Using a Dictionary Production System to impose a WordNet on a Dictionary. A Software Presentation

Holger Hvelplund[1] and Allan Ørsnes[2]
Ingénierie Diffusion Multimédia (IDM) – www.idm.fr

In this demonstration, we will make a practical presentation of how a semantic web – a WordNet – can be imposed on a dictionary using a Dictionary Production System. Using a WordNet structure makes a lot of sense as there are already several WordNets for different languages which can be used freely.
We will demonstrate how we first impose the structure in a rather crude way allowing for lots of ambiguities and then use the various tools in the Dictionary Production System to target the dubious areas and refine them through manual intervention, either as part of a usual editing of a new edition or in a separate run.
Finally we will demonstrate how the data can be extracted and made available to a Dictionary Publishing System.

Introduction

When publishing a dictionary electronically, it is highly desirable that users (and search engines) can explore the dictionary content in different ways. Users (and search engines) must of course be able to look up the meaning of a given word, but it also makes sense to allow users (and search engines) to explore the semantically related words to a given word: either for finding an alternative to a word when encoding or simply as a way of browsing the dictionary.

The Dictionary Production & Publishing Systems

The Dictionary Production and Publishing Systems demonstrated in the presentation are based on *international standards* XML, XSL, DTD, CSS, XHTML and Unicode.

The Dictionary Production System consists of: a) a Dictionary Entry Editor; and b) a centralized Content Management System. Using the Dictionary Entry Editor multiple users can work simultaneously on one or many dictionary projects and the content they produce is stored in a centralized database in the Content Management System. A locking system ensures that only one user at the time can edit a given dictionary entry.

The Dictionary Production System is *open* and *scalable*. The system administrator uses configuration files to *customize* the Dictionary Production System for the requirements of each individual dictionary project and even for each individual user, task and /or delivery channel in a given project. If and when new requirements are identified during the course of the project, the system administrator can in a stepwise process revise and update the configuration files without interrupting the content production.

An intelligent versioning system keeps track of dependencies between configuration files and can help the system administrator in re-using configuration files so that configuration files for new dictionary projects can be quickly produced in a way where knowledge-sharing, consistency and compliance with other projects is optimized.

[1] hvelplund@idm.fr

[2] orsnes@idm.fr

The Dictionary Publishing System is also open and scalable: a) content of one or more dictionary projects can flow seamlessly – via the fulfilment component of the Dictionary Production System – into the Dictionary Publishing System ensuring that changes to the content can be published quickly and in a controlled way; b) the templating system in the Dictionary Publishing System makes it easy for the web team responsible for the web site to adapt and optimize the web site for demands and opportunities in the market – for example delivery of dictionary content via many delivery channels such as applications for desktops, netbooks, smart phones, e-books; and c) the features of the Dictionary Publishing System gives end-users quick and easy access to information when they read text on the screen, when they produce text and when they want to test and/or expand their vocabulary.

Step 1

Usually a WordNet is stored in a relational database so the first step is to convert the WordNet data to XML (the format used by Dictionary Production System) and implementing the relations using a *cross-reference mechanism* in the Dictionary Production System. The Synsets are the entities we describe, setting up the relations between them. For clarity we concentrate here on the inheritance hierarchy only, the hyperonym relation which completely identifies this. As will be evident any relation encoded in the WordNet will be possible to impose.

```
<SYNSET ID=”xxx” NAME=”adolescent” POS=”noun”>
  <DEF>a juvenile between the onset of puberty and maturity</DEF>
  <MEMBERS>
    <MEMBER ID=”teenager_noun_1” FREQ=””>
      <L-EXPR>teenager</L-EXPR>
      <EX>Example sentence with "teenager"</EX>
    </MEMBER>
    <MEMBER ID=”adolescent_noun_1” FREQ=””>
      <L-EXPR>adolescent</L-EXPR>
      <EX>Example sentence with "adolescent"</EX>
    </MEMBER>
  </MEMBERS>
  <REL NAME=”hyperonym” IDREF=”juvenile” />
</SYNSET>
```

After import, the Dictionary Production System can be used to maintain the WordNet data; create cross lingual links between Synsets in WordNets – for example between Synsets in WordNets for different languages; and to impose the structure on all – English in this case – dictionaries edited in the Dictionary Production System.

Step 2

The in-built processing feature of the Dictionary Production System which allows running of scripts over the data will:

Extract the data from the WordNet and the dictionary to be linked

Using the lexical expressions and part of speech information in WordNet Synsets and the headword and part of speech information in the dictionary entries it creates link in the dictionary entries to the relevant Synset (or to a given member inside a Synset).

At the time of publishing, it can then be decided if and how this should be displayed to the user.

Upload the new version of the dictionary data to the Dictionary Production System

Run the cross referencing feature in the Dictionary Production System – it allows cross-referencing between projects such as between entries in the WordNet and the dictionary database.

We now have a first version of the dictionary with the structure of the Synsets imposed (indirectly) via the links to the Synsets which are themselves parts of a structure.

Step 3

We will here use the *XML search* component of Dictionary Production System to retrieve entries with:

Several links to Synsets: the links must be checked and distributed over the senses within the entry.

No links to the Synsets: these will have to be considered and will either result in adding new Synsets or linking to an already existing one.

One link and one sense: we will assume these are OK and will automatically transfer the entry-level link to the sense level using the in-built processing feature of the Dictionary Production System (we will need to have some statistics here saying how big each of these groups are in different standard dictionaries).

Step 4

Step 4 demonstrates prototypical editing tasks using the *dictionary entry editor* and the *cross reference tool* of the Dictionary Production System.

Step 5

Extracting the data is done using the fulfillment feature of the Dictionary Production System: A process is set up which will extract the Synsets along with the dictionary data. Reverse links from the Synsets to the dictionary data (entry senses) are inserted as well as the hyperonym link as supplemented with the reverse hyponym link. This allows users of the dictionary exploring the semantically related words to a particular entry.

Step 6

The demonstration presents: a) tools, and procedures that can be used for electronic publication of content produced in the Dictionary Production System on desktops, netbooks, smart phones and e-books and for making adaptations to the user interface; b) features that end users have access to when they read text, write text or learn vocabulary; c) how online and offline access to content can be combined in for example products for smart phones.

Discussion

The procedure outlined requires some work for the first dictionary, but after that most is reusable:

- The WordNet as uploaded in step 1
- The process for the initial mapping in step 2
- The process for exporting the data in step 5
- The electronic publishing of the dictionary data in step 6

The only thing that has to be done from project to project is not surprisingly the refinement of the linking – steps 3 and 4.