NLP Use of Multiple Sources of Information Available In A Spanish Machine-Readable Dictionary

Marisa Jiménez
Microsoft Research
One Microsoft Way
WA 98052
USA
marialj@microsoft.com

Abstract

This paper discusses how multiple sources of information available in a Spanish machine-readable dictionary (MRD) have provided valuable linguistic information to our Spanish NLP system. In particular we discuss how we have used the definitions, codes and headwords of the MRD to obtain morphological and semantic information, which is used by the Spanish morphology and generation modules in our system. In the first part of the paper we describe the mapping process of morphological information contained in the MRD codes, definitions, and headwords into the inflectional paradigms used by our Spanish morphological analyzer. In the second part we discuss how we used the MRD definitions to create links between Spanish noun/verb pairs; these links are used by our Spanish generation component.

Introduction

In the last fifteen years extensive research has been published on several methods to extract linguistic information from machine-readable dictionaries (MRDs) [Amsler 1980; Markowitz et al. 1986; Jensen and Binot 1987; Byrd et al. 1988; Boguraev et al. 1989; Calzolari 1990; Montemagni 1991; Dolan et al. 1993; Pentheroudakis and Vanderwende 1993, among others]. Although MRDs have long been acknowledged as a complete source of linguistic information, most attention has been given to information contained in their definitions. Most work reported on the use of Spanish MRDs has also focused mainly on information extraction out of dictionary definitions [Castellón and Martí 1990; Martí et al. 1990; Ageno et al. 1991; Castellón et al. 1991; Martí et al. 1998; Gonzalo et al. 1999, among others].

Our Spanish NLP system has greatly benefited from the use of linguistic information obtained from multiple sources of a popular Spanish MRD, the VOX dictionary. We have used the dictionary definitions, the headword, and almost every field and code of this dictionary to obtain morphological, semantic, syntactic and even dialectal information. In this paper we will discuss how we used information contained in the VOX dictionary codes, definitions and headwords to obtain morphological and semantic information that has been used by our Spanish morphological and generation modules.

Use of MRDs at Microsoft Research

Since its formation, the Natural Language Group at Microsoft Research has been using MRDs. Our approach has always been to use as many sources of information from a MRD as possible to extract linguistic information. The information that we have extracted from

these dictionaries ranges from lemmas, senses, syntactic and semantic features, part of speech probabilities, definitions, and morphological information [Dolan et al. 1993; Pentheroudakis and Vanderwende 1993; Coughlin 1996, among others].

Our Spanish system has been using an electronic version of the VOX dictionary as a source of linguistic information. The VOX dictionary was one of the first MRDs to be available in Spanish, and has been used as a lexical resource by several projects, among them Acquilex I and II and the Spanish EuroWordnet. Most work reported on these projects used the VOX definitions as the main source of information and mostly with the purpose of creating taxonomies and knowledge bases [Castellón and Martí 1990; Ageno et al. 1991; Gonzalo et al. 1999, among others]. On the other hand, we were interested in using all sources of information available in VOX to obtain linguistic information that could be used by our Spanish NLP system. Our plan was to automatically convert the VOX to a database that could be used by our system, and to obtain as much linguistic information as possible in the process. During this dictionary conversion we used the dictionary definitions, headwords, and all dictionary codes and fields to extract morphological, semantic, syntactic and even dialectal information.

When we purchased the electronic version of the VOX definition dictionary, we already had a Spanish monolingual dictionary that contained around 95,000 entries. After the conversion process, we merged all the entries extracted from VOX with that of our existing Spanish dictionary; the result is our current Spanish dictionary, which contains 140,664 entries. Merging both dictionaries had many challenges, among them converting plural noun entries and multiple words entries in VOX into valid dictionary entries in our system.

Extraction of inflectional mappings from multiple sources in VOX

Prior to acquiring the VOX dictionary, all nouns, adjectives and verbs in our existing monolingual dictionary contained automatically coded inflectional information that was used by our finite-state morphological analyzer. We had 17 inflectional paradigms for nouns, 14 for adjectives, and around 100 verbal paradigms. An example of an inflectional paradigm for nouns would be *Noun-casa*. All entries coded with this paradigm would be nouns, either masculine or feminine, that form the plural by adding the ending –s. Another example for adjectives would be *Adj-trabajador*, used for adjectives ending in –r that take –a as the feminine singular ending, -es as the masculine plural ending, and –as as the feminine plural ending.

The VOX dictionary did not contain much explicit reference to how an entry inflects, but it was full of implicit morphological information scattered throughout different parts of the dictionary entries. Nouns and adjectives in VOX were full of morphological information, contrary to verbs, where most inflectional information was provided in an appendix table only available in the printed version of the dictionary. There were around 28,000 nouns and adjectives in VOX that were not already part of our existing dictionary. As these entries were rich in implicit inflectional information, we thought that it was worth the effort to map that information into the inflectional paradigms used by our system.
Other work on mapping morphological information from MRDs has been reported in the literature. Tuells [1998] describes the semiautomatic mapping of inflectional information

from a Catalan MRDs to a morphological analyzer, but his work involved the manual coding of all plural entries. Pentheroudakis and Vanderwende [1993] describe a fully automated way to extract morphological information from the LDOCE English dictionary; the focus of their paper, though, is on derivational morphology. Our approach did not involve manual coding, and focused only on inflectional morphology.

We used three different dictionary sources of information to create mappings to our inflectional paradigms. The first one was the dictionary's *Users Guide,* which provided 20 different morphological codes for nouns and adjectives. Most of these codes refer to the gender and part of speech of the word (e.g. *f* for feminine nouns and *m* for masculine nouns), and they were pretty straightforward. As dictionary codes alone were not enough to guess the inflectional class of nouns and adjectives, we used two other sources of morphological information: the dictionary definitions and the headword field.

The headword field contained useful information about morphological alternations in the entry. If a noun or adjective contained masculine/feminine alternations, VOX made this alternation explicit. Some examples of the masculine/feminine alternations used in VOX are *administrador, -ra, actor,-riz, alcalde -esa,* among others. There were close to 200 masculine/feminine alternations in the VOX entries for nouns and adjectives, which we mapped to the 11 masculine/feminine alternations used by our Spanish morphology. The third source of information that we used was definitions.

In figure 1 we provide an example of a dictionary entry where we used these three sources of information, which we have highlighted in bold:

ENTR = administrador, -ra
ACEP = 1
CATG = **adj.-s.**
SIGN = Que administra. -
ACEP = 2
CATG = **m.,**
CATG = **f.**
SIGN = **Persona** que administra bienes ajenos.

Figure 1: VOX entry for *administrador*

We used the definitions, the morphological alternation *r, -ra* in the headword field, and the dictionary codes *f, m,* and *adj.-s* to identify this type of entries as belonging to a close inflectional class of masculine and feminine human nouns and adjectives (e.g.: *trabajador* 'worker', *director* 'director', *profesor* 'teacher').

In figure 2 we provide another example of a VOX entry to illustrate another type of mapping. In this case, we used the fact that the headword ended in *–a,* together with the code *com.* and the dictionary definition to determine that this type of entries belonged to a class of human nouns that have the same masculine and feminine form (e.g.: *artista* 'artist', *comunista* 'comunist'):

ENTR = artista
ACEP = 1
CATG = **com.**
SIGN = **Persona** que ejercita alguna arte bella.

Figure 2: VOX entry for *artista*

Out of all the 28,351 new nouns and adjectives unique to VOX, we were able to find morphological mappings for 80% of those entries without any manual coding. The other 20% were mainly proper nouns and multiple word entries. As for accuracy, we manually reviewed a sampling of 300 entries, in which the morphological mappings that we extracted were 95% accurate.

Extraction of links between Spanish noun/verb pairs from VOX definitions

It has long been acknowledged that MRD definitions are full of semantic patterns that can be very useful for natural language processing. Early work such as Markowitz et al. [1989] already brought attention to significant semantic patterns found in dictionary definitions, such as *part, cause, agent,* etc. The VOX dictionary definitions have been used as a source of semantic relations for NLP purposes. As pointed out by Gonzalo et al. [1999], most projects on extracting semantic relations from Spanish MRDs definitions have focused on words with the same part of speech, while relations between words with a different part of speech, such as that between nouns and verbs, have received less attention.

One semantic relation between noun/verb pairs that can be found in the VOX definitions is the *action/effect* relation. Some of these noun/verb pairs are also morphologically derived from each other. An example would be *enamoramiento/enamorar* 'love/to love', where the noun *enamoramiento* is morphologically derived from the verb *enamorar* through the use of the productive suffix *–miento*. Not all noun/verbs pairs with a *cause/effect* relation are morphologically derived, though. For example, pairs such *abertura/abrir* 'opening/to open' may have been morphologically related in a previous stage of the language, but nowadays there is no productive Spanish morphological rule that links them. Gonzalo et al. [1999] discusses a semiautomatic way to extract semantic relations between Spanish noun/verb pairs, but their work only concentrates on those that are morphologically derived.

All Spanish noun/verb pairs that are semantically related via an *action/effect* relation, and also morphologically derived from each other (e.g.: *evaporación/evaporar* 'evaporation/to evaporate') already had links to each other in our Spanish monolingual dictionary. These links had been extracted automatically along the lines described in Pentheroudakis and Vanderwende (1993). The other subclass of these noun/verb pairs, those that were **not** morphologically derived from each other, were not linked yet. We decided to use the VOX dictionary definitions with the purpose of creating links between noun/verb pairs that belong to this second class.
We used a pattern matching technique to go automatically over all the VOX noun definitions where an *action* or *effect* relation between the head noun and a verb was expressed (excluding those that were morphologically derived because they were already linked in our

dictionary), and extracted the corresponding verb in the definition. In figure 3 we exemplify this process by providing the VOX definition for one of these entries, the noun *juicio* 'judgement':

ENTR = juicio.
SIGN = **Acción** de juzgar.

Figure 3: definition for noun *juicio* in the VOX dictionary

We looked in the definition for an exact match of the words *acción/efecto* 'action/effect' followed by the preposition *de* 'of' and a verb. In order to make sure that the verb in the definition was really related to the head noun, we required that the noun and the verb had similar-enough lemmas. For example, in the case illustrated in figure 3, the noun ***juicio*** 'judgement' and the verb ***juzgar*** 'to open' share the first two letters of their lemmas. By making this strict requirement, we avoided picking cases where the noun and the verb in the VOX definition were not at all related, such as in the case of the definition for *acceso* 'access', illustrated in figure 4, where *acceso* is not related to the verb *llegar* 'to arrive' in the definition:

ENTR = **acceso.**
SIGN = **Acción** de **llegar** o acercarse.

Figure 4: definition for noun *acceso* in the VOX dictionary

After extracting a list of noun/verb pairs, we manually reviewed it for possible errors, and used the reviewed list to automatically create links between these pairs in our Spanish dictionary. To do so we used a similar approach to that described in Pentheroudakis and Vanderwende [1993] for clearly morphological related pairs. In figure 5 we provide an example of one of these links; it is in the entry for *juicio* 'judgement' in our monolingual dictionary. We keep the link to the verb *juzgar* 'to judge' inside the attribute **Bases**, and the feature **N_lex** is used to express that this noun/verb pair is semantically linked, but not morphologically derived.

```
Segtype  NOUN
Lemma    "juicio"
Bases
        {Lemma    "juzgar"
         Bits     N_lex
         Cat      Verb } }
```

Figure 5: links to *juzgar* in the entry for *juicio*

Using our method, we were able to extract 1,326 links between Spanish noun/verb pairs with an *action/effect* relation. Out of the almost 7,000 noun/verb pairs in VOX where this kind of semantic relation is present, the links we extracted are a small percentage (18%), but because we used a strict pattern matching technique we got almost 100% accuracy. In the future we

are planning to extend the coverage of our technique. Finally, we would like to point out that these dictionary links are actively used by our Spanish generation component, which is currently part of an English-Spanish machine translation application that is being developed by our group. As giving the details of how our Spanish generation uses these links would fall beyond the scope of this paper, we refer you to Aikawa et al. [2001] for details.

Conclusions

In this paper we have argued that using different sources of information available in the VOX MRD entries has provided valuable linguistic information to our Spanish NLP system. In particular, we have discussed how we used dictionary codes, headwords, and definitions to obtain morphological and semantic information, which has been very valuable for our Spanish morphology and generation modules.

Acknowledgements

We would like to express our gratitude to Lucy Vanderwende, Deborah Coughlin, and Maite Melero for their help during the development of this paper. We also thank Gary Kacmarcik for his help during the process of automatically extracting the information in the MRD.

References

[Ageno et al. 1991] Ageno, A., S. Cardoze, I. Castellón, M.A. Martí, G. Rigau, H. Rodríguez, M. Taule, M.F. Verdejo, 1991. An Environment for the Management and Extraction of Taxonomies from On-line Dictionaries, in: *ESPRIT BRA-3030 ACQUILEX WP No. 020.*

[Aikawa et al. 2001] Aikawa, T., M. Melero, L. Schwartz, and A. Wu, 2001. Sentence Generation for Multilingual Machine Translation, in: *Proceedings of the MT Summit VIII,* Santiago de Compostela, Spain.

[Amsler 1980] Amsler, R.A, 1980. *The Structure of the Merriam-Webster Pocket Dictionary,* Doctoral Dissertation, University of Texas, Austin.

[Boguraev et al. 1989] Boguraev, B., R. Byrd, J. Klavans, and M. Neff, 1989. From Structural Analysis of Lexical Resources to Semantics in a Lexical KnowledgeBase, in: *Proceedings of the Workshop on Lexical Acquisition, International Joint Conference on Artificial Intelligence (IJCAI),* Detroit, Michigan.

[Byrd et al. 1988] Byrd, Roy J, Nicoletta Calzolari, Martin S. Chodorow, Judith L. Klavans, Mary S. Neff, Omneya Rizk., 1988. Tools and Methods for Computational Lexicology, in: *Computational Linguistics: Special Issue on the Lexicon.*

[Calzolari 1990] Calzolari, N., 1990. Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge Base, in: U. Zernik (ed.), *Lexical Acquisition: Exploting Online Resources to Build a Lexicon,* Laurence Erlbaum.

[Castellón and Martí 1990] Castellón, I. and M.A. Martí, 1990. Gramática para el Análisis del Diccionario Vox, in: *Proceedings of the 6th Annual Meeting of SEPLN,* San Sebastián, Spain.

[Castellón et al. 1991] Castellón, I., G. Rigau, H. Rodríguez, M.A. Martí, M.F. Verdejo, 1991. Loading the MRD into the LDB, in: Characteristics of the Vox Dictionary. *ESPRIT BRA-3030 ACQUILEX WP No. 019.*

[Coughlin 1996] Coughlin, D. A., 1996. Deriving Part-of-Speech Probabilities from a Machine-Readable Dictionary, in: *Proceedings of the Second International Conference on New Methods in Natural Language Processing,* Ankara, Turkey. pp. 37-44.

[Dolan et al. 1993] Dolan, W., L. Vanderwende and S.D. Richardson, 1993. Automatically Deriving Structured Knowledge Bases from Online Dictionaries, in: *Proceedings of the First Conference of the Pacific Association for Computational Linguistics,* 5-14.

[Dorr et al. 1998] Dorr, B., M.A. Martí, I. Castellón, 1998. Evaluation of EuroWordNet and LCS-Based Lexical Resources for Machine Translation, in: *First International Conference on Language Resources & Evaluation*, Granada, Spain, pp. 393-99.

[Gonzalo et al. 1999] Gonzalo, J., F. Verdejo and I. Chugur, 1999. Using EuroWordNet in a Concept-Based Approach to Cross-Language Text Retrieval, in: *Applied Artificial Intelligence* 13(7), *Special Issue on Multilinguality in the Software Industry: the AI contribution.*

[Jensen and Binot 1987] Jensen, K. and J.-L. Binot, 1987. A Semantic Expert Using an Online Standard Dictionary. Reprinted as Chapter 11 in *Natural Language Processing: The PLNLP Approach*, ed. K. Jensen, G.E. Heidorn and S.D. Richardson, Kluwer 1992.

[Markowitz et al. 1986] Markowitz, J., T. Ahlswede and M. Evens, 1986. Semantically Significant Patterns in Dictionary Definitions, in: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 112-119.

[Martí et al. 1990] Martí, M.A., I. Castellón, M. Taulé, 1990. Algunas Puntualizaciones sobre Lingüística Computacional, in: *Limits*, n. 8, Barcelona, Spain.

[Martí et al. 1998] Martí, M.A., I. Castellón, A. Fernández, 1998. Extracción de Información de Corpus Diccionariales, in: *Novàtica*, Junio 98.

[Montemagni 1998] Montemagni, S., 1991. Specializing a Broad Coverage Grammar for the Analysis of Dictionary Definitions, in: *ESPRIT BRA-3030 ACQUILEX WP No.023.*

[Pentheroudakis and Vanderwende 1993] Pentheroudakis, J.P., and L. Vanderwende, 1993. Automatically Identifying Morphological Relations in Machine-Readable Dictionaries, in: *Proceedings of the Ninth Annual Conference of the UW Center for the NewOED and Text Research*, pp 114-131.

[Tuells 1998] Tuells, T., 1998. Constructing and Updating the Lexicon of a Two-Level Morphological Analyzer from a Machine-Readable Dictionary, in: *First International Conference on Language Resources & Evaluation*, Granada, Spain, pp. 847-52.