International Journal of Information Science 2012, 2(3): 8-12
DOI: 10.5923/j.ijis.20120203.01

Words Polysemy Analysis: Implementation of the Word Sense Disambiguation Algorithm Based on Magnini Domains

Francis C. Fernández-Reyes[*], **Exiquio C. Leyva Pérez**, **Rogelio Lau Fernández**

Artificial Intelligence and Infrastructure of Informatics Systems Department, Superior Polytechnic Institute "José Antonio Echeverría" (CUJAE), La Habana, Cuba

Abstract This paper presents an analysis of the lexical resources used in Word Sense Disambiguation (WSD) process by methods based on Magnini domains. At the same time, the characteristics of two algorithms that use Magnini domains are shown and we define the implementation of Word Domain Disambiguation (WDD) algorithm as defined in [1]. Later on, we proceed designing the experiments to test the algorithm and we arrived to different conclusions.

Keywords Word Sense Disambiguation Based On Lexical Resources, Magnini Domains, Wordnet Extension

1. Introduction

Linguistics knowledge constitutes the theoretical foundation to develop a great range of technological applications, which have become very important for the informatics society. Among the multiple systems which make use of this knowledge we have[2]: knowledge management and search, natural language interfaces between computers and users, automatic translation, among others.

Word Sense Disambiguation is the process which its main objective is to assign to each word, given in a context, a definition or meaning (predefined sense or not), which is different from others than it could have. Every natural language processing application, from the simplest application until the most complex one, which needs to understand the sense of a term, in a certain context, requires the use of techniques for semantic ambiguity resolution, that is the case of the orthographic correctors before the sentence: *The rough indicates a cold.* As you can appreciate, it is not enough, in order to correct the orthography, find the lemma of the word rough, also we have to know, in certain degree, the sense in the text, it is easy to appreciate that the more proper word is *cough* on the previous example.

The information retrieval task –was previously carried today it includes, in some measure, the sense of the text in which we are searching and it provides mechanisms that offers the users an effective response to their search[3], for example, if we are searching for *pig*, not only recover the documents that contain this word, but also include those documents that refer to *pork, animal, pet.* Despite the importance that word sense disambiguation methods involves, they are an ongoing problem, still without a complete solution[2,4].

There are several methods which try to disambiguate correctly a word; those are classified into so many ways, depending exclusively on the author´s criteria. The knowledge based methods use, mainly, great lexical resources and heuristics based on linguistics notions. The most used lexical resources for the disambiguation task are[5]: WordNet (lexical database) and SemCor (annotated corpus for automatic learning), also the corpus Senseval, the Magnini Domain and the eXtended WordNet are used. The heuristics used in this work, exploit some linguistics and mathematical hypothesis, and apply the WordNet semantics relations.

The present paper has as a main objective, the analysis of the Magnini domains behaviour on the Word sense disambiguation process, and we design several experiments using the task *English lexical sample* from collection Senseval 2 in order to accomplish it. We develop different tests to the implemented algorithm varying the context of the disambiguation objective words. In a first section the resources used by the algorithm (WordNet and Magnini domains) are defined, in a second section, the design and implementation of the defined algorithm for decreasing analysis of words polysemy appears. Finally, we proceed to analyze the results obtained from the experiments and arrived to conclusions.

2. Resources Used by the Algorithm

The first step for linguistic knowledge informatics processing is the formal representation of the given knowledge.

[*] Corresponding author:
ffernandez@ceis.cujae.edu.cu (Francis de la C. Fernández Reyes)

Published online at http://journal.sapub.org/ijis

Multiple resources have been created in order to represent the information of natural language, for instance, lexical databases, corpus, thesaurus and ontologies, etc. Other resources are based on the use of informatics tools that perform specific analysis focused on some particular task of natural language processing, for example, for grammatical analysis we can find a series of programs that allow text tokenization, identify grammatical categories ("Part of Speech", POS) and establish the lemma of a word.

Different lexical resources in one way or another access the algorithm are shown briefly, in order to determine the correct sense of a word inside a given context through Magnini domains.

2.1. Lexical Database: WordNet

A lexical database is a collection of linguistics information that is organized according to a specific model and facilitates the storage, recovery and modification of its own data. The data model, could follow a hierarchical net or relational structure[6].

The lexical databases are used in Linguistic as information sources that are reused in other resources, for example, a computational lexicon or a terminology database[6].

WordNet is a designed lexical database based on psycholinguistics theories of mental lexicon[7] with the aim of speeding up the searches on the English language online dictionaries, later on, with the EuroWordNet project[8] the search was expanded to others languages, such as: Spanish, Dutch, etc.

This lexical database is being built based on syntactic categories of noun, verb, adjective and adverb as well as semantic relations of hyponym, hyperonym, meronym, holonym, synonym, antonym, coordinated terms and troponym. The equivalent of synonym and antonym relations, on natural language, is expressed by means of synonym and antonym of the words, respectively. The relations hyponym and hyperonym expresses relations "kind of", it means, two terms given: *tree* and *Pine*; *Pine* is a "kind of" *tree*, or *Pine* is hyponym of *tree*, while *tree* is hyperonym of *Pine*. The coordinated terms are based on hyperonym, two terms are always coordinated when they share a common hyperonym, for example, *Pine* and *Oak* shares *tree* as common hyperonym so they are coordinated terms. The meronym and holonym relations express relations "part of", it means, two terms given: *car* and *wheel*; *wheel* is "part of" *car*, or *wheel* is meronym of *car*, while, *car* is holonym of *wheel*. The troponomy is a relation that appears on verbs, two verbs are troponyms if one of them is activated inside the other one in some way, for example, *whisper* and *talk*.

WordNet combines characteristics of other linguistics resources due to the addition of definitions – or glosses – of terms on each of it senses, just as a dictionary and define sets of synonyms words, with different semantic relations between them, just as a thesaurus. Besides, it constitutes a resource of wide use by word sense disambiguation algorithms and its distribution is free[9]. WordNet is based on theoretical assumption of lexical matrix, integrated by elements word and meaning. At lexical matrix, columns correspond to words and rows correspond to concepts or meanings. In addition, concepts are represented by a word list that can be used to express the sense definition, which means, by all the elements that belong to a same row and constitute synonyms. This word lists are called "synsets" (term that comes from **syn**onym **sets**).

2.2. Magnini Domains

There are some labels which serve to mark the senses (synsets) of WordNet in order to provide information about categories and appropriated semantic levels. Those labels are "Subject Field Codes" (SFC), as they are habitually known and represent relevant words sets for a specific domain. The better approximation of SFCs are the labels used on the dictionaries (Medicine, Architecture, and so on), even if their purpose is restricted to word uses belonging to a specific terminological domain.

They are also known, sometimes, as IRST domain and this name is due to the institute where they were developed ("Istituto per la Ricerca Scientifica e Tecnologica"), or Magnini domains because their creators are Magnini and Cavaglia[1]. The SFC can include senses of different syntactic categories, for example, Medicine encompasses noun senses (doctor, hospital) and verb ones (operate).

The SFCs are organized into hierarchies and families. This hierarchy has different levels of specialization, if the level is deeper, the specialization degree is bigger. There is a SFC Factotum for the generic senses which are very hard to classify due to their association with highly polysemous words.

Each WordNet synset has one or various Magnini domains associated to it, just as we had presented before. For instance, word chair has four synsets at WordNet 1.6 and every one of them has an only domain associated to it, as table 1 shows. Often, the same domains can be associated to synsets of different grammatical categories. This information, integrated to WordNet 1.6 allows establishing several relations between words belonging to different sub-hierarchies and includes, within the same domain, various senses of a same word.

2.3. Collection English Lexical Sample of Senseval 2

The english lexical sample task of Senseval 2 competition, is equivalent to a set of texts taken from Brow Corpus and manually tagged with WordNet senses. This collection of texts has several paragraphs, where each of them, have three or four sentences and in some of them, one target word appears.

Table 1. Domains Associated to Synsets of Word Chair.

Synset	Domain	Noun
02418562	furniture	chair#1
00393476	pedagogy	chair#2
07496412	person	chair#3
02626821	law	chair#4

This set of texts has 73 words to disambiguate, but each of them has more than one instance or occurrence, therefore there are 4328 test instances, divided into 29 nouns, 29 verbs and 15 adjectives. The results presented in this paper were obtained after the execution of the algorithm on the word chair that appears in this task with 69 occurrences.

This task provides also a key file where each target word is associated to its correct sense obtained from WordNet (sometimes appears more than one correct sense). The word is associated to an identifier inside the set of texts that allows finding its appropriated sense in this key file.

3. Word Sense Disambiguation Algorithm Based on Magnini Domains

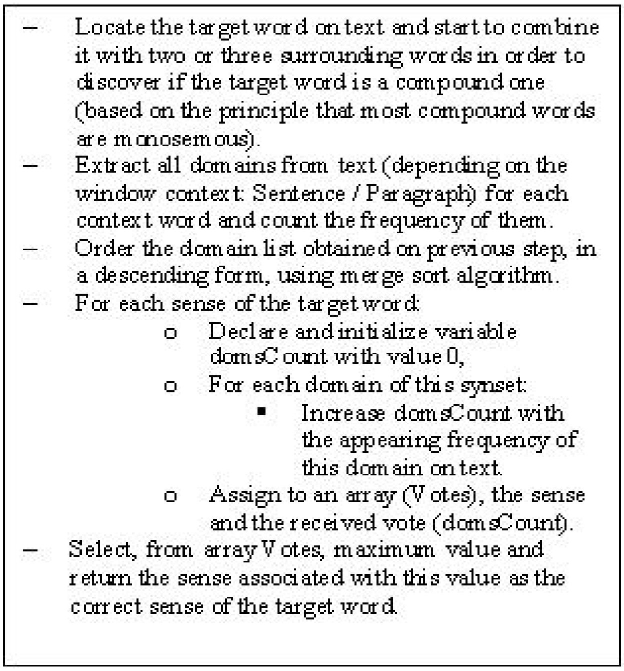
The interesting aspect of domain-driven disambiguation as well as methods for determining word sense dominance is that they shift the focus from the linguistic understanding to a domain-oriented type-based vision of sense ambiguity [4]. In the last three decades, a large body of work has been presented that concerns the development of automatic methods for the enrichment of existing resources such as WordNet. These include proposals to extract semantic information from dictionaries, approaches using lexico-syntactic patterns, heuristic methods based on lexical and semantic regularities, taxonomy-based ontologization. Other approaches include the extraction of semantic preferences from sense-annotated and raw corpora, as well as the disambiguation of dictionary glosses based on cyclic graph patterns. Other works rely on the augmenting of WordNet by means of an annotated hierarchy of domains labels[10].

In[1] work the influence that domains have on the word sense disambiguation process is shown and an algorithm to find the correct sense of a word in a given context is proposed, when applications didn't require a fine granularity for sense distinction. This algorithm is named Word Domain Disambiguation (WDD).

WDD assigns to each word of the text a domain label instead of a sense label. Domain is understood as the set of words that are strongly semantic related.

The fundamental idea of applying the domains to word sense disambiguation process is to provide relevant information to establish semantic relations between word senses. For instance, word *bank* has 10 senses from WordNet 1.6 but three of them *bank#1, bank#3* and *bank#6* can be grouped under *economy* domain, while *bank#2* and *bank#7* belongs to *geography* and *geology* domains. In order to apply WDD we need a lexical resource where senses of words are associated with domains. Thus, the resource used in this work was the WordNet extension, WordNet Domains[1], which has all the synsets tagged with one or more domains.

The work of[11] presents an algorithm that uses Magnini domains to reduce the polysemy of words according to N principal text domains where the target word appears, therefore, we achieve an implementation of an algorithm which basic idea is reduced to a comparing process between domains that appears on the context and domains of the different senses of the target word. The correct sense of the target word is computed with the more frequently domain on text that corresponds with some sense of the target word, if more than one sense is associated with it, all senses are given back. Figure 1 shows the pseudo-code of the algorithm based on Magnini domains.



```
– Locate the target word on text and start to combine
  it with two or three surrounding words in order to
  discover if the target word is a compound one
  (based on the principle that most compound words
  are monosemous).
– Extract all domains from text (depending on the
  window context: Sentence / Paragraph) for each
  context word and count the frequency of them.
– Order the domain list obtained on previous step, in
  a descending form, using merge sort algorithm.
– For each sense of the target word:
      o Declare and initialize variable
        domsCount with value 0,
      o For each domain of this synset:
            ▪ Increase domsCount with
              the appearing frequency of
              this domain on text.
      o Assign to an array (Votes), the sense
        and the received vote (domsCount).
– Select, from array Votes, maximum value and
  return the sense associated with this value as the
  correct sense of the target word.
```

Figure 1. Pseudo-code of an algorithm based on Magnini Domains

4. Discussion of Preliminary Experimental Results

Collection english lexical sample from competition Senseval 2 was used in the experiments. This text collection, extracted from Brown Corpus, contains target words instances. In order to realize the process of semantic tagging, each target word occurrence is labelled with one or more senses acquired from WordNet, which are the correct ones according to the context in which instance appears.

The test dataset consists of 73 tasks, where each one contains several occurrences of a target word. Leave out some exceptions, all words inside a task are used with the same grammatical category, for instance, task *art* has 98 occurrences of word *art*, all belonging to noun category. Each word occurrence is named instance and consists on a sentence that has the target word, as well as, one to three surrounding sentences that supply the context of this one.

Tasks are grouped according to the grammatical category on three sets – nouns, verbs and adjectives. Tables 2, 3 and 4 shows all tasks that appear in this collection divided into nouns, verbs and adjectives. For each grammatical category we find the word, as well as the number of instances inside the task and the quantity of senses finding in WordNet and in test collection.

Table 2. Composition of Noun Task for Collection English Lexical Sample

Word	Instance Quantity	WN Senses	Test Data Senses
art	98	4	15
authority	92	7	8
bar	151	13	18
bum	45	4	6
chair	69	4	8
channel	73	7	11
child	64	4	5
church	64	3	5
circuit	85	6	16
day	145	10	12
detention	32	2	7
dyke	28	2	4
facility	58	5	5
fatigue	43	4	6
feeling	51	6	6
grip	51	7	7
hearth	32	3	5
holiday	31	2	5
lady	53	3	8
material	69	5	10
mouth	60	8	11
nation	37	4	4
nature	46	5	7
post	79	8	10
restraint	45	6	9
sense	53	5	12
spade	33	3	6
stress	39	5	7
yew	28	2	4
TOTAL	1754		
AVERAGE		5.1	8.2

Table 3. Composition of Verb Task for Collection English Lexical Sample

Word	Instance Quantity	WN Senses	Test Data Senses
begin	280	10	7
call	66	28	17
carry	66	39	20
collaborate	30	2	2
develop	69	21	14
draw	41	35	22
dress	59	15	12
drift	32	10	9
drive	42	21	13
face	93	14	6
ferret	1	3	1
find	68	16	17
keep	67	22	20
leave	66	14	10
live	67	7	9
match	42	9	7
play	66	35	20
pull	60	18	25
replace	45	4	4
see	69	24	13
serve	51	15	11
strike	54	20	20
train	63	11	8
treat	44	8	5
turn	67	26	26
use	76	6	6
wander	50	5	5
wash	12	12	7
work	60	27	18
TOTAL	1806		
AVERAGE		16.4	12.2

Table 4. Composition of Adjective Task for Collection English Lexical Sample

Word	Instance Quantity	WN Senses	Test Data Senses
blind	55	3	6
colorless	35	2	3
cool	52	6	7
faithful	23	3	3
fine	70	9	14
fit	29	3	3
free	82	8	13
graceful	29	2	2
green	94	7	14
local	38	3	4
natural	103	10	23
oblique	29	2	3
simple	66	7	5
solemn	25	2	2
vital	38	4	4
Total	768		
Average		4.7	7.1

It is essential to notice that quantities of senses not necessarily must be equals; it is explained because just only a subset of all synsets from WordNet is used to annotate words of the collections, since others are extremely rare and the data didn’t contain examples of their use, this is particular truly for verbs, that´s why the average frequency of WordNet senses is higher than the Test data senses. On the other hand, also it is possible that a word be tagged with a sense that is not possible according to WordNet, such is the case of compound words. On the noun and adjective task appear several compounds words, which mean that average frequency on Test data set is higher than the WordNet sense inventory.

Table 5. Domains and Synsets associated with some Tasks

Domain	Sense
art	art#1
art	art#2
art	art#3
publishing	art#4
furniture	chair#1
pedagogy	chair#2
person	chair#3
law	chair#4
telecommunication	channel#1
transport	channel#2
geography	channel#3
telecommunication	channel#4
anatomy	channel#5
factotum	channel#6
religion	church#1
Buildings, religion, town_planning	church#2
religion	church#3
color	colorless#1
factotum	colorless#2
quality	colorless#3
color	colorless#4
quality	colorless#5

In order to make the experiments, we select from the 73 tasks of the collection, those that didn´t have labeled all synsets with factotum domain, and execute the algorithm varying the context from Sentence to Paragraph. Table 5 shows the domains and synsets, obtained from WordNet 1.6, for some of the tasks we take under consideration, they were selected randomly in order to show some of the conclusions arrived. Besides, table 6 shows the results obtained by the algorithm for those tasks, as well as the computed evaluation metrics.

Table 6. Results Obtained from Developed Algorithm (All the Metrics were Calculated for All the Tasks)

Task (Domains Quantity)	Domain according to appearing frequency	Correct Answers (Sentence/ Paragraph)
art (2)	art, 0.75 publishing, 0.25	44 / 64
chair (4)	furniture, 0.25 pedagogy, 0.25 person, 0.25 law, 0.25	25 / 12
channel (6)	telecommunication, 0.33 transport, 0.17 geography, 0.17 anatomy, 0.17 factotum, 0.17	29 / 23
church (3)	buildings, 0.2 religion, 0.6 town_planning, 0.2	33 / 52
colorless(5)	color, 0.4 quality, 0.4 factotum, 0.1	20 / 33
Precision	(45.51% / 48.89%)	
Recall	(33.45% / 47.53%)	
Coverage	(70.59% / 92.76%)	
F1 Measure	(38.56% / 48.20%)	

As it can be appreciated, the algorithm for task art, obtains an increment on the computed quantity of correct disambiguated words when the context is changed from Sentence to Paragraph; however, it didn’t behaves at same way for task chair when the context is incremented to Paragraph, much noise is inserted in the algorithm and its precision decreases.

Another remarkable aspect results the compound words identification process, in case of task art, the algorithm detects 10 compound words from 12 that total exists and for task chair, it identified 2 compound words from 4 existents. This result affects directly the precision of algorithm, and loses a 25% of recall due to the 4 compound words which don’t detect. Generally, Paragraph context is better than the Sentence one.

5. Conclusions

Magnini domains represent a way to reduce the polysemy of words, however, still it has influence the appearing frequency inside the synset who is analyzed, such is the case of word *art* which have a same domain with an appearing frequency of 0.75, three of four senses had labelled with domain *art*.

One way to obtain a vote over a synset corresponds with the idea of selecting the synset which domain is more represented inside the context, under this hypothesis we developed the algorithm presented and the experiments accomplished on it.

The correct determination of compound words increases precision and recall of the word sense disambiguation process over the test dataset.

REFERENCES

[1] B. Magnini y G. Cavaglia. "Integrating Subject Field Codes into WordNet". en Second International Conference on Language Resources and Evaluation (LREC). Athens, Greece. 2000

[2] A. Gelbukh y G. Sidorov, "Procesamiento automático del español con enfoque en recursos léxicos grandes". México. 2006.

[3] E. Tello Leal, I. Lopez-Arevalo, y V. Sosa-Sosa, Prototipo para desambiguación del sentido de las palabras mediante etiquetado de palabras y relaciones semánticas, in Revista Avances en Sistemas e Informática. 2010: Medellín. p. 27-32.

[4] R. Navigli, "Word sense disambiguation: A survey". ACM Comput. Surv.Vol. 41, No. 2, pp. 1-69. 2009.

[5] R. Navigli, K.C. Litkowski, y O. Hargraves. "SemEval-2007 Task 07: Coarse-Grained English All-Words Task". en 4th International Workshop on Semantic Evaluations (SemEval). Prague, Czech Republic: Association for Computational Linguistics. 2007

[6] S. Arano, La ontología: una zona de interacción entre la Lingüística y la Documentación in Hipertext.net. 2003, Sección Científica de Ciencias de la Documentación · Departamento de Periodismo y de Comunicación Audiovisual, Universidad Pompeu Fabra.

[7] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, y K. Miller, "Introduction to WordNet: An on-line lexical database". International Journal of Lexicography.Vol. 3, No., pp. 235-244. 1990.

[8] P. Vossen, "EuroWordNet: Building a Multilingual WordNet Database with Semantic Relations between Words". Procesamiento del lenguaje natural.Vol. 18, No., pp. 145-158. 1996.

[9] [9] A. Pons Porrata, Desarrollo de algoritmos para la estructuración dinámica de información y su aplicación a la detección de sucesos., in Departamento de Lenguajes y Sistemas Informáticos. 2004, Universidad Jaume I: Castellón. p. 132.

[10] S. Paolo Ponzetto y R. Navigli. "Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems". en Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics. 2010

[11] A. Pons-Porrata, R. Berlanga-Llavori, y J. Ruiz-Shulcloper. "Un nuevo método de desambiguación del sentido de las palabras usando WordNet." en X Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA). Universidad Politécnica de Madrid. 2003