SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

# Word Sense Disambiguation based on Term to Term Similarity in a Context Space

**Javier Artiles**
Dpto. Lenguajes y
Sistemas Informáticos
UNED, Spain
javart@bec.uned.es

**Anselmo Peñas**
Dpto. Lenguajes y
Sistemas Informáticos
UNED, Spain
anselmo@lsi.uned.es

**Felisa Verdejo**
Dpto. Lenguajes y
Sistemas Informáticos
UNED, Spain
felisa@lsi.uned.es

## Abstract

This paper describes the exemplar based approach presented by UNED at Senseval-3. Instead of representing contexts as bags of terms and defining a similarity measure between contexts, we propose to represent terms as bags of contexts and define a similarity measure between terms. Thus, words, lemmas and senses are represented in the same space (the context space), and similarity measures can be defined between them. New contexts are transformed into this representation in order to calculate their similarity to the candidate senses. We show how standard similarity measures obtain better results in this framework. A new similarity measure in the context space is proposed for selecting the senses and performing disambiguation. Results of this approach at Senseval-3 are here reported.

## 1 Introduction

Word Sense Disambiguation (WSD) is the task of deciding the appropriate sense for a particular use of a polysemous word, given its textual or discursive context. A previous non trivial step is to determine the inventory of meanings potentially attributable to that word. For this reason, WSD in Senseval is reformulated as a classification problem where a dictionary becomes the class inventory. The disambiguation process, then, consists in assigning one or more of these classes to the ambiguous word in the given context. The Senseval evaluation forum provides a controlled framework where different WSD systems can be tested and compared.

Corpus-based methods have offered encouraging results in the last years. This kind of methods profits from statistics on a training corpus, and Machine Learning (ML) algorithms to produce a classifier. Learning algorithms can be divided in two main categories: Supervised (where the correct answer for each piece of training is provided) and Unsupervised (where the training data is given without any answer indication). Tests at Senseval-3 are made in various languages for which two main tasks are proposed: an all-words task and a lexical sample task. Participants have available a training corpus, a set of test examples and a sense inventory in each language. The training corpora are available in a labelled and a unlabelled format; the former is mainly for supervised systems and the latter mainly for the unsupervised ones.

Several supervised ML algorithms have been applied to WSD (Ide and Véronis, 1998), (Escudero et al., 2000): Decision Lists, Neural Networks, Bayesian classifiers, Boosting, Exemplar-based learning, etc. We report here the exemplar-based approach developed by UNED and tested at the Senseval-3 competition in the lexical sample tasks for English, Spanish, Catalan and Italian.

After this brief introduction, Sections 2 and 3 are devoted, respectively, to the training data and the processing performed over these data. Section 4 characterizes the UNED WSD system. First, we describe the general approach based on the representation of words, lemmas and senses in a Context Space. Then, we show how results are improved by applying standard similarity measures as cosine in this Context Space. Once the representation framework is established, we define the criteria underlying the final similarity measure used at Senseval-3, and we compare it with the previous similarity measures. Section 5 reports the official results obtained at the Senseval-3 Lexical Sample tasks for English, Spanish, Italian and Catalan. Finally, we conclude and point out some future work.

## 2 Data

Each Lexical Sample Task has a relatively large training set with disambiguated examples. The test examples set has approximately a half of the number of the examples in the training data. Each example offers an ambiguous word and its surrounding context, where the average context window varies from language to language. Each training example gives one or more semantic labels for the ambiguous word corresponding to the correct sense in that context.

Senseval-3 provided the training data and the test data in XML format. The XML tagging conventions provides an excellent ground for the corpora processing, allowing a simple way for the data browsing and transformation. However, some of the XML well-formedness constraints are not completely satisfied. For example, there is no XML declaration and no root element in the English Lexical Sample documents. Once these shortcomings are fixed any XML parser can normally read and process the data.

Despite the similarity in the structure of the different corpora at the lexical sample task in different languages, we had found a heterogeneous vocabulary both in the XML tags and the attributes, forcing to develop 'ad hoc' parsers for each language. We missed a common and public document type definition for all the tasks.

Sense codification is another field where different solutions had been taken. In the English corpus nouns and adjectives are annotated using the WordNet 1.7.1. classification[1] (Fellbaum, 1998), while the verbs are based on Wordsmyth[2] (Scott, 1997). In the Catalan and Spanish tasks the sense inventory gives a more coarse-grained classification than WordNet. Both tasks have provided a dictionary with additional information as examples, typical collocations and the equivalent synsets at WordNet 1.5. Finally, the Italian sense inventory is based on the MultiWordnet dictionary[3] (Pianta et al., 2002). Unlike the other mentioned languages , the Italian task doesn't provide a separate file with the dictionary.

Besides the training data provided by Senseval, we have used the SemCor (Miller et al., 1993) collection in which every word is already tagged in its part of speech, sense and synset of WordNet.

## 3 Preprocessing

A tokenized version of the Catalan, Spanish and Italian corpora has been provided. In this version every word is tagged with its lemma and part of speech tag. This information has been manually annotated by human assessors both in the Catalan and the Spanish corpora. The Italian corpus has been processed automatically by the TnT POStagger[4] (Brants, 2000) including similar tags.

The English data lacked of this information, leading us to apply the TreeTagger[5] (Schmid, 1994) tool to the training and test data as a previous step to the disambiguation process.

Since the SemCor collection is already tagged, the preprocessing consisted in the segmentation of texts by the paragraph tag, obtaining 5382 different fragments. Each paragraph of Semcor has been used as a separate training example for the English lexical sample task. We applied the mapping provided by Senseval to represent verbs according to the verb inventory used in Senseval-3.

## 4 Approach

The supervised UNED WSD system is an exemplar based classifier that performs the disambiguation task measuring the similarity between a new instance and the representation of some labelled examples. However, instead of representing contexts as bags of terms and defining a similarity measure between the new context and the training contexts, we propose a representation of terms as bags of contexts and the definition of a similarity measure between terms. Thus, words, lemmas and senses are represented in the same space, where similarity measures can be defined between them. We call this space the Context Space. A new disambiguation context (bag of words) is transformed into the Context Space by the inner product, becoming a kind of abstract term suitable to be compared with singular senses that are represented in the same Context Space.

### 4.1 Representation

The training corpus is represented in the usual two-dimension matrix A as shown in Figure 1, where

- $c_1, ..., c_N$ is the set of examples or contexts in the training corpus. Contexts are treated as bags of words or lemmas.
- $lem_1, ..., lem_T$ is the set of different words or lemmas in all the training contexts.

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

[2] http://www.wordsmyth.net

[3] http://multiwordnet.itc.it/

[4] http://www.coli.uni-sb.de/ thorsten/tnt/

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

- $w_{i,j}$ is the weight for $lem_i$ in the training context $c_j$.

A new instance q, represented with the vector of weights $(w_{1q}, ..., w_{iq}, ..., w_{Tq})$, is transformed into a vector in the context space $\vec{q} = (q_1, ..., q_j, ..., q_N)$, where $\vec{q}$ is given by the usual inner product $\vec{q} = q \cdot A$ (Figure 1):

$$q_j = \sum_{i=1}^{T} w_{iq} w_{ij}$$

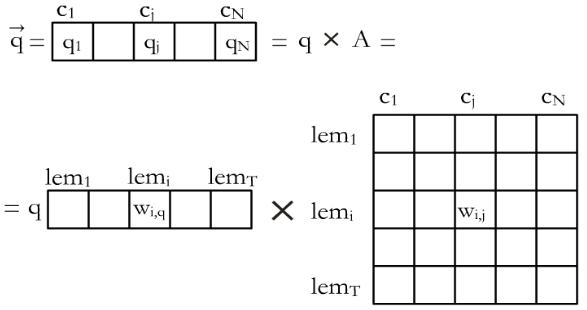

Figure 1: Representation of terms in the Context Space, and transformation of new instances.

If vectors $c_j$ (columns of matrix A) and vector $q$ (original test context) are normalized to have a length equal to 1, then $q_j$ become the cosine between vectors $q$ and $c_j$. More formally,

$$\vec{q} = q.A = (cos(q, c_1), ..., cos(q, c_j), ..., cos(q, c_N))$$

where

$$cos(q, c_j) = \sum_{i=1}^{T} \frac{w_{iq}}{\|q\|} \frac{w_{ij}}{\|c_j\|}$$

and

$$\|x\| = \sqrt{\sum_i x_i^2}$$

At this point, both senses and the representation of the new instance $\vec{q}$ are represented in the same context space (Figure 2) and a similarity measure can be defined between them:

$$sim(s\vec{en}_{ik}, \vec{q})$$

where $sen_{ik}$ is the $k$ candidate sense for the ambiguous lemma $lem_i$. Each component $j$ of $s\vec{en}_{ik}$ is set to 1 if lemma $lem_i$ is used with sense $sen_{ik}$ in the training context $j$, and 0 otherwise.

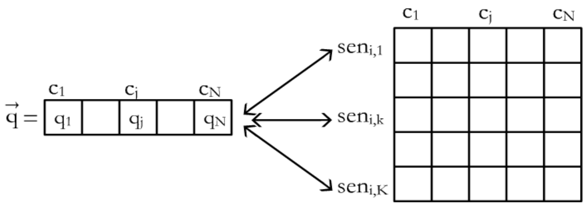

Figure 2: Similarity in the Context Space.

For a new context of the ambiguous lemma $lem_i$, the candidate sense with higher similarity is selected:

$$argmax_k \quad sim(s\vec{en}_{ik}, \vec{q})$$

### 4.2 Bag of words versus bag of contexts

Table 1 shows experimental results over the English Lexical Sample test of Senseval-3. System has been trained with the Senseval-3 data and the SemCor collection. The Senseval training data has been lemmatized and tagged with TreeTagger. Only nouns and adjectives have been considered in their canonical form.

Three different weights $w_{ij}$ have been tested:

- Co-occurrence: $w_{ij}$ and $w_{iq}$ are set to {0,1} depending on whether $lem_i$ is present or not in context $c_j$ and in the new instance q respectively. After the inner product $q \cdot A$, the components $q_j$ of $\vec{q}$ get the number of co-occurrences of different lemmas in both q and the training context $c_j$.
- Term Frequency: $w_{ij}$ is set to $tf_{ij}$, the number of occurrences of $lem_i$ in the context $c_j$.
- $tf.idf$: $w_{ij} = (1 + \log(tf_{ij})) \cdot (\log(\frac{N}{df_i}))$, a standard $tf.idf$ weight where $df_i$ is the number of contexts that contain $lem_i$.

These weights have been normalized $(\frac{w_{ij}}{||c_j||})$ and so, the inner product $q \cdot A$ generates a vector $\vec{q}$ of cosines as described above, where $q_j$ is the cosine between $q$ and context $c_j$.

Two similarity measures have been compared. The first one (maximum) is a similarity of $q$ as bag of words with the training contexts of sense $sen$. The second one (cosine) is the similarity of sense $sen$ with $\vec{q}$ in the context space:

- Maximum: $sim(s\vec{en}, \vec{q}) =$
$= Max_{j=1}^{N} \quad (sen_j \cdot q_j) =$
$= Max_{\{j/sen \in c_j\}} q_j =$
$= Max_{\{j/sen \in c_j\}} cos(q, c_j)$

| Weight | Similarity | Nouns | Adjectives | Verbs | Total |
|---|---|---|---|---|---|
| Co-occurrences | Maximum | **60.76%** | 35.85% | 60.75% | 59.75% |
| (normalized) | Cosine | 59.99% | **55.97%** | 63.88% | **61.78%** |
| Term frequency | Maximum | 56.83% | 50.31% | 56.85% | 56.58% |
| (normalized) | Cosine | **60.76%** | 53.46% | 63.83% | **62.01%** |
| tf.idf | Maximum | 59.82% | 48.43% | 59.94% | 59.42% |
| (normalized) | Cosine | 60.27% | 53.46% | **64.29%** | **62.01%** |
|  | Most frequent<br>(baseline) | 54.01% | 54.08% | 56.45% | 55.23% |

Table 1: Bag of words versus bag of contexts, precision-recall

Similarity with sense *sen* is the highest similarity (cosine) between $q$ (as bag of words) and each of the training contexts (as bag of words) for sense *sen*.

- Cosine: $sim(\vec{sen},\vec{q}) = cos(\vec{sen},\vec{q}) =$ $= \sum_{\{j/sen\in c_j\}} \frac{sen_j}{||\vec{sen}||} \cdot \frac{cos(q,c_j)}{||\vec{q}||}$

  Similarity with sense *sen* is the cosine in the Context Space between $\vec{q}$ and $\vec{sen}$

Table 1 shows that almost all the results are improved when the similarity measure (cosine) is applied in the Context Space. The exception is the consideration of co-ocurrences to disambiguate nouns. This exception led us to explore an alternative similarity measure aimed to improve results over nouns. The following sections describe this new similarity measure and the criteria underlying it.

### 4.3 Criteria for the similarity measure

Co-occurrences behave quite good to disambiguate nouns as it has been shown in the experiment above. However, the consideration of co-occurrences in the Context Space permits acumulative measures: Instead of selecting the candidate sense associated to the training context with the maximum number of co-occurrences, we can consider the co-occurences of q with all the contexts. The weights and the similarity function has been set out satisfying the following criteria:

1. Select the sense sen$_k$ assigned to more training contexts c$_i$ that have the maximum number of co-occurrences with the test context q. For example, if sense $sen_1$ has two training contexts with the highest number of co-occurrences and sense $sen_2$ has only one with the same number of co-occurrences, $sen_1$ must receive a higher value than $sen_2$.
2. Try to avoid label inconsistencies in the training corpus. There are some training examples where the same ambiguous word is used with the same meaning but tagged with different sense by human assessors. Table 2 shows an example of this kind of inconsistencies.

### 4.4 Similarity measure

We assign the weights $w_{ij}$ and $w_{iq}$ to have $\vec{q}$ a vector of co-occurrences, where $q_j$ is the number of different nouns and adjectives that co-occurr in q and the training context $c_j$. In this way, $w_{ij}$ is set to 1 if $lem_i$ is present in the context $c_j$. Otherwise $w_{ij}$ is set to 0. Analogously for the new instance q, $w_{iq}$ is set to 1 if $lem_i$ is present in q and it is set to 0 otherwise.

According to the second criterium, if there is only one context $c_1$ with the higher number of co-occurrences with $q$, then we reduce the value of this context by reducing artificially its number of co-occurrences: Being $c_2$ a context with the second higher number of co-occurrences with $q$, then we assign to the first context $c_1$ the number of co-occurrences of context $c_2$.

After this slight modification of $\vec{q}$ we implement the similarity measure between $\vec{q}$ and a sense $sen_k$ according to the first criterium:

$$sim(\vec{sen},\vec{q}) = \sum_{j=1}^{N} sen_j \cdot N^{q_j}$$

Finally, for a new context of $lem_i$ we select the candidate sense that gives more value to the similarity measure:

$$argmax_k \quad sim(\vec{sen}_k,\vec{q})$$

| <answer instance=”grano.n.1” senseid=”grano.4”/> |
|---|
| <previous> La Federacin Nacional de Cafeteros de Colombia explic que el nuevo valor fue establecido con base en el menor de los precios de reintegro mnimo de grano del pas de los ltimos tres das, y que fue de 1,3220 dlares la libra, que fue el que alcanz hoy en Nueva York, y tambin en la tasa representativa del mercado para esta misma fecha (1.873,77 pesos por dlar). </previous> <target> El precio interno del caf colombiano permaneci sin modificacin hasta el 10 de noviembre de 1999, cuando las autoridades cafetaleras retomaron el denominado ”sistema de ajuste automtico”, que tiene como referencia la cotizacin del <head>grano</head> nacional en los mercados internacionales. </target> |
| <answer instance=”grano.n.9” senseid=”grano.3”/> |
| <previous> La carga qued para maana en 376.875 pesos (193,41 dlares) frente a los 375.000 pesos (192,44 dlares) que rigi hasta hoy. </previous> <target> El reajuste al alza fue adoptado por el Comit de Precios de la Federacin que fijar el precio interno diariamente a partir de este lunes tomando en cuenta la cotizacin del <head>grano</head> en el mercado de Nueva York y la tasa de cambio del da, que para hoy fueron de 1,2613 dlares libra y1.948,60 pesos por dlar </target> |

Table 2: Example of inconsistencies in human annotation

| Weight | Similarity | Nouns | Adjectives | Verbs | Total |
|---|---|---|---|---|---|
| Co-occurrences | Without criterium 2 | 65.6% | 45.9% | 62.5% | 63.3% |
| (not normalized) | With criterium 2 | 66.5% | 45.9% | 63.4% | 64.1% |

Table 3: Precision-recall for the new similarity measure

Table 3 shows experimental results over the English Lexical Sample test under the same conditions than experiments in Table 1.

Comparing results in both tables we observe that the new similarity measure only behaves better for the disambiguation of nouns. However, the difference is big enough to improve overall results. The application of the second criterium (try to avoid label inconsistencies) also improves the results as shown in Tables 3 and 4. Table 4 shows the effect of applying this second criterium to all the languages we have participated in. With the exception of Catalan, all results are improved slightly (about 1%) after the filtering of singular labelled contexts. Although it is a regular behavior, this improvement is not statistically significative.

|  | With<br>Criterium 2 | Without<br>Criterium 2 |
|---|---|---|
| Spanish | 81.8% | 80.9% |
| Catalan | 81.8% | 82.0% |
| English | 64.1% | 63.3% |
| Italian | 49.8% | 49.3% |

Table 4: Incidence of Criterium 2, precision-recall

## 5 Results at Senseval-3

The results submited to Senseval-3 were generated with the system described in Section 4.4. Since one sense is assigned to every test context, precison and recall have equal values. Table 4 shows official results for the Lexical Sample Task at Senseval-3 in the four languages we have participated in: Spanish, Catalan, English and Italian.

|  | Fine<br>grained | Coarse<br>grained | Baseline<br>(most frequent) |
|---|---|---|---|
| Spanish | 81.8% | - | 67% |
| Catalan | 81.8% | - | 66% |
| English | 64.1% | 72% | 55% |
| Italian | 49.8% | - | - |

Table 5: Official results at Senseval-3, precision-recall

Differences between languages are quite remarkable and show the system dependence on the training corpora and the sense inventory.

In the English task, 16 test instances have a correct sense not present in the training corpus. Since we don't use the dictionary information our system was unable to deal with none of them. In the same way, 68 test instances have been tagged as ”Unasignable” sense and again the system was unable to detect none of them.

## 6 Conclusion and work in progress

We have shown the exemplar-based WSD system developed by UNED for the Senseval-3 lexical sample tasks. The general approach is based on the definition of a context space that becomes a flexible tool to prove quite different similarity measures between training contexts and new instances. We have shown that standard similarity measures improve their results applied inside this context space. We have established some criteria to instantiate this general approach and the resulting system has been evaluated at Senseval-3. The new similarity measure improves the disambiguation of nouns and obtains better overall results. The work in progress includes:

- the study of new criteria to lead us to alternative measures,
- the development of particular disambiguation strategies for verbs, nouns and adjectives,
- the inclusion of the dictionary information, and
- the consideration of WordNet semantic relationships to extend the training corpus.

## Acknowledgements

Special thanks to Julio Gonzalo for the lending of linguistic resources, and to Víctor Peinado for his demonstrated sensibility.

This work has been supported by the Spanish Ministry of Science and Technology through the following projects:

- Hermes (TIC-2000-0335-C03-01)
- Syembra (TIC-2003-07158-C04-02)
- R2D2 (TIC 2003-07158-104-01)

## References

Thorsten Brants. 2000. Tnt - a statistical part-of-speech tagger. In *In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*.

G. Escudero, L. Màrquez, and G. Rigau. 2000. A comparison between supervised learning algorithms for word sense disambiguation. In *In Proceedings of the 4th Computational Natural Language Learning Workshop, CoNLL*.

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

N. Ide and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*.

G. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance. In *In Procedings of the 3rd DARPA Workshop on Human Language Technology*.

E. Pianta, L. Bentivogli, and C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.

M. Scott. 1997. Wordsmith tools lexical analysis software for data driven learning and research. Technical report, The University of Liverpool.