Liang, S.F., Stevens, R., Scott, D. and Rector, A. (2011). Automatic Verbalisation of SNOMED Classes Using OntoVerbal. Proceedings of the 13th Conference on Artificial Intelligence in Medicine (AIME'11). Bled, Slovenia.

Automatic Verbalisation of SNOMED Classes Using OntoVerbal

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, UK M13 9PL
{Fennie.Liang, Robert.Stevens, Rector}@cs.man.ac.uk
[2]School of Informatics, University of Sussex, Falmer, Brighton, BN1 9QH, UK
D.R.Scott@sussex.ac.uk

Abstract. SNOMED is a large description logic based terminology for recording in electronic health records. Often, neither the labels nor the description logic definitions are easy for users to understand. Furthermore, information is increasingly being recorded not just using individual SNOMED concepts but also using complex expressions in the description logic ("post-coordinated" concepts). Such post-coordinated expressions are likely to be even more complex than other definitions, and therefore can have no pre-assigned labels. Automatic verbalisation will be useful both for understanding and quality assurance of SNOMED definitions, and for helping users to understand post-coordinated expressions. *OntoVerbal* is a system that presents a compositional terminology expressed in OWL as natural language. We describe the application of *OntoVerbal* to SNOMED-CT, whereby SNOMED classes are presented as textual paragraphs through the use of natural language generation technology.

Keywords: ontology verbalisation, natural language generation, describing ontologies.

1. Introduction

SNOMED-CT, managed by the International Health Terminology Standards Development Organisation (IHTSDO), is now mandated as a terminology for use in electronic health records in numerous countries including the USA, UK, Canada, Australia, several countries in continental Europe, and beyond. SNOMED describes diagnoses, procedures, and the necessary anatomy, biological processes (morphology) and the relevant organisms that cause disease. The goal is for terms from SNOMED to form a controlled vocabulary for filling electronic health records, with the controlled usage of terms, coupled with the hierarchy of SNOMED, enabling extensive querying to be made and statistics to be gathered.

SNOMED is developed using a Description Logic (DL) [2]. The DL structure allows compositional descriptions to be made of entities in a domain; that is, entities are described in terms of other entities, but this comes at the cost of cognitive

complexity and unfamiliar notation. For example, the rendering of the concept heart disease in the Web Ontology Language (OWL) is:

Class: Heart disease
EquivalentTo: Disorder of cardiovascular system
and RoleGroup some (Finding site some Heart structure)

While such descriptions are explicit, and can aid automated reasoners in building the terminology, they can be hard for humans to understand.

Other terminologies often include natural language definitions corresponding to the logical definitions. These should be easier to understand, especially when in a style of natural language used by the community in question. For instance, the above example could be verbalised as "*A heart disease is a disease that is found in a heart structure*". Such natural language definitions are, however, time-consuming to produce by hand. We have built a natural language verbaliser, *OntoVerbal*, to help automate the process of making OWL ontologies such as these more transparent. Automatic generation of natural language from knowledge representations such as OWL is known as "verbalisation" [3].

There is clearly an intuitive correlation between axioms and sentences, and between groups of related axioms and paragraphs. In generating such paragraphs we need to ensure that they are more than simply collections of individual sentences, each expressing a given axiom in the ontology; instead, they need to be structured, coherent and fluent. This can be achieved through four main operations: (a) grouping axioms together based on a shared topic; (b) aggregating axioms sharing a common pattern; (c) organising these as sentences according to theories of discourse structure, such as Rhetorical Structure Theory (RST) [7]; and (d) making use of linguistic devices designed to make the text hang together in a meaningful and organised manner — e.g., conjunctions [9], discourse markers [4] and punctuation.

2. The OntoVerbal System

OntoVerbal starts by grouping axioms relating to a designated class according to their relations to that class and the complexity of the grouped axioms' arguments. Axioms are defined as having *direct* relations to a class if the class is the first class that appears in its argument. For example an axiom like "A SubClassOf B" is in a *direct* relation to the class A, but is in an *indirect* relation to the class B. This example also represents a *simple* axiom, as only named classes appear in its argument. Axioms are classified as *complex* when they contain not only named classes, but also properties, cardinalities or value restrictions, or combinations of named classes in anonymous class expressions — e.g.: "A EquivalentTo B and hasRestriction some R" where R is the argument and contains another class expression.

Distinguishing between *direct* and *indirect* axioms allows all of the information about a given class to be brought together and presented in a rhetorically coherent way with a single topic; it also provides a framework for indicating to the reader when a topic has changed, by using linguistic devices for maintaining coherence of the text. We achieve this coherence through the application of RST, a theory of discourse

coherence that addresses issues of semantics, communication and the nature of the coherence of texts, and which plays an important role in computational methods for generating natural language texts [10].

The *OntoVerbal* system has three main processes for axioms collected from a designated class.

The first process verbalises all *simple* axioms, where*by* axioms of the form "A SubClassOf B" are expressed as "*an* A *is a* B*"*. Where there are more subclass axioms, for example, "A SubClassOf C" and "A SubClassOf D" we treat this as a case requiring aggregation [6] to produce, for example, *"an* A *is a* B*, a* C *and a* D*"*. While SubClassOf relations are expressed through "is a" (and other semantically equivalent expressions), we treat EquivalentClass relations as definitions. So, for example, if the SubClassOf relation in the previous example were instead an EquivalentClass relation, the resulting text would be *"an* A *is defined as a* B*, a* C *and a* D*"*. In the cases mentioned so far, the relations are *direct* ones. Axioms in *indirect* relations will need to be inverted so as to be directed to the designated class. For example, in a context where the topic is B, the axiom we saw earlier, "A SubClassOf B", would be more properly expressed as *"a* B *includes an* A*"*, if the thread of discourse is to be maintained (i.e., for the text to "hang together"). In this context, the alternative, earlier, rendition would introduce a disfluency through the sudden shift of topic from B to A [11], and thus place an additional cognitive load on the reader [5].

The second process verbalises *complex*, but *direct* axioms. *Complex* axioms are necessarily longer than *simple* ones, and we ensure the maintenance of fluency and coherence between sentences through the use of relative clauses, discourse markers and punctuation. In our approach, *complex* SubClassOf axioms are expressed as "*an* A *is a* B *that …*" and *complex* EquivalentClass axioms are expressed as *"an* A *is defined as a* B *that …*". We use the discourse marker "additionally" to connect sentences from the outputs of the *simple* to the *complex* (*direct*) axioms process. This leads to results such as "*An* A *is a* B*, which includes a* C*, and both an* X *and a* Y *are an* A*. Additionally, an* A *is an* X *that …, and is defined as a* Y *that …*". However, if no sentence occurs from the *simple* axioms process then the discourse marker will be omitted.

The third process verbalises *complex*, but *indirect* axioms. In cases where *complex* axioms are in *indirect* relation to the designated class, this feature must be signalled in the generated text, since a change of topic (i.e., a new subject) is introduced. Without such signalling, the text will lack coherence and fluency and be harder to understand. We introduce these axioms through the use of key phrases, such as: *"Another relevant aspect of an* A *is that"* or *"Other relevant aspects of an* A *include the following:"*.

3. Discussion

Our attempts at verbalising SNOMED have so far focused on generating an appropriate rhetorical structure that conveys the basic meaning of the concept. There are obvious issues with articles and plurality in our verbalisations, and these will be addressed as *OntoVerbal* progresses. There are, however, several more substantial

issues in verbalisation to tackle, some of which are generic to DL ontologies and some of which are peculiar to SNOMED.

For example, many SNOMED expressions are, in fact, redundant, and so *OntoVerbal*'s near-literal verbalisations can also seem redundant. Consider, for example, the generated paragraph:

> *A* Heart disease *includes a* Disorder of cardiac function*, and both an* Acute disease of cardiovascular system *and a* Heart disease *are an* Acute heart disease*. Additionally, a* Heart disease *is defined as a* Disorder of cardiovascular system *that has a* Finding site *some* Heart structure*. Other relevant aspects of a* Heart disease *include the following:* Hypertensive heart disease *is a* Heart disease *that has a* Finding site *some* Heart structure *and is* Associated with *some* Hypertensive disorder*;* a Chronic heart disease *is defined as a* Chronic disease of cardiovascular system *that is a* Heart disease *and has a* Clinical course *some* Chronic*; a* Structural disorder of heart *is defined as a* Heart disease *that has an* Associated morphology *some* Morphologically abnormal structure *and a* Finding site *some* Heart structure*; a* Disorder of cardiac ventricle *is defined as a* Heart disease *that has a* Finding site *some* Cardiac ventricular structure*.*

where the underlined sentence is clearly redundant in its multiple references to "chronic". This is a problem that combines issues for both the description logic representation and the verbalisation.

Similarly, because SNOMED's representation lacks a disjunction operator, it makes awkward use of complex intersections and this leads to infelicities of the sort seen here:

> *A* Lower body part structure *is a* Lower body structure*, which includes a* Pelvis and lower extremities. *An* Abdominal structure *is a* Chest and abdomen*, an* Abdomen and pelvis*, a* Structure of subregion of trunk *and a* Lower body part structure.

which should more properly read as the following, with the inclusion of the underlined words:

> *A* Lower body part structure *is a* Lower body structure*, which includes a* Pelvis and lower extremities *and an* Abdominal structure. *An* Abdominal structure *is a* Chest and abdomen*, an* Abdomen and pelvis*, a* Structure of subregion of trunk *and a* Lower body part structure.

We will be addressing these problems, refining our currently rudimentary treatment of plurals and articles, and exploring the use of layout (e.g., bulleted lists) [8] in future versions of the system.

There are two aspects of SNOMED that lead us to exceptionally depart for our stated goal of faithfulness of the generated output to the SNOMED input. The first involves the RoleGroup construct that is supposed to group aspects together. In reality, however, most of them appear to have only one role group, and therefore the RoleGroup is redundant in these cases. Our approach to this in *OntoVerbal* is to simply ignore such constructs.

The second aspect relates to multiple terms for class IDs. Most SNOMED class IDs have several associated terms, such as "preferred term" (that is expected to be used most commonly in medical records and interfaces), and "fully specified term" (that is intended to be completely unique and self explanatory) [1]. Given that there are thus many synonyms referring to a single concept, we have decided to let *OntoVerbal* in all cases use the stated "preferred term" to express SNOMED concepts.

4. Conclusions

OntoVerbal currently produces near-literal verbalisation of SNOMED concepts in well structured natural language. This addresses the problems of comprehension of complex logical descriptions of medical concepts in terminologies such as SNOMED, but also other DL based terminologies and ontologies. The need for automatically-generated verbalisations is especially important when post co-ordination is used, as without it there is no possibility of the provision of natural language versions of the concepts. Such verbalisations can provide much-needed documentation for artifacts like SNOMED, thereby making their content more accessible to users.

OntoVerbal's output currently lacks some linguistic polish, but having addressed the basic issues of grouping, aggregation and rhetorical structure in the verbalisations, other features of the output can be addressed. Then the role of automatic verbalisation of complex axiomatic descriptions in error detection, facilitation of users' comprehension, and creation of innovative presentations for artefacts such as SNOMED, can be explored.

Acknowledgments. 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.

References

1. SNOMED-CT User Guide, **http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_CT/About_SNOMED_CT/Use_of_SNOMED_CT/SNOMED_CT_User_Guide_20090731.pdf**
2. Baader, F., Horrocks, I., and Sattler, U.: Description logics as ontology languages for the semantic web. Lecture Notes in Artificial Intelligence. 2605, 228-248. (2005)
3. Baud, R.H., Rodrigues, J.-M., Wagner, J.C. *et al*.: Validation of concept representation using natural language generation. Journal of the American Medical Informatics Association, 841. (1997)
4. Callaway, C.B.: Integrating discourse markers into a pipelined natural language generation architecture. 41st Annual Meeting on Association for Computational Linguistics. 1, 264-271. (2003)
5. Clark, H.H.: Psycholinguistics. MIT Press, (1999)
6. Dalianis, H.: Aggregation as a subtask of text and sentence planning. In: Florida AI Research Symposium, FLAIRS-96, pp. 1-5. J.H.Stewman, (1996)
7. Mann, W.C., and Thompson, S.A.: Rhetorical Structure Theory: toward a functional theory of text organisation. Text. 8, 243-281. (1988)
8. Power, R., Scott, D., and Bouanyad-Agha, N.: Document structure. Computational Linguistics. 29, 211-260. (2003)
9. Reape, M., and Mellish, C.: Just what is aggregation, anyway? In: European Workshop on Natural Language Generation. (1999)
10. Scott, D., and Souza, C.S.d.: Getting the message across in RST-based text generation. In: Current research in natural language generation, Mellish, C., Dale, R. and Zock, M., eds., pp. 31-56: Academic Press, (1990)
11. Walker, M.A., Joshi, A.K., and Prince, E.F.: Centering Theory in Discourse. Oxford University Press, (1998)