A Case Study in Automatic Building of Wordnets

Eduard Barbu
Graphitech

Trento, 38050, Italy
eduard.barbu@graphitech.it

Verginica Barbu Mititelu
Research Institute for Artificial Intelligence
Romanian Academy
Bucharest, 050711, Romania
vergi@racai.ro

Abstract

In this paper we 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 Princeton WordNet).

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).

In what follows we present a methodology that can be used for automatically building wordnets strictly aligned (that is, using only the EQ_SYNONYM relation) with an already available wordnet. We have started our experiment with the study of nouns, so the data presented here are valid only for this grammatical category.

The methodology we present has two phases. In the first one 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 paper has the following organization. Firstly we state the implicit assumptions in building a wordnet strictly aligned with other wordnets. Then we shortly describe the resources that one needs in order to apply the heuristics, and also the criteria we used in selecting the Source language test synsets to be implemented. Finally, we state the problem to be solved in a more formal way, the heuristics employed will be presented and their success evaluated against a case study (automatically building a Romanian wordnet using PWN 2.0).

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 (Carroll, 1964) hypothesis (the principle of linguistic relativity).
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 with to conceptual space they reflect we will follow (Sowa, 1992) who consider 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 are the parts (if any) that could be safely transferred in the target language.

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.

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. This is dictated by the methodology we have adopted (automatic mapping and automatic relation import). 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[1].

For fulfilling both criteria we chose a subset of noun concepts from the Romanian Wordnet 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 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 (Levițchi and Bantaş, 1992) dictionary and other dictionaries available on web.

- The second component is (Levițchi and Bantaş, 1992) dictionary.

[1] 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 formal) a manually or semi-automatically built ontology is much better than an automatically built one.

The other 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 $\mathbf{T_L}$ we denote the target lexicon. In our experiment T_L will contain Romanian words (nouns). $T_L = \{ rw_1, rw_2, \ldots rw_m \}$ where rw_i, with i=1..m, denotes a target word.
2. By $\mathbf{S_L}$ we denote the source lexicon. In our case S_L will contain English words (nouns). $S_L = \{ ew_1, ew_2, \ldots ew_n \}$ where ew_j, with j=1..n, denotes a source word.
3. $\mathbf{W_T}$ and $\mathbf{W_S}$ 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. $\mathbf{B_D}$ is a bilingual dictionary which acts as a bridge between $\mathbf{S_L}$ and $\mathbf{T_L}$. $B_D = (\mathbf{S_L}, \mathbf{T_L}, M)$ is a 3-tuple, where M is a function that associates to each word in $\mathbf{S_L}$ a set of words in $\mathbf{T_L}$. For an arbitrary word $ew_j \in S_L$, $M(ew_j) = \{ rw_1, rw_2, \ldots rw_k \}$.

A bilingual dictionary formally maps words and not word senses. If word senses had been mapped, then building W_T from a W_S would have been trivial.

Our heuristics is that by increasing the number of interconnections between words in both languages we can uniquely characterize the meaning of a word (synset) as the set of its relations with other synsets or other words. In the source wordnet a set of relations already exist. Other useful relations can be derived in the source wordnet or can be imposed by an external resource with which the wordnet is linked (ontology, domains). In the target language the useful relations can be derived using resources like corpuses, monolingual dictionaries, already classified sets of documents. 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 imposes an equivalence class on word senses.

Let $EnSyn = \{ ew_{j_{11}}^{i_{11}}, ew_{j_{12}}^{i_{12}} \ldots ew_{j_{1n}}^{i_{1n}} \}$ (where $ew_{j_{11}}$, $ew_{j_{12}}$, $ew_{j_{1n}}$ are the words in synset and superscripts denote their sense numbers) be a S_L 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}.

The B_D translations of the words in the synset will be:

$M(ew_{j_{11}}) = \{ rw_{i_{11}}, \ldots rw_{i_{1m}} \}$

$M(ew_{j_{12}}) = \{ rw_{i_{21}}, \ldots rw_{i_{2k}} \}$

....

$M(ew_{j_{1n}}) = \{ rw_{i_{n1}}, \ldots rw_{i_{nt}} \}$

We build the corresponding $\mathbf{T_L}$ synset as

$M(ew_{j_{ik}})$ if $\exists\ ew_{j_{ik}} \in EnSyn$ such that $NoSenses(ew_{j_{ik}}) = 1$

$M(ew_{j_{11}}) \cap M(ew_{j_{12}}) \ldots \cap M(ew_{j_{1n}})$ otherwise.

Words belonging to the same synset in S_L should have a common translation in $\mathbf{T_L}$. Above we distinguished two cases:

1. At least one of the words in a synset is monosemous.
2. All words in the synset are polysemous.

In the first case we build the $\mathbf{T_L}$ synset as the set of translations of the monosemous word. In the second case the corresponding $\mathbf{T_L}$ synset will be constructed by the intersection of all $\mathbf{T_L}$ translations of the S_L words in the synset.

Taking the actual RoWN as 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:

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

The generated synset has all literals of the correct synset and at least one more (Over-generation, OG).

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

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

The generated synset has no literals in common with the correct one (Disjoint,DJ).

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[2].
The evaluation of the first heuristics is given in Table 1.

NMS	PM	Error types			Correct		PE
		OG	OP	DJ	UG	ID	
8493	87	210	0	0	300	7983	2

Table 1. The results of the first heuristic.

The NMS column represents the number of synsets mapped by the heuristic, the Percents mapped (PM) column contains the percents of the synsets mapped by the heuristics from the total number of the synsets (9716). The Percent errors (PE) 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 component 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 the hyperonymy relation can be interpreted as an IS-A relation[3]. 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 $EnSyn_1$={ $ew_{j_{11}}^{i_{11}}$, $ew_{j_{12}}^{i_{12}}$ ... $ew_{j_{1t}}^{i_{1t}}$ } and $EnSyn_2$={ $ew_{j_{21}}^{i_{21}}$, $ew_{j_{22}}^{i_{22}}$ ... $ew_{j_{2s}}^{i_{2s}}$ } be two S_L synsets such that $EnSyn_1$ HYP $EnSyn_2$, meaning that $EnSyn_1$ is a hyperonym of $EnSyn_2$. Then we generate the translation lists of the words in the synsets. The intersection is computed as before:

T_L $EnSyn_1$= $M(ew_{j_{11}}) \cap M(ew_{j_{12}}) \ldots \cap M(ew_{j_{1t}})$

T_L $EnSyn_2$ = $M(ew_{j_{21}}) \cap M(ew_{j_{22}}) \ldots \cap M(ew_{j_{2s}})$

The generated synset in the target language will be computed as

T_L Synset = T_L $EnSyn_1$ $\cap$ T_L $EnSyn_2$

Given the above consideration, it is possible that a hyponym and its hyperonym have the same translation in 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 list of translations for all words in the hyperonym and hyponym synsets and then constructs the T_L synsets by intersecting their translations. In case the intersection is not empty the created synset will be assigned to both S_L language synsets.

Because the procedure generates autohyponym synsets this could be an indication that the T_L created synsets could be clustered in the T_L.

The results of the second heuristic are presented in Table 2. 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.

NMS	PM	Error types			Correct		PE
		OG	OP	DJ	UG	ID	
1028	10	213	0	150	230	435	35

Table 2. The results of the second heuristic.

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 and Cavaglia, 2000).
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

[2] The Under-generation case means that the resulted synset is not reach enough; it does not mean that it is incorrect.

[3] This not entirely true because in PWN the hyperonym relation can also be interpreted as an INSTANCE-OF relation, as in PWN there are also some instances included (e.g. *New York*, *Adam*, etc.).

is very probable that the distinction is preserved in the target language by the same word.

For our task we label every word in the B_D dictionary with its domain label. For English words the domain is automatically generated from the English synset labels. For labeling Romanian words 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. After POS tagging and lemmatizing the documents we selected as features the most relevant nouns. For this we used the well known χ^2 statistic. The selected nouns were labeled with the corresponding document categories.
The following entry is a B_D dictionary entry augmented with domain information:

$M(ew_1\ [D_{1,...}]) = rw_1\ [D_1, D_2 \ldots],\ rw_2\ [D_1, D_3 \ldots],$
$rw_i\ [D_2, D_4 \ldots]$

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

Let again $EnSyn_l = \{ ew_{j_{11}}^{i_{11}},\ ew_{j_{12}}^{i_{12}} \ldots\ ew_{j_{1n}}^{i_{1n}} \}$ be an S_L synset and D_i the associated domain. Then the $\mathbf{T_L}$ synset will be constructed as follows:

$\mathbf{T_L}$ Synset $= \bigcup_{m=j_{11}..j_{1n}} M(ew_m)$, where each $rw_i \in M(ew_m)$ has the property that its domain matches the domain of $EnSyn_l$.

For each synset in the S_L we generated all the translations of its literals in the $\mathbf{T_L}$. Then the $\mathbf{T_L}$ synset is built using only those $\mathbf{T_L}$ literals whose domain matches the S_L synset domain. The results of this heuristic are given in Table 3.

NMS	PM	Error types			Correct		PE
		OG	OP	DJ	UG	ID	
7520	77	689	0	0	0	6831	9

Table 3. The results of the third heuristic.

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. After we lemmatized and tagged all the definitions of the synsets in the S_L and all the definitions of target words that are translations of S_L words we performed a gloss match. The target definitions were translated using B_D and compared with the source definitions. The comparison procedure counted the number of nouns common to both definitions.
As one can see in Table 4 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.

NMS	PM	Error types			Correct		PE
		OG	OP	DJ	UG	ID	
3527	36	25	0	78	547	2877	3

Table 4. The results of the fourth heuristic.

9 Combining results

For choosing the final synset we devised a set of meta-rules by evaluating the pro and con of each heuristic rule. For example, given a high quality dictionary the probability that the first heuristic will fail is very low. So the synsets obtained using it will be automatically 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 his members is 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.

NMS	PM	Error types			Correct		PE
		OG	OP	DJ	UG	ID	
9610	98	615	0	250	635	8110	9

Table 5. The combined results of the heuristics.

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 seeing which of them can be safely transferred in the target wordnet. The conceptual relations can be safely transferred because they hold between concepts. The only lexical relation that holds between nouns and that was subject of 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 $\mathbf{T_L}$ synsets only if R is declared as transitive (R+, unlimited number of compositions, e.g. hypernym) 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 Conclusions and future work

Other experiments of automatically building wordnets that we are aware off 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 Romanian Wordnet that contains 9610 synsets and 11969 relations with 91% accuracy.

The results obtained encourage us to develop other heuristics for automatically building a target Wordnet from a source Wordnet. 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 two anonymous reviewers for helping us in improving the final version of the paper.

References

J. Atserias, S. Clement, X. Farreres, G. Rigau, H. Rodríguez. 1997. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In *Proceedings of the International Conference on Recent Advances in Natural Language*.

J. B. Carroll (Ed.). 1964. *Language, Thought and Reality Selected writings of Benjamin Lee Whorf*, The MIT Press, Cambridge, MA.

Dicţionarul explicativ al limbii române. 1996. 2nd edition, Bucureşti, Univers Enciclopedic.

Ch. Fellbaum (Ed.). 1998. *WordNet: An Electronical Lexical Database*, MIT Press.

A. Kilgarriff. 1997. *I don't believe in word senses*. In Computers and the Humanities, 31 (2), 91-113.

C. Lee, G. Lee, J. Seo. 2000. Automatic WordNet mapping using Word Sense Disambiguation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 2000)*, Hong Kong.

L. Levitchi and A. Bantaş. 1992. *Dicţionar englez-român*, Bucureşti, Teora.

B. Magnini and G. Cavaglia. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000*, Athens, Greece.

J. F. Sowa. 1992. Logical Structure in the Lexicon. In J. Pustejovsky and S. Bergler (Eds.) *Lexical Semantics and Commonsense Reasoning*, LNAI 627, Springer-verlag, Berlin, 39-60.

D. Tufiş (Ed.). 2000. Special Issue on the BalkaNet Project of *Romanian Journal of Information Science and Technology*, vol. 7, no. 1-2.

P. Vossen.1998. *A Multilingual Database with Lexical Semantic Networks,* Dordrecht, Kluwer.