Using WordNet for linking UWs to the UNL UW System

Luis Iraola

Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Madrid, Spain
luis@opera.dia.fi.upm.es

Abstract. This paper presents the work done with the Spanish-UNL dictionary compiled at the Spanish Language Centre in order to enrich the universal words it contained with the supplementary semantic information required to produce a master entries dictionary. Focusing on a subset of the Spanish-UNL dictionary, namely on the substantives it contains, the work has consisted in automatically enrich the universal word associated with each substantive with the semantic information required to link the universal word to the Universal Word System. For this process, WordNet has been employed as an external source of semantic information and used in addition to semantic features already present in the dictionary. The results achieved are not final and further work is required for a fully automatic, high quality semantic enrichment of the current entries. However, the work done shows the fruitfulness of the approach and its outcome has contributed to the creation of a master entries dictionary.

1   Introduction

A UNL dictionary in which language entries are associated with universal words (UW for short) can be viewed as a repository of UWs and as such does not organise its contents in any way. It links a set of UWs with lexical items of a specific language, each entry having no relation with any other. The necessity of establishing certain relations between UWs arises when considering several desirable features of the UNL system:

– Setting the combinatory possibilities of each UW with respect to any other UW regarding the conceptual relations that may link them and the attributes they may accept.
– Enabling a "fall-back" generation mechanism for those UWs that are not linked with head words in a given language at a given time. Those UWs would be replaced with semantically close but linked UWs so allowing generation to continue.

In order to support these features, a network with the set of UWs as nodes and semantic relations as arcs has been proposed. Such network is called the UNL UW System [1, 2]. Therefore, and in order to build the UW System, UNL Language Centres have to modify their respective UNL language dictionaries for including such new information. Once modified, the new master entries dictionaries will be the repository from which current language dictionaries will be produced as well as the UNL Knowledge Base will be created.

The current UW System consists of several hierarchies to which UWs are linked via inclusion relations ('icl') with broader meaning UWs. At the top level, the UW System distinguishes between entities ('thing'), actions originated by an agent ('do'), actions that happen without the intervention of an agent ('occur'), states ('be'), modifications of actions, events or states ('how') and modifications of entities. Each of these maximally general concepts ('thing', 'do', 'occur', 'be' and 'how') is the root of a hierarchy.

The hierarchy under 'thing' is by far the most elaborated one, containing distinctions between concrete and abstract things, functional and spatial entities and so on. Every UW denoting and entity must be located somewhere under the "thing" concept, and for doing this expert knowledge of the UW System and of the lexical meaning of the UW to be linked is required. Experts in the UW system and in English lexicography may manually establish the appropriate semantic links between each UW and the UW System. However, given the amount of entries that need to be processed (in the order of tens of thousands) any alternative that allows us to automate at least part of the task deserves to be explored.

2   Using Wordnet For Linking Uws

In order to automate the task of locating under "thing" the UWs associated with Spanish substantives, a procedure involving the use of WordNet [3] has been devised and put into practice. The procedure relies onto two central insights:

–   WordNet 1.6 covers the practical totality of the English lexicon. English substantives in particular are organised in a hierarchy using the hyponym relation, which has been found very similar to the relation of semantic inclusion ('icl') employed by UNL. Besides, a first inspection found substantial similarities between the WordNet hierarchy and the most general concepts employed in the UW System regarding the organisation of the meanings of substantives.
–   Given the polysemous nature of most English substantives, these words appear in more than one synset. However, these synsets frequently share a common hypernym in the hyponym hierarchy and this common ancestor can be related to a concept under 'thing' in the UW System.

2.1 Some Examples

The UW 'arrogance' (associated to the Spanish substantive 'arrogancia') has been located in the WordNet 1.6 hyponymy hierarchy in the following place:

```
1 sense of arrogance
Sense 1
arrogance, haughtiness, lordliness
   => pride
    => trait
     => attribute
      => abstraction
```

The synset identified with the meaning of 'arrogance' is linked in a chain of hypernyms with different nominal concepts until we reach the root node 'abstraction'. The distinction between concrete and abstract things plays also a key role in the UW System, and besides that, the intermediate node 'attribute' is also an organising concept in the UNL hierarchy. Therefore, using the hyponymy relation defined in Word-Net we can automatically link the UW 'arrogance' with 'attribute'.

The UW 'conquest' (associated with the Spanish substantive 'conquista') has been registered in WordNet 1.6 as polysemous, each of its three senses located in a different place under the hyponymy hierarchy:

```
3 senses of conquest
Sense 1
conquest, conquering, subjection, subjugation
  => capture, gaining control, seizure
   => acquiring, getting
    => deed, feat, effort, exploit
     => accomplishment, achievement
      => action
       => act, human action, human activity
Sense 2
conquest
  => success
   => attainment
    => accomplishment, achievement
      => action
        => act, human action, human activity
Sense 3
seduction, conquest
  => success
   => attainment
    => accomplishment, achievement
     => action
      => act, human action, human activity
```

The three senses of 'conquest' share common hypernyms from the node 'accomplishment, achievement' upward. Immediately upon this common node we find 'action', which is also a concept employed in the UW System for organising entities. Therefore, we can locate 'conquest' under 'action' since whatever the exact sense of 'conquest' is the intended one for its association with the Spanish headword, it must be under 'action' necessarily.

Linking 'conquest' with 'action' is certainly a high level link, one that does not precise the specific kind of action we are dealing with. However, this high level distinction is all that we need at the present since the UW System does not make finer distinctions under 'action'.

The word 'book', when considered a substantive, is highly polysemous according to WordNet 1.6:

```
8 senses of book
Sense 1
book
  => publication
   => work, piece of work
```

```
     => product, production
      => creation
       => artifact, artefact
        => object, physical object
         => entity, something
Sense 2
book, volume
  => product, production
   => creation
    => artifact, artefact
     => object, physical object
      => entity, something
Sense 3
record, recordbook, book
  => fact
   => information, info
    => message, content, subject matter, substance
     => communication
      => social relation
       => relation
        => abstraction
Sense 4
script, book, playscript
  => dramatic composition, dramatic work
   => writing, written material
    => written communication, written language
     => communication
      => social relation
       => relation
        => abstraction
Sense 5
ledger, leger, account book, book of account, book
  => record
   => document
    => communication
     => social relation
      => relation
       => abstraction
Sense 6
book
  => section, subdivision
   => writing, written material
    => written communication, written language
     => communication
      => social relation
       => relation
        => abstraction
Sense 7
daybook, book, ledger
  => journal
   => book, volume
    => product, production
     => creation
      => artifact, artefact
```

```
        => object, physical object
         => entity, something
Sense 8
book
  => product, production
   => creation
    => artifact, artefact
     => object, physical object
      => entity, something
```

From these eight nominal senses, those numbered 1, 2, 7 and 8 share hypernyms from the node 'product, production' upwards, while those numbered 3, 4, 5 and 6 do so from the node 'communication'. Given that there is no common hypernym for the eight senses of 'book', we can not link as easily as in the two previous examples the UW 'book' to the UW System. In this case, it is required to disambiguate which of the two chains of hypernyms should be used to link the UW or else if both of them are pertinent.

2.2 Linking Procedure

To summarise, and according to the previous examples, using WordNet for linking UWs associated with substantives requires the completion of the following steps:

1. To pair the high level concepts employed by the UW System with those present in the higher levels of the hyponym hierarchy of Wordnet.
2. To search the UWs in WordNet synsets and to analyse the hypernyms to which the synsets are linked to. In this process, four cases may arise:

   (a) The UW is not present in any synset. In this case the use of WordNet is of no help and the UW should be linked to UW System by other means. Given the large coverage of WordNet, this case should not be frequent.

   (b) The UW is monosemous. In this case, the chain of hypernyms is traversed until a node paired with an UW System concept is found.

   (c) The UW is polysemous and all its synsets share a common ancestor in their respective hypernym chains. In this case we proceed from that common node as in the previous case.

   (d) The UW is polysemous but there is not a single common ancestor for all the hypernym chains. In this case, extra information is needed for deciding which synset or set of synsets is chosen for linking the UW to the UW System.

2.3 Pairing WordNet 1.6 with the UW System

All first and second level concepts placed under 'thing' in the UW System have been paired with their counterparts in the WordNet 1.6 hyponym hierarchy. Frequently, the pairing is biunivocal, e.g. the UNL concept 'abstract thing', is paired with the WordNet node 'abstraction'. Occasionally, a UNL concept is paired with several synsets in WordNet. For instance, the UNL concept 'information' has been found related to several of the senses catalogued in WordNet for this word:

```
5 senses of information
Sense 1
information, info
  => message, content, subject matter, substance
   =>  communication
    => social relation
      => relation
       => abstraction
Sense 2
data, information
  => collection, aggregation, accumulation, assemblage
    => group, grouping
Sense 3
information
  => cognition, knowledge
    => psychological feature
Sense 4
information, selective information, entropy
  => measure, measurement
   => magnitude
    => property
     => attribute
       => abstraction
Sense 5
information
  => accusation, accusal
   => charge, complaint
    => pleading
     => allegation, allegement
      => claim
       => assertion, averment, asseveration
        => declaration
         => statement
          => message, content, subject matter
           => communication
            => social relation
             => relation
              => abstraction
```

At least the first four senses of 'information' are related to the UNL concept, so it has been paired with the four synsets. The pairing has been done manually. The entity hierarchy proposed in [1] has been examined and WordNet counterparts have been found for all UNL concepts placed on the first and second levels below 'thing'. A total of 73 high level concepts of the current UW System have been paired with their corresponding synsets in the WordNet substantive hierarchy.

All UWs associated with Spanish substantives in the Spanish-UNL dictionary have been automatically annotated with information coming from WordNet 1.6. Specifically, for each UW, its basic UW has been annotated with the set of synset identifiers in which the basic UW appears as a substantive. The set is empty for those basic UWs not found in WordNet, it contains a single identifier for monosemous basic UWs or several identifiers for polysemous basic UWs. For each synset identifier, the chain of hypernyms linking the synset with a top node of the WordNet hierarchy of substantives has been also retrieved. In the case of polysemous basic UWs, hypernym chains

sharing a common ancestor have been collapsed into a single chain starting from the common ancestor.

If more than one hypernym chain remains after collapsing chains with common ancestors, the UW is considered semantically ambiguous and extra information is required for selecting a single chain. Two information sources are exploited: the semantic restrictions that may occur along with the basic UW for creating the UW and certain semantic features that may be present in the Spanish substantive.

2.4 Disambiguation by Means of Semantic Restrictions

If the ambiguous UW has semantic restrictions, we may disambiguate it processing the restrictions in very much the same way as the basic UW: we annotate the restrictions with their corresponding hypernym chains and look for a common ancestor between one of these chains and one of those resulting from annotating the basic UW.

Example. The UW 'Malay(icl>language)' is ambiguous after annotating its basic UW 'Malay'. Two hypernym chains ending in 'Malay' share no common ancestor:

1. abstraction>relation>social_relation>communication>   language>natural_language>Austronesian>Malayo-Polynesian>Western_Malayo-Polynesian>Malay
2. entity>life_form>person>person_of_color>Asian> Malay

If we now take the basic UW employed as restriction ('language') and annotate it with its hypernym chains, we end up with two chains after grouping chains with common ancestors:

(a) abstraction>relation
(b) psychological feature>cognition

Chain (1) shares a common ancestor ('relation') with chain (a), while chain (2) shares no ancestor neither with (a) nor with (b). Therefore, we can select chain 1) for locating 'Malay(icl>language)' in the UW System and discard chain (2).

2.5 Disambiguation by Means of Semantic Features

When cataloguing Spanish substantives, two semantic features have been set for all of them because they are correlated with certain syntactic phenomena. The features 'human' and 'animate' are set to true for those substantives referring to human beings and animate entities respectively. This information is employed for reducing the number of hypernym chains in the following way: if the 'human' feature is set to true, all chains that do not include the node 'person' are discarded, if the 'animate' feature is set to true, all chains that do not include the node 'living thing' are discarded. In addition to these semantic features, the syntactic feature 'countable noun', employed for distinguishing mass nouns, is also taken into account: if 'countable noun' is set to true, all chains that do not include the node 'physical object' are discarded, if it is set to false, then all chains not including the nodes 'abstraction' or 'substance' are discarded.

Example. The UW 'translator' has been initially annotated with the hypernym chains:

1. abstraction>relation>social_relation>communication>   written_communication>writing>coding_system>code>software>program>translator
2. entity>life_form>person>communicator>negotiator> mediator>interpreter

Given that its associated Spanish substantive ('traductor') sets the feature 'human' to true, we can disambiguate and select chain 2) because it contains the node 'person'.

As for the order in which these information pieces is employed for disambiguation, semantic restriction are explored first, and only when they do not render a single hypernym chain the semantic features 'human' and 'animate' are taken into account. Eventually, the syntactic feature 'countable noun' is considered as a last resort.

2.6 Locating the UWs in the UW System

All UWs annotated with a single hypernym chain have been located in the UW System by linking them to the most specific chain node that is paired with a UNL concept.

Example. The UW 'Indonesian' is annotated with the following hypernym chain: entity > life_form > person > person_of_color > Asian > Indonesian.

Starting from its most specific node and moving upwards, the intermediate node 'person' is the first one that is paired to its homonymous UNL concept. Therefore, 'Indonesian' is located in the UW System by the following link: 'Indonesian'—icl→ 'person'

3   Results

14,911 UWs associated to Spanish substantives have been processed by the method just described. The initial annotation of these UWs with hypernym chains produced the following results:

Table1. Initial annotation results.

UWs not found in WordNet 1.6	1,447 (9.7%)
UWs Annotated with a single chain	7,863 (52.7%)
UWs Annotated with several chains	5,601 (37.5%)

The disambiguation mechanisms have been able to resolve 2,480 UWs (44,2%) from the total of 5,601 initially ambiguous UWs. Every non-ambiguous UW (7,863 plus 2,480) has been located in the UW System by linking it with one of the 73 high level concepts.

The final figures concerning the task of locating by automatic means the UWs associated to Spanish substantives in the Spanish-UNL dictionary rendered these final results:

Table 2. Final linking results.

UWs linked to the UW System	10,343 (69.3%)
UWs not linked because of ambiguity	3,121 (20.9%)
UWs not linked because not found	1,447 (9.7%)

3.1   Analysis of the results

Approximately ten percent of the UWs associated to Spanish substantives have not been found in WordNet 1.6. An analysis of this ten percent shows that most of these UWs fall into the following categories:

1. Proper names ('Alphonso', 'Louis', 'IAS')
2. Discrepancies in capitalisation ('Internet' versus 'internet' in WordNet 1.6)
3. Discrepancies in the use of separators ('leather jacket' and 'sister-in-law' versus 'leatherjacket' and 'sister_in_law' in WordNet 1.6).
4. Use of inflected forms as UWs: 'begs', 'studies', 'gratting'.
5. Use of phrases as UWs: 'garden wall', 'day pupil', 'small tail' o 'student music group'.

UWs included in categories 2 y 3 may be easily solved by a more flexible searching mechanism. Phrasal UWs may be syntactically analysed and their heads used instead of the whole phrase for linking purposes. Proper names require other resources such as lists of personal proper names and institutions while UWs included in category number 4 require careful examination.

Generally speaking, the UWs that remained ambiguous lack of semantic restrictions or have a very general restriction such as 'icl>thing'. These UWs may be disambiguated manually or their restrictions completed or make more precise. As for the quality performance of the disambiguation mechanisms, an initial inspection of the results allows to put forward the following considerations:

1. Using the semantic restrictions (when they are not extremely general) and the semantic features 'human' and 'animate' largely produces the selection of the correct hypernym chain.
2. Disambiguation based on the 'countable noun' feature is less reliable.

Disambiguation based on grouping chains sharing a common ancestor may lead to very general links in the UW System, since in the worst case the common ancestor is the top concept 'thing' and then the UW is linked to this general concept instead than to a more specific one, which is contrary to the main goal of the UW System.

4   Conclusions

This paper has presented a simple and effective way of using a well-known, freely available lexical resource such as WordNet for automating at least partially the creation of the UW System. Taking advantage of the conceptual similarities between WordNet and the UW System, we have mapped the upper levels of the UNL entity

hierarchy to the upper levels of the hyponym-hypernym relation defined in WordNet. This has open the possibility of automatically link a substantial part of the UWs associated with substantives in the Spanish-UNL dictionary with the concepts of the UW System.

The results obtained encourage a further development of this approach, deepening the mapping between UNL concepts and the WordNet hierarchy and exploring novel ways of disambiguating hypernym chains.

References

1. Uchida, I. "Master Dictionary specifications. Version 1.0", *UNDL Foundation*, October 2000.
2. Uchida, I. "The UNL UW System. Version 1.0", *UNDL Foundation*, January 2001.
3. Fellbaum, C. *WordNet: An Electronic Lexical Database*, MIT Press, 1998.