Extracting WordNet-like Top Concepts from Explanatory Dictionaries*

Hiram Calvo and Alexander Gelbukh

[1] Center for Computing Research, National Polytechnic Institute,
Av. Juan de Dios Bátiz s/n, esq. Av. Mendizábal, México, D.F., 07738. Mexico
hiramcalvo@gmail.com, www.gelbukh.com

Abstract: Correct interpretation of the text frequently requires knowledge of semantic categories of nouns, especially in languages with free word order. For example, in Spanish the phrases *pintó un cuadro un pintor* (lit. *painted a picture a painter*) and *pintó un pintor un cuadro* (lit. *painted a painter a picture*) mean the same: 'a painter painted a picture'; with the only way to tell the subject from the object being by knowing that *pintor* 'painter' is causal agent *cuadro* is a thing. We present a method for extracting semantic information of this kind from existing machine-readable human-oriented explanatory dictionaries. First, we extract from the dictionary an is-a hierarchy and manually mark the categories of a few top-level concepts. Then, for a given word, we follow the hierarchy upward until finding a concept whose semantic category is known. Application of this procedure to two different human-oriented Spanish dictionaries gives additional information as compared with using solely Spanish EuroWordNet. In addition, we show the results of an experiment conducted to evaluate the similarity of word classification with this method.

1. Introduction

Accurate translation implies correct interpretation of the text, which in turn requires determining the logical relations between the entities mentioned in the natural language texts, as well as the semantic roles that these entities play in the sentences they are mentioned in.

In some cases, knowledge-poor techniques involving only the general language knowledge such as word order, but not the knowledge about specific words, can work. However, determining the function of a noun phrase in a sentence cannot rely solely on word order, particularly for languages that have a rather free order of constituents, such as Spanish. For example, the following three Spanish phrases convey the same meaning 'a painter painted a picture':

(1)	*pintó un cuadro un pintor*	lit. 'painted a picture a painter'
(2)	*pintó un pintor un cuadro*	lit. 'painted a painter a picture'
(3)	*un pintor pintó un cuadro*	lit. 'a painter painted a picture'

* Work done under partial support of Mexican Government (CONACyT, SNI). A preliminary version of this work was presented at CICLing-2004 conference.

To determine the function of a noun phrase, a syntactic analyzer needs additional information, e.g. the semantic category of the noun. In (1) and (2), *pintor* 'painter' belongs to the semantic category of ***causal agents*** and thus is the subject, while *cuadro* 'picture' belongs to the category of ***things*** (inanimate objects) and thus is the object. There are two reasons for such a conclusion: First, the verb *pintar* 'to paint' subcategorizes for a causal agent subject and a generally inanimate object. Second, even without subcategorization information, in Spanish animate objects are introduced by the preposition *a* lit. 'to', so that the meaning 'a picture painted a painter', where painter is the object (should one wanted to say this), would be expressed in Spanish differently, even keeping the same word order:

(1a)	*pintó un cuadro a un pintor*	lit. 'painted a picture to a painter'
(2a)	*pintó a un pintor un cuadro*	lit. 'painted to a painter a picture'
(3a)	*a un pintor (lo) pintó un cuadro*	lit. 'to a painter (it) painted a picture'

Thus, the absence of *a* in (1) to (3) in contrast with (1a) to (3a) indicates that the animate element is the subject and not object. Only in (3) the subject can be (arguably) guessed, without knowing its animity, by the absence of *lo* in contrast to (3a); however, this *lo* is (arguably) optional. Whether we rely on subcategorization or syntactic rules, we need the semantic category of the noun, in our example, animity.

The semantic category information for nouns can be used not only to tell the subject from the object, but also to determine other functions the noun phrase can have in a sentence, such as indirect object or circumstantial complement. In Spanish a noun phrase in a sentence can have, for example, such functions as subject, object, indirect object, spatial, temporal, causal, instrumental, comitative (the one who accompanies), possessive, or opposite. Given the semantic category of a noun phrase and the preposition preceding it, hypotheses on the possible functions of the noun phrase in a sentence can be automatically formed as shown in Table 1 (in the table, Ø stands for no preposition); see Section 3 for the discussion of the categories.

Existing sources providing semantic information in a formal way usable for automatic text processing are incomplete and/or difficult to find, especially for languages other than English. This paper presents a method for acquiring semantic categories of nouns from a machine-readable human-oriented explanatory dictionary (hereafter abbreviated as HOED).

The paper is organized as follows. In Section 2, we consider related work on extracting semantic categories from explanatory dictionaries. In Section 3, we show how we obtain semantic categories useful for syntactic disambiguation from a HOED. Section 4 describes an experiment conducted to evaluate the quality of the semantic categories extracted from two HOEDs by comparing them against those obtained from Spanish EuroWordNet. Finally, Section 5 concludes the paper.

2. Related Work

The first work that pursued the construction of a taxonomy from a HOED was Amsler's [1]. He worked manually with the Merriam-Webster Pocket Dictionary. Subsequently, several studies were carried out on other dictionaries using more automatic

Table 1. Examples of functions of a Spanish noun phrase corresponding to a combination of preposition and category.

Function	Possible expressions: **preposition** {category}	Approximate English glosses
Subject	Ø {any category}	
Object	Ø {not animate}, **a** {animate}	to
indirect object	**a** {animate}, **para** {animate}	to, for
spatial-place	**a** {place}, **en** {place}, **dentro de** {place}, **sobre** {place, thing}, **bajo** {place, thing}, **ante** {place, thing}, **tras** {place, thing}, **por** {place}, **entre** {place, thing} **y** {place, thing}, **entre** {place, thing (plural)}	to, in, inside, on, below, in front of, beyond, near, between ... and, between
spatial-trajectory	**a** {place}, **de** {place}, **hacia** {place}, **hasta** {place}, **por** {place}, **para** {place}	to, from, towards, until, by, for
temporal	Ø {time}, **a** {time, numeral}, **en** {time}, **por** {time, numeral}, **para** {time, numeral}, **hasta** {time, numeral}, **hacia** {time, numeral}, **de** {time, numeral} **a** {time, numeral}, **desde** {time, numeral}, **tras** {action, state}, **entre** {time, numeral} **y** {time, numeral}, **dentro de** {time}	at, on, by, for, until, towards, from ... to, from, through, between ... and, within
causal-finality	**para** {action}	in order to
causal-reason	**por** {action}	because of
Instrumental	**con** {instrument}	with
Comitative	**con** {animate}	with
Possessive	**de** {animate}	of
Opposite	**contra** {thing, animate}	against

methods. Chodorow *et al*. [2] worked with Webster's Seventh New Collegiate Dictionary, whereas both Guthrie *et al*. [3] and Vossen [4] used the Longman Dictionary of Contemporary English (LDOCE) [5]. LDOCE's strict organization and notation makes it suitable for extracting semantic information automatically. Most dictionaries are not organized in such a precise way, so that semi-automatic methods have been used to extract semantic information from them. In particular for the Spanish VOX Dictionary, Ageno *et al*. [6] have created an environment facilitating extraction of semantic information from HOEDs. In this environment, the user has to select manually the correct hypernym sense amongst those proposed by the system.

In other fields, there are works devoted to the enrichment of WordNet with semantic information extracted from HOEDs, e.g., Montoyo *et al*. [7] and Nastase *et al*. [8].

In general, the purpose of the work done on extracting semantic information from HOEDs differs from ours in that these works attempt to extract a whole taxonomy from a HOED, while our purpose is only to determine the semantic category of a noun out of a set of predefined categories selected for the task of determining the

function(s) of a noun phrase in a sentence. As we show in the next section, this task can be done in an automated manner.

3. Acquiring Semantic Categories from a Dictionary

An explanatory dictionary (for example, [9]) gives definitions such as the following:

abeto s m 1 *Árbol* del género Abies de la familia de las pináceas, de hojas perennes, resinoso…

'fir n<oun> m<asculine> 1 *Tree* of the genus Abies of the family Pinaceae, with evergreen leaves, resinous…'

In short, our method consists in following the is-a chain formed by word definitions, until a word with a known (manually assigned) category is reached; the word in question inherits this category. For example, for the word ***abeto*** 'fir' we have:

abeto $\overset{\text{is-a}}{\rightarrow}$ *árbol* $\overset{\text{is-a}}{\rightarrow}$ *planta* $\overset{\text{is-a}}{\rightarrow}$ *ser* 'fir $\overset{\text{is-a}}{\rightarrow}$ tree $\overset{\text{is-a}}{\rightarrow}$ plant $\overset{\text{is-a}}{\rightarrow}$ being',

where *ser* 'being' has the category *life_form* assigned to it manually, thus giving this same category for the initial word ***abeto*** 'fir'.

Some heuristics are used in the process of building the is-a chain. Usually we consider the first noun in the definition as the defined word's hypernym, which in the example above is *árbol* 'tree'. To determine if the word is a noun, we use the syntactic categories found in the dictionary definitions themselves; see the mark *s* (*sustantivo* 'noun') in the example above.

To build the chains of nouns, a simple lemmatizer was used to find the definitions of the nouns that appeared in the text of the dictionary in plural. For example consider the definition of ***abarrotes:*** *mercancías* 'packings: merchandise$_{\text{plural}}$'. Here, *mercancías* was lemmatized to *mercancía* to find its definition in the same dictionary.

A complication of the process of building the such chains is that sometimes they have cycles: a word is (indirectly) defined through another word that in its turn is defined through the first one. This probably happens due to a number of reasons, such as definitions through synonyms, imperfections in the definitions, or imperfections in our generalization algorithm and heuristics. In fact, as Gelbukh and Sidorov point out in [10], cycles in the system of definitions are inevitable in any dictionary in which all words, even such general ones as *thing* or *something*, have definitions. To break the cyclic chains, some (few) words are to be chosen as top concepts, whose categories are assigned manually. The algorithm does not try to further generalize these concepts, which ends the chain. Gelbukh and Sidorov [10] give an algorithm to select a minimal set of such top concepts whose categories are to be assigned manually.

The set of categories we have chosen comprise the 25 unique beginners for WordNet nouns described in [11]. Table 2 shows these categories along with the top concepts manually selected to which they have been assigned.

Table 2. Top concepts corresponding to the semantic categories of nouns

Category	Top concepts	English glosses
activity	*acción, acto, actividad*	action, act, activity
animal	*animal*	animal
life_form	*vida, organismo, órgano*	life, organism, organ
phenomenon	*fenómeno*	phenomenon
thing	*instrumento, objeto, cosa, aparato*	instrument, object, thing, device
causal_agent	*ser, persona, humano*	being, person, human
attribute	*propiedad, cualidad, color*	property, quality, color
flora	*planta, fruto, flor*	plant, fruit, flower
cognition	*conocimiento, palabra, abstracción*	knowledge, word, abstraction
process	*proceso*	process
event	*evento, acontecimiento*	event, happening
feeling	*sentimiento, emoción*	feeling, emotion
form	*figura, forma, línea*	figure, form, line
food	*comida, comestible*	food, comestible
state	*estado*	state (condition)
grouping	*conjunto, grupo, serie*	set, group, series
substance	*sustancia, energía, líquido, fibra*	substance, energy, liquid, fiber
place	*espacio, lugar, distancia, territorio*	space, place, distance, territory
time	*tiempo, periodo*	time, period
part	*parte, miembro, extremidad*	part, member, limb
possession	*acumulación, asignación*	accumulation, assignation
motivation	*afán, deseo, aliciente, causa*	eagerness, desire, incentive, cause

4. Experimental Results

In order to evaluate the quality of the categories of words found through our procedure, we considered two HOEDs: Lara [9] and Anaya. The first dictionary (Lara) contains approximately 12,500 entries, of which 8,000 are nouns. The second dictionary (Anaya) has nearly 33,000 nouns.

We applied our method to both HOEDs and then we compared the categories found with those of Spanish EuroWordNet[1] (henceforth abbreviated as S-EWN). As in the case of HOEDs, in S-EWN the semantic categories of nouns were defined by

[1] S-EWN was jointly developed by the University of Barcelona (UB), the National University of Distance Education (UNED), and the Polytechnic University of Catalonia (UPC), Spain.

the construction of is-a chains. Table 3 shows the comparison between the is-a chains in S-EWN and in Lara for the word *abeto* 'fir'.

Table 3. Chains from S-EWN and the dictionary for the word *abeto* 'fir'

S-EWN	*abeto*_1 → *conífera*_1 → *árbol_gimnospermo*_1 → *árbol*_2 → *planta_leñosa*_1 → *planta_vascular*_1 → *flora*_1 → *forma_de_vida*_1 → *entidad*_1	'fir → conifer → gymnospermous tree → tree → ligneous plant → vascular plant → flora → life form → entity'
Lara	*abeto* → *árbol* → *planta* → *orgánico*	'fir → tree → plant → organic'

In Table 4 totals for the two HOEDs and S-EWN are presented. Base forms are the number of nouns without considering their different senses. As not every noun leads to a top concept, there are some words for which no category could be found. Particularly for S-EWN, 30 nouns lead to the main concept *entidad_1* 'entity_1' without traversing any other concept that might yield a category.

Table 4. Totals for the two HOEDs and S-EWN

	Lara	**%**	**Anaya**	**%**	**S-EWN**	**%**
Nouns	8009	100.00%	32944	100.00%	40563	100.00%
Base forms	7857	98.10%	20529	62.31%	26335	64.92%
Classified	7400	92.40%	29027	88.11%	40533	99.92%
Not classified	608	7.59%	3916	11.89%	30	0.07%
Cycles	64	0.80%	2292	6.96%	0	0.00%

We measured three aspects of similarity of the categories yielded by the three dictionaries comparing pairs of dictionaries. These aspects are:

f1c1(a,b): nouns found in both dictionaries (a and b), with matching classification,

f1c0(a,b): nouns found in both dictionaries, but the classification in the first dictionary (a) doesn't match any of the second (b), and

f0c0(a,b): nouns classified in the first dictionary (a) that are not found in the second dictionary (b).

Lara is a small dictionary, whereas Anaya and Wordnet are four or five times greater, respectively. To compensate for this difference, results are normalized considering the sum of classifications of the two dictionaries being compared. That is, if we are comparing for example Lara and Anaya, results will be divided by 7400+29027=36427. Table 5 shows the results of comparing every possible pair of dictionaries. la stands for Lara, an for Anaya, and wn for S-EWN.

Table 5. Pair-wise comparison of dictionaries.

f1c1						
a	**b**	**f1c1(a,b)**	**f1c1(b,a)**	**total classif.**	**%**[2]	
la	**an**	3427	3427	36427	18.82%	
an	**wn**	7243	7243	69171	20.94%	
la	**wn**	2830	2830	47544	11.90%	
					17.22%	← average
f1c0						
a	**b**	**f1c0(a,b)**	**f1c0(b,a)**	**total classif.**	**%**[3]	
la	**an**	2853	7172	36427	27.52%	
an	**wn**	13501	15332	69171	41.68%	
la	**wn**	3204	8686	47544	25.01%	
					31.40%	← average
f0c0						
a	**b**	**f0c0(a,b)**	**f0c0(b,a)**	**total classif.**	**%**[4]	
la	**an**	1390	18428	36427	54.40%	
an	**wn**	8283	17569	69171	37.37%	
la	**wn**	1366	28628	47544	63.09%	
					51.62%	← average

On average, **17.22%** of the nouns were classified equally amongst the three dictionaries, **31.40%** are found but their classification doesn't match, and **51.62%** are different nouns. If we consider only the nouns that are found amongst the three dictionaries (that is, 100 – 51.62% = **48.38%**), we find that **35.60%** are classified equally, and **64.91%** are classified differently. In other words, little more than a third part of the classifications matches amongst the three dictionaries in average.

5. Conclusions and Future Work

We have presented a method to extract automatically semantic categories of nouns from machine-readable human-oriented explanatory dictionaries (HOED). Using a HOED, semantic categories can be determined for nouns absent from Spanish EuroWordNet (S-EWN). However, the quality of classifications was not as good as expected. The agreement amongst the classifications yielded by three dictionaries, two of them HOEDs, and the other S-EWN, has an average of 35.60% of the total number of words classified by the three dictionaries. This is possibly due to the lack of a word sense disambiguation module, as well as the different schemas adopted by the three dictionaries.

In the future, a word sense disambiguation module should be added to the procedure of chain construction, and the heuristics used to extracting the hypernym for a

[2] ([f1c1(a,b)+f1c1(b,a)] / total_clasif) × 100%
[3] ([f1c0(a,b)+f1c0(b,a)] / total_clasif) × 100%
[4] ([f0c0(a,b)+f0c0(b,a)] / total_clasif) × 100%

word from its definition should be refined. For example, those involving 'part of', 'set of', 'member of', etc. In addition, the repertoire of the semantic categories and the set of words to which they are assigned manually are to be revised.

Finally, although the HOEDs we used are adequate for our purposes, their use for automatically adding entries to S-EWN may be hampered by the fact that the information they provide is neither sufficiently detailed nor as systematical as it is required by S-EWN, even as to the hypernym relation.

References

1. Amsler, R. A. *The structure of the Merriam-Webster Pocket Dictionary*. Ph. D. Dissertation, University of Texas, 1980.
2. Chodorow, Martin, Roy J. Byrd and George E. Heidorn, Extracting Semantic Hierarchies from a Large On-Line Dictionary. In *Proceedings of the 23rd Annual Meeting of the* ACL, pp. 299-304, 1985.
3. Guthrie, Louise, Brian Slator, Yorick Wilks, and Rebecca Bruce. Is there content in empty heads? In *Proceedings of the 13th International Conference on Computational Linguistics,* COLING90, 1990.
4. Vossen, Piek, The end of the chain: where does decomposition of lexical knowledge lead us eventually? Esprit BRA-3030 ACQUILEX WP 010. English Department, University of Amsterdam, The Netherlands, 1990
5. Proctor, Paul (Ed.). *The Longman Dictionary of Contemporary English* (LDOCE). London, 1978
6. Ageno, Alicia, Irene Castellón, M. A. Martí, Francesc Ribas, German Rigau, Horacio Rodríguez, Mariona Taulé, Felisa Verdejo. SEID: An environment for extraction of Semantic Information from on-line dictionaries. In *Proceedings of 3th conference on Applied Natural Language Processing*. Trento, Italia, 1992.
7. Montoyo, Andrés, Manuel Palomar and German Rigau. WordNet Enrichment with Classification Systems, in *Proceedings of NAACL 2001, Workshop on WordNet and Other Lexical Resources,* Pittsburgh, PA, USA, 2001
8. Nastase, Vivi, and Stan Szpakowicz. Augmenting WordNet's Structure Using LDOCE. In Alexander Gelbukh (ed): *Computational Linguistics and Intelligent Text Processing,* CICLing 2003, Lecture Notes in Computer Science 2588: 281–294, Springer-Verlag, 2003.
9. Lara, Luis Fernando. *Diccionario del español usual en México*. Digital edition. Colegio de México, Center of Linguistic and Literary Studies, 1996.
10. Gelbukh, Alexander, and Grigori Sidorov. Selección automática del vocabulario definidor en un diccionario explicativo. *Procesamiento del Lenguaje Natural* 29: 55–62, 2002.
11. Miller, George. Nouns in WordNet: a Lexical Inheritance System, *International Journal of Lexicography*, Volume 3. num. 4, pp. 245-264, 1994.