The Semantic Web as a *Linguistic* Resource: Opportunities for Natural Language Generation

Chris Mellish * , Xiantang Sun

Dept of Computing Science
University of Aberdeen
King's College
ABERDEEN AB24 3UE, UK

Abstract

This paper argues that, because the documents of the semantic web are created by human beings, they are actually much more like natural language documents than theory would have us believe. We present evidence that natural language words are used extensively and in complex ways in current ontologies. This leads to a number of dangers for the semantic web, but also opens up interesting new challenges for natural language processing. This is illustrated by our own work using natural language generation to present parts of ontologies.

Key words: Ontologies, Semantic web, Natural language processing

1 Preamble

The work described in this paper originated as part of a project to generate natural language from ontologies for the semantic web. On the face of it, this seems like a fairly straightforward application of natural language generation (NLG) research, which studies the task of generating appropriate linguistic material from originally non-linguistic inputs. Indeed, there is a growing amount of work on NLG from semantic web material, given the practical importance of presenting such material to knowledge engineers and users.

* Corresponding author.
Email addresses: cmellish@csd.abdn.ac.uk (Chris Mellish), xsun@csd.abdn.ac.uk (Xiantang Sun).

Standardly, building an NLG system requires constructing a mapping (sometimes called a lexicon) between the terms of the input representation and natural language words and phrases [11]. This mapping is always domain-dependent since, although different applications may share the same basic *syntax* for their input representations (e.g. perhaps some syntax for first order logic), nevertheless each domain has its own idiosyncratic repertoire of predicates, concepts, etc. and these map onto natural language in ways that correspond to specific language use in the domain [6].

When we started to look seriously at semantic web ontologies as the domain for NLG, it became increasingly clear that viewing human-written ontologies as just another non-linguistic form of input was missing the point in many respects. We were forced to question the traditional view of how ontologies relate to natural language and therefore what NLG from ontologies should or could be.

2 The Relation between Ontologies and Natural Language: In Theory

The semantic web relies on the representation and exchange of knowledge using agreed terms. These terms are listed and further specified in *ontologies*, logical theories specifying conceptualisations of parts of the world. These conceptualisations are simplified models of the world developed for particular purposes. In the end, successful exchange of knowledge relies on an agreement on the semantics of the terms in an ontology and in the use of these terms in ways consistent with this semantics. This is usually only possible if the intended use is similar to that envisaged when the ontology was developed.

The terms in an ontology are different in kind from natural language words, although often they appear to be related to actual words in English or some other human language [5]:

(1) Terms in an ontology are given a precise formal, but shallow, description, whereas natural language word senses can only be defined informally and in a way that relies on deep human knowledge. For instance, the difference between the English words "mistake", "error", "blunder" and "slip" involves subtleties (about amount of criticism expressed and assumed accidentalness of the described event) that could not easily be stated in current ontology definition languages.
(2) The whole point of ontologies is to ensure that there is exactly one meaning for each ontology term; this contrasts with the situation in natural language where there are complex word-meaning relations. For instance, the English word "leg" is ambiguous between (at least) part of a piece

of furniture, part of an animal or part of a journey. Conversely, different words like "mistake" and "blunder" might be considered to have very similar meanings for some purposes.

(3) Ontologies are designed to be complete and minimal for specific applications; human languages are open-ended and idiosyncratic, with gaps and duplications. For instance, Spanish has no word for the concept of "stab".

(4) Terms in an ontology are carefully chosen for relevance in some domain or for some intended use. Natural language words, on the other hand, reflect the world view/culture of the language users and the historical development of this view/culture in much more complex ways.

In summary, ontologies can be regarded as a formalisation of some kind of ideal, "good practice" in natural language word use, where communication can be precise and successful every time. This is much the same as the way in which formal logic arose as an attempt to formalise "good practice" in natural language argumentation. In both cases, the formalisation captures some elements of the real world but also makes many simplifying assumptions. So it is necessary to distinguish between ontology terms and natural language words – they are very different sorts of things.

3 The Relation between Ontologies and Natural Language: In Practice

In some domains (particularly parts of Medicine) ontologies make a strict distinction between the ontology terms and natural language words that can be used to express them. In addition, ontology definition languages provide facilities (e.g. the RDFS `label` construct) to express natural language words separately from the ontology terms. However, in practice in many cases ontology designers choose versions of natural language phrases as their formal terms. In logical terms, it makes no difference whether a concept is labelled in a way that can be "understood" by humans, e.g. *Leg*, rather than as an arbitrary identifier, e.g. $C40274$. It is therefore natural for ontology designers to choose mnemonic names for their concepts and properties. However, there are dangers in this. Current ontology languages are extremely simple logics, with very low expressive power. So a set of axioms about the term *Leg* cannot possibly capture more than a tiny part of what (some sense of) "leg" means. The formal definition that an ontology provides for a term dramatically under-specifies what the term means. But if the ontology designer labels their concept *Leg*, unless they can turn off their in-built natural language understanding (and disambiguation) capabilities they can easily have the illusion that they have captured the exact sense that they require. Similarly a user of the ontology can easily get a false sense of security in using this concept,

simply because of its name. This is an instance of the problem of "wishful mnemonics" discussed by McDermott [7]. McDermott described how inappropriate use of natural language terms for programming constructs, program modules and symbols in knowledge representation languages can mislead, in terms of the actual problem solved and the power and sophistication of the approach:

> *"A good test for the disciplined programmer is to try using gensyms in key places and see if he still admires his system"* [7]

And yet AI practitioners seemed to be happy to create these illusions, or unaware of what they were doing (and maybe they still are ...).

In terms of the semantic web endeavour, such problems could represent a real threat to progress. Not least, using natural language names could easily lead to an ontology designer failing to express axioms which are "obviously true" but in fact very necessary in order to make necessary distinctions for a computational agent (or a native speaker of a different language). For the purposes of this paper, however, the main conclusion to be drawn from this discussion is that the semantic web – its ontologies and knowledge bases – could actually be a lot more like natural language documents than the theory says they should be.

4 Linguistic Structures in Real Ontologies

To investigate the extent to which existing ontologies make use of natural language terms, we carried out an experiment to see what structures are present in the names used in actual ontologies available on the internet. In this experiment, we concentrated on OWL ontologies [8], filtering out ontologies (such as WordNet) that are specifically designed to represent linguistic information.

We wrote a Java program using the Google API to help us look for online ontologies coded in OWL. We did this using the keywords "owl filetype: owl", which indicates that our desired ontologies must contain the string "owl" as well as having owl as their file extension (filetype:owl indicates the file type, and every legal OWL ontology must contain the string "owl"). Using these key words, theoretically all online ontologies found by Google should be returned. Actually we obtained around five thousand links; however only some of these links were able to provide us with real ontology files, because firstly, Google limits its API not to be accessed more than one thousand times per day, and also because some of the links were not available. In total we collected 882 ontology files coded in OWL (111 Mb) as our corpus.

In analysing these ontologies, we were interested in two kinds of names, names of classes and names of properties, and wanted to know what these names consisted of. In order to detect the English words contained in these names, we used the WordNet [10] API to help us recognise English words occurring in these names. Because there are no agreed rules for how to name concepts, people use various ways of giving names, e.g., $PostgraduatePhD$, $International\_Student$, red_wine, $Redwine2004$, $hasProducer$, $part\_of$ and even meaningless strings like ABC, $ED009$ etc. In our approach we can detect multiple English words joined together if there is any separator between them, such as a capital letter, a underline or a number. For instance, $PostgraduatePhD$ will be recognised as two English words. Each name is associated with a *pattern* recording its analysis as a sequence of parts of speech. Formally, a pattern is a string in the language:

$$L = (\texttt{Noun}|\texttt{Adj}|\texttt{Verb}|\texttt{Prep}|\texttt{Adv}|\texttt{Than}|\texttt{Or}|\texttt{Un}|\texttt{C})^*$$

where `Noun` etc. name standard parts of speech (`Than` and `Or` being used for particular closed class words that appeared in the corpus), `Un` names an unknown word and `C` names a capital letter not starting a recognised word (sequences of capital letters were not further analysed). Thus, for instance, $PostgraduatePhD$ is analysed as `NounNoun` and $ICD10$ is represented as `CCUn`. Patterns can be surprisingly long: for instance, the pattern

```
NounNounPrepNounNounPrepNounUnNoun
```

is the analysis of the class name made up of the words "Muscle Layer of Secondary Duct of Left Coagulating Gland" joined by underscores. In the situation of handling a word which can be recognised as a noun and also as a verb (e.g., "work"), our system treats the word as a noun when it analyses names of classes, and treats it as a verb when analysing names of properties, because we believe that nouns have a higher possibility than verbs to occur in names of classes, while verbs have a higher possibility to occur in names of properties. In addition, our system can do simple morphological analysis including detecting plural nouns and verbs in present, past or passive tense by using two sets of linguistic rules and applying them to every input name. When the above two cases occur together (e.g., "works" may be a plural noun and also a present verb), the system gives the rules for handling plural nouns higher priority when it analyses names of classes and gives the rules for handling verbs higher priority when analysing names of properties. For instance, "works" is seen as a noun when the system analyses names of classes, but as a verb when analysing names of properties. There may be some special cases of names that our system cannot recognise, because firstly our rules may not cover all possibilities and also the WordNet API has limits on the words it can recognise. However, the analysis was enough to enable us to determine useful information about the general forms of names that people have used in

Pattern	Frequency	Percentage
`Noun`	5084	14%
`NounNoun`	4092	11%
`UnUn`	1837	5%
`Un`	1755	5%
`AdjNoun`	1528	4%
`NounNounNoun`	1378	4%
`UnNoun`	1366	4%
`AdjNounNoun`	681	2%
`NounUn`	577	2%
`UnNounNoun`	482	1%
...	...	...
TOTAL	37260	100%
(.)*`Noun`	26708	72%
`Noun`*	11011	30%
(`C`|`Un`)*	5266	14%

Fig. 1. Frequencies of patterns in class names

defining OWL ontologies.

Figure 1 shows the frequencies of some of the patterns that applied to the 37260 different class names found. These frequencies are the frequencies that classes were *defined* in ontologies (first introduced; not the number of times they were *used*). If the same class name was used in more than one ontology, it is counted several times. There were 3003 different patterns found, and the first ten are listed in order of frequency in the figure. Below are frequencies for selected meta-patterns (regular expressions over patterns). From these it can be seen that 72% of the class names ended with recognised nouns. Also 30% consisted entirely of strings of nouns (up to 7). Finally, only 14% of the class names contained no recognised word (i.e. are composed of entirely of unknown words and capital letters). So there is clearly a considerable amount of linguistic material in these names.

Figure 2 shows similar frequencies for property names. This time, for technical reasons, multiple uses of the same name in different ontologies were counted as just one occurrence. Although the numbers are smaller (and the popular patterns now include adverbs and prepositions), there are many similarities

Pattern	Frequency	Percentage
`Verb`	132	10%
`VerbVerb`	129	10%
`Noun`	80	6%
`VerbPrep`	73	5%
`VerbNoun`	72	5%
`VerbVerbVerb`	50	4%
`NounPrep`	36	3%
`VerbVerbPrep`	35	3%
`Un`	32	2%
`VerbVerbNoun`	31	2%
`NounVerb`	31	2%
`VerbAdv`	23	2%
`VerbAdjVerb`	23	2%
`VerbUn`	21	2%
`VerbNounVerb`	21	2%
`VerbNounPrep`	16	1%
...	...	...
TOTAL	1354	100%
`Verb`(.)*	885	65%
`Noun`(.)*	216	16%
(.)*`Verb`	571	42%
(.)*`Prep`	262	19%
(`C`|`Un`)*	43	3%

Fig. 2. Frequencies of patterns in property names

(especially when one considers that ambiguous verb/nouns will have been classed as nouns for the class names and verbs for the property names).

These figures give a striking picture of the extent of linguistic material in existing ontologies, and also of its relative complexity.

5 The Semantic Web as a Linguistic Resource

That the semantic web is partly a linguistic resource is implicitly acknowledged by applications such as ontology reconciliation [4] and ontology search [14]. Such applications assume that, in general:

(1) The names of concepts and properties matter and
(2) The names of concepts and properties are meaningful to a human user.

Such applications would not be able to work without making these assumptions, which basically amount to requiring that concept and property names make use of natural language words.

If the documents of the semantic web are at least partially linguistic in nature, then we can apply variations of natural language processing operations to them. Indeed, NLP techniques may be *needed* in order to fully understand what is actually stated in these documents. For instance, word sense disambiguation techniques may be required to handle concepts with unsufficiently specific defining axioms; machine translation might be required to translate ontologies into different languages. Some signs of this are beginning to be seen, for instance in work to measure the similarity between ontologies and natural language texts using an adaptation of "bag of words" models [3].

That complex NLP may be needed for significant uses of documents is disappointing news for the semantic web, but offers many interesting tasks for NLP researchers to develop existing techniques in the context of a version of semi-structured natural language.

6 Opportunities for Natural Language Generation

As we discussed in the preamble, a significant cost in developing an NLG system for a new domain is the production of a lexicon for the domain, relating concepts in the domain to natural language words that can be used to denote them. This means that NLG systems are in practice domain-dependent. Indeed, it is a significant challenge to the field to produce portable systems or even system components [9].

If semantic web documents are largely already filled with words in the desired natural language, then there is the prospect of building NLG systems very cheaply, because the lexicon comes "for free". Indeed, one can envisage domain-*independent* NLG systems for the semantic web, which have no specific domain resources but merely access to generic linguistic resources which

enable them to decode the linguistic material already present in the input. NLG in such a situation avoids many of the problems of traditional NLG (specifically lexical choice) and is more like reconstituting natural language sentences from linguistic fragments – an extreme form of the kind of flexible NLG from existing phrases used in multi-document summarisation [1].

As yet, however, although many researchers have sought to produce domain-independent *frameworks* for building NLG systems for the semantic web, to our knowledge there has been no proposal to construct a single domain-independent *system* for producing language from semantic web material. Indeed, our experiment shows that some technical problems need to be addressed for this vision to become a reality:

- Concept and property names can be made from multiple words. Also abbreviations can be used. Some simple natural language analysis is necessary to handle these cases and also unknown words, which may be names.
- Morphological analysis is needed to recognise, for instance, plural nouns, present and past participles. There is also part of speech ambiguity.
- Translating from property names to appropriate realisations may be non trivial. For instance, if a concept X has the value Y for the property *contains*, does this mean that "X contains Y" or that "X is among the things that Y contains"?

It may well be possible to find appropriate engineering solutions to the above problems. But there are also dangers:

- Words may be used in unnatural technical senses, which means that referring to a concept by its name may actually mislead.
- It may not be clear which words in the NLG output are the terms of the ontology and which are informal NL words - which words are being used to explain which others?
- There may be serious consequences of cases of inaccurate analysis. E.g. a word might have a particular quite specific interpretation in this ontology, but may be used in the language as if it has another sense.

7 Our own work

Our current research is addressing the problem of presenting parts of OWL DL ontologies in natural language. This will extend existing approaches to generating from simpler DLs (e.g. [12]) by taking into account the fact that in a language like OWL DL a concept is described more by a set of constraints than by a frame-like definition. Hence information about a concept cannot be presented in a single sentence but requires an extended *text* with multiple

sentences, the overall structure having to be planned so as to be coherent as a discourse. The work is also different from other work which generates text about *individuals* described using ontologies [13,2], in that it presents the ontology *class axioms* themselves.

Following our experiments, our initial approach is to see how much can be achieved with no restrictions on the ontology (as long as it is expressed in legal OWL DL) and only generic linguistic resources (such as WordNet). This is also motivated both because

- there is a practical need to present parts of arbitrary current ontologies (which often come with no consistent commenting or linguistic annotations) and also because
- if we can determine the main deficiencies of such an approach then we can then make informed recommendations about what kinds of extra annotations or naming conventions would be valuable in the ontologies of the future.

So we aim to maximise the use made of the existing linguistic material in an ontology, even though there could also be dangers in doing so.

The following example shows the kind of text we are currently able to generate (assuming some manual postprocessing for capitalisation and punctuation):

> *What is a MEA?*
> A MEA is a kind of Actuality which contains exactly 1 thing, which is a Cathode, an Anode and an Electrolyte. Everything a FuelCell contains is a MEA. Only something which is a FuelCell, a MEA, an Electrode or a Catalyst contains something.

Although there are no agreed principles for naming concepts and properties in ontologies, it is encouraging that a large percentage of these names include English words which can be recognised by WordNet. This gives us a chance to interpret the syntax of these names and help us produce more fluent natural language. For instance, for the constraint:

```
restriction(Onproperty(hasProducer)
        allValuesFrom(French))
```

we can say "has a producer who is French", instead of something like "has a property, hasProducer, which must have as its value, something that is in

the class French". The above example gains greatly from this - in this case, WordNet is able to provide all the relevant part of speech information, except for MEA (and we have provided the information that "MEA" is a noun by hand).

Our current approach to realising an axiom in English involves a search through multiple rules matching against structural patterns in OWL axioms and attempting to exploit part of speech information about the names where this can be inferred. This search may yield several possible realisations. We currently choose between these according to how closely they come to having an "ideal" sentence length. This parameter can be set in advance according to text requirements.

In the future, we would like to find generic rules for how ontology builders name concepts and properties, and how these can be exploited in realisation, by doing further analysis of our existing corpus. The aim is to get elegant natural language without requiring domain-dependent resources.

8 Conclusions

Semantic web documents contain a surprising amount of complex linguistic material. The reliance of knowledge engineers on this leads to dangers of inadequate formalisation. It also leads to a number of interesting and challenging tasks for natural language processing. In particular, there is a prospect of building domain-independent NLG tools for presenting semantic web material.

Acknowledgments

This work is supported by EPSRC research grant GR/S62932.

References

[1] R. Barzilay, K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557. Association for Computational Linguistics, 1999.

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

[3] G. Burek, M. Vargas-Vera, and E. Moreale. Indexing student essays paragraphs using lsa over an integrated ontological space. In *Proceedings of International Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*, Geneva, 2004.

[4] A. Hameed, A. Preece, and D. Sleeman. Ontology reconciliation. In S. Staab and R. Studer, editors, *Handbook on Ontologies in Information Systems*, pages 231–250. Springer Verlag, 2003.

[5] Graeme Hirst. Ontology and the lexicon. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 209–230. Springer Verlag, 2004.

[6] Richard Kittredge, Tanya Korelsky, and Owen Rambow. On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314, 1991.

[7] D. McDermott. Artificial intelligence meets natural stupidity. In J. Haugeland, editor, *Mind Design*, pages 143–160. Bradford Books, 1981.

[8] D. L. McGuinness and F. van Harmelen. Owl web ontology language overview. `http://www.w3.org/TR/owl-features/`, 2004.

[9] C. Mellish, M. Reape, D. Scott, L. Cahill, R. Evans, and D. Paiva. A reference architecture for generation systems. *Natural Language Engineering*, 10(3/4):227–260, 2004.

[10] G. Miller. Wordnet: A lexical database for english. *CACM*, 38(11):39–41, 1995.

[11] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.

[12] J. Wagner, J. Rogers, R. Baud, and J-R. Scherrer. Natural language generation of surgical procedures. *Medical Informatics*, 53:175–192, 1999.

[13] G. Wilcock. Talking owls: Towards an ontology verbalizer. In *Human Language Technology for the Semantic Web and Web Services, ISWC-2003*, pages 109–112, Sanibel Island, Florida, 2003.

[14] Y. Zhang, W. Vasconcelos, and D. Sleeman. Ontosearch: An ontology search engine. In *Proceedings of the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer Verlag, 2004.