BDCon: A Spanish knowledge database

Adán Cassán, Sergi Cervell, Mireia Colom, Rafael Marín,
Josep M. Merenciano, Gema Pérez, Lluís Valentín

Department of Computational Linguistics, Planeta Actimedia
C/ Aribau, 198, 5ª planta, 08036, Barcelona, Spain
{rmarin, jmmerenciano, gperez, lvalentin}@planeta-actimedia.es

Abstract

In this paper we describe a knowledge base that has been built using the partially structured knowledge from encyclopaedias. The BDCon (from the Spanish: *Base de Datos de Conocimiento*) is a general ontology built around an extended Spanish lexicon extracted from two encyclopaedias and a cartographic database. It is composed by a number of interconnected knowledge structures, each of them covering a different aspect of world knowledge. The purpose of the BDCon is to classify the contents of various reference works publishing companies, as well as to support a set of advanced linguistic tools.

1. Introduction

The BDCon (from the Spanish *Base de Datos de Conocimiento*) is a general ontology, or knowledge base, created using two complete Spanish encyclopaedias and a general cartographic database as a knowledge source. It is one of the results of a project to create an all-purpose contents bank for the Spanish encyclopaedia and reference works publishing company "Grupo Planeta".

The main objective of this system is to enable the reuse of the contents that are traditionally part of the company's patrimony; to organize these contents so that they are readily available for the editors, and also to support a number of Natural Language Processing (NLP) applications within the fields of document and information retrieval, error checking, automatic SGML or XML mark-up, information extraction, automatic hypertext mark-up, question answering systems, and so on (Gaizauskas and Wilks, 1998; Mandala *et al.*, 1998).

2. A Contents Bank for the management of encyclopedias and reference works

The system that has been created to cover these needs is called "El Banco de Contenidos Planeta" ("Planeta Contents Bank"), and will, when totally operative, be able to manage the entire collection of reference works of all the companies in the Planeta Communication Group (Trotzig, 2000).

The Contents Bank has two main parts: the document collection and the Semantic Classification system. The latter is composed of the BDCon and the linguistic tools of the system.

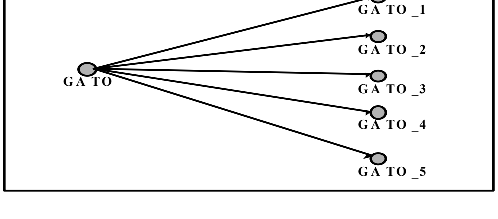


Figure 1. Relation between the part-of-speech tagger dictionary and the word senses in the BDCon.

The most basic of those is a part-of-speech tagger that can be considered the bottom layer of the linguistic applications (Wilks, 1998).

As can be seen in Figure 1, the tagger also acts as an interface between the knowledge base and the written documents of the Contents Bank since the lemmas in it are associated with their corresponding entries in the knowledge base lexicon.

2.1. The knowledge base (BDCon)

The BDCon is a general ontology organized around an extended and constantly growing Spanish lexicon extracted and compiled from two encyclopaedias and a cartographic database. Currently it contains about 115,000 common nouns (word senses), 30,000 adjectives, 12,000 verbs and more than 500,000 entity names.

The ontology is composed of a series of interconnected knowledge structures each of them covering a specific aspect of the concepts involved.

The general idea behind the use of encyclopaedias as a knowledge source is that the knowledge found there is already quite structured, i.e. the definitions and articles associated with the entry terms already contain much of the information necessary for the knowledge base. This, in turn, allows for a certain degree of automatization of the knowledge base construction process.

The purpose at the beginning of the project was to customize the Spanish version of EuroWordnet (Gonzalo *et al.*, 1998), and use it as the basis for the knowledge base, associating each node or synset (set of synonyms) (Miller, 1990) with its corresponding word sense in our lexical thesaurus. A detailed explanation of why this was done only partially, and of the problems involved, are outside the scope of this paper. Nor is a detailed description of the set of linguistic tools that have been implemented for use with the BDCon. Some mentions of these factors will, however, be included in the description of the structure of the BDCon, and how it was constructed.

The BDCon has two major functions within the Contents Bank. The first is to work as an advanced information retrieval system to enable publishers to find the documents they need with the maximum precision and recall, whether they be textual, photographic, audiovisual or of any other sort. The other function is to work as a general knowledge based system supporting the linguistic tools that are being developed.

The BDCon can be seen as a system of coordinates where each one of the roughly 700,000 terms in the lexicon has a finite set of parameter types that identifies it uniquely. These terms are linked through a specific relation to a collection of documents of the Contents Bank, in which there are mentioned or to which they can be related.

We have distinguished between six basic parameter types; each of which constitutes a specific aspect of world knowledge, and is represented by its own specific knowledge structure, requiring its own specific types of relations to the terms of the lexicon. The basic parameters and the structures in which they are organized are as follows:

- The **type of entity.** Each type constitutes a node in a typological structure (ATIP) similar but not identical to the Wordnet hypernym/hyponym structure.
- A **topic** or knowledge area to which the entity can be said to belong. Each topic constitutes a node in the topic structure (ATEM).
- The **place** to which concrete entities can be associated. Each place (physical, political, historical or geographic area) constitutes a node in the geographical structure (EGEO).
- The **time period** to which concrete entities can be associated. The nodes (historical periods associated with explicit or implicit dates, a beginning and end) are organized as a chronological structure (ECRON).
- A **general classification type** that categorize the words and documents of the Contents Bank according to the type of information they hold. Each type is a node in a loosely organized structure (ASOP).
- General **semantic relations** that cover possible relations between terms e.g. the relation between the painting “La Gioconda” and the artist “Leonardo Da Vinci”, the relation between the river “Thames” and “London”, etc. This information is organized in the semantic knowledge structure (RelCon).

In Figure 2, it is shown a general schema of the BDCon:

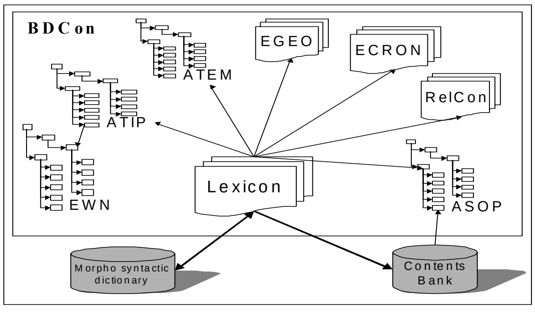


Fig. 2. A general schema of the BDCon.

The way in which the knowledge structures are connected with each other is through the terms of the lexicon, which can thus be said to constitute the backbone of the system.

2.2. The lexicon

The BDCon lexicon can be thought of as the super set of all possible lexicons, encyclopaedias or other, that can be classified in the Contents Bank. This means that it must be conceived, together with the rest of the BDCon, as a dynamic and constantly growing entity.

The lexicon contains both single words (“coche”, “barco”, “Antonio”,…) and multi lexical entities (“agua de cal”, “Francisco de Goya”, “La guerra del Golfo”, ...). These words include common nouns, verbs, adjectives, etc., as well as entity names. The lexical words have an entry for each word sense, as shown in Table 1.

agua	Cuerpo líquido a temperatura y presión ordinarias.
agua	Cualquiera de los licores que se obtienen por infusión.
agua de cepas	Vino.
agua de olor	La que está compuesta con sustancias aromáticas.
…	…

Table 1. Some of the word senses of the entry “agua” (‘water’).

As the terms in the lexicon originally come from the encyclopaedia world, their different senses are grouped around a main entry which constitutes the encyclopaedia entry.

In order to be able to distinguish between the different senses of a word, these meanings are classified according to whether they are the main sense of the entry (e.g. for the entry “agua” (water), the main sense would be the liquid H_2O), and according to their use, that is, each sense has been associated with a label that states whether, of all the possible sense of a word, the current one is *very frequently used*, *not frequently* or *very infrequently used*. This information is used as key element by the contextual word sense disambiguation application to filter those word senses that have a low probability of appearance.

2.3. The typological structure (ATIP)

The typological structure (ATIP) contains about 115,000 nodes related mainly through an IS_A relation. The nodes are conceptual entities that are represented by a set of synonyms in the same way as in Wordnet (Miller, 1990), where the idea was taken from. The relation between the nodes is transitive and permits only a single parent node.

The decision to keep the relations to one single parent was taken in order to guarantee that the structure would be able to grow indefinitely and still be manageable, i.e. that the transitive quality of the relations between the nodes would not degenerate with size.

In some cases, however, where a single parent is not enough, a complementary intransitive relation (CR) can be created. In practice, this possibility is used very sparingly (around 6% of nodes) and only when there seems to be no other option, e.g. “aliment <-IS_A- milk <-IS_A- dry milk; liquid <-CR- milk". Obviously dry milk is not a liquid although it's a kind of milk.

Figure 3 shows another example of a complementary relation:

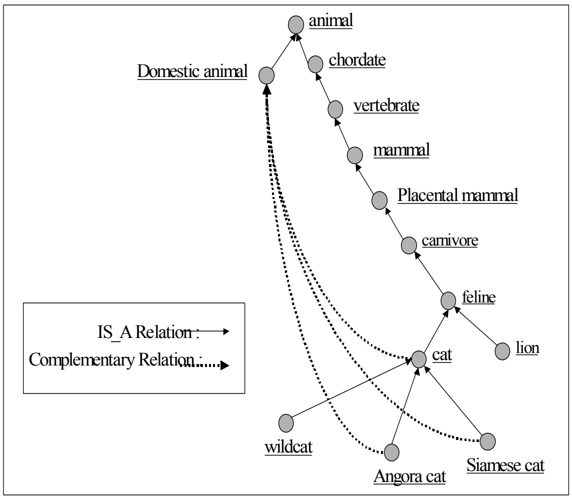


Fig.3. Example of a complementary relation.

The ATIP was built by a team of linguists using the Spanish EuroWordnet (EWN) top ontology as a starting point. When we started to design the system we considered using the Spanish Wordnet, i.e. the Spanish version of EWN, as the base for our typological structure. Spanish Wordnet is part of the EWN project (Vossen, 1998), that involved the translation of the English WordNet into various European languages while keeping the links to the original English synsets in the Princeton Wordnet. During our work with the Spanish Wordnet, however, we saw that the thesaurus it was based upon did not fit our specific needs, so we finally decided to use it as an aid to build our ontology.

In order not to loose our basic relation to the Wordnet, however, we kept a connection between the ATIP node identifiers and the EWN synset identifiers where such correspondence was possible.

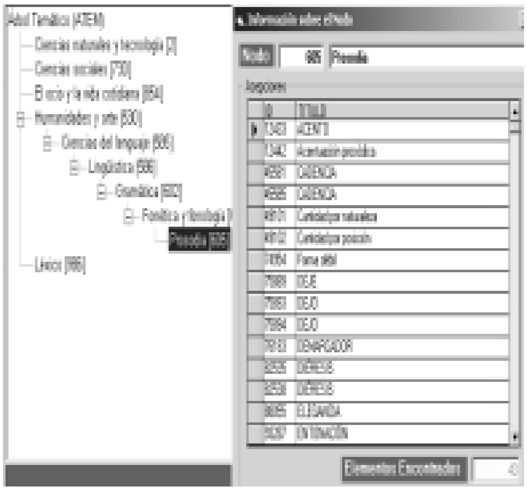
Thus, starting off with roughly 115,000 word senses from the encyclopaedia lexicon, the team, using the knowledge conveyed by the textual definitions of each word sense, and aided by the Wordnet structure, dynamically created a typological structure that covers the major part of lexical words in Spanish. A strict set of norms was defined from the beginning in order to ensure that the result would adapt both to the needs of a general encyclopaedia classification system, and to the needs of a knowledge base system.

2.4. The topic structure (ATEM)

The topic structure (ATEM) contains nodes that represent the "areas of human knowledge" from an encyclopaedic point of view. The ATEM nodes constitute a more or less finite set of around 800 semantic labels that have the following two functions: 1) To help the publishers of a specific encyclopaedia to organize the work in a structure that is intuitive and easy to use; 2) to relate concepts in such a way that the context of the words associated to them in a text can be defined and then used for word sense disambiguation. A typical use of this kind of disambiguation in the encyclopaedia context is the automatic creation of hypertext.

There can be various types of relations between ATEM nodes, and these are defined according to their use. The relations are transitive but the nodes, rather than describing classes of objects, only describe those topic areas that have been estimated, from a pragmatical point of view, to be useful to include.

The typical types of relations that can be found in the ATEM structure are hypernyms, meronyms, and also other more general relations like *belongs_to* or *is_related_to*.


Árbol Temático (ATEM)
- Ciencias naturales y tecnología [3]
- Ciencias sociales [730]
- El ocio y la vida cotidiana [864]
- Humanidades y arte [530]
  - Ciencias del lenguaje [586]
    - Lingüística [596]
      - Gramática [602]
        - Fonética y fonología [
          - Prosodia [605]
- Léxico [996]

Información sobre el Nodo

Nodo 605 Prosodia

Acepciones

ID	TITULO
13403	ACENTO
13442	Acentuación prosódica
46581	CADENCIA
46585	CADENCIA
49071	Cantidad por naturaleza
49072	Cantidad por posición
71954	Forma débil
77889	DEJE
77983	DEJO
77984	DEJO
79133	DEMARCADOR
82535	DIÉRESIS
82538	DIÉRESIS
86095	ELISANDA
89297	ENTONACIÓN

Elementos Encontrados

Fig. 4. A fragment of the ATEM topic tree.

The ATEM was built keeping in mind the traditional encyclopaedia needs, i.e. to organize the topics of the works, as well as the specific requirements of the knowledge base -that is, document classification and indexing- and to support a word sense disambiguation application.

2.5. The geographic structure EGEO

The EGEO is composed of three geographic structures: 1) geophysical: contains geological formations and so on (e.g. "The Pacific Ocean" or "The Sea of Bearing"); 2) geopolitical: includes countries and regions such as "France" or "Barcelona"; 3) geohistorical: contains historical areas ("The Roman Empire" or "The USSR").

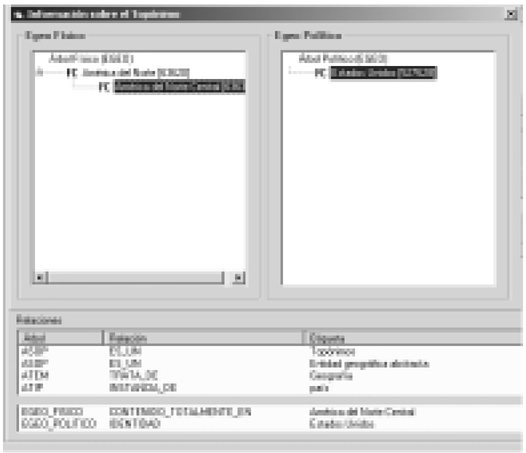
Through the use of the EGEO structure those entities that refer to concrete objects (people, events, etc.), or to classes of objects (animals, plants, etc.), can be associated to a specific geographic location.

The structure was conceived to cover two different needs; to be able to organize place names hierarchically in order to obtain a geographical localization system, and to classify or associate other entities according to their location.

The creation of the geophysical and geopolitical EGEO was made in a semi automatic fashion. The origin of the data is a well known cartographic base with around 300,000 place names, a third of which are geological formations, bodies of water, etc., and the rest are political place names.

During the extraction process from the cartographic database, we were able to keep the information about: 1) The parent element of each entity which is used to create the structures: Spain<-Catalunya<-Barcelona; 2) the link of each entity into the ATIP: city, country, lake, river, an so on.

The place names that coincide with encyclopaedia entries are associated to their corresponding EGEO nodes, and the information of the articles, when structured into identifiable patterns, is extracted with the help of an automaton that uses a pattern matching approach designed for this type of information extraction.



Atrib	Relación	Objeto
ASOP	ES_UN	Topónimos
ASOP	ES_UN	Entidad geográfica abstracta
ATEM	TRATA_DE	Geografía
ATIP	INSTANCIA_DE	país
EGEO_FISICO	CONTENIDO_TOTALMENTE_EN	América del Norte Central
EGEO_POLITICO	IDENTIDAD	Estados Unidos

Figure 5. A fragment of the EGEO structure.

With the parts of the EGEO structure that deal with Spain and certain parts of South America, and therefore are treated in more detail, the links to the official national cartographical institutes are kept so that all types of knowledge involved, such as the number of inhabitants or size, can be easily updated.

2.6. The chronological structure ECRON

This structure contains information about dates and chronological periods to which concrete entities can be associated (dates of birth and death, historical periods, etc.). The backbone of this knowledge structure is a tree of historical periods ranging from geological times to the present days. Each node in the tree has a label, a starting date and an ending date, which can be represented either by an exact year or by any of the time periods in the tree.

In addition, each date is associated with a modifier that states how the date should be interpreted; the exact or approximative year, the beginning, middle or end of the stated period, etc. All concrete entities in the lexicon, proper names, place names, etc., must be associated with the ECRON in one way or another.

2.7. The general classification type structure ASOP

The elements of this structure represent the types or categories of the words and documents of the Contents Bank. The structure includes part of speech categories; noun, verb, adjective, etc., certain semantic categories that distinguish for instance between proper names and place names, and a distinction between types of documents i.e. textual documents like word definitions or biographies, photographs, multimedia documents, etc. This general categorization, although sometimes redundant with the typological classification, is nevertheless interesting from the document retrieval point of view.

The ASOP categories are also part of the basic structure of the BDCon as they are used to designate the record structures or frames that connect the lexicon items with the different knowledge structures of the BDCon.

2.8. Frames

Apart from the slots, that contain the uniquely identifying parameters, these frames can also contain slots with factual information, e.g. number of inhabitants, the area of a country, length of a river, etc. Here are some examples of basic information frames:

Slot	**Value**	**Relation**	**BDCon structure**
Name	Alfred Adler		thesaurus
Nationality	Austria	NATION_OF	EGEO
Type	medical doctor/ psychologist/ man	INSTANCE_OF	ATIP
Topic	PSYCHOLOGY	RELATED_TO	ATEM
Born in	Vienna	BIRTH_PLACE	EGEO
Born year	1870	BIRTH_YEAR	ECRON
Died in	Aberdeen	DEATH_PLACE	EGEO
Died year	1937	DEATH_YEAR	ECRON
Category	anthroponym	KIND_OF	ASOP

Table 2. Proper name frame

Slot	**Value**	**Relation**	**BDCon structure**
Name	Barcelona		thesaurus
Location	Spain	IS_IN	EGEO
Type	City	INSTANCE_OF	ATIP
Topic	GEOGRAPHY	RELATED_TO	ATEM
Category	Place name	KIND_OF	ASOP
Inhabitants	1,500,000		
Extension	90 km2		

Table 3. Place name frame

Slot	**Value**	**Relation**	**BDCon structure**
Name	mountain		thesaurus
Type	geological formation	IS_A	ATIP
Category	Com. noun	IS_A	ASOP
Topic	GEOMORPHOLOGY	RELATED_TO	ATEM
Use	Very usual		C.N.Frame

Table 4. Common noun frame

Since the basic information in biographies and other similar encyclopaedia articles is reasonably well structured, the proper nouns are entered in the system in a semi automatic fashion, i.e. extracted automatically from the text, and validated by hand. The information extraction is carried out by an application that uses lexical and semantic information obtained from the BDCon to look for feature structure patterns in the corresponding encyclopaedia definition texts. The need of lexical and semantic information for the information extraction decided the order of input of material into the system; first the typological structure with lexical words was entered, then the time periods and place names, and afterwards, the biographies, events, and other entities that require both geographical and chronological information.

3. Conclusions and future work

The use of partially structured information such as can be found in encyclopaedias, together with the help of the world knowledge extracted from EuroWordNet, has been a decisive element in the success of the construction of the BDCon. Once the basic knowledge base has been built, the task of acquiring further knowledge to be fed into the system becomes easier and can be done with less and less human interaction. One of the many steps in the future development of the BDCon is the application of a clustering algorithm that, working on previously disambiguated text and using the power of the existing knowledge structures, especially the typological and topic structures, will be able to create weighted semantic relations between terms with little or even no human interaction. The resulting knowledge will then be used to improve the performance of the linguistic indexing and knowledge extraction tools, making them more and more useful for the publishers who need them for their work.

4. References

Gaizauskas, R. & Wilks, Y. (1998). Information extraction: beyond document retrieval. *Journal of Documentation*, 54: 70-105.

Gonzalo, J., Verdejo, F., Chugur, I. & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal.

Mandala, Rila, Tokunaga Takenobu & Tanaka Hozumi (1998). The use of WordNet in information retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal.

Miller, G.A. (1990). WordNet: An on-line lexical database. *International journal of Lexicography*, 3(4): 235-312.

Trotzig, D. (2000). Proyecto SCASEM: un sistema de catalogación semántica. *Procesamiento del Lenguaje Natural*, 24: 251-252.

Uschold, M. & Gruninger, M. (1996). Ontologies: Principles, Methods and Applications. *The Knowledge Engineering Review*, 11(2): 93-136.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Wilks, Y. & Stevenson, M. (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering*, 4(2): 135-144.