Generating Multilingual Descriptions from Linguistically Annotated OWL Ontologies: the NaturalOWL System

Dimitrios Galanis and **Ion Androutsopoulos**
Department of Informatics
Athens University of Economics and Business
Patission 76, GR-104 34 Athens, Greece

Abstract

We introduce NaturalOWL, an open-source multilingual natural language generator that produces descriptions of instances and classes, starting from a linguistically annotated ontology. The generator is heavily based on ideas from ILEX and M-PIRO, but it is in many ways simpler and it provides full support for OWL DL ontologies with RDF linguistic annotations. NaturalOWL is written in Java, and it is supported by M-PIRO's authoring tool, as well as an alternative plug-in for the Protégé ontology editor.

1 Introduction

In recent years, considerable effort has been devoted to the Semantic Web (Antoniou and van Harmelen, 2004), which can be thought of as an attempt to establish mechanisms that will allow computer applications to reason more easily about the semantics of the Web's resources (documents, services, etc.). Domain ontologies play a central role in this endeavour: in effect, they establish domain-dependent semantic vocabularies (classes of entities; particular entities, called instances; properties of instances; axioms governing their use) that can be used to publish on the Web knowledge in shared machine-readable representations, and to annotate other resources (e.g., documents, videos) with machine-readable metadata describing aspects of their semantics.

In the case of natural language documents, some semantic annotations can be produced automatically via ontology-aware information extraction (Bontcheva and Cunningham, 2003); but information extraction can currently provide reliably only relatively simple types of semantic information, mostly by identifying and classifying named entities, and, less reliably, relations between them. When texts are generated automatically from formal knowledge bases, however, the generator can easily annotate the texts with much richer information, including full representations of their semantics, expressed in machine-readable markup. In fact, an entire strand of work in natural language generation (NLG) has focused on generating textual descriptions of an ontology's classes or instances. A well-known example of such work is ILEX (O'Donnell et al., 2001), which was demonstrated mostly with ontologies of museum exhibits. More recently, the M-PIRO project (Androutsopoulos et al., 2007) developed a multilingual extension of ILEX, which was tested in several domains, including museum exhibits and computing equipment. In this type of work, the ontology's role is no longer simply to provide a semantic vocabulary; the ontology acts as a repository of knowledge, and parts of the knowledge (e.g., information pertaining to particular instances or classes) can be rendered automatically in multiple natural languages or in a machine-readable form that carries the same semantic content. For example, M-PIRO's generator can deliver a description like the following to a human on-line shopper,

> A110: This is a laptop, manufactured by Toshiba. It has a Centrino Duo processor, 512 MB RAM, and an 80 GB hard disk. Its speed is 1.7 GHz and it costs 850 Euro.

and the following semantically equivalent formal representation to a software agent.[1] Alternatively, each individual sentence of the text could be marked up with a machine readable semantic representation.

```
<Laptop rdf:ID="A110">
 <manufacturedBy rdf:resource="#toshiba" />
 <hasProcessor rdf:resource="#centrinoDuo" />
 <hasMemory rdf:datatype="...#string">512 MB</memory>
 <hasDisk rdf:datatype="...#string">80 GB</disk>
 <speed rdf:datatype="...#string">1.7 GHz</speed>
 <cost rdf:datatype="...#string">850 Euro</cost>
</Laptop>
```

The standard formalism for publishing ontologies on the Semantic Web (SW) is currently OWL.[2] There are application domains (e.g., on-line shops) where one can envisage future SW sites that will maintain and publish their content entirely in the form of OWL ontologies. Then, NLG technology, embedded in server or browser plug-ins, could be used to render parts of the OWL ontologies in multiple natural

[1] For simplicity, we show scalar values as strings that include the units of measurements. More principled, language-independent representations of these values are also possible.

[2] Consult `http://www.w3.org/TR/owl-features/`.

languages or equivalent machine-readable representations on demand (Androutsopoulos et al., 2007). NLG can, thus, be seen as a key technology of the SW, which makes knowledge accessible to both humans and computers, a major target of the SW.

In this paper, we introduce NaturalOWL, a prototype open-source natural language generator intended to demonstrate what NLG can offer to the SW.[3]will be announced in the final version of this paper. NaturalOWL is heavily based on ideas from ILEX and M-PIRO, but unlike its predecessors it provides full support for OWL DL, the most principled version of OWL that corresponds to description logic (Baader et al., 2002); many NLG researchers will be familiar with this form of logic. Our previous attempts to support OWL in M-PIRO's generator ran into problems, because of incompatibilities between OWL and M-PIRO's ontological model (Androutsopoulos et al., 2005). Compared to ILEX and M-PIRO, NaturalOWL is also simpler; for example, it is entirely template-based, as opposed to the Systemic Grammars its predecessors employed.[4] Although future work may enhance some of NaturalOWL's components, the simplicity of the current system makes it easier to explain to SW researchers, who may not be familiar with NLG. It also simplifies the task of extending the system to support additional natural languages. NaturalOWL currently supports English and Greek.

It has been argued (Mellish and Sun, 2006) that in most OWL ontologies, classes and properties are given names that are either English words (e.g., `Laptop`, `cost`) or concatenations of English words (e.g., `manufacturedBy`, `hasMemory`). Based on this observation, Sun and Mellish (2006) generate texts from RDF descriptions, RDF being the description formalism on which OWL is based, without any domain-dependent linguistic resources. They use WordNet to tokenize the names of classes and properties, as well as to assign part-of-speech (POS) tags to tokens, and this allows them to guess that a class name like `Laptop` above is in fact a noun that can be used to refer to that class, or that `<manufacturedBy rdf:resource="#toshiba" />` should be expressed in English as "[This laptop] was manufactured by Toshiba". Hewlett *et al.* (2005) adopt very similar techniques. This approach, however, is problematic when texts have to be generated in multiple languages. Even in the monolingual case, there are significant problems: for example, a POS-tagger is often needed to distinguish between noun and verb uses of the same token, and morphological or even syntactic analysis is needed (especially in highly inflected languages) to extract tokens from class and property names and convert them into grammatical phrases; in effect, this re-introduces the need to interpret texts. Furthermore, our experience is that generating high-quality texts often requires linguistic information that is not present, not even indirectly, in OWL ontologies, nor can it be embedded conveniently in them (Androutsopoulos et al., 2005).

We, therefore, propose to annotate OWL ontologies with stand-off RDF markup that associates elements of the ontologies (e.g., classes, properties) with domain-dependent linguistic resources (e.g., lexicon entries, templates). We believe that this kind of linguistic annotation should be a standard part of ontology engineering for the SW; apart from allowing parts of the ontology to be presented to end-users in natural language, it facilitates presenting ontologies to domain experts for validation; and the annotations can also be useful when querying or extending ontologies via natural language (Katz et al., 2002; Bernstein and Kaufmann, 2006). NaturalOWL's RDF linguistic annotations will hopefully contribute towards a discussion in the NLG community on how to annotate OWL ontologies with linguistic information, and this may eventually produce standards that will allow alternative NLG components to render OWL ontologies in natural language, in the same way that alternative browsers can be used to view HTML pages.

Below we present briefly NaturalOWL's processing stages and its annotations of OWL ontologies. Following Wilcock (2003), the processing stages communicate in XML, but they are implemented in Java, instead of XSLT, and there is a clearer separation between processing code and linguistic resources.

2 Document planning

When instructed to produce a natural language description of an instance, NaturalOWL first selects from the ontology all the logical facts that are directly relevant to that instance; for example, when describing the laptop of the first page, it would select the fact that the instance is a `Laptop`, the fact that its manufacturer is Toshiba, etc.[5] NaturalOWL may be instructed to include facts that are further away in a graph representation of the ontology, up to a maximum (configurable) distance; setting the distance to two when describing a statue, for example, would also include in the selected facts information about the statue's sculptor (e.g., the country and year they were born in). This is very similar

[3]NaturalOWL and its Protégé plug-in can be downloaded from `http://www.aueb.gr/users/ion/publications.html`.

[4]Consult `http://www.ltg.ed.ac.uk/methodius/` for another offspring of M-PIRO's generator that uses CCG grammars.

[5]To save space, we restrict the discussion to descriptions of instances. NaturalOWL can also describe classes, but it conveys only information that is explicit in the ontology, unlike the work of Mellish and Sun (2005), where class descriptions also convey inferred facts. There are also separate stand-off annotations that specify how interesting each type of fact is per user type, and other user modelling information, much as in ILEX and M-PIRO. The ordering annotations could also be made sensitive to user type and target language.

to ILEX's content selection, but without employing rhetorical relations.

The selected facts of distance one are then ordered by consulting ordering annotations (see `owlnl:order` below), which specify a partial order of properties (e.g., that the manufacturer should be mentioned first, followed by the processor, memory and disk in any order, and then the price); is-a facts are always mentioned first. Second distance facts are always placed right after the corresponding directly relevant facts, producing texts like "This is a statue. It was sculpted by Nikolaou, who was born in Athens in 1968. This statue is made of marble and it...". In the application domains we have considered, this ordering scheme was adequate, although in other domains more elaborate text planning approaches may be needed; consult, for example, Bontcheva and Wilks (2004) for an application of text schemata to NLG from ontologies.

3 Microplanning, surface realisation

For each property, one or more micro-plans need to be specified per language. NaturalOWL's micro-plans are templates, each consisting of a sequence of slots. Each slot can be filled by an expression referring to the owner of the property (the laptop, in the case of `manufacturedBy`), the value (filler) of the property (Toshiba), or a string. The following RDF annotations refer to the `manufacturedBy` property.[6] After setting the property's order, they define an English micro-plan, according to which `<manufacturedBy rdf:resource="#toshiba" />` should be rendered in English as a phrase starting with (first slot) a nominative expression referring to the owner (the laptop). The `owlnl:retype` element of the first slot allows the system to select automatically among using the owner's name in natural language (if it has one), a noun phrase (e.g., "this laptop"), or a pronoun to refer to the owner, depending on context. The second slot will be filled by the string "was manufactured", which is marked up as being a past passive verb form; this additional markup is needed when aggregating phrases to form longer sentences. The third and fourth slots will be filled by the string "by" and an accusative automatically selected referring expression corresponding to the filler, respectively. The micro-plan may generate, for example, a phrase like "It was manufactured by Toshiba".[7]

[6] The linguistic annotations are kept in separate files from the OWL ontology, but they refer to its elements via their unique identifiers; we abbreviate the identifiers to save space.

[7] Although OWL properties (and other elements of OWL ontologies) can be associated with strings in multiple languages via `rdfs:label` tags, this mechanism is inadequate for template micro-plans; for example, it provides no principled way to indicate positions where referring expressions should be placed, or to annotate sub-strings with syntactic categories.

```
<owlnl:property rdf:about="...#manufacturedBy">
  <owlnl:order>1</owlnl:order>
  <owlnl:EnglishMicroplans ...>
    <owlnl:microplan ...>
      <owlnl:aggrAllowed>true</owlnl:aggrAllowed>
      <owlnl:slots ...>
        <owlnl:owner>
          <owlnl:case>nominative</owlnl:case>
          <owlnl:retype>re_auto</owlnl:retype>
        </owlnl:owner>
        <owlnl:verb>
          <owlnl:voice>passive</owlnl:voice>
          <owlnl:tense>past</owlnl:tense>
          <owlnl:val>was manufactured</owlnl:val>
        </owlnl:verb>
        <owlnl:text>
          <owlnl:val>by</owlnl:Val>
        </owlnl:text>
        <owlnl:filler>
          <owlnl:case>accusative</owlnl:case>
          <owlnl:retype>re_auto</owlnl:retype>
        </owlnl:filler>
      </owlnl:slots>
    </owlnl:microplan>
  </owlnl:EnglishMicroplans>
  <owlnl:GreekMicroplans ...>
  ...
</owlnl:Property>
```

NaturalOWL currently employs a very simple algorithm for generating referring expressions: once the instance being described has been introduced by mentioning its class (e.g., "This is a statue."), it uses pronouns to refer to that instance (e.g., "It was sculpted by Nikolaou."), until the focus moves to another instance ("Nikolaou was born in Athens. He was born in 1968."). Then, when the focus returns to the original instance, a demonstrative is used ("This statue is made of..."). As in M-PIRO, some properties may contain canned strings, and there are special annotations to flag canned strings that change the focus. Again, more elaborate referring expression generation algorithms can be added.

To be able to generate expressions like "this is a statue" or "this laptop", NaturalOWL requires OWL classes to be associated with noun phrases (more precisely n-bars). This is illustrated in the RDF statements below, where the class `Laptop` is associated with a noun phrase entry `laptop-NP` of NaturalOWL's domain-dependent multilingual lexicon, also expressed in RDF. The lexicon entry lists the various forms of the noun phrase, provides information on gender etc. The nominative singular form in Greek would be "φορητός υπολογιστής" (portable computer). We have considered associating classes with WordNet (or EuroWordNet) synsets, but the domain ontologies we have experimented with contain highly technical concepts, which are not covered by WordNet. Nevertheless, we plan to consider emerging standards for linguistic annotations (Ide and Romary, 2004), especially regarding the lexicon.

The phrases of the micro-plans are then aggregated in longer sentences, using roughly M-PIRO's

aggregation rules (Melengoglou, 2002). This produces the final text, and, hence, there is no separate surface realization phase, apart from adding presentation markup, markup for speech synthesizers etc.

```
<owlnl:owlClass rdf:about="...#Laptop">
  <owlnl:hasNP rdf:resource="#laptop-NP"/>
</owlnl:owlClass>

<owlnl:NP rdf:ID="laptop-NP">
  <owlnl:LanguagesNP ...>
    <owlnl:EnglishNP>
      <owlnl:gender>nonpersonal</owlnl:gender>
      <owlnl:singular ...>laptop</owlnl:singular>
      <owlnl:plural ...>laptops</owlnl:plural>
    </owlnl:EnglishNP>
    <owlnl:GreekNP>
      <owlnl:gender>masculine</owlnl:gender>
      <owlnl:singularForms>
        <owlnl:nominative ...>...</owlnl:nominative>
        <owlnl:genitive ...>...</owlnl:genitive>
        <owlnl:accusative ...>...</owlnl:accusative>
      </owlnl:singularForms>
      ...
</owlnl:NP>
```

4 Source authoring

NaturalOWL is supported by M-PIRO's authoring tool (Androutsopoulos et al., 2007), which has been extended by NCSR "Demokritos" to be compatible with OWL DL. The tool helps "authors" port NaturalOWL to new application domains, including the tasks of ontology construction, defining micro-plans, creating the domain-dependent lexicon, etc. NaturalOWL is also accompanied by a plug-in for Protégé, an ontology editor most SW researchers are familiar with.[8] The plug-in provides the same functionality as M-PIRO's authoring tool. Consult also Bontcheva (2004) for related work on authoring tools.

5 Conclusions and further work

We introduced NaturalOWL, an open-source natural language generator for OWL DL ontologies that currently supports English and Greek. The system is intended to demonstrate the benefits of adopting NLG techniques in the Semantic Web, and to contribute towards a discussion in the NLG community on relevant annotation standards. NaturalOWL was partly developed and is being extended in project Xenios, where it is used by mobile robots acting as museum guides, an application that requires, among others, extensions to generate spatial expressions.[9]

References

I. Androutsopoulos, S. Kallonis, and V. Karkaletsis. 2005. Exploiting OWL ontologies in the multilingual generation of object descriptions. In *10th European Workshop on* NLG, pages 150–155, Aberdeen, UK.

I. Androutsopoulos, J. Oberlander, and V. Karkaletsis. 2007. Source authoring for multilingual generation of personalised object descriptions. *Natural Language Engineering*. In press, available on-line.

G. Antoniou and F. van Harmelen. 2004. *A Semantic Web Primer*. MIT Press.

F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider, editors. 2002. *The Description Logic handbook: theory, implementation and application*. Cambridge University Press.

A. Bernstein and E. Kaufmann. 2006. GINO – a guided input natural language ontology editor. In *5th Int. Semantic Web Conference*, pages 144–157, Athens, GA.

K. Bontcheva and H. Cunningham. 2003. The Semantic Web: a new opportunity & challenge for Human Language Technology. In *Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd Int. Semantic Web Conf.*, Sanibel Island, FL.

K. Bontcheva and Y. Wilks. 2004. Automatic report generation from ontologies: the MIAKT approach. In *9th Int. Conf. on Applications of Natural Language to Information Systems*, pages 324–335, Manchester, UK.

K. Bontcheva. 2004. Open-source tools for creation, maintenance, and storage of lexical resources for language generation from ontologies. In *4th Conf. on Language Resources & Evaluation*, Lisbon, Portugal.

D. Hewlett, A. Kalyanpur, V. Kolovski, and C. Halaschek-Wiener. 2005. Effective NL paraphrasing of ontologies on the Semantic Web. In *Workshop on End-User Semantic Web Interaction, 4th Int. Semantic Web conference*, Galway, Ireland.

N. Ide and L. Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3/4):211–225.

B. Katz, J. Lin, and D. Quan. 2002. Natural language annotations for the Semantic Web. In *International Conference on Ontologies, Databases, and Application of Semantics*, University of California, Irvine.

A. Melengoglou. 2002. Multilingual aggregation in the M-PIRO system. Master's thesis, School of Informatics, University of Edinburgh, UK.

C. Mellish and X. Sun. 2005. Natural language directed inference in the presentation of ontologies. In *10th European Workshop on* NLG, Aberdeen, UK.

C. Mellish and X. Sun. 2006. The Semantic Web as a linguistic resource: opportunities for Natural Language Generation. *Knowledge Based Systems*, 19:298–303.

M. O'Donnell, C. Mellish, J. Oberlander, and A. Knott. 2001. ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3):225–250.

X. Sun and C. Mellish. 2006. Domain independent sentence generation from RDF representations for the Semantic Web. In *Combined Workshop on Language-Enabled Educational Technology and Development and Evaluation of Robust Spoken Dialogue Systems, European Conference on AI*, Riva del Garda, Italy.

G. Wilcock. 2003. Talking OWLs: towards an ontology verbalizer. In *Workshop on Human Language Technology for the Semantic Web, 2nd Int. Semantic Web Conf.*, pages 109–112, Sanibel Island, FL.

[8] See `http://protege.stanford.edu/`.

[9] Xenios is funded by the European Union and the Greek General Secretariat for Research and Technology; consult `http://www.ics.forth.gr/xenios/` for further information.