Natural Language Generation from Ontologies

Thomas Smith
Electronics and Computer Science
University of Southampton
taes1g09@ecs.soton.ac.uk

ABSTRACT

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Natural language*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language generation*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*

General Terms

Algorithms, Design, Documentation, Theory

Keywords

Semantic web, natural language generation, natural language interface

1. INTRODUCTION

provide an overview of the state of the art in approaches to NLG from SW-data.

2. BACKGROUND

convert a given input into a natural language output - output is dependent on the goal of the communication and the context. The format/characteristics of the input, and the requirements of the output goven the architecture of the generator. Methods range from simple pre-built templates to complex rule- and projection-based realisations THe 5 stages of NLG:

content selection: (cite [22]) open planning is an informed search of the most relevant nodes, most useful when the goal is to say everything there is to say about an object of the domain

discourse and information structuring : derivation of a coherent discourse structure for the content that is to be rendered in natural language. Popular approach is *Rhetorical Structure Theory* (RST)(cite [82]), due to clear distinction between main (nucleus) and supporting (satellite) arguments. Can be undertaken at the same time as content selection to aid in ensuring that the nodes selected form a cohenrent structure, and can be packaged into units of content known as *Elementary Discourse Units* (EDUs) (cite [35]).

aggregation

lexicalisation Use a language-oriented ontology to map a semantic representation onto an abstract linguistic representation *Generalised Upper Model* (GUM) (cite [10])

realisation and linearisation

3. NLG AND THE SEMANTIC WEB

The use of Semantic Web technologies can greatly simplify the traditional NLG process in a number of ways.

A common problem in NLG is the projection of the domain information onto the laguage information used for the NL rendering. Often solved by using a single bespoke combination knowledge base, however for SW data OWL/RDF provides support for modularisation of knowledge into separate layered ontologies that can represent the connections between domain and communication knwledge whilst still retaining a degree of portabilitly for the communication layer. The communication layer may also contain supplementary information about the context or user profile, for the determination of particularly relevant content during the NLG process. Alternatively, some approaches do not make use of a separate communication layer ontology, and instead annotate the domain classes and individuals directly with lexical information such as noun phrases, gender and plural forms, using the same Semantic Web formalisms as the original data. One such system is NaturalOWL (cite [56]), which uses traditional reasoning to help annotate OWL ontologies in a multilingual fashion. Finally, the labels of entities have been found by Power et al. [**?**] to be largely meaningful natural language words, and so domain-independent lexicon creation can be performed cheaply in exchange for a certain degree of fluency.

The NLG process can also benefit from the availability of standard SW reasoners in order to perform inference on the domain data and deliver natural language content in a more fluent and natural manner <come back to this with CNL>. The content selection process in many NLG techniques can be made to scale up to the large quantities of data available in SW contexts - many approaches extract fragments of interest based upon a maximum distance from an original node [?]. The increasing standardisation of formats across domains in SW data means that a NLG system developed for one purpose has the potential to be made protable across domains or applications. Task-specific knowledge may be layered separately from the domian knowledge, for example using the GUM [?], and in order to further reduce the amount of task-specific knowledge, the generated language constructions may be reduced to a controlled subset known as a *Controlled Natural Language* (CNL). Generic CNL-grammars are defined in terms of a formal language (i.e., OWL) and linguistic expressions, and for each new domain only lexicons need be defined [?]. NLG techniques are also able to make use of existing linked data concepts in SW-data, for example by using existing provenance models or FOAF information to supplement information available directly from the domain of interest.

Natural Language Generation can be used to present Semantic Web concepts and data in a user-friendly and easily understandable manner. Therefore, there are two classes of users that can most benefit from the application of NLG techniques to SW technologies: those who create and populate ontologies, and end-users that make use of data published and communicated via the Semantic Web. The first group would make use of a *Natural Language Interface* (NLI) for ontology engineering - typically they might be domain experts but not SW experts. The second group would be seeking specific information from the SW, and NLG techniques would be used to present that information in an appropriate format.

3.1 NLG for Ontology Engineering

In a NLI for ontology engineering, NLG techniques may be used to support and guide an author through the development of ontologies and the construction of valid queries. Providing natural language paraphrases of input definitions can help to highlight trivially satisfiable and other incorrect restrictions [1]. CNLs are useful in this context as they can provide unabiguous translations between SW concepts and natural language. As the NLI is presented during the creation of the SW resource, lexical annotations are often unavailable and so linguistic resources at this stage are typically derived automatically from the entity and relation labelling information available in the ontology. The content selection and discourse structing steps typically present in NLG are largely unnecessary for the NLI, as it can be assumed that the user will want access to all of the data in the same structure as the source ontology. Some basic aggregation is generally possible in the form of redundancy elimination and restriction combination (e.g., min-cardinality and max-cardinality may be replaced by “between ... and ...”). Finally the ontology may be lexicalised and realised via a CNL such as *Attempto Controlled English* (ACE), using the Attempto Parsing Engine [?].

3.2 NLG for Publication

The use of NLG for publication of SW data is a more complex problem than the generation of an authoring NLI, as the scope of communication goals is far broader. Depending on the context and requirements of the end user, the desired output may be anything from the answer for a specific question to a comprehensive summary of the complete ontology. For example, NaturalOWL [?] generates personalised multilingual descriptions of individuals and classes from the domain of interest. The NLG tasks rely on annotations to the input OWL ontology that specify, among other things, and item's interest for a given type of user. During content selection, the system orders information about items according to the properties of the current user profile, and then constructs a discourse structure using only a number of the most interesting facts. Items are annotated with the approprate lexical information, and after aggregation into suitable complex sentences presented to the end user. Not all SW systems rely on prior annotation of the data however. ArtEquAKT is a system that harvests biographical information about artists in order to provide a personalised dynamic biography in response to user queries [?]. Candidate fragments are selected and analysed, knowledge triples are extracted, consolidated and stored, and a personalised biography is generated according to the information gathered, the core ontology and a set of templates

4. STANDARDISATION

Generalised Upper Model [10] is in OWL-DL

5. FUTURE WORK

6. CONCLUSIONS

7. REFERENCES

[1] A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, and C. Wroe. Owl pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *Engineering Knowledge in the Age of the Semantic Web*, pages 63–81. Springer, 2004.