# Concept sense disambiguation in concept maps using WordNet

## (Technical Report)

Alfredo Simón[1], Luigi Ceccaroni[2] and Alejandro Rosete[1]

[1] Centro de Estudios de Ingeniería de Sistemas (CEIS)
Instituto Superior Politécnico "José Antonio Echeverría"
Ave. 114, No. 11901, Marianao, C. Habana, Cuba
{asimon, rosete}@ceis.cujae.edu.cu
[2] Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politècnica de Catalunya (UPC), Campus Nord, Edif. Omega, C. Jordi Girona, 1-3, 08034 Barcelona, Spain
luigi@lsi.upc.edu

**Abstract.** In this report an unsupervised and knowledge-based algorithm for concept sense disambiguation in concept maps is proposed. Concept maps are graphical tools for organizing and representing knowledge, based on concepts and labeled interconnections among them, forming propositions. The disambiguation process is carried combining Magnini's domain, context information and the gloss. It's supported in the Spanish WordNet lexical database and the lexical relations *hypernyms-hyponyms*, *meronyms-holonyms* and *instance*.

**Keywords:** concept maps, concept sense disambiguation, Magnini's domain and WordNet

## 1 Concept maps

CMs are a graphically rich technique for organizing and representing knowledge that emerged within the pedagogical science. They were proposed by Novak et al. [3], who defined them as a "*technique that simultaneously represents a strategy of learning, a method to grasp the most significant aspect of a topic and a schematic resource included in one structure of propositions*". They include *concepts*, *linking-words* (that specify the relationship between concepts) and *propositions* (that contain two or more concepts connected using linking words to form a meaningful statement). Fig. 1 shows an example of CM that describes the nitrogen cycle. As it can be observed, CMs are a kind of semantic network, but one that is more flexible and informal, oriented to be used and interpreted by humans.

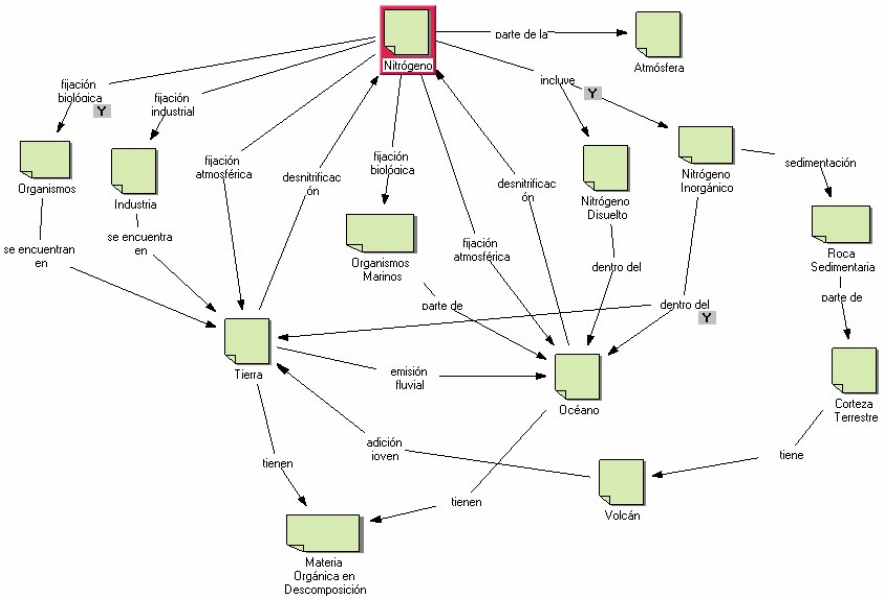


**Fig 1.** Concept map of the nitrogen cycle

## 2 WordNet

*WordNet* is a lexical knowledge created at the University of Princeton [2], whose basic structure is the *synset,* composed by a set of words related by *synonym*, an *identifier*, a *gloss* (description of meaning) and tagged with some domains, such as: Medical, Sport and Geology. The *synsets* are distributed in form of a semantic net and interconnected among themselves by several types of lexical relations, such as: *hyperonyms*, *hyponyms*, *meronyms*, *holonyms*, *antonym*, *role*, *instante* and *causes*. The *synset* also defines the meanings and the sense of a word, and its can be found in one or various *synsets,* when the polysemy take place. WordNet is one of the most extensively used lexical knowledge based in natural language processing campus and it can be used as an ontology if its lexical links are interpreted according to a formal semantics [1].

## 3 Algorithm

Input: PL.
Output: synset associated to each concept in the CM
Variables: $c_i$ is a concept; $s_{ij}$ is a synset of $c_i$ and $S(C)$ is a set of $s_{ij}$ for a set of concepts C; $D_{mc}$ is a set of domain of the CM; $Context_{mc}(c_i, r)$ is a set of concepts inside the context in the CM created using $c_i$ as center and including all concepts in a radio $r$ (arcs between two concept); $Context_{wn}(s_{ij})$ is a context in WordNet in witch $s_{ij}$

is a center and include all path former by *hyperonyms, meronyms* and *instance* relations between $s_{ij}$ and some synsets of the concepts include in $Context_{mc}(c_i, r)$; $l$ the length of this path (arcs between two synsets) and $\alpha$ the quantity of concept in $Context_{mc}(c_i, r)$ that have any $s_{ij}$ in this path.

Procedure:

1: All $s_{ij}$ of each $c_i \in$ CM are identified in WN;
2: Two lists are created: CD (concepts with only one synset) and CND (concepts with more than one synset);
3: The CM domains are identified and stored in $D_{mc}$. The domains with major occurrence in the synsets of the concepts (more than 45 %) and the domains of the *principal-concept* of the CM are selected;
4: i = 1;
5: For each $c_i \in$ CND
6: Disambiguation by domain. $c_i$ is disambiguate if:
   a. have only one $s_{ij}$ tagged with at least one domain of $D_{mc}$;
   b. have any $s_{ij}$ tagged with child domain of some domain of $D_{mc}$;
   c. have any $s_{ij}$ tagged with a domain that shares the same immediate ancestor that any domain of $D_{mc}$;
7: The $c_i$ disambiguated is adding to CD and eliminated of CND;
8: If CND ≠ [ ] then r is started in 2 and the disambiguation by context is carried out:
   a. The $Context_{mc}(c_i, r)$ is created;
   b. The $Context_{wn}(s_{ij})$ for associated to $Context_{mc}(c_i, r)$ is created;
   c. If the $Context_{wn}(s_{ij})$ isn't created them
      r ++ and the $Context_{mc}(c_i, r)$ is created again (go to 8.a) ; Else
   d. For each $s_{ij}$ the weight is calculated by the expression:

$$S_c = \max_{s_{ij} \in S(\{c_i\})} \left\{ w\left(Context_{wn}(s_{ij}), Context_{mc}(c_i, r)\right) = \sum_{j=1}^{n} \alpha * \frac{1}{l_j} \right.;$$

   e. If $S_c$ is unitary then $s_{ij}$ is selected for concept disambiguate and $c_i$ is add to CD and eliminated of CND; Else
      If length ($Context_{mc}(c_i, r)$) < length (CM) then r ++ and go to 8.a; Else
      If ∃ s' | s' is an immediate common hyperonym among any $s_{ij} \in S_c$ and any $s_{kj} \in S(Contexto_{mc}(c_i, r))$ then $Sc = s_{ij}$ and go to 8.e;
9: If CND ≠ [ ] and ∃ $c_i \in$ CND with gloss then the disambiguation by gloss is do
   a. r is started in 2;
   b. The $Context_{mc}(c_i, r)$ is created;
   c. The quantity of concepts present in $Context_{mc}(c_i, r)$ and in the gloss of each $s_{ij}$ is calculated. If the result for each $s_{ij}$ is cero then r = length(CM) and go to 9.b
10: The $s_{ij}$ with best result is selected, $c_i$ is adding to CD and eliminated of CND;

## 4 Experimental Results

The algorithm presented was proven with a total of 31 polysemic concepts organized in four CMs of Environment Domain. CMs selected should be well constructed,

theoretically and have significant amount of concepts with synset in WN and of their more than one synset. They were revised by an expert in this domain. In Tables 1-3 the results of all tests is showed.

**Table 1**. Characteristics of CMs tested.

| Concept Maps | Concepts | Links | Concepts in WN | Polysemic concepts | Synset* concept | Domains |
|---|---|---|---|---|---|---|
| Nitrógeno | 13 | 22 | 9 | 5 | 3.8 | 14 |
| Geología | 12 | 13 | 12 | 3 | 5 | 10 |
| Plantas | 16 | 17 | 16 | 10 | 3.1 | 22 |
| Agua | 21 | 21 | 21 | 13 | 3.53 | 24 |
| Total: | 62 | 73 | 58 | **31** | **3.85** | **17.5** |

**Table 2**. Concepts disambiguation by domain and context results.

| CMs/Heuristics | Precision | | Recall | | Coverage | |
|---|---|---|---|---|---|---|
| | Domain | Context | Domain | Context | Domain | Context |
| Nitrógeno | 0.500 | 0.666 | 0.300 | 0.400 | 0.600 | 0.600 |
| Geología | 0.833 | 0.666 | 0.833 | 0.666 | 1 | 1 |
| Plantas | 0.937 | 1 | 0.750 | 0.600 | 0.800 | 0.600 |
| Agua | 0.750 | 0.937 | 0.346 | 0.750 | 0.461 | 0.800 |
| **Ave.** | **0.755** | **0.817** | **0.557** | **0.604** | **0.715** | **0.750** |

**Table 3**. Concepts sense disambiguation results (Integrations of domain, context and gloss).

| Concept Maps | Precision | Recall | Coverage |
|---|---|---|---|
| Nitrógeno | 0.660 | 0.400 | 0.600 |
| Geología | 0.833 | 0.833 | 1 |
| Plantas | 1 | 0.800 | 0.800 |
| Agua | 0.954 | 0.807 | 0.846 |
| **Ave.** | **0.861** | **0.710** | **0.811** |

## References


[1] Gangemi, A., Navigli, R. and Velardi, P.: The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, LNCS 2888, Spinger-Berlang (2003) 820-838.

[2] Miller G.A, Beckwidth, R., Fellbaum, C., Gross, D., and Miller, K.J.: Introduction to WordNet: An On-line Lexical Database. In Proceedings of the International Journal of Lexicography (five papers) Vol 3, No.4 (1990) 235-244.

[3] Novak, J. y Gowin, D.: Learning how to learn, Cambridge Press. New Cork (1984)