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] Department of Computer Science and Engineering, Chung-Ang University,
221 Huksuk-Dong, DongJak-Ku, Seoul, 156-756, 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, PIFI-IPN, CGEPI-IPN), Korean Government (KIPA Professorship for Visiting Faculty Positions in Korea), ITRI of Chung-Ang University, and RITOS-2. The second author is currently on Sabbatical leave at Chung-Ang University.