Extracting Semantic Categories of Nouns for Syntactic Disambiguation from Human-Oriented Explanatory Dictionaries*

Hiram Calvo [1] and Alexander Gelbukh [1,2]

[1] Center for Computing Research, National Polytechnic Institute,
Av. Juan de Dios Bátiz s/n, esq. Av. Mendizábal, México, D.F., 07738. México
hcalvo@sagitario.cic.ipn.mx, gelbukh@cic.ipn.mx; www.gelbukh.com

[2] Chung-Ang University, Seoul, Korea

Abstract: Syntactic disambiguation frequently requires knowledge of the 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'. The only way to tell the subject from the object is by knowing that *pintor* 'painter' is a causal agent and *cuadro* is a thing. We present a method for extracting semantic information of this kind from existing machine-readable human-oriented explanatory dictionaries. 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 classifications using this method.

1. Introduction

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 consider the following sentences: (1) *La señora llevó a la niña a la calle,* lit. 'The woman took to the girl to the street' and (2) *La señora llevó a la calle a la niña*, lit. 'The woman took to the street to the girl'. Both sentences convey the same meaning: 'The woman took the girl to the street'. In Spanish, a noun preceded by the preposition *a* 'to' has the role of direct object if it is animate, or indirect object or circumstantial complement if it is not animate. Without semantic information, a system is not able to determine the syntactic functions of *a la niña* and *a la calle* in a sentence. When information on the semantic categories of *niña* 'girl' (causal_agent) and *calle* 'street' (place) is considered, it is possible to determine automatically that *la señora* 'the woman' is the subject, *a la niña* is the direct object and *a la calle* is a circumstantial complement of place.

* Work done under partial support of Mexican Government (CONACyT, SNI), IPN (PIFI, CGEPI), and RITOS-2. The second author is currently on Sabbatical leave at Chung-Ang University.

Existing sources providing semantic information of this kind 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 human-oriented explanatory dictionaries (hereafter, HOED).

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 automatic methods. Chodorow *et al*. [2] worked with Webster's New Collegiate Dictionary, whereas both Guthrie *et al*. [3] and Vossen [4] used the Longman Dictionary of Contemporary English (LDOCE) [5]. 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 WordNet enrichment 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 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.

2. Acquiring Semantic Categories from a Dictionary

In short, our method consists in following the is-a chain formed by the nouns in 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 $\xrightarrow{\text{is-a}}$ *árbol* $\xrightarrow{\text{is-a}}$ *planta* $\xrightarrow{\text{is-a}}$ *ser* 'fir $\xrightarrow{\text{is-a}}$ tree $\xrightarrow{\text{is-a}}$ plant $\xrightarrow{\text{is-a}}$ being',

where *ser* 'being' has the category *life_form* assigned to it manually (see Table 1), thus giving this same category for the initial word *abeto* 'fir'.

A complication of the process of building 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. Gelbukh and Sidorov point out in [10] that 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 generalize further these concepts, which ends the chain.

The set of categories we have chosen comprise the 25 unique beginners for WordNet nouns described by Miller in [11]. Table 1 shows these categories along with the top concepts manually selected to which they have been assigned.

3. Experiment

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)

Table 1. Top concepts corresponding to the semantic categories of nouns

Category	Top concepts	Category	Top concepts
activity	action, act, activity	feeling	feeling, emotion
animal	animal	form	figure, form, line
life_form	life, organism, being	food	food, comestible
phenomenon	phenomenon	state	state, condition
thing	instrument, object, thing	grouping	set, group, series
causal_agent	being, person, human	substance	substance, energy, liquid, fiber
place	space, place, distance	attribute	property, quality, color
flora	plant, fruit, flower	time	time, period
cognition	knowledge, abstraction	part	part, member, limb
process	process	possession	accumulation, assignation
event	event, happening	motivation	desire, incentive, cause

Table 2. 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.**	**%**	
	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.**	**%**	
	la	**an**	1390	18428	36427	54.40%	
	an	**wn**	8283	17569	69171	37.37%	
	la	**wn**	1366	28628	47544	63.09%	
						51.62%	← average

contains approximately 8,000 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 S-EWN). As in the case of HOEDs, in S-EWN the semantic categories of nouns were defined by the construction of is-a chains.

We measured three aspects of similarity of the categories yielded by the three dictionaries comparing pairs of dictionaries: **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). Table 3 shows the results of comparing each possible pair of dictionaries. **la** stands for Lara, **an** for Anaya, and **wn** for S-EWN. The **total classif.** column shows the sum of the number of nouns classified in dictionary (a) plus those of (b). This way results can be normalized given the difference among dictionary sizes.

[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.

[2] ([f1c1(a,b)+f1c1(b,a)] / total_clasif) x 100

In average, **17.22%** of the nouns were classified equally amongst the three dictionaries, **31.40%** are found but their classification does not 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.

4. Conclusions and Future Work

Using a HOED, semantic categories can be determined for nouns absent from Spanish EuroWordNet (S-EWN). However, the agreement of classifications among the three dictionaries studied, two of them HOEDs, and the other S-EWN, was lower than expected. An average of 35.60% of the total number of words classified by the three dictionaries agrees in classification. This is possibly due to the lack of a WSD module, as well as the different definition schemes adopted by the three dictionaries.

In the future, a WSD module can be added to the procedure of chain construction, and the heuristics used to extracting the hypernym for a word from its definition, including the words chosen as top concepts, should be revised.

Finally, an evaluation of syntactic analysis using the semantic categories provided by this method is convenient to determine the degree to which the semantic categories extracted from HOEDs enhance syntactic analysis disambiguation when the noun classifications among different dictionaries vary.

References

1. R. A. Amsler, *The structure of the Merriam-Webster Pocket Dictionary*. Ph.D. Dissertation, U. of Texas, 1980.
2. M. Chodorow, R. J. Byrd and G. E. Heidorn, Extracting Semantic Hierarchies from a Large On-Line Dictionary. In *Proc. of the 23rd Meeting of the* ACL, pp. 299-304, 1985.
3. L. Guthrie, B. Slator, Y. Wilks, and R. Bruce. Is there content in empty heads? In *Proc. of the 13th Intl. Conf. on Comp. Linguistics,* COLING90, 1990.
4. P. Vossen, The end of the chain: where does decomposition of lexical knowledge lead us eventually? ACQUILEX WP 010. English Department, U. of Amsterdam, 1990.
5. P. Proctor, (Ed.). *The Longman Dictionary of Contemporary English*. London, 1978
6. A. Ageno, I. Castellón, M. A. Martí, F. Ribas, G. Rigau, H. Rodríguez, M. Taulé, F. Verdejo. SEID: An environment for extraction of Semantic Information from on-line dictionaries. In *Proc. of 3th conf. on Applied NLP.* Trento, It., 1992.
7. A. Montoyo, M. Palomar and G. Rigau. WordNet Enrichment with Classification Systems, in *Proc. of NAACL 2001,* Pittsburgh, PA, USA, 2001.
8. V. Nastase and S. Szpakowicz, Augmenting WordNet's Structure Using LDOCE. In A. Gelbukh (ed): CICLing 2003, LNCS 2588: 281–294, Springer-Verlag, 2003.
9. L. F. Lara, *Diccionario del español usual en México*. Digital edition. Colegio de México, Center of Linguistic and Literary Studies, 1996.
10. A. Gelbukh and G. Sidorov. Selección automática del vocabulario definidor en un diccionario explicativo. *Procesamiento del Lenguaje Natural* 29: 55–62, 2002.
11. G. Miller. Nouns in WordNet: a Lexical Inheritance System, *International Journal of Lexicography*, Volume 3. num. 4, pp. 245-264, 1994.