Automatic Building of Wordnets

Eduard Barbu* and Verginica Barbu Mititelu†

* Graphitech Italy
2, Salita Dei Molini
38050 Villazzano, Trento, Italy
eduard.barbu@graphitech.it

†Romanian Academy Research Institute for Artificial Intelligence
13 Calea 13 Septembrie,
Bucharest 050711, Romania
vergi@racai.ro

Abstract

In what follows we will present a two-phase methodology for automatically building a wordnet (that we call target wordnet) strictly aligned with an already available wordnet (source wordnet). In the first phase the synsets for the target language are automatically generated and mapped onto the source language synsets using a series of heuristics. In the second phase the salient relations that can be automatically imported are identified and the procedure for their import is explained. The assumptions behind such methodology will be stated, the heuristics employed will be presented and their success evaluated against a case study (automatically building a Romanian wordnet using PWN).

1. Introduction

The importance of a wordnet for NLP applications can hardly be overestimated. The Princeton WordNet (PWN) (Fellbaum 1998) is now a mature lexical ontology which has demonstrated its efficiency in a variety of tasks (word sense disambiguation, machine translation, information retrieval, etc.). Inspired by the success of PWN many languages started to develop their own wordnets taking PWN as a model (cf. http://www.globalwordnet.org/gwa/wordnet_table.htm). Furthermore, in both EuroWordNet (Vossen 1998) and BalkaNet (Tufiş 2004) projects the synsets from different versions of PWN (1.5 and 2.0) were used as ILI repositories. The created wordnets were linked by means of interlingual relations through this ILI repository.

The rapid progress in building a new wordnet and linking it with an already tested wordnet (usually PWN) is hindered by the amount of time and effort needed for developing such a resource. To take a recent example, the development of core wordnets (of about 20000 synsets, as is the case with the Romanian wordnet) for Balkan languages took three years (2001-2004).

2. Assumptions

The assumptions that we considered necessary for automatically building a target wordnet using a Source wordnet are the following:

1. There are word senses that can be clearly identified. This assumption is implicit when one builds a wordnet aligned or not with other wordnets. This premise was extensively questioned among others by (Kilgarriff 1997) who thinks that word senses have not a real ontological status, but they exist only relative to a task. We will not discuss this issue here.

2. A rejection of the strong reading of Sapir-Whorf (Caroll 1964) hypothesis (the principle of linguistic relativity). Simply stated, the principle of linguistic relativity says that

---

1 In both projects there was a number of synsets expressing language specific concepts added to the ILI repository.
language shapes our thought. There are two variants of this principle: strong determinism and weak determinism. According to the strong determinism language and thought are identical. This hypothesis has today few followers if any and the evidence against it comes from various sources among which the possibility of translation in other language. However, the weak version of the hypothesis is largely accepted. One can view the reality and our organization of reality by analogy with the spectrum of colors which is a continuum in which we place arbitrary boundaries (white, green, black, etc.). Different languages will “cut” differently this continuous spectrum. For example, Russian and Spanish have no words for the concept blue. This weak version of the principle of linguistic relativity warns us, however, that a specific source wordnet could not be used for automatically building any target wordnet. We further discuss this bellow.

3. The acceptance of the conceptualization made by the source wordnet By conceptualization we understand the way in which the source Wordnet “sees” the reality by identifying the main concepts to be expressed and their relationships. For specifying how different languages can differ with respect to conceptual space they reflect we will follow (Sowa 1992) who considers three distinct dimensions:

- **accidental.** The two languages have different notations for the same concepts. For example the Romanian word *măr* and the English word *apple* lexicalize the same concept.
- **systematic.** The systematic dimension defines the relation between the grammar of a language and its conceptual structures. It deals with the fact that some languages are SVO or VSO, etc., some are analytic and other agglutinative. Even if it is an important difference between languages, the systematic dimension has little import for our problem

- **cultural.** The conceptual space expressed by a language is determined by environmental, cultural factors, etc. It could be the case for example, that concepts that define the legal systems of different countries are not mutually compatible. So when someone builds a wordnet starting from a source wordnet he/she should ask himself/herself what the parts (if any) that could be safely transferred in the target language are. More precise what the parts that share the same conceptual space are.

The assumption that we make use of is that the differences between the two languages (source and target) are merely accidental: they have different lexicalizations for the same concepts. As the conceptual space is already expressed by the Source wordnet structure using a language notation, our task is to find the concepts notations in the target language.

When the Source wordnet is not perfect (the real situation), then a drawback of the automatic mapping approach is that all the mistakes existent in the source wordnet are transferred in the target wordnet: consider the following senses of the noun *hindrance* in PWN:

1. hindrance, deterrent, impediment, balk, baulk, check, handicap -- (something interferes with or delays action or progress)
   \[\Rightarrow\] cognitive factor -- (something immaterial that interferes with or delays action or progress)
2. hindrance, hitch, preventative, preventative, encumbrance, incumbrance, interference -- (any obstruction that impedes or is burdensome)
   \[\Rightarrow\] artifact, artefact -- (a man-made object taken as a whole).

We listed the senses 1 and 2 of the word *hindrance* together with one of their hyperonyms. As one can see, in PWN a distinction is made between *hindrance* as a cognitive factor and *hindrance* as an artefact, so that these are two different senses of the word *hindrance*. According to this definition a speed bump can be classified as a hindrance because it is an artifact, but a stone that stays in the path of someone cannot be one, because it is not a man made object.

Another possible problem appears because the previously made assumption about the sameness of the conceptual space is not always true as the following example shows:

- mister, Mr -- (a form of address for a man)
- sir -- (term of address for a man)

In Romanian both *mister* and *sir* in the listed senses are translated by the word *domn*. But in Romanian it would be artificial to create two distinct synsets for the word *domn*, as they are not different, not even in what their connotations are concerned.

3. Selection of concepts and resources used

When we selected the set of synsets to be implemented in Romanian we followed two criteria.

The first criterion states that the selected set should be structured in the source wordnet (i.e. every selected synset should be linked by at least one semantic relation with other selected synsets). This is dictated by the methodology we have adopted (automatic mapping and automatic relation import). If we want to obtain a wordnet in the target language and not just some isolated synsets, this criterion is self-imposing.

The second criterion is related to the evaluation stage. To properly evaluate the built wordnet, it should be compared with a “golden standard”. The golden standard that we use will be the Romanian Wordnet (RoWN) developed in the BalkaNet project

---

2 One can argue that this Romanian wordnet is not perfect and definitely incomplete. However, PWN is neither perfect. Moreover, it is indisputable that at least in the case of ontologies (lexical or
For fulfilling both criteria we chose a subset of noun concepts from the RoWN that has the property that its projection on PWN 2.0 is closed under the hyperonym and the meronym relations. Moreover, this subset includes the upper level part of the PWN lexical ontology. The projection of this subset on PWN 2.0 comprises 9716 synsets that contain 19624 literals.

For the purpose of automatic mapping of this subset we used an in-house dictionary built from many sources. The dictionary has two main components:

- The first component consists of the above-mentioned 19624 literals and their Romanian translations. We must make sure that this part of the dictionary is as complete as possible. Ideally, all senses of the English words should be translated. For that we used the (Leviţchi & Bantaş 1992) dictionary and other dictionaries available on web.
- The second component of the dictionary is (Leviţchi & Bantaş 1992) dictionary. Some dictionaries (in our case the dictionary extracted from the already available Romanian wordnet) also have sense numbers specified, but, from our experience, this information is highly subjective, does not match the sense as defined by PWN and it is not consistent over different dictionaries, so we chose to disregard it.

The second resource used is the Romanian Explanatory Dictionary (EXPD 1996) whose entries are numbered to reflect the dependencies between different senses of the same word.

4. Notation introduction

In this section we introduce the notations used in the paper and we outline the guiding idea of all heuristics we used:

1. By $T_I$ we denote the target lexicon. In our experiment $T_L$ will contain Romanian words (nouns).

2. By $S_i$ we denote the source lexicon. In our case $S_i$ will contain English words (nouns).

3. $W_T$ and $W_S$ are the wordnets for the target language and the source language, respectively.

4. $w_j^k$ denotes the $k^{th}$ sense of the word $w_j$.

5. $B_B$ is a bilingual dictionary which acts as a bridge between $S_i$ and $T_L$. $B_B = (S_i, T_L, M)$ is a 3-tuple, where $M$ is a function that associates to each word in $S_i$ a set of words in $T_L$. For an arbitrary word $w_j^k \in S_i$, $M(w_j^k) = \{w_j^1, w_j^2, \ldots, w_j^n\}$.

Formally the bilingual dictionary maps words and not word senses. If word senses had been mapped, then building $W_T$ from a $W_S$ would have been trivial. If we ignore the information given by the definitions associated with word senses, then, formally a sense of a word in the PWN is distinguished from other word senses only by the set of relation it contacts in the semantic network. This set of relations defines what it is called the position of a word in the semantic network. Ideally, every sense of a word should be unambiguously identified by a set of connections; it should have a unique position in the semantic net. Unfortunately this is not the case in PWN. There are many cases when different senses of a word have the same position in the semantic network (i.e they have precisely the same connections with other word senses).

The idea of our heuristics could be summed up in three points:

1. Increase the number of relations in the Source wordnet to obtain a unique position for each word sense. For this an external resource can be used to which the wordnet is linked, such as Wordnet Domains.
2. Try to derive useful relations between the words in the target language. For this one can use corpuses, monolingual dictionaries, already classified set of documents etc.
3. In the mapping stage of the procedure take profit of the structures built at points 1 and 2.

We have developed so far a set of four heuristics and we plan to supplement them in the future.

5. The first heuristic rule

The first heuristic exploits the fact that synonymy enforces equivalence classes on word senses.

Let $EnSyn = \{w_{1j}^i, w_{2j}^i, \ldots, w_{nj}^i\}$ where $w_{1j}^i, w_{2j}^i, w_{nj}^i$ are the words in synset and the superscripts denote their sense numbers (be a $S_i$ synset and length(EnSyn) > 1). We impose the length of a synset to be greater than one when at least one component word is not a variant of the other words. So we disregard synsets such as {artefact, artifact}. For achieving this we computed the well known Levenshtein distance between the words in the synset. The $B_B$ translations of the words in the synset will be:

$M(w_{1j}^i) = \{w_{1j}^1, \ldots, w_{1j}^n\}$

$M(w_{2j}^i) = \{w_{2j}^1, \ldots, w_{2j}^n\}$

$\ldots$

$M(w_{nj}^i) = \{w_{nj}^1, \ldots, w_{nj}^n\}$

We build the corresponding $T_L$ synset as

1. $M(w_{ja}^i)$ if $\exists w_{ja}^i \in EnSyn$ such that the number of senses $NoSenses (w_{ja}^i) = 1$
2. \[ M(\text{ew}_{j1}) \cap M(\text{ew}_{j2}) \ldots \cap M(\text{ew}_{jn}) \]

Words belonging to the same synset in \( S_i \) should have a common translation in \( T_L \). Above we distinguished two cases:

1. At least one of the words in a synset is monosemous. In this case we build the \( T_L \) synset as the set of translations of the monosemous word.

2. All words in the synset are polysemous. The corresponding \( T_L \) synset will be constructed by the intersection of all \( T_L \) translations of the \( S_i \) words in the synset.

Taking the actual RoWNAs a gold standard we can evaluate the results of our heuristics by comparing the obtained synsets with those in the RoWN. We distinguish five possible cases:

1. The synsets are equal (this case will be labeled as Identical).

2. The generated synset has all literals of the correct synset and some more. (Over-generation).

3. The generated synset and the golden one have some literals in common and some different (Overlap)

4. The generated synset literals form a proper subset of the golden synset (Under-generation)

5. The generated synset have no literals in common with the correct one (Disjoint).

The cases Over-generation, Overlap and Disjoint will be counted as errors. The other two cases, namely Identical and Under-generation, will be counted as successes.

The evaluation of the first heuristics is given in Table 1, at the end of section 9.

The percents mapped column contains the percents of the synsets mapped by the heuristics from the total number of the synsets (9716). The percent errors column represents the percent of synsets from the number of mapped synsets wrongly assigned by the heuristics. The high number of mapped synsets proves the quality of the first part of the dictionary we used. The only type of error we encountered is Over-generation.

6. The second heuristic rule

The second heuristic draws from the fact that, in the case of nouns, the hyperonymy relation can be interpreted as an IS-A relation. It is also based on two related observations:

1. A hyperonym and his hyponyms carry some common information.

2. The information common to the hyperonym and the hyponym will increase as you go down in the hierarchy.

Let \( \text{EnSyn}_1 = \{ \text{ew}_{j1}^{w1}, \text{ew}_{j2}^{w2}, \ldots, \text{ew}_{jn}^{wn} \} \) and \( \text{EnSyn}_2 = \{ \text{ew}_{j1}^{w1}, \text{ew}_{j2}^{w2}, \ldots, \text{ew}_{jn}^{wn} \} \) be two \( S_i \) synsets such that \( \text{EnSyn}_1 \) HYP \( \text{EnSyn}_2 \), meaning that \( \text{EnSyn}_1 \) is a hyperonym of \( \text{EnSyn}_2 \). Then we generate the translation lists of the words in the synsets. The intersection is computed as:

\[ \text{T}_L \text{EnSyn}_1 = M(\text{ew}_{j1}^{w1}) \cap M(\text{ew}_{j2}^{w2}) \ldots \cap M(\text{ew}_{jn}^{wn}) \]

\[ \text{T}_L \text{EnSyn}_2 = M(\text{ew}_{j1}^{w1}) \cap M(\text{ew}_{j2}^{w2}) \ldots \cap M(\text{ew}_{jn}^{wn}) \]

The generated synset in the target language will be computed as

\[ \text{T}_L \text{Synset} = \text{T}_L \text{EnSyn}_1 \cap \text{T}_L \text{EnSyn}_2 \]

Given the above consideration, it is possible that a hyponym and its hyperonym have the same translation in the other language and this is more probable as you descend in the hierarchy. The procedure formally described above is applied for each synset in the source list. It generates the lists of common translations for all words in the hyperonym and hyponym synsets and then constructs the \( T_L \) synsets by intersecting these lists. In case the intersection is not empty the created synset will be assigned to both \( S_i \) language synsets.

Because the procedure generates autohyperonym synsets this could be an indication that created synsets could be clustered in \( T_L \).

It is possible that a \( T_L \) synset be assigned to two different source pair synsets as in the figure below. So we need to perform a clean-up procedure and choose the assignment that maximizes the sum of depth level of the two synsets.

In the figure common information is found between the middle synset and the upper synset (its hyperonym) and also between the middle synset and the lower synset (its hyponym). Our procedure will prefer the second assignment.

The results of the second heuristic are presented in Table 2, at the end of section 9. The low number of mapped synsets (10%) is due to the fact that we did not find many common translations between hyperonyms and their hyponyms.

7. The third heuristic

The third heuristics takes profit of an external relation imposed over the wordnet. At IRST PWN 1.6 was augmented with a set of Domain Labels, the resulting resource being called Wordnet Domains (Magnini & Cavaglia 2000). PWN 1.6 synsets have been semi-automatically linked with a set of 200 domain labels taken from Dewey Decimal classification, the world most widely used library classification system. The domain labels are hierarchically organized and each synset received one or more domain labels. For the synsets that cannot be labeled unambiguously the default label “factotum” has been used.

Because in the BalkaNet project the RoWN has been aligned with PWN 2.0, we performed a mapping...
between PWN 1.6 and PWN 2.0. By comparison with PWN1.6, PWN 2.0 has new additional synsets and also the wordnet structure is slightly modified. As a consequence not all PWN 2.0 synsets can be reached from PWN 1.6 either because they are completely new or because they could not be unambiguously mapped. This results in some PWN 2.0 synsets that have not domains. For their labelling we used the three rules below:

1. If one of the direct hyperonym of the unlabeled synsets has been assigned a domain, then the synset will automatically receive the father’s domain, and conversely, if one of the hyponyms is labelled with a domain and father lacks domain, then the father synset will receive the son’s domain.
2. If a holonym of the synset is assigned a specific domain, then the meronym will receive the holonym domain and conversely.
3. If a domain label cannot be assigned, then the synset will receive the default “factotum” label.

The idea of using domains is helpful for distinguishing word senses (different word senses of a word are assigned to different domains). The best case is when each sense of a word has been assigned to a distinct domain. But even if the same domain labels are assigned to two or more senses of a word, in most cases we can assume that this is a strong indication of a fine-grained distinction. It is very probable that the distinction is preserved in the target language by the same word.

We labelled every word in the B3 dictionary with its domain label. For English words the domain is automatically generated from the English synset labels. For labelling Romanian words we used two methods:

1. We downloaded a collection of documents from web directories such that the categories of the downloaded documents match the categories used in the Wordnet Domain. The downloaded document set underwent a pre-processing procedure with the following steps:
   a. Feature extraction. The first phase consists in finding a set of terms that represents the documents adequately. The documents were POS tagged and lemmatized and the nouns were selected as features.
   b. Features selection. In this phase the features that provide less information were eliminated. For this we used the well known \( \chi^2 \) statistic. \( \chi^2 \) statistic checks if there is a relationship between being in a certain group and a characteristic that we want to study. In our case we want to measure the dependency between a term \( t \) and a category \( c \). The formula for \( \chi^2 \) is:
   \[
   \chi^2 (t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}
   \]
   Where:
   - \( A \) is the number of times \( t \) and \( c \) co-occur
   - \( B \) is the number of times \( t \) occur without \( c \)
   - \( C \) is the number of times \( c \) occurs without \( t \)
   - \( D \) is the number of times neither \( c \) nor \( t \) occurs
   - \( N \) is the total number of documents
   For each category we computed the score between that category and the noun terms of our documents. Then, for choosing the terms that discriminate well for a certain category we used the formula below (where \( m \) denotes the number of categories):
   \[
   \chi^2_{\text{max}} (t) = \max_m (\chi^2 (t, c_i))
   \]
   2. We took advantage of the fact that some words have already been assigned subject codes in various dictionaries. We performed a manual mapping of these codes onto the Domain Labels used at IRST. The Romanian words that could not be associated domain information were associated with the default factotum domain.

The following entry is a B3 dictionary entry augmented with domain information:

\[
\text{M} (\text{ew} \_1, \text{ew} \_2, \ldots) = \{ \text{rw} \_1 \_1, \text{rw} \_1 \_2, \text{rw} \_2 \_1, \text{rw} \_2 \_2, \ldots \}
\]

In the square brackets the domains that pertain to each word are listed.

Let again \( \text{EnSyn}_1 = \{ \text{ew} \_1 \_1, \text{ew} \_1 \_2, \ldots \} \) be an \( S_L \) language synset and \( D \) the associated domain. Then the \( T_L \) synset will be constructed as follows:

\[
\text{T}_L \ Synset = \bigcup_{m=1}^{m} \text{M}(\text{ew}_m), \quad \text{where each} \quad \text{rw}_i \in \text{M}(\text{ew}_m)
\]

\( \text{rw}_i \) has the property that its domain “matches” the domain of \( \text{EnSyn}_1 \) that is: either is the same as the domain of \( \text{EnSyn}_1 \), subsumes the domain of \( \text{EnSyn}_1 \) in the IRST domain labels hierarchy or is subsumed by the domain of \( \text{EnSyn}_1 \) in the IRST domain labels hierarchy.

For each synset in the \( S_L \) we generated all the translations of its literals in the \( T_L \). Then the \( T_L \) synset is built using only those \( T_L \) literals whose domain “matches” the \( S_L \) synset domain.

The results of this heuristic are given in Table 3 at the end of section 9.

8. The fourth heuristic rule
The fourth heuristics takes advantage of the fact that the source synsets have a gloss associated and also that target words that are translations of source words have associated glosses in EXPD. As with the third heuristic the procedure comprises a preprocessing phase. We preprocessed both resources (PWN and EXPD):

1. We automatically lemmatized and tagged all the glosses of the synsets in the \( S_L \).
2. We automatically lemmatized and tagged all the definitions of the words that are translations of \( S_L \) words.
3. We chose as features for representing the glosses the set of nouns. The target definitions were automatically translated using the bilingual dictionary. All possible source definitions were generated by translating each lemmatized noun word in the Tl definition. Thus, if a Tl definition of one Tl word is represented by the following vector \([r_w_1, r_w_2, \ldots r_w_p]\), then the number of Sl vectors generated will be: \(N = n_d \times t_{w_1} \times t_{w_2} \times \ldots t_{w_p}\), where \(n_d\) is the number of definitions the target word has in the monolingual dictionary (EXPD), and \(t_{w_i}\), with \(k=1..p\), is the number of translations that the noun \(w_i\) has in the bilingual dictionary.

By \(R_{\text{Gloss}}\) we denote the set of Sl representation vectors of Tl glosses of a Tl word: \(R_{\text{Gloss}} = \{T_1, T_2, \ldots T_n\}\). By \(S_v\) we denote the vector of Sl synset gloss.

The procedure for generating the Tl synset is: for each Sl synset we generate the Tl list of the translation of all words in the synset. Then for each word in the Tl list of translation we compute the similarity between \(S_v\) and its \(R_{\text{Gloss}}\). The computation is done in two steps:

1. We give the vectors in \(S_v\) and \(R_{\text{Gloss}}\) a binary representation. The number (m) of positions a vector has will be equal to the number of distinct words existent in the \(S_v\) and in all vectors of \(R_{\text{Gloss}}\). The presence of 1 in the vector means that a word is present and the existence of 0 means that a word is absent from the vector.

2. For each \(T_i\) vector in \(R_{\text{Gloss}}\) we compute the product: \(S_v \bullet T_i = \sum_j s_j * t_j\). If there exists at least one such that \(\sum_j s_j * t_j \geq 2\) we compute \(\max(S_v \bullet T_i)\) and we add the word to the Tl synset.

Notice that by using this heuristic rule we can automatically add a gloss to the Tl synset.

As one can see in Table 4 at the end of section 9 the number of incomplete synsets is high. The percent of mapped synsets is due to the low agreement between the glosses in Romanian and English.

### 9. Combining results

For choosing the final synsets we devised a set of meta-rules by evaluating the pro and con of each heuristic rule. For example, given the high quality dictionary the probability that the first heuristic will fail is very low. So the synsets obtained using it will automatically be selected. A synset obtained using the other heuristics will be selected and moreover will replace a synset obtained using the first heuristic, only if it is obtained independently using the heuristics 3 and 2, or by using the heuristics 3 and 4. If a synset is not selected by the above meta-rules will be selected only if it is obtained by the heuristics number 3 and the ambiguity of its members is at most equal to 2. Table 5 at the end of this section shows the combined results of our heuristics.

As one can observe there, for 106 synsets in PWN 2.0 the Romanian equivalent synsets could not be found. There also resulted 635 synsets smaller than the synsets in the RoWN.

### Table 1: The results of the first heuristic

<table>
<thead>
<tr>
<th>Number of mapped synsets</th>
<th>Percents mapped</th>
<th>Error types</th>
<th>Correct</th>
<th>Percent errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Over-generation</td>
<td>Overlap</td>
<td>Disjoint</td>
</tr>
<tr>
<td>8493</td>
<td>87</td>
<td>210</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 2: The results of the second heuristic

<table>
<thead>
<tr>
<th>Number of mapped synsets</th>
<th>Percents mapped</th>
<th>Error types</th>
<th>Correct</th>
<th>Percent errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Over-generation</td>
<td>Overlap</td>
<td>Disjoint</td>
</tr>
<tr>
<td>1028</td>
<td>10</td>
<td>213</td>
<td>0</td>
<td>150</td>
</tr>
</tbody>
</table>
10. Import of relations
After building the target synsets an investigation of the nature of the relations that structure the source wordnet should be made for establishing which of them can be safely transferred in target wordnet. As one expects the conceptual relations can be safely transferred because these relations hold between concepts. The only lexical relation that holds between nouns and that was subject to scrutiny was the antonym relation. We concluded that this relation can also be safely imported. The importing algorithm works as described bellow.

If two source synsets $S_1$ and $S_2$ are linked by a semantic relation $R$ in $W_S$ and if $T_1$ and $T_2$ are the corresponding aligned synsets in the $W_T$, then they will be linked by the relation $R$. If in $W_S$ there are intervening synsets between $S_1$ and $S_2$, then we will set the relation $R$ between the corresponding $T_L$ synsets only if $R$ is declared as transitive (R+, unlimited number of compositions, e.g. hyponym) or partially transitive relation (R$k$ with $k$ a user-specialized maximum number of compositions, larger than the number of intervening synsets between $S_1$ and $S_2$). For instance, we defined all the holonymy relations as partially transitive ($k=3$).

11. Conclusion and future work
Other experiments of automatically building wordnets that we are aware of are (Atserias et al., 1997) and (Lee et al., 2000). They combine several methods, using monolingual and bilingual dictionaries for obtaining a Spanish Wordnet and, respectively, a Korean one starting from PWN 1.5.

However, our approach is characterized by the fact that it gives an accurate evaluation of the results by automatically comparing them with a manually built wordnet. We also explicitly state the assumptions of this automatic approach. Our approach is the first to use an external resource (Wordnet Domains) in the process of automatically building a wordnet.

We obtained a version of RoWN that contains 9610 synsets and 11969 relations with 91% accuracy.

The results obtained encourage us to develop other heuristics. The success of our procedure was facilitated by the quality of the bilingual dictionary we used.

Some heuristics developed here may be applied for the automatic construction of synsets of other parts of speech. That is why we also plan to extend our experiment to adjectives and verbs. Their evaluation would be of great interest in our opinion.

Finally we would like to thank the three anonymous reviewers for helping us in improving the final version of the paper.

References:

---

Table 3: The results of the third heuristic

<table>
<thead>
<tr>
<th>Number of mapped synsets</th>
<th>Percents mapped</th>
<th>Error types</th>
<th>Correct</th>
<th>Percent errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Over-generation</td>
<td>Overlap</td>
<td>Disjoint</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7520</td>
<td>77</td>
<td>689</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: The results of the fourth heuristic

<table>
<thead>
<tr>
<th>Number of mapped synsets</th>
<th>Percents mapped</th>
<th>Error types</th>
<th>Correct</th>
<th>Percent errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Over-generation</td>
<td>Overlap</td>
<td>Disjoint</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3527</td>
<td>36</td>
<td>25</td>
<td>0</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 5: The combined results of the heuristics

<table>
<thead>
<tr>
<th>Number of mapped synsets</th>
<th>Percents mapped</th>
<th>Error types</th>
<th>Correct</th>
<th>Percent errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Over-generation</td>
<td>Overlap</td>
<td>Disjoint</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9610</td>
<td>98</td>
<td>615</td>
<td>0</td>
<td>250</td>
</tr>
</tbody>
</table>


