Automating class definitions from OWL to English

Automating Generation of Textual Class Definitions from OWL to English

*Robert Stevens***, James Malone**, Sandra Williams* and Richard Power**

** Department of Computing, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK*

*** European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK*

**** School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, UK*

ABSTRACT

Text definitions for entities within bio-ontologies are a cornerstone of the effort to gain a consensus in understanding and usage of those ontologies. Writing these definitions is, however, a considerable effort and there is often a lag between specification of the entities in the ontology and the development of the text-based definitions. As well as these text definitions, there can also be logical descriptions and definitions of an ontology's entities. The goal of natural language generation (NLG) from ontologies is to take the logical description of entities and generate fluent natural language. We should be able to use NLG to automatically provide text-based definitions from an ontology that has logical descriptions of its entities and thus avoid the bottleneck of authoring these definitions by hand.

In this paper we present some early work in using NLG to provide such text definitions for the Experimental factor Ontology (EFO). We present our results, discuss issues in generating text definitions, and highlight some future work.

1 INTRODUCTION

The Experimental Factor Ontology (EFO) is an application ontology used to describe experimental variables in functional genomics data (Malone et al., 2010). EFO uses OWL to produce a rich, axiomatic description of classes in the domain; see, for example, the Gene Expression Atlas (Kapushesky et al., 2010). The aim of EFO was three-fold: i) to provide coverage for functional genomics data by importing reference ontology classes or creating new classes, ii) add user-friendly labels and synonyms to these classes, iii) create axiom-rich class descriptions in OWL. As a consequence of this prioritization, text definitions of many of the classes in EFO are not present.

Capturing logical definitions is powerful as it enables automated consistency checking and complex querying, however such definitions can be confusing to a user not familiar with OWL.

Text definitions are an important factor in the usability of ontologies, so there is a need for both forms of description within EFO and other ontologies. As a result we would like both logical and text definitions for the classes in EFO; authoring both is time consuming and there is an issue of keeping the two in step. Given that we have a logical representation of the entity, we should be able to have a text version of the same definition automatically generated with no problems of synchronization and no delay in provision of the important text definition.

The task of generating texts from ontologies has been called ‘ontology verbalisation’ (see Smart, 2008). A number of verbalisers for OWL (Web Ontology Language) have been developed, with varying aims and limitations: for instance, some are concerned only with ABox verbalisation (e.g., Hielkema 2009; Galanis and Androutsopoulos, 2007); others produce only separate sentences, one for each OWL axiom (e.g., Kalijurand, 2007). Our system for describing classes[1] has much in common with this work, but differs in two ways: first, we cover at present only a subset of OWL (the simple description logic EL++); second, instead of realising axioms one by one, we apply some rules for organisation and aggregation, using generic methods applicable to any ontology, so as to provide coherent descriptions for each class (or individual or property).

From a computational linguistics perspective, ontology verbalisation has some unusual features. Most applications in natural language generation aim to produce high-quality text in restricted domains for which specialised text-planners, grammars and lexicons have been developed. In verbalising an ontology we aim for texts that are useful and understandable but not necessarily of the highest quality, using methods that are domain-general. The challenge is thus to find generic techniques for (a) grouping related axioms on the same class; (b) realising logical patterns in English; (c) ag-

[*] To whom correspondence should be addressed.

[1] We focus here for simplicity on descriptions of atomic classes, but the program actually generates descriptions for individuals and properties as well.

gregating axioms sharing a common pattern, such as use of the same property, so that they can be expressed efficiently in a single sentence, and (d) inferring lexical entries for atomic entities (classes and properties) from identifiers and labels in the ontology, with due attention to details like correct parts of speech and plural forms.

What we present here is a prototype for doing this using EFO as our text-bed. The results already look promising and presentation of generated text definitions to users has suggested ways in which our techniques can be improved.

2 DESCRIPTION GENERATOR

The description generator accepts as input an ontology encoded in OWL/XML format, and produces as output a text file that lists the atomic entities, in alphabetical order of their English names, accompanied by descriptions in English sentences. The descriptions are produced by a program that collects all the axioms relating to a given class, groups them according to common structure, realises each group through an English sentence, and assembles the resulting sentences into a paragraph. Sentence generation is accomplished using a generic grammar based on logical patterns in OWL (limited at present to EL++), together with a lexicon for realising atomic entities. A provisional lexicon is derived automatically from the identifier names or annotation labels in the input ontology; if desired it can be corrected by hand.

At present this is a prototype system with several limitations. First, as already mentioned, it is restricted to a fragment of OWL − albeit one that is commonly used in practice. Second, it is implemented in SICStus Prolog, a language well suited to fast prototyping, but not to large-scale applications: accordingly it cannot deal with input files larger than 2 Mb. The production version of the converter will not have this limitation, but it was found acceptable for the rapid prototyping it allowed. Thirdly, the methods used for deriving lexical entries from identifiers and labels are rudimentary (and of course they assume that these names are based on English). Finally, the grammar for realising logical patterns is mostly based on intuition (either our own, or that of previous researchers); as yet there are no systematic empirical studies on the best linguistic formulations.

The process of generating descriptions has five phases:

(1) Transcoding from OWL/XML to Prolog.

(2) Constructing a lexicon for atomic entities.

(3) Selecting the axioms relevant for describing each class.

(4) Aggregating axioms with a similar structure.

(5) Generating sentences from (possibly aggregated) axioms.

The input to the generator is an ontology in OWL/XML format; this is transcoded to a Prolog format analogous to OWL Functional Syntax. All identifiers for atomic entities are then listed, and for each identifier a provisional lexical entry is computed. To do this, the program first checks whether a label is provided in an annotation assertion; if so, the lexical entry is based on this label, otherwise it is based on the identifier itself. To obtain the lexical entry from an identifier, the program discards the namespace, then splits the remaining string into words on the assumption that word boundaries are indicated by underline characters or capital letters; some simple heuristics are then applied to massage the resulting word string into a plausible English phrase. It is assumed that the syntax of each phrase will be severely constrained as follows: individuals are expressed by proper names; classes by common nouns (with singular and plural forms); and properties by transitive verbs (simple or compound) with slots for a subject and an object. Lexical entries are saved as Prolog terms with four arguments: identifier, part of speech, singular form, and plural form (if relevant):

```
lex(class(EFO_0000322),noun, 'cell line', 'cell
lines').
lex(class(EFO_0002095),noun,'22rv1','22rv1s').
```

As can be seen, the lexicon is reliant on the names/labels provided by the ontology builder, and uses no other source of evidence. It therefore assumes for instance that ‘22rv1’ will be an English common noun, and derives the regular plural form by adding -s.

Once the lexicon has been built, the ontology is searched for axioms that describe each class, property and individual in the lexicon (i.e., each atomic entity). For example, to describe the atomic class *EFO_0002095* the algorithm retrieves all axioms in which this class occurs as a top-level argument (e.g., *A* or *B* if the axiom is *subClassOf(A,B)*) obtaining the following set:

```
subClassOf(class(EFO_0002095),
  class(EFO_ 0000322)).
subClassOf(class(EFO_0002095),
  objectSomeValuesFrom(
    objectProperty(#bearer_of),
    class(EFO_0001663))).
subClassOf(class(EFO_0002095),
  objectSomeValuesFrom(
    objectProperty(#derives_from),
    class(#NCBITaxon_9606))).
subClassOf(class(EFO_0002095),
  objectSomeValuesFrom(
    objectProperty(#derives_from),
    class(EFO_0000858))).
```

At this stage, we could simply generate a sentence for each axiom; however, the resulting text would contain many repetitions; for example, for the set of axioms for cell line 22RV1 we would obtain:

```
A 22rv1 is a cell line.
A 22rv1 is something that is bearer of a prostate carcinoma.
A 22rv1 is something that derives from a homo sapiens.
A 22rv1 is something that derives from a prostate.
```

To obtain more fluent descriptions, our algorithm combines axioms that share a common pattern and differ in only one constituent. Thus in the example we are considering, it finds three axioms having the following abstract form:

```
subClassOf(Class,
  objectSomeValuesFrom(Property, Class)).
```

These are combined to obtain the following aggregated axiom in which the varying constituent is replaced by a list:

```
subClassOf(class(EFO_0002095),
  [objectSomeValuesFrom(
    objectProperty(#bearer_of),
      class(EFO_0001663)),
    objectSomeValuesFrom(
      objectProperty(#derives_from),
        class(#NCBITaxon_9606)),
    objectSomeValuesFrom(
      objectProperty(l#derives_from),
        class(EFO_0000858))]).
```

The grammar can then realise the aggregated axiom by a single sentence rather than several sentences.

The final stage is to generate a sentence for each axiom, (or aggregated axiom), thus obtaining a description of the class (or other atomic entity). This is done by feeding each axiom to a Definite Clause Grammar with rules for each logical pattern in EL++; this grammar will consult the lexicon whenever it needs to express an atomic entity. As an example, here is the rule used for realising a two-argument statement with the functor *equivalentClasses*; as can be seen, it presupposes a further rule for realising classes by indefinite noun phrases:

```
s(equivalentClasses(Class1,Class2), Lexicon) -->
  np(a, Class1, Lexicon),
  [is], [defined], [as],
  np(a, Class2, Lexicon).
```

At present we have no heuristics for ordering axioms within a description, so the sentences are assembled into a paragraph following the same order in which the axioms were originally retrieved from the ontology.

For examples from the output for the EFO ontology, see Table 1.

3 METHOD AND RESULTS

We verbalised a subset of 50 cell lines from EFO. These included 45 without (and 5 with) text definitions; 10 also had necessary and sufficient conditions; 45 had only necessary conditions from just a subclass axiom to several restrictions. The cells covered a range of human and mouse cells, some of which exhibited diseases. Table 1 provides some examples of text definitions. (Supplementary information can be found at `http://mcs.open.ac.uk/nlg/SWAT/bio-ontologies.html.`)

Table 1. Example of natural language definitions extracted from corresponding OWL axioms. NB, all cell lines shown have the 'subclass of cell line' axiom. *note these subclass relations are placed on the subclasses but we illustrate them here for context.

Class label	OWL axioms (Manchester syntax)	Natural Language Definition Extracted
22rv1	bearer_of some 'prostate carcinoma' derives_from some 'Homo sapiens' derives_from some prostate	A 22rv1 is a cell line. A 22rv1 is all of the following: something that is bearer of a prostate carcinoma, something that derives from a homo sapiens, and something that derives from a prostate.
HeLa	bearer_of some 'cervical carcinoma' derives_from some 'Homo sapiens' derives_from some cervix derives_from some 'epithelial cell'	A he la is a cell line. A he la is all of the following: something that is bearer of a cervical carcinoma, something that derives from a homo sapiens, something that derives from an epithelial cell, and something that derives from a cervix.
Ara-C-resistant murine leukemia	has subclass b117h* has subclass b140h*	A ara c resistant murine leukemia is a cell line. A b117h, and a b140h are kinds of ara c resistant murine leukemias.
GM18507	derives_from some 'Homo sapiens' derives_from some lymphoblast has_quality some male	A gm18507 is all of the following: something that has as quality a male, something that derives from a homo sapiens, and something that derives from a lymphoblast.

A sample of 10 of the 50 verbalisations were selected based on the widest range of axioms (i.e. number and type on each class) and an on-line survey was created. Users of the ontology interest group at EBI and the Functional Genomics group at EBI were asked to rate on a scale of 1 to 5 how much they thought the definitions were readable such that their intention could be understood. Participants were also able to add specific comments to each definition.

Table 2. Summary of survey results on natural language definitions. Judgements range from 1 (understandable) to 5 (not understandable). The survey was completed by 21 people (questions did not require an answer).

Judgements	1	2	3	4	5
Totals	25.7% (48)	32.1% (60)	27.3% (51)	9.1% (17)	5.9% (11)

4 DISCUSSION

Perhaps the most interesting outcome of the process was that the new natural language definitions exposed an oddity in one of the EFO classes that had not been previously identified. The definition for ‘Ara-C-resistant murine leukemia’ indicated that the subclasses b117h and b140h were both types of this class which implied that they were diseases rather than cell lines. Ontologically, the classes are subtypes of cell line; however, it is clear that the label for this class is incorrect and would be better served by, for example, appending ‘cell line’ to the end of the class label.

The survey results also revealed an interesting trend towards simplicity in definitions. The class definition that was deemed most understandable was BDCM (described in Table 1) which only asserts that the class is a cell line. The most common remark left was for class GM18507 (also described in Table 1). Here, participants commented that the line ‘has as quality a male’ was confusing. Similarly, some comments were also made on the language of ‘bearer of’ in the context of a disease; such relationships come from using the relation ontology (Smith et al., 2005) as part of the OBO process.

Overall, the modal answer given was the 2nd highest rank, which appears to indicate in this limited response that answers were at least some way to conveying an understandable meaning.

5 CONCLUSION

We have presented an early prototype for generating text definitions from logical descriptions of classes. We verbalised a selection of cell line classes from EFO and undertook an informal survey. Whilst it is not possible to draw statistically significant conclusions from this kind of survey, it has suggested that the text definitions we generated are understandable and useful within the context of an ontology with sparse use of text definitions.

Suggestions for improvements in the English realization of the definitions have been gathered. Our initial verbalisations made the OWL semantics explicit (for example, by saying “Every cell line is ….). This was found to be obstructive to understanding and we replaced it with a simple “A cell line is….”. Similarly, explicit verbalisations of all relationships was seen to reduce understanding; for example qualities of cells. Such dependent entities could become adjectival forms of the independent entities in which they inhere (cell has quality female becomes female cell). Similarly, the formal ontological nature of some relationships reduced understanding; this suggests that alternative wording be found that is closer to the user’s domain without loss of precision. In addition, a variety of output styles is possible, with some being closer to domain language, some making more of OWL’s semantics explicit whilst others preserve more of the ontology’s form. In the short term we will continue to generate EFO text definitions and improve their quality for that user group. Overall, however, we do need a systematic survey of appropriate verbalisations of definitions to inform such renderings.

Whilst there remains much to do to improve our verbalisations, we are encouraged by the reactions to these early attempts; the providers of EFO are now including these generated text definitions in their latest release (version 2.3). We foresee that generic tools for verbalisation of ontologies from logical descriptions will be both possible and useful.

ACKNOWLEDGEMENTS

Sandra Williams, Richard Power and Robert Stevens are funded by the SWAT project (EPSRC grants EP/G033579/1 and EP/G032459/1); James Malone is funded by EMBL and EMERALD (project number LSHG-CT-2006-037686). Thanks to EBI Functional Genomics group for comments on generated definitions.

REFERENCES

Galanis, D. and I. Androutsopoulos (2007) Proceedings of the 11th European Workshop on Natural Language Generation, Schloss Dagstuhl, Germany, 143-146.

Hielkema, F (2009) Using Natural Language Generation to Provide Access to Semantic Metadata. PhD Thesis, University of Aberdeen.

Kaljurand, K..(2007) Attempto Controlled English as a Semantic Web Language. PhD thesis, Faculty of Mathematics and Computer Science, University of Tartu.

Kapushesky, M. et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. Nucleic Acids Res. 38:D690–D698.

Malone, J. et al. (2010) Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 26(8), 1112-1118.

Smart, P. (2008) Controlled Natural Languages and the Semantic Web. Technical Report ITA/P12/SemWebCNL, School of Electronics and Computer Science, University of Southampton.

Smith, B., W. Ceusters, B. Klagges, J. Kӧhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L Rector and C. Rosse (2005) Relations in Biomedical Ontologies. Genome Biology, 6:5,R46.