An Experiment on Automatic Semantic Tagging of Dictionary Senses.

German Rigau.

Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.
Pau Gargallo 5, 08028 Barcelona. Spain. g.rigau@lsi.upc.es

Abstract.

We present a methodology to enrich a dictionary sense with semantic tags. The method relies on the use of a wide-coverage noun taxonomy and the notion of conceptual distance among concepts. Some experimental results about the performance of the method are provided.

Keywords: Lexicon, Electronic Dictionaries, Word Sense Disambiguation, Machine Translation.

1 Introduction

Lexicon is generally considered as a major 'bottleneck' in NLP. The use of machine-readable versions of conventional dictionaries (MRDs) in the acquisition of lexical knowledge has been widely extended because they provide substantial quantities of lexical information. Although some dictionaries, e.g. LDOCE, are strictly organised and codified, allowing the use of very precise procedures for extracting semantic information (e.g. disambiguated taxonomies [Copestake 1990], [Bruce & Guthrie 92]), this is not the case of most dictionaries. In our experimentwe have used an MRD version of VOX [Biblograf 87] (about 160.000 senses). There are several features in VOX that prevent the application of the approaches used in LDOCE:

- the vocabulary in the definition field is not restricted.
- there are no semantic codes.
- there are no box-codes.
- the explicit semantic information does not appear systematically.

Semi-automatic procedures for extracting disambiguated taxonomies have been applied to this kind of less organised dictionaries. Within Acquilex Project[1] we developed an environment, SEISD [Ageno et al. 92] in order to extract implicit semantic information from MRDs. One of the main tasks carried out by the SEISD environment is the construction of taxonomies from the VOX dictionary definitions. In SEISD the user selects the correct hyperonym sense amongst those proposed by the system using a top-down strategy. This procedure cannot achieve the desired level of efficiency and semantic coverage. For example, we find that a great amount of dictionary senses which belong to the semantic class <food> (as for instance *queso* ~ <cheese>) do not appear in the semi-automatically generated taxonomy of <food>. Such problems would be solved if dictionary senses bore explicit semantic tags. In the taxonomy creation task, for instance, the explicit semantic tags assigned to each dictionary sense could be a helpful information to automatically discriminate the correct hypernym. Furthermore, for each semantic tag taken as a top of a taxonomy, those dictionary senses not included in the taxonomy could be detected and proposed as possible new tops.

Following this introduction, section 2 presents the semantic knowledge sources used by the system. Section 3 is devoted to the definition of the Conceptual Distance measure applied. Section 4 shows the methodology used in the experiment. In section 5, we show the performance of the system for a subset of the dictionary and finally in the last section some conclusions are stated.

2 Using massive Lexical Knowledge Sources

Currently, the combined use of several linguistic resources plays a central role in the lexical research community. With monolingual resources, [Yarowsky 92] combines the use of the Grolier encyclopaedia as a training corpus with

[1] Acquilex ESPRIT BRA-3030.

the categories of the Roget's International Thesaurus to create a statistical model for the word sense disambiguation problem; [Ribas 94] extracts selectional restrictions from the phrasally analyzed Penn Treebank Corpus using WordNet [Miller 90]. With bilingual resources, [Ageno et al. 94] have developed an environment to generate lexical cross-relations between English and Spanish; [Knight & Luk 94] construct a large-scale ontology to support semantic processing in the Pangloss knowledge-based translation system, merging several multilingual resources.

We have developed and implemented a method in order to select a semantic tag for a particular dictionary sense automatically, using several massive semantic knowledge sources.

The system needs to know how words are clustered in semantic classes, and how semantic classes are hierarchically organised. For this purpose, we have used a broad semantic taxonomy for English, WordNet [Miller et al. 90] (about 62,000 synonym sets or synsets).

The problem of connecting each Spanish word from one dictionary sense to a set of synsets of WordNet is solved by means of another wide knowledge source, the VOX-Harrap's Spanish/English bilingual dictionary [Biblograf 92] (about 32,000 senses). For example, the entry *masa* contains the following information:

> **masa** **1** *f* mass **2** *f* CULIN dough. **3***f* (mortero) mortar. **4** *f* ELEC ground. **5** *f* (de cosas) volume. **6** *f* (multitud) crowd of people. **7** *f* AM (pastel) cake.

where the headword and the sense numbers are in bold, the part-of-speech code in italic , one of the 43 predefined semantic fields in capital letters, in brackets an indicator (usually a related Spanish word useful for the readers to disambiguate the correct sense in English) and in plain text the translations in English.

The use of the bilingual dictionary to link the Spanish words of a VOX dictionary sense with the semantic classes represented in WordNet obviously produces an explosion of ambiguity. For instance, while the word *masa* in the VOX dictionary has 13 noun senses, the bilingual entry with 7 translations can be connected (without disambiguation using the semantic field or the indicator context) to 29 synsets. Note that some of these connections can be wrong; for instance, the religious sense of mass is not a correct translation of the Spanish word *masa*.

3 Conceptual Distance

Thus, because of the lack of a broad semantic taxonomy in Spanish and the possibility of noise produced by the translation process, we need a fine grained measure of semantic distance [Miller & Teibel 91] among concepts in the hierarchical semantic net of WordNet to discover the more reliable lexical cohesion of a given set of words in Spanish.

Several measures of relatedness among words based in the coocurrence of them in a text have been described; mutual information, t-test, etc. [Church et al. 91], in Context Space the cosine function [Schütze 92], conditional probability, etc. [Wilks et al. 93]; but less attention has been paid to measures of relatedness among concepts in a structured hierarchical net.

Conceptual Distance tries to provide a basis for determining closeness in meaning among words in a structured hierarchical net. The Conceptual Distance among concepts in a hierarchical semantic net is defined in [Rada et al. 89] as the length of the shortest path that connects them. Besides applying conceptual distance in a medical bibliographic retrieval system and merging several semantic nets, they demonstrate that Conceptual Distance is a metric. In a similar approach, [Sussna 93] defines a complex weighting mechanism to balance the heterogeneous nature of the relations among synsets in WordNet. Following these ideas, the Conceptual Distance described in [Agirre et al. 94] is

$$CD(c_1,c_2) = \sum_{i \in shortestpath(c_1,c_2)} \frac{1}{depth(c_i)}$$

expresses that the Conceptual Distance between two concepts depends on the length of the shortest path that connects them and the specificity of the concepts in the path. That is to say, the lower the concepts are in a hierarchy, the closer they seem to be.

4 The Method of Tagging

Based on the ideas described above, we have implemented a system to label VOX dictionary senses with semantic tags.

```
<entity >
    <object, inanimate object, physical object>
        <substance, matter>
            <food, nutrient>
                <aliment, nourishment, nutrient, nutriment, sustenance,  victuals>
                    <dish>
                        <patty, cake>                     (magdalena, masa)
                    <baked good, baked goods>
                        <cake>                            (magdalena, masa)
                            <sponge cake>                 (bizcocho)
                        <bread, breadstuff, staff of life>
                            <bun, roll>                   (magdalena, bollo)
                <foodstuff>
                    <concoction, mixture>
                        <dough>                           (masa)
```

Figure 1: Partial view of magdalena_X_I_1 lattice.

To illustrate the process consider as an example the VOX monolingual dictionary definition:

> **magdalena_X_I_1** , bollo pequeño de forma preferentemente redonda y bombeada, hecho de masa de bizcocho.
>
> (≈ little round-shaped bun made of sponge cake)

1) Selection of those words which appear as nouns in the bilingual dictionary.

In our case, the program finds {magdalena, bollo, pequeño, forma, redonda, hecho, masa, bizcocho} as nouns in the bilingual. Note that no part-of-speech disambiguation has been performed and only {magdalena, bollo, forma, masa, bizcocho} are nouns in this definition.

2) Taking the translations of each previously selected word, the program extracts all the possible hypernym synsets in WordNet. All the concepts represented by these synonym sets are placed in a lattice.

For *magdalena_X_I_1,* a partial view of the lattice built by the program is shown in figure 1. The indentation represents the hypo/hypernym relation between synsets (e.g. the concept <cake> is an hyponym of <baked good, baked goods>). This partial lattice is the result of the following translations from the bilingual dictionary:

magdalena_S_1: bun, cake.
bollo_S_1: bun, roll.
masa_S_2: dough.
masa_S_7: cake.
bizcocho_S_1: sponge cake.

3) The formula applied in order to compute the conceptual distance for each synset is

$$CD(c_i) = \sum_{j \in hyponyms\_in\_the\_lattice(c_i)} \frac{1}{depth(c_j)}$$

The hyponyms of a concept (or synonym set in the lattice) are a subset of those that appear in WordNet. Each concept is associated to the number of different Spanish words the synset is related to. The program separates the synsets in the lattice by the number of different words in Spanish. For each number (or level) it selects the synset with minor conceptual distance.

4) Each synset in WordNet is assigned a file number indicating to which semantic field it is related. There are 44 files or semantic fields in WordNet. For example all the <food> nouns come from file 13. We consider this number as the semantic tag of the synset. In order to label a monolingual VOX dictionary sense we take either the most voted semantic tag among the different levels or the one with a minor distance. The execution for the sense *magdalena_X_I_1* is presented below:

Word magdalena in Monolingual has 1 senses.

```
magdalena_1  (magdalena bollo pequeño forma
preferentemente redonda bombeada hecho masa
bizcocho).
• magdalena: 1 Bilingual senses, 2 synsets.
((bun cake))
• bollo: 3 Bilingual senses, 7 synsets.
((bun roll)(dent)(bump))
• pequeño: 1 Bilingual senses, 12 synsets.
((child))
• forma: 4 Bilingual senses, 22 synsets.
((way)(form)(manners social_conventions)
(curves))
• redonda: 2 Bilingual senses, 4 synsets.
((region)(semibreve))
• hecho: 2 Bilingual senses, 3 synsets.
((fact)(event  incident))
• masa: 7 Bilingual senses, 29 synsets.
((mass)(dough)(mortar)(ground)(volume)
(crowd)(cake))
• bizcocho: 1 Bilingual senses, 1 synsets.
((sponge_cake))
------Words : 2
Distance : 0
Synset : 3735786, (bun roll)
Words: (bollo magdalena)
Tag :  13, FOOD.
------Words : 3
Distance : 0.089285714373456
Synset : 3697820, (cake)
Words : (magdalena masa bizcocho)
Tag: 13, FOOD.
------Words : 4
Distance : 0.166666666666667
Synset  : 3693472, (baked_good baked_goods)
Words : (magdalena bollo masa bizcocho)
Tag: 13, FOOD.
------Words : 5
Distance  : 1.25
Synset : 7642, (object inanimate_object
physical_object)
Words : (magdalena bollo forma masa bizcocho)
Tag: 17, OBJECT.
------Words: 6
Distance: 1.7
Synset: 7642, (object inanimate_object
physical_object)
Words : (magdalena bollo forma redonda masa
bizcocho)
Tag: 17, OBJECT.
```

In this case, the most voted tag is 13, FOOD.

5 Experimental Results

Our experiment has been performed within the <food> domain. The set of 191 dictionary senses containing the word *masa* in its definition field in the monolingual VOX dictionary has been selected. Note that this word might not be the genus term of the dictionary sense. In order to test the classification process we tagged by hand the 191 senses, finding 50 dictionary senses to belong to the <food> domain. The results are the following:

H\P	<food>	¬ <food>	?	total
<food>	42	6	2	50
¬ <food>	2	119	20	141
				191

Where H stands for human and P for program. Note that 22 dictionary senses (about 10%) have not been tagged by the program. Therefore, the recall is 84% and the precision 95%.

6 Conclusions

This paper has described an approach to automatic semantic tagging of dictionary senses in Spanish using massive lexical knowledge sources. This method could also be used directly to tag or check (previously tagged) dictionary senses in English with even better results, because of the lower degree of ambiguity produced by the translation. Better results could be obtained with a larger bilingual dictionary (as the lack of a lot of headword translations in the current bilingual dictionary is obvious). This method disambiguates the appropriate sense with very few words and without any training process. Also, the results are surprisingly good taking into account the lack of any syntactic analysis.

Acknowledgement The author was supported by a grant of the Ministerio de Educación y Ciencia, 92-BOE-16392.

References

[Ageno et al. 92] Ageno A., Castellón I., Martí M.A., Ribas F., Rigau G., Rodríguez H., Taulé M., Verdejo F., *SEISD: An environment for extraction of Semantic Information from on-line dictionaries.* Proceedings of 3th Conference on Applied Natural Language Processing. Trento. Italia. 1992.

[Ageno et al. 94] Ageno A., Castellón I., Ribas F., Rigau G., Rodríguez H., Samiotou A., *TGE: Tlink Generation Environment.* In Proceedings of the 16th International Conference on Computational Linguistics (Coling'94). Kyoto, Japan.

[Agirre et al. 94] Agirre E., Arregi X., Artola X., Díaz de Ilarraza A. and Sarasola K., *Conceptual Distance and Automatic Spelling Correction*, submitted to the workshop on Computational Linguistics for Speech and Handwriting Recognition, Leeds, 1994.

[Biblograf 87] Diccionario General Ilustrado de la Lengua Española VOX. Ed. Biblograf S.A. Barcelona, 1987.

[Biblograf 92] VOX Harrap´s Diccionario esencial Inglés-Español, Español-Inglés. Segunda Edición. Biblograf S.A. Barcelona, 1992.

[Bruce & Guthrie 92] Bruce R. and Guthrie L., *Genus disambiguation: A study in weighted preference.* In Proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France.

[Church et al. 91]Church K., Gale W., Hanks P. and Hindle D., *Using Stadistics in Lexical Analisys*, in Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Zernik U. Ed. Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Copestake 90] Copestake A., *An approach to building the hierarchical element of a lexical kwoledge base from a machine readble dictionary.* in Proceedings of the Fisrt Intenational Workshop Inheritance in NLP (Tliburg, the Netherlands), 19-29, 1990.

[Knight & Luk 94] Knight K. and Luk S., *Building a Large-Scale Knowledge Base for Machine Translation.* In Proceedings of the AAAI'94.

[Miller 90] Miller G., *Five papers on WordNet*, Special Issue of International Journal of Lexicogrphy 3(4). 1990.

[Miller & Teibel 91] Miller G. and Teibel D., *A proposal for Lexical Disambiguation*, in Proceedings of DARPA Workshop on Speech and Natural Language, 395-399, Pacific Grave, California, February, 1991

[Rada et al. 89] Rada R., Mili H., Bicknell E. and Blettner M., *Development an Applicationof a Metric on Semantic Nets*, in IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, 17-30. 1989.

[Ribas 94] Ribas F., *An Experiment on Learning Appropriate Selectional Restrictions from Parsed Corpus.* In Proceedings of the 16th International Conference on Computational Linguistics (Coling'94). Kyoto, Japan.

[Schütze 92] Schütze H., *Context Space*, in Workshop Notes of Fall Session of Statistically-Based Natural Language Processing Techniques, AAAI'92.

[Sussna 93] Sussna M., *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network*, in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia USA. 1993.

[Wilks et al. 93] Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B., *Providing Machine Tractablle Dictionary Tools*, in Semantics and the Lexicon (Pustejowsky J. ed.), 341-401, 1993.

[Yarowsky 92] Yarowsky, D. *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.* In Proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France.