UMCC-DLSI: Integrative resource for disambiguation task

Yoan Gutiérrez and **Antonio Fernández**
DI, University of Matanzas
Autopista a Varadero km $3^{1/2}$
Matanzas, Cuba
yoan.gutierrez,antonio.fernandez@umcc.cu

Andrés Montoyo and **Sonia Vázquez**
DLSI, University of Alicante
Carretera de San Vicente S/N
Alicante, Spain
montoyo,svazquez@dlsi.ua.es

Abstract

This paper describes the UMCC-DLSI system in SemEval-2010 task number 17 (All-words Word Sense Disambiguation on Specific Domain). The main purpose of this work is to evaluate and compare our computational resource of WordNet's mappings using 3 different methods: Relevant Semantic Tree, Relevant Semantic Tree 2 and an Adaptation of k-clique's Technique. Our proposal is a non-supervised and knowledge-based system that uses Domains Ontology and SUMO.

1 Introduction

Ambiguity is the task of building up multiple alternative linguistic structures for a single input (Kozareva et al., 2007). Word Sense Disambiguation (WSD) is a key enabling-technology that automatically chooses the intended sense of a word in context. In this task, one of the most used lexical data base is WordNet (WN) (Fellbaum, 1998). WN is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. Due to the great popularity of WN in Natural Language Processing (NLP), several authors (Magnini and Cavaglia, 2000), (Niles and Pease, 2001), (Niles and Pease, 2003), (Valitutti, 2004) have proposed to incorporate to the semantic net of WN, some taxonomies that characterize, in one or several concepts, the senses of each word. In spite of the fact that there have been developed a lot of WordNet's mappings, there isn't one unique resource to integrate all of them in a single system approach. To solve this need we have developed a resource that joins WN[1], the SUMO Ontology[2], WordNet Domains[3] and WordNet Affect[4]. Our purpose is to test the advantages of having all the resources together for the resolution of the WSD task.

The rest of the paper is organized as follows. In Section 2 we describe the architecture of the integrative resource. Our approach is shown in Section 3. Next section presents the obtained results and a discussion. And finally the conclusions in Section 5.

2 Background and techniques

2.1 Architecture of the integrative resource

Our integrative model takes WN 1.6 as nucleus and links to it the SUMO resource. Moreover, WordNet Domains 2.0 (WND) and WordNet Affect 1.1 (WNAffects) are also integrated but mapped instead to WN 2.0. From the model showed in Figure 1, a computational resource has been built in order to integrate the mappings above mentioned.

The model integrator's proposal provides a software that incorporates bookstores of programming classes, capable to navigate inside the semantic graph and to apply any type of possible algorithm to a net. The software architecture allows to update WN's version.

In order to maintain the compatibility with other resources mapped to WN, we have decided to use WN 1.6 version. However, the results can be offered in anyone of WN's versions.

[1] http://www.cogsci.princeton.edu/ wn/

[2] http://suo.ieee.org

[3] http://wndomains.fbk.eu/

[4] http://wndomains.fbk.eu/wnaffect.html

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 427–432,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

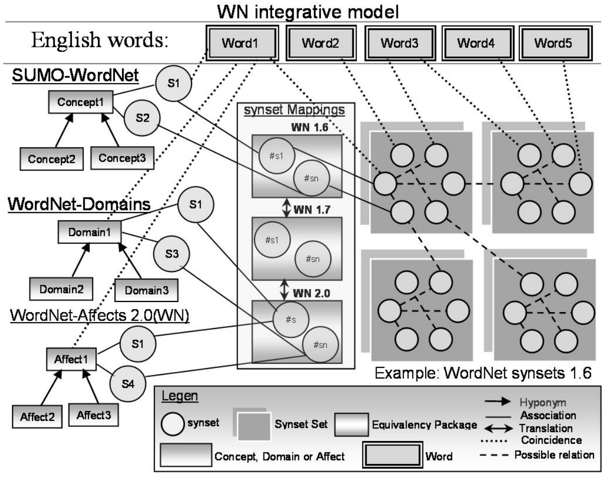

Figure 1: WordNet integrative model

2.2 The k-clique's Technique

Formally, a clique is the maximum number of actors who have all possible ties presented among themselves. A "Maximal complete sub-graph" is such a grouping, expanded to include as many actors as possible.

"A k-clique is a subset of vertices C such that, for every $i, j \in C$, the distance $d(i, j)_k$. The 1-clique is identical to a clique, because the distance between the vertices is one edge. The 2-clique is the maximal complete sub-graph with a path length of one or two edges". (Cavique et al., 2009)

3 The Proposal

Our proposal consists in accomplishing three runs with different algorithms. Both first utilize the domain's vectors; the third method utilizes k-cliques' techniques.

This work is divided in several stages:

1. Pre-processing of the corpus (lemmatization with Freeling) (Atserias et al., 2006).

2. Context selection (For the first (3.1), and the third (3.3) run the context window was constituted by the sentence that contains the word to disambiguate; in the second run the context window was constituted by the sentence that contains the word to disambiguate, the previous sentence and the next one).

3. Obtaining the domain vector, this vector is used in first and the second runs (when the lemma of the words in the analyzed sentence is obtained, the integrative resource of WordNet's Mappings is used to get the respective senses from each lemma).

4. Obtaining the all resource vector: SUMO, Affects, and Domain resource. This is only for the third run (3.3).

5. Relevant Semantic Tree construction (Addition of concepts parents to the vectors. For the first (3.1) and second (3.2) runs only Domain resource is used; for the third (3.3) run all the resources are used).

6. Selection of the correct senses (the first and the second runs use the same way to do the selection; the third run is different. We make an exception: For the verb "be" we select the sense with the higher frequency according to Freeling.

3.1 Relevant Semantic Tree

With this proposal we measure how much a concept is correlated to the sentence, similar to Reuters Vector (Magnini et al., 2002), but with a different equation. This proposal has a partial similarity with the Conceptual Density (Agirre and Rigau, 1996) and DRelevant (Vázquez et al., 2004) to get the concepts from a hierarchy that they associate with the sentence.

In order to determine the Association Ratio (RA) of a domain in relation to the sentence, the Equation 1 is used.

$$RA(D, f) = \sum_{i=1}^{n} RA(D, f_i) \qquad (1)$$

where:

$$RA(D, w) = P(D, w) * \log_2 \frac{P(D, w)}{P(D)} \qquad (2)$$

f: is a set of words w.

f_i: is a i-th word of the phrase f.

$P(D, w)$: is joint probability distribution.

$P(D)$: is marginal probability.

From now, vectors are created using the Senseval-2's corpus. Next, we show an example:

For the phrase: "But it is unfair to dump on teachers as distinct from the educational establishment".

By means of the process *Pres-processing* analyzed in previous stage 1 we get the lemma and the following vector.

Phrase [unfair; dump; teacher, distinct, educational; establishment]

Each lemma is looked for in WordNet's integrative resource of mappings and it is correlated with concepts of WND.

Vector	
RA	**Domains**
0.9	Pedagogy
0.9	Administration
0.36	Buildings
0.36	Politics
0.36	Environment
0.36	Commerce
0.36	Quality
0.36	Psychoanalysis
0.36	Economy

Table 1: Initial Domain Vector

After obtaining the Initial Domain Vector we apply the Equation 3 in order to build the Relevant Semantic Tree related to the phrase.

$$DN(CI, Df) = RA\_CI - \frac{MP(CI, Df)}{TD} \quad (3)$$

Where DN: is a normalized distance

CI: is the Initial Concept which you want to add the ancestors.

Df: is Parent Domain.

RA_CI: is a Association Ratio of the child Concept.

TD: is Depth of the hierarchic tree of the resource to use.

MP: is Minimal Path.

Applying the Equation 3 the algorithm to decide which parent domain will be added to the vector is shown here:

```
if $(DN(CI, Df) > 0)$
{
if ( $Df$ not exist)
    $Df$ is added to the vector with $DN$ value;
else
    $Df$ value = $Df$ value + $DN$;
}
```

As a result the Table 2 is obtained.

This vector represents the Domain tree associated to the phrase.

Vector	
RA	**Domains**
1.63	Social_Science
0.9	Administration
0.9	Pedagogy
0.8	RootDomain
0.36	Psychoanalysis
0.36	Economy
0.36	Quality
0.36	Politics
0.36	Buildings
0.36	Commerce
0.36	Environment
0.11	Factotum
0.11	Psychology
0.11	Architecture
0.11	Pure_Science

Table 2: Final Domain Vector

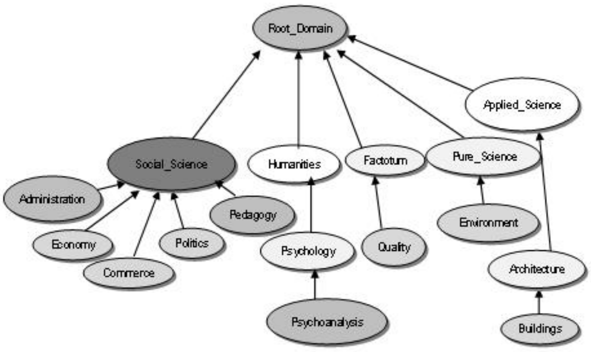

Figure 2: Relevant semantic tree

After the Relevant Semantic Tree is obtained, the Domain Factotum is eliminated from the tree. Due to the large amount of WordNet synsets, that do not belong to a specific domain, but rather they can appear in almost all of them, the Factotum domain has been created. It basically includes two types of synsets: Generic synsets, which are hard to classify in a particular domain; and Stop Senses synsets which appear frequently in different contexts, such as numbers, week days, colors, etc. (Magnini and Cavaglia, 2000), (Magnini et al., 2002). Words that contain this synsets are frequently in the phrases, therefore the senses associated to this domain are not selected.

After processing the patterns that characterize the sentence, the following stage is to determine the correct senses, so that the next steps ensue:

1. Senses that do not coincide with the grammatical category of Freeling are removed.

2. For each word to disambiguate all candidate senses are obtained. Of each sense the relevant vector are obtained using the Equation 4, and according to the previous Equation 3 parent concepts are added.

$$RA(D, s) = P(D, s) * \log_2 \frac{P(D, s)}{P(D)} \quad (4)$$

   where s: is a sense of word.

   $P(D, s)$: is joint probability distribution between Domain concept D and the sense s.

   $P(D)$: is marginal probability of the Domain concept.

3. The one that accumulates the bigger value of relevance is assigned as correct sense. The following process is applied:

   For each coincidence of the elements in the senses' domain vector with the domain vector of the sentence, the RA value of the analyzed elements is accumulated. The process is described in the Equation 5.

$$AC(s, VRA) = \frac{\sum_k VRA[Vs^k]}{\sum_{i=1} VRA_i} \quad (5)$$

   where AC: The RA value accumulated for the analyzed elements.

   VRA: Vector of relevant domains of the sentence with the format: VRA [domain — value RA].

   Vs: Vector of relevant domain of the sense with the format: Vs [domain].

   Vs_k: Is a k-th domain of the vector Vs.

   $VRA[Vs^k]$: Represents the value of RA assigned to the domain Vs^k for the value VRA.

   The $\sum_{i=1} VRA_i$ term normalizes the result.

3.2 Relevant Semantic Tree 2

This run is the same as the first one with a little difference, the context window is constituted by the sentence that contains the word to disambiguate, the previous sentence and the next one.

3.3 Adaptation of k-clique's technique to the WSD

They are applied, of the section 3, the steps from the 1 to the 5, where the semantic trees of concepts are obtained.

Then they are already obtained for all the words of the context, all the senses discriminated according to Freeling (Atserias et al., 2006).

Then a sentence's net of knowledge is built by means of minimal paths among each sense and each concept at trees. Next the k-clique's technique is applied to the net of knowledge to obtain cohesive subsets of nodes.

To obtain the correct sense of each word it is looked, as proposed sense, the sense belonging to the subset containing more quantities of nodes and if it has more than a sense for the same word, the more frequent sense is chosen according to Freeling.

4 Results and Discussion

The conducted experiments measure the influence of the aforementioned resources in the disambiguation task. We have evaluated them individually and as a whole. In the Table 3 it is represented each one of the inclusions and combinations experimented with the Relevant Semantic Tree method.

Resources	Precision	Recall	Attempted
WNAffect	0.242	0.237	97.78%
SUMO	0.267	0.261	98.5%
WND	0.328	0.322	98.14%
WND & SUMO	0.308	0.301	97.78%
WND & SUMO & WNAffect	0.308	0.301	97.78%

Table 3: Evaluation of integrated resources

As it can be observed, in the evaluation for specific domain corpus the best results are reached when only domain resource is used. But this is not a conclusion about the resources inclusion because the use of this method for global domain, for example with the task English All words from Senseval-2 (Agirre et al., 2010), the experiment adding all the resources showed good results. This is due to the fact that the global domain includes information of different contexts, exactly what is representing in the mentioned resources. For this reason, in the experiment with global domain and the inclusion of all the resource obtained better results than using this method with specific domain, 42% of recall and 45% of precision (Gutiérrez, 2010).

For example, with the k-clique's technique, utilizing the English All word task from Senseval-2´s corpus, the results for the test with global dominion were: with single domain inclusion 40 % of precision and recall; but with the three resources 41.7 % for both measures.

Table 4 shows the obtained results for the test data set. The average performance of our system is 32% and we ranked on 27-th position from 27 participating systems. Although, we have used different sources of information and various approximations, in the future we have to surmount a number of obstacles.

One of the limitations comes from the usage of the POS-tagger Freeling which introduces some errors in the grammatical discrimination. Representing a loss of 3.7% in the precision of our system.

The base of knowledge utilized in the task was WordNet 1.6; but the competition demanded the results with WordNet 3.0. In order to achieve this we utilized mappings among versions where 119 of 1398 resulting senses emitted by Semeval-2 were did not found. This represents an 8.5%.

In our proposal, the sense belonging to the Factotum Domain was eliminated, what disabled that the senses linked to this domain went candidates to be recovered. 777 senses of 1398 annotated like correct for Semeval-2 belong to domain Factotum, what represents that the 66% were not recovered by our system. Considering the senses that are not correlated to Factotum, that is, that correlate to another domains, we are speaking about 621 senses to define; The system would emit results of a 72,4%. Senses selected correctly were 450, representing a 32%. However, 189 kept on like second candidates to be elected. This represents a 13.5%. If a technique of more precise decision takes effect, the results of the system could be increased largely.

5 Conclusion and future works

For our participation in the Semeval-2 task 17 (All-words Word Sense Disambiguation on Specific Domain), we presented three methods for disambiguation approach which uses an integrative resource of WordNet mappings. We conducted an experimental study with the trail data set, according to which the Relevant Semantic Tree reaches the best performance. Our current approach can be improved with the incorporation of more granularities in the hierarchy of WordNet Domains. Because it was demonstrated that to define correct senses associated to specific domains an improvement of 72.4% is obtained. At this moment, only domain information is used in our first and second method. Besides was demonstrated for specific domains, the inclusion of several resources worsened the results with the first and second proposal method, the third one has been not experimented yet. Despite the fact that we have knowledge of SUMO, WordNet-Affect and WordNet Domain in our third method we still not obtain a relevant result.

Methods	Precision	Recall	Attempted
Relevant Domains Tree	0.328	0.322	98.14%
Relevant Semantic Tree 2	0.321	0.315	98.14%
Relevant Cliques	0.312	0.303	97.35%

Table 4: Evaluation results

It would be convenient to enrich our resource with other resources like Frame-Net, Concept-Net or others with the objective of characterizing even more the senses of the words.

Acknowledgments

This paper has been supported partially by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educació - Generalitat Valenciana (grant no. PROMETEO/2009/119 and ACOMP/2010/288).

References

Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistic (COLING´96)*, Copenhagen, Denmark.

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-kai Hsieh, Maurizio Tesconi, Monica

Monachini, Piek Vossen, and Roxanne Segers. 2010. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010), Association for Computational Linguistics.*

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA.*

Luís Cavique, Armando B. Mendes, and Jorge M. Santos. 2009. An algorithm to discover the k-clique cover in networks. In *EPIA '09: Proceedings of the 14th Portuguese Conference on Artificial Intelligence*, pages 363–373, Berlin, Heidelberg. Springer-Verlag.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Yoan Gutiérrez. 2010. Resolución de ambiguedad semántica mediante el uso de vectores de conceptos relevantes.

Zornitsa Kozareva, Sonia Vázquez, and Andrés Montoyo. 2007. Ua-zsa: Web page clustering on the basis of name disambiguation. In *Semeval I. 4th International Wordshop on Semantic Evaluations*.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating subject field codes into wordnet. In *Proceedings of Third International Conference on Language Resources and Evaluation (LREC-2000)*.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. Comparing ontology-based and corpus-based domain annotations in wordnet. In *Proceedings of the First International WordNet Conference*, pages 146–154.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *FOIS*, pages 2–9.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *IKE*, pages 412–416.

Ro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Sonia Vázquez, Andrés Montoyo, and German Rigau. 2004. Using relevant domains resource for word sense disambiguation. In *IC-AI*, pages 784–789.