OntoVerbal: a Protégé plugin for verbalising ontology classes

Shao Fen Liang [1]*, Robert Stevens [1], Donia Scott [2] and Alan Rector [1]

[1]School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[2]School of Engineering and Informatics, University of Sussex, Falmer, Brighton, BN1 9QH, UK

ABSTRACT

OntoVerbal attempts to reduce the difficulties that non-ontology experts face in 'reading' ontologies, and the burden that ontology authors face in writing natural language definitions of classes. It does this by verbalising (i.e., automatically generating as natural language) the axioms of OWL classes. Its method relies on presenting, through the use of natural language generation (NLG), naturalistic descriptions of ontology classes as textual paragraphs. OntoVerbal has been implemented as a Protégé plugin that can offer an alternative 'English' view of a class and graphical views provided by various other Protégé plugins. The plugin provides automatic RDF label generation for ontology entities and a natural language description for each class, both for the asserted and 'inferred' forms of the class. We have made OntoVerbal, version 1.0, available for Protégé 4.1 via `http://swatproject.org/demos.asp`.

1 INTRODUCTION

Ontology development involves at least two 'hard' authoring activities: creating axioms in a new ontology and editing existing axioms. Thus it is fundamental to using an ontology that the author is able to understand its content. As a consequence, managing ontologies is a highly skilled task that tends to be carried out by specialists. A richly axiomatised ontology can be hard to read, either in a native OWL syntax or in some graphical presentation. Given the growing importance and proliferation of ontologies in the biomedical and other fields, the lack of ready access to their content is a major stumbling block to wider use. Also, natural language descriptions are a desirable feature of ontologies and mandated by the Open Biomedical Ontologies consortium, and as these are time-consuming to write, support for their production can be valuable (Stevens *et al.* (2011)).

OntoVerbal has been developed to help address these problems (Liang *et al.* (2011a)). It has applied methods from linguistics, psycholinguistics and computational linguistics to achieve its language generation (Liang *et al.* (2011b)). In particular, OntoVerbal deploys axioms of a selected class into a discourse structure. So axioms can be transformed into a set of sentences and then into a structured and well ordered paragraph that represents the class. OntoVerbal's aim is not perfect natural language, but a generic approach to producing acceptable English for a class' axioms.

2 ONTOVERBAL IN PROTÉGÉ

OntoVerbal generates a natural language paragraph for any selected class. To illustrate examples in this paper, we use the heart ontology[1] that describes the anatomy of a human heart. The ontology's axioms relating to the *Vavle* class are :

(<AnatomicalCavity>DisjointClasses <Valve>)
(<TricuspidValve>SubClassOf <Valve>)
(<PartialValve>SubClassOf <Valve>)
(<Valve>SubClassOf <AnatomicalConcept>)
(<SemiLunarValve>SubClassOf <Valve>)
(<VestigialCardiacValve>SubClassOf <Valve>)
(<MitralValve>SubClassOf <Valve>)
(<AtrioVentricularValve>EquivalentTo (<Valve>and (<hasValveInput> some <AtriumCavity>) and (<hasValveOutput>some <VentricularCavity>))
)

OntoVerbal has structured and ordered these axioms into a English paragraph (Figure 1) according to Rhetorical Structure Theory as

> A valve is a kind of anatomical concept. More specialised kinds of valve are mitral valve, partial valve, semi lunar valve, tricuspid valve and vestigial cardiac valve. Also, a valve is different from an anatomical cavity. Another relevant aspect of a valve is that an atrio ventricular valve is defined as a valve that has valve input an atrium cavity and has valve output a ventricular cavity.

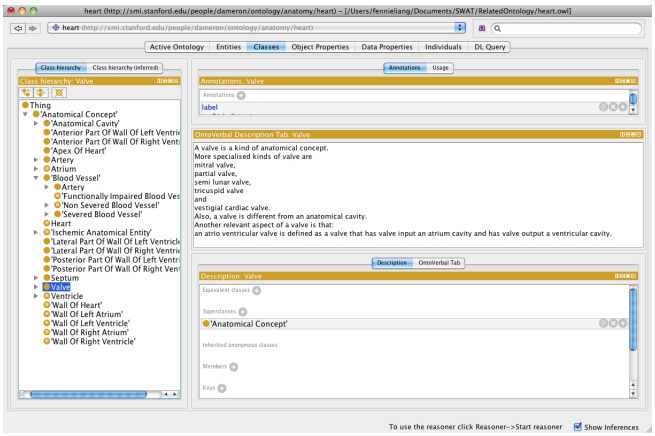

Fig. 1. the OntoVerbal description of Valve `http://www.swatproject.org/publications/Valve.jpg`

The names of the ontology's classes provide much of the lexical content of the generated English, and so if the ontology

*To whom correspondence should be addressed: fennie.liang@cs.man.ac.uk

[1] retrieved from `http://owl.cs.manchester.ac.uk/repository/download?ontology=http://smi.stanford.edu/people/dameron/ontology/anatomy/heart\&format=RDF/XML` downloaded April 2012.

does not use well formed labels, but only URI fragments, this will have a detrimental effect on OntoVerbal's verbalisation; for example, instead of reading *A valve is a kind of anatomical concept* the reader will confronted with *A <URI#Valve> is a kind of <URI#AnatomicalConcept>*. In the latter case, OntoVerbal will also lose some of its abilities in the paragraph generation, such as putting articles in right places. For this reason, OntoVerbal will make its own labels from URI fragments. The natural language generation engine will supply labels for ontology classes, object properties, data properties and individuals. It breaks entity URI fragments such as CamelCase, Under_score or a mixture of both into separate words. The message "RDF labels have been generated successfully" will popup after the generation has completed. One thing to note is that the generator only provides labels for those entities without pre-existing RDF labels (i.e. rdfs:label). Therefore, if some classes in the active ontology have already got their RDF labels then OntoVerbal only produces labels for labelless classes.

OntoVerbal can also provide descriptions for classes after reasoning. The description for the Valve class after running a reasoner becomes:

> A valve is a kind of anatomical concept. A more specialised kind of valve is partial valve. Also, a valve is different from a left atrium cavity, a coronary artery, a vestigial cardiac valve, a conus artery, a right marginal artery, a pulmonary valve, ... an apex of heart, an anterior part of wall of right ventricle, a valve of coronary sinus, a left circumflex artery and an aorta. Another relevant aspect of a valve is that an atrio ventricular valve is defined as a valve that has valve input an atrium cavity and has valve output a ventricular cavity.

After reasoning, much more is known about the class and this obviously has an effect on the verbalisation. However, a reader needs verbalisation of both views at different times—as is provided in tools such as Protégé. The inferred description (figure 2), in fact, contains 64 disjoint classes, and some of them are omitted in this paper. The red coloured classes showed in Protégé are inconsistent, but OntoVerbal's descriptions has ignored the red colour and still generate descriptions using the inferred axioms.

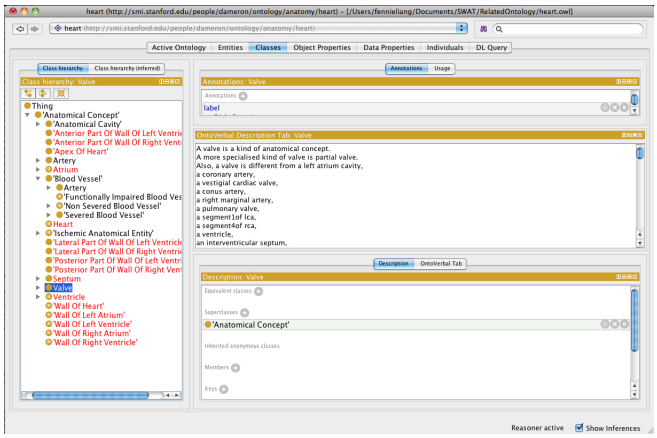

Fig. 2. inferred OntoVerbal description of Valve
`http://www.swatproject.org/publications/InferredValve.jpg`

3 DISCUSSION

Currently, OntoVerbal uses lightweight linguistic approaches for its NL paragraph generation. The main reason is that OntoVerbal is intended as a real time application and employing heavy linguistic methods will slow down its performance. Also, since the aim is not to produce perfect English, but rather English that is acceptable for the purpose of revealing clearly the content of the ontology, the output of OntoVerbal will at times include incorrect articles and/or plurals, and be clumsy in places. Since OntoVerbal is intended to be faithful to its input, in contexts where the selected class contains many related axioms, it will sometimes produce excessively long paragraphs. Given that our aim is for rapid generation of coherent English text for any class, we feel that these compromises are acceptable.

The OntoVerbal Description tab can generate paragraphs for classes without RDF labels, but the text will be of reduced quality compared to those with hand-crafted labels. The OntoVerbal Description tab can also provide more specific descriptions for classes if a reasoner is used. OntoVerbal will not replace hand-crafted natural language descriptions, but can provide a substitute in their absence. It also provides an alternative view to an ontology's axioms in a reasonably familiar natural language form that seeks to 'ease' access to often complex ontologies.

ACKNOWLEDGEMENTS

This work is part of the Semantic Web Authoring Tool (SWAT) project (see www.swatproject.org), which is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/G032459/1, to the University of Manchester, the University of Sussex and the Open University. We are grateful for the comments received from our colleagues on the project.

REFERENCES

Liang, S. F., Stevens, R., Scott, D., and Rector, A. (2011a). Automatic verbalisation of SNOMED classes using ontoverbal. *Proceedings of the 13th Conference on Artificial Intelligence in Medicine, AIME 2011*, pages 338–342.

Liang, S. F., Scott, D., Stevens, R., and Rector, A. (2011b). Unlocking medical ontologies for non-ontology experts. *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011*, pages 174–181.

Stevens, R., Malone, J., Williams, S., Power, R., and Third, A. (2011). Automating generation of textual class definitions from OWL to English. *Journal of Biomedical Semantics*, **2**(Suppl 2), S5.