The production of documents from ontologies

Dana Dannélls[1]

Abstract. The production of documents from an ontology is a challenging task which requires a significant effort from a natural language generator. Addressing this problem involves a careful examination of how the knowledge formalized in an ontology can be verbalized and realized. We have started to exploit the abilities of generating natural language texts from a Web Ontology Language (OWL) and to examine how the content of the ontology can be rendered in natural language texts that support reader and listener preferences. In this paper we present our line of research and exemplify some of the difficulties we encountered while attempting to generate fragments of texts from a domain specific ontology.

1 Introduction

A major challenge for a language generator developer who wishes to make use of Semantic Web ontologies is how to alter the input knowledge-base, so as to verbally express contents that describe a concept in an ontology. This task becomes even harder when the user preferences such as the preferred language, text length and syntax must be computed.

Our research project aims to adapt the presentation of a text content for a specific readership from Web ontologies. As a primary step towards accomplishing this aim we utilized a domain specific Web Ontology Language (OWL) and started to exploit how natural language texts may be produced from this expressive language. Below we outline a number of steps which we believe are significant for the quality of the produced text:

1. Selection of the axioms describing a concept;[2]
2. Presentation order of the selected axioms;
3. Verbalization and realization of the selected and ordered axioms.

In this paper we focus on the third step and show that given the selected ontology content, verbalization and realization of the relationships and classes describing a concept exhibit great variations, which depend on the context in which they appear. We illustrate some of these variations and discuss their implications for text production.

The remainder of this paper is structured as follows. Section 2 provides an overview of previous work on generation from ontologies and discusses a number of the advantages and challenges that Web ontology languages pose to language generators. Section 3 provides a description of the domain ontology and the domain ontology language. Section 4 exemplifies the difficulties in verbalizing the knowledge contained in the ontology which we came across while attempting to produce coherent and cohesive texts. Section 5 ends up with conclusions and main directions for future research.

[1] Natural Language Processing Research Unit, Department of Swedish Language, University of Gothenburg, email: dana.dannells@svenska.gu.se

[2] An axiom is an ontology statement which states the relationships among concepts.

2 Background

There are many definitions for the term *ontology* [11]. In this context, an ontology is defined as a structured framework for modeling the concepts and relationships of some domain expertise, which provides the structural and semantic ground for computer based processing of domain knowledge. To allow better use of ontologies in applications, traditional ontology language standards such as DAML+OIL and OWL[3] have been specified by the World Wide Web Consortium (W3C). One of the purposes of these established standards is to enable better communication between humans and machines in which information is given a well defined meaning.

2.1 Generating from ontologies

Generation techniques deal with the process of converting semantic representation into surface form in a particular language. The features of the text produced are normally chosen with respect to a particular target reader group. There have been successful attempts to develop natural language generation tools that generate texts from Web ontology languages [1, 2, 12, 13].

Wilcock [12] presents an approach in which the concepts defined in the ontology are employed for generating the lexicon. Bontcheva and Wilks [2] concentrate on the semantic representations encoded in Semantic Web standards and discuss how these can be exploited to generate text summaries. They point out the content of the ontology itself as a major factor for the quality of the output. Gawronska and Erlendsson [4] show how biological ontologies as Kyoto Encyclopedia of Genes and Genomes, may be utilized for generating graphs representing the essential contents of biomedical scientific articles.

Mellish and Sun [8] describe the large extent of linguistic material in existing Web ontologies and its complexity. They exemplify how an extended text with multiple sentences can be generated from class axioms.

Similarly to [13, 2], this work is concerned with generating textual descriptions of concepts from a domain-specific ontology. As opposed to [8], this approach deals with individuals and requires manual input of the lexicon. In contrast to [13] who uses templates to produce texts, we intend to utilize a grammar-based surface realiser to enhance linguistic variations in the generated texts.

2.2 Opportunities and challenges

As pointed out by many authors, there are several advantages which make Web ontology languages such as OWL particularly suitable to generate from. For example, axioms can be seen as forming a

[3] http://www.w3.org/TR/

graph in which routes between axioms correspond to different possible transitions in a coherent text [7]; axioms can be used to accommodate a generation system to different contextual degrees and user needs; the use of multiple-inheritance converts the class hierarchy into a directed graph and not a tree structure.

Web ontologies provide implicit information about a domain. This is an advantage that has been exploited by a number of Natural Language Generation (NLG) systems [9] who utilize the domain background knowledge base to complete generation related tasks. In many domain ontologies the ontology concepts used to express classes and relationships are similar to their lexical entry, which in many aspects facilitate the generation tasks. However, natural languages are ambiguous and even ontologies which do not make a distinction between the ontology concepts and natural language words that describe them, contain ambiguities that need to be resolved.

To reveal implicit information about a concept, inferences have to be drawn. These inferences that are mostly based on DL [10], might render in different axiom sets, depending on the axiom selection constraints, such as constraints that are set due to the user preferences. Furthermore, it is necessary to fully understand what the knowledge in the selected axiom set actually states before natural language words can be expressed. The content and knowledge formalized in an ontology can lead to ambiguous content interpretations, and can also bring up problems during the process of verbalization. This has brought with it an awareness of the need to encode linguistic knowledge about concepts directly into ontologies [5].

3 The domain ontology model

The work described in this paper is based on the CIDOC Conceptual Reference Model (CRM) ontology,[4] which is an initiative to construct an ontology within the Cultural Heritage (CH) domain. The CIDOC ontology consists of 81 relations and 244 concepts and is available in various formats, among which is OWL. It contains facts about concepts (sets of objects) and roles (binary relations) and provides a conceptual model that subscribes an object-centred view of the CH domain.

3.1 Population and maintenance

Since the CIDOC-CRM ontology does not contain information about individuals (single objects), populating the ontology was a necessary step. We enhanced the ontology with additional lexical entries, as well as new concepts and relationships.

On the task of ontology population, most of the work that has been carried out relates to information extraction from unstructured natural language text or semi-structured HTML pages [6]. In our work, the process of ontology population was conducted manually, it is based on a small corpus of CH texts that we have collected from internal museum repositories. Following the guidelines given by the reference document [3] for filling in concept-values along with a thorough analysis of the information content, we have so far enriched the ontology with a total of 150 new concepts. Each concept was assigned with its lexical lemma that links to a lexical string-name.

3.2 The ontology terminology

An OWL ontology (lite or DL) has a description logic based semantics which consists of a set of axioms. Axioms assert facts about concepts (Tbox) and facts about individuals (Abox). Roles are usually asserted in the form of inclusion axioms.

As with any representation of an OWL ontology, the CIDOC CRM ontology contains classes (concepts) that define a group of individuals that belong together because they share some properties (roles). A subclass is a class that is a specialization of another class (its superclass). According to the CRM documentation, *specialization* means: (1) all instances of the subclass are also instances of its superclass; (2) the intension of the subclass extends the intension of its superclass; (3) the subclass inherits the definition of all of the properties declared for its superclass in addition to having one or more properties of its own.

Properties serve to define relationships of a specific kind between two classes. A property can have a subproperty which is a specialization of another property (its superproperty). A property must be defined with reference to both its domain and range. The term *specialization* in the context of properties has similar meaning as for classes with additional restrictions, i.e: (4) the domain of the subproperty is the same as the domain of its superproperty or a superclass of that domain; (5) the range of the subproperty is the same as the range of its superproperty or the subclass of that range.

4 Realization of a concept in the ontology

In the semantics of OWL, a given axiom may be expressed in several ways and may have more than one realization possibilities. In this section we exemplify some of the discussed challenges (see section 2.2) which are related to realization of concepts in the CIDOC-CRM ontology.

4.1 A concept representation

The following example, taken from our ontology, describes the class *EdelfeltProduction*. This particular class comprises a set of productions that has been carried out by Albert Edelfelt.[5]

The example present an ontology content that describes the concept *EdelfeltPortraitProduction* formulated in an RDF language. The knowledge it conveys is that a production of a portrait took place in France and was made by Albert Edelfelt between 1880 and 1890.

```
<museum:EdelfeltProduction rdf:about="#EdelfeltPortraitProduction">
 <crm:P14F.carried_out_by>
  <crm:E21.Person rdf:about="#AlbertEdelfelt"/>
  </crm:P14F.carried_out_by>
  <crm:P12F.occurred_in_the_presence_of>
  <crm:E21.Person rdf:about="#AlbertEdelfelt"/>
  </crm:P12F.occurred_in_the_presence_of>
 <crm:P7F.took_place_at>
  <crm:E48.Place_Name rdf:about="#France"/>
  </crm:P7F.took_place_at>
 <crm:P4F.has_time_span>
<crm:E49.Time_Appellation rdf:about="#1880-1890"/>
  </crm:P4F.has_time_span>
</crm:E12.Production>
```

The class *EdelfeltProduction* is a subclass of *E12.Production*. *E12.Production* has multiple subclasses, i.e. *E11.Modification* and

[4] http://cidoc.ics.forth.gr/

[5] According to the CRM reference document: "a production can present activities, that are designed to, and succeed in, creating one or more new items".

E63.Beginning_of_Existence, this is shown below.[6]

```
<owl:Class rdf:about="&crm;E12.Production">
 <rdfs:subClassOf rdf:resource="&crm;E11.Modification />
 <rdfs:subClassOf rdf:resource="&crm;E63.Beginning_of_Existence
/>
</owl:Class>
```

E11.Modification is a subclass of *E7.Activity* and *E63.Beginning_of_Existence* is a subclass of *E5.Event*, hence the inferred relation *P12F.occurred_in_the_presence_of*.

4.2 Surface realization

Given an ontology, populated by individuals, given some user preferences, the task is to verbalize and realize the selected ontolgy content. A direct realization of the selected information describing the concept *EdelfeltPortraitProduction* may result in the following text:

> This Edelfelt portrait production was carried out by Albert Edelfelt. The Edelfelt portrait production occurred in the presence of Albert Edelfelt. The Edelfelt portrait production took place in France. The Edelfelt portrait production has time span 1880-1890.

Inferred knowledge Inferred relationships may have distinguished interpretations, therefore in order to resolve their meaning knowledge about the domain and the context in which a concept appears are required. For example, following the above ontology fragment, we interpretate that the inferred relationship *P12F.occurred_in_the_presence_of* carries out redundant information within the context of the concept *EdelfeltPortraitProduction*, and thus does not contribute with new information. As a result of this interpretation, the inferred relationship could be eliminated, or "selected" and verbalized instead of the relationship *carried_out_by*. On the other hand, when a production describes an activity which has resulted in a movie production, e.g. within the context of the concept *TheLordOfTheRingMovieProduction*, the inferred relationship *P12F.occurred_in_the_presence_of* will not provide redundant information but rather contribute with new knowledge.

Verbalization The choice of the lexical entry encoding a relationship is both domain and user dependent, for example, the relationship *carried_out_by* could be verbalized as either "painted by" or "created by" depending on the concepts it describes. Furthermore, the choice between synonyms for the relationship *created_by* are various: "produce by", "bring out by", "develop by", "acquire by", etc. Some differences in categorisations or internal makeup must be present if the difference in information content is to be consequential.

When verbalizing the description about the concept *EdelfeltPortraitProduction* we want to establish a text which is more similar to the following:

> This portrait production was carried out by Albert Edelfelt. The production took place in France. It covers the period 1880-1890.

Humans can recognize that semantic representations are intimately linked, this realization process could be also automated rather easily. However, the problem of how words and other linguistic phenomena might be integrated with the internal representations that support reasoning is yet to be explored.

[6] The notation &crm; is used as a shortcut for the complete URL to the CIDOC-CRM ontology.

5 Conclusion and future work

We presented an ongoing research and illustrated the problems we encountered while attempting to generate coherent and cohesive texts from a Web ontology language. This research work is based on the domain specific CIDOC-CRM ontology. Text planning follows the ontology axioms structure; the assertional part of the ontology is developed manually; both the terminological part and the assertional part are applied to present parts of the ontology.

This paper showed that even-though OWL provides powerful reasoning opportunities for natural language generators, it poses difficulties to language generators that need to be resolved. We highlighted the problem of inferable relationships that are necessary or unnecessary in a particular context, a task which is not trivial for machines. Relationships might have a particular, quite specific interpretation depending on the context in which they appear and the concept they describe. This invokes a difficulty on choice of a lexical entry encoding a relationship.

Our research work is only in its early stages. Exploiting OWL for realization purposes and finding general, domain-independent solutions requires a considerable amount of work. In the near future we are planning to address issues related to content selection and lexical determination of relationships between concepts, a task which depends on the chosen semantic content, the concept it describes, the class hierarchy that is utilized to represent the concept, and the target language.

REFERENCES

[1] K. Bontcheva, 'Generating tailored textual summaries from ontologies.', in *ESWC*, pp. 531–545, (2005).

[2] K. Bontcheva and Y. Wilks, 'Automatic report generation from ontologies: the miakt approach', in *Nineth International Conference on Applications of Natural Language to Information Systems*, (2004).

[3] N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff, *Definition of the CIDOC Conceptual Reference Model*, the version 4.2.4 of the reference document edn., March 2008. http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.4_March2008_.pdf.

[4] B. Gawronska and B. Erlendsson, 'Syntactic, semantic and referential patterns in biomedical texts: towards in-depth text comprehension for the purpose of bioinformatics.', in *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science NLUCS*, pp. 68–77, Miami, USA, (May 2005).

[5] J. Judgem, M. Sogrin, and A. Troussov, 'Galaxy:ibm ontological network miner', *In Sören Auer, Christian Bizer, Claudia Müller, and Anna V. Zhdanova,*, **113**, 157–160, (2007).

[6] V. Karkaletsis, A Valarakos, and C.D. Spyropoulos, 'Populating ontologies in biomedicine and presenting their content using multilingual generation', in *AIME*, pp. 256–265, (2005).

[7] C. Mellish and J. Z. Pan, 'Natural language directed inference from ontologies', *Artificial Intelligence*, **172**(10), 1285–1315, (2008).

[8] C. Mellish and X. Sun, 'The semantic web as a linguistic resource: Opportunities for natural language generation', *Knowl.-Based Syst.*, **19**(5), 298–303, (2006).

[9] S. Daniel Paiva, 'A survey of applied natural language generation systems', Technical Report ITRI-98-03, Information Technology Research Institute(ITRI), University of Brighton, UK, (1998).

[10] E. Reiter and C. Mellish, 'Using classification to generate text', in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL92)*, (1992).

[11] S. Staab and R. Studer, *Handbook on Ontologies*, International Handbooks on Information Systems, Springer, 2004.

[12] G. Wilcock, 'Talking owls: Towards an ontology verbalizer', in *Human Language Technology for the Semantic Web and Web Services*, pp. 109–112, (2003). Sanibel Island, Florida.

[13] G. Wilcock and K. Jokinen, 'Generating responses and explanations from rdf/xml and daml+oil.', in *Knowledge and Reasoning in Practical Dialogue Systems IJCAI-2003*, pp. 58–63, (2003). Acapulco.