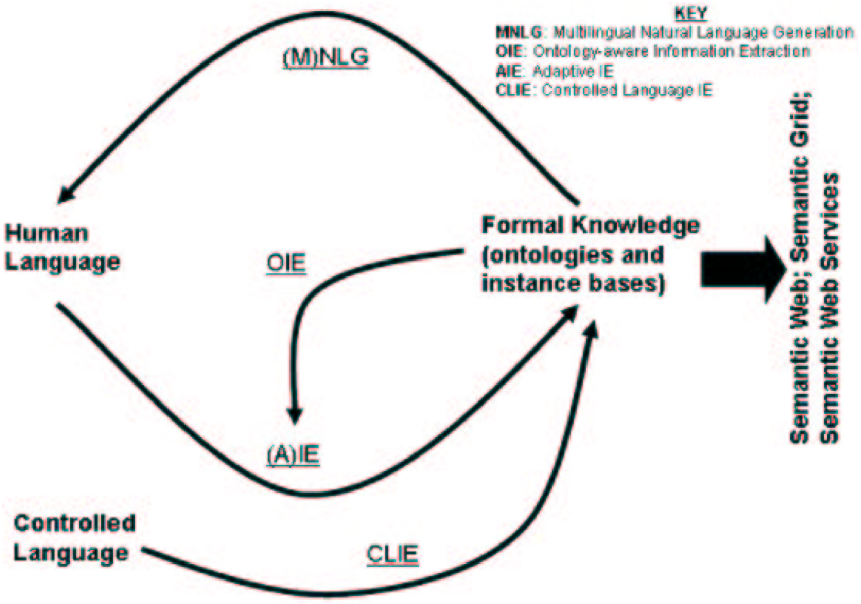
KEY
MNLG: Multilingual Natural Language Generation
OIE: Ontology-aware Information Extraction
AIE: Adaptive IE
CLIE: Controlled Language IE
(M)NLG
Human Language
OIE
Formal Knowledge (ontologies and instance bases)
Semantic Web; Semantic Grid; Semantic Web Services
(A)IE
Controlled Language
CLIE

Fig. 1. Closing the language loop

type in a list of relevant words and receive in return a set of documents (e.g. newspaper articles) which contain likely matches. The user would then read the documents and extract the requisite information themselves. They might then enter the information in a spreadsheet and produce a chart for a report or presentation. In contrast, an IE system user could, with a properly configured application, automatically populate their spreadsheet directly with the names of companies and the price movements. The new challenge for IE is to populate ontologies and generate metadata.

Natural Language Generation (NLG) is the inverse of IE: from structured data in a knowledge base NLG techniques produce natural language text, tailored to the presentational context and the target reader[2]. NLG techniques use and build models of the context and the user and use them to select appropriate presentation strategies. For example, deliver short summaries to the user's WAP phone or a longer multimodal text if the user is using their desktop. Similarly, NLG techniques can use simpler terminology and explain unknown terms to the naive user, while different terminology and text style is used for the expert user. The new challenge for NLG is to generate texts from ontologies and metadata, which requires the development of new NLG methods allowing easy portability between domains, based on machine learning.

[2] For an introduction to NLG see [12].

3 From Language to Knowledge

3.1 Ontology-aware Information Extraction

Recently there has been work on using Information Extraction (IE) to help users annotate (semi-)automatically Web pages with semantic content e.g., [10, 8]. The user trains the IE tools by annotating manually some pages, until the system can start suggesting annotations automatically. Then the user can continue to train the system by correcting its errors and/or annotating missed information. These annotation tools however do not provide the user with a way to customise the integrated language technology directly. While many users would not need or want such customisation facilities, users who already have ontologies with rich instance data will benefit if they can make this data available to the IE components.

The more serious problem however, as discussed in [8], is that there is often a gap between the annotations and their types produced by IE and the classes and properties in the user's ontology. The proposed solution is to write some kind of rules, such as logical rules, to achieve this. For example, an IE system would typically annotate London and UK as locations, but extra rules are needed to specify that there is a containment relationship between the two (for other examples see [8]). However, rule writing of the proposed kind is too difficult for most users and a new solution is needed to bridge this gap.

Therefore, the outstanding challenge is to develop tools to provide the user with a way to customise the integrated language technology directly by connecting the IE components to their ontology to make the tools sensitive to future changes in the model and to bridge the gap between IE results and ontology classes. This ontology-aware IE can be configured to provide a service that will annotate any page relative to a particular ontology, so that software agents can use IE services to find instances of concepts from their own models. This removes some need to map between ontologies: the annotator extracts directly to the user's own ontology. The work will need to go beyond state-of-the-art by:

1. Developing support for learning with unlabeled data, adopting recent techniques from within Data Mining, to extract maximum information from the minimal manual input.
2. Developing hybrid adaptive IE tools, combining rule-based and machine learning approaches and using reasoning services, to perform entity tracking within and across documents.

3.2 Controlled Language IE (CLIE)

Creating formal data is a high initial barrier to entry for small organisations and individuals wishing to make data available to semantic knowledge technology. Part of the answer is in authoring tools, but it is also possible that the definition of a controlled language for formal data description will lower this barrier significantly. Building on controlled language MT work, IE for controlled language analysis could achieve the high levels of accuracy necessary to make this viable.

3.3 Semantic Reference Disambiguation

IE systems currently recognise particular entities and relations, but do not resolve them with respect to a given ontology of classes and instances as needed for the Semantic Web. For instance, they recognise Cambridge as an entity of type Location or City, but do not disambiguate it with respect to which real-world entity it is, i.e., Cambridge in the UK or the US or some other new instance not present in the ontology.

Therefore, existing coreference methods need to be extended with new algorithms for semantic reference disambiguation. A variety of techniques can be explored here. First, vector-space models can be used to detect whether the entity in the text occurs in the same context as an instance in the ontology, as has been done in work on cross-document coreference [2]. Another approach could be to apply work on communities of practice from knowledge management [1] and treat the problem as ensuring referential integrity of ontologies. A useful baseline approach is to disambiguate to the most frequent instance as determined by a reference corpus.

3.4 Quantitative Evaluation: Data, Tools and Metrics

An integral part of the development of machine learning approaches for IE is the ability to perform automatic quantitative evaluation in order to measure differences between different versions of the system and also allow comparative evaluation with other approaches. Automatic quantitative evaluation of IE for the Semantic Web requires: an annotated corpus, an evaluation metric and a scoring tool implementing this metric. Existing corpora and evaluation metrics for IE (e.g., those created for the Message Understanding Conferences [13]) are not suitable for evaluating IE tools in the Semantic Web context, because these corpora and metrics only detect very coarse-grained types of entities, without a specific ontology, and without creating a reference between the entities and events in the documents and those that occur in the target ontology.

The challenge is to create corpora and metrics suitable for evaluating the performance of the IE tools specifically on annotating content relative to ontologies. This will include evaluation along several dimensions:

- Detection of entities and events, given a target ontology of the domain.
- Disambiguation of the entities and events from the documents with respect to instances in the given ontology. For example, measuring whether the IE correctly disambiguated "Cambridge" in the text to the correct instance: Cambridge, UK vs Cambridge, MA.
- Decision when a new instance needs to be added to the ontology, because the text contains a new instance, that does not already exist in the ontology.

In order to achieve this, an evaluation corpus, annotated with the correct ontological class and instance, is needed. The corpus needs to consist of two parts – testing and evaluation part, so that the testing part can be used for system development and testing, while the evaluation one will be used as a gold-standard for evaluation only.

In addition, new metrics for scoring are needed, in order to take into account the nature of the task: for example, the use of ontologies means that correctness is more of a scalar issue, rather than a binary one. The scoring tool needs to automatically compare the system results with the human-annotated standard and produce quantitative measures. In addition, there needs to be a regression testing tool that enables tracking of the system's performance over time, which takes into account relations and distances in the ontology.

4 From Knowledge to Language

NLG can be applied to provide automated documentation of ontologies and knowledge bases. Unlike human-written texts, an automatic approach will constantly keep the documentation up-to-date which is vitally important where knowledge is dynamic and is updated frequently. The NLG tools will also allow generation in multiple languages without the need for human or automatic translation.

The main challenge posed for NLG by the Semantic Web is to provide tools and techniques that are extendable and maintainable (the majority of existing NLG applications can only be modified and extended by specialists). The most promising avenue seems to be the development of novel approaches that combine machine learning with advanced interactive tools for non-specialist users, in order to enhance the adaptivity of NLG.

In addition, the NLG tools can provide context aware and personalised profile-sensitive delivery using state-of-the-art methods for generation of personalised presentations, based on the automatically built user profiles [5]. These methods can effectively summarise knowledge at the appropriate level of granularity and present it in natural language.

Finally, the quantitative evaluation of some NLG methods also poses a challenge due to the lack of corpora, metrics, and evaluation tools [3].

5 Infrastructures, interoperability, and support

Existing HLT infrastructures, such as GATE [7, 6], while offering powerful capabilities, are oriented towards specialists. However, HLT take-up in other fields, like bioinformatics or knowledge technologies, is dependent on tools that offer targetted support for non-experts to customise language processing facilities for their specific domains and tasks.

In addition, a number of HLT fields, e.g., Information Extraction, can also benefit from tools and resources developed in relation to these other fields. For example, ontologies and reasoning services from the Semantic Web can be used as part of the IE task, in order to produce Semantic Web content that is automatically derived from existing data. Also, unsupervised Machine Learning methods for Information Extraction need digital library resources such as gazetteers and thesauri as a source of readily available training data. Therefore another challenge is to provide interoperation with these infrastructures and services, which in combination will offer far more than any of them on their own.

Finally, infrastructural support for delivery of language processing technology over the Grid and with Web services is needed, in order to parallelise slow operations and to enable embedding of HLT in diverse Semantic Web applications.

The first steps towards providing interoperability between Semantic Web and HLT infrastructures have been carried out as part of the open-source GATE HLT infrastructure [4]. GATE has been extended recently to provide support for importing, accessing and visualising ontologies as a new type of resource available to language processing applications, such as IE. Much of this functionality is provided through the integration of the Protégé editor [11] within the GATE visual environment. Ontology import/export is provided from/to DAML+OIL and the formats supported by Protégé . In addition, the results of any IE application can be exported for the Semantic Web in DAML+OIL format.

Another recent effort in this area is KIM – a Knowledge and Information Management platform [4]. KIM offers an RDF(S) repository for storage and management of both language and Semantic Web data, reasoning services, ontology editing and browsing, semantic query interface, and a browser plug-in for document viewing/annotation.

However, a number of open issues are yet to be solved in this area, the most important of which are helping non-expert users to customise the language technology embedded in their applications and the delivery of HLT as Semantic Web services.

6 Conclusion

This position paper motivated the need for Semantic Web enabled Human Language Technology tools and discussed the major outstanding challenges in this area. It introduced the idea of a “language loop” and showed how HLT can be used to bridge the gap between the current web of language and the Semantic Web. We also argued for a closer integration between HLT and Semantic Web tools and infrastructures.

Progress in the development of the Information Society has seen a truly revolutionary decade. Dot com crash notwithstanding, all our lives have been radically changed by the advent of widespread public networking. We believe that a new social revolution is imminent, involving the transition from Information Society to Knowledge Society. SEKT aims to contribute to this revolution, and to embed language technology at its heart.

References

1. H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. O’Hara, and N. Shadbolt. Managing Reference: Ensuring Referential Integrity of

Ontologies for the Semantic Web. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 317–334, Siguenza, Spain, 2002.

2. A. Bagga and A. W. Biermann. A methodology for cross-document coreference. In *Proceedings of the Fifth Joint Conference on Information Sciences (JCIS 2000)*, pages 207–210, 2000.
3. K. Bontcheva. Reuse and problems in the evaluation of nlg systems. In *Proceedings of EACL03 Workshop on Evaluation Initiatives*, Budapest, Hungary, 2003.
4. K. Bontcheva, A. Kiryakov, H. Cunningham, B. Popov, and M. Dimitrov. Semantic web enabled, open source language technology. In *EACL workshop on Language Technology and the Semantic Web: NLP and XML*, Budapest, Hungary, 2003.
5. Kalina Bontcheva. Tailoring the Content of Dynamically Generated Explanations. In M. Bauer, P. Gmytrasiewicz, and J. Vassileva, editors, *User Modelling 2001*, volume 2109 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, Berling Heidelberg, 2001.
6. H. Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254, 2002.
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. S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM — Semi-automatic CREAtion of Metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 358–372, Siguenza, Spain, 2002.
9. M. Klein, D. Fensel, A. Kiryakov, and D. Ognyanov. Ontology Versioning and Change Detection on the Web. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 197–212, Siguenza, Spain, 2002.
10. E. Motta, M. Vargas-Vera, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 379–391, Siguenza, Spain, 2002.
11. N.F. Noy, M. Sintek, S. Decker, M. Crubzy, R.W. Fergerson, and M.A. Musen. Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16(2):60–71, 2001.
12. E. Reiter and R. Dale. Building Natural Language Generation Systems. *Journal of Natural Language Engineering*, Vol. 3 Part 1, 1999.
13. SAIC. Proceedings of the Seventh Message Understanding Conference (MUC-7). `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html`, 1998.