Computación y Sistemas Vol. 5 No.4 pp. 288 - 295
© 2002, CIC - IPN. *ISSN 1405-5546* Impreso en México

PH. D. THESIS ABSTRACT

Domain Membership Degrees and Classification Methods

Grados de Pertenencia a Dominios y Métodos de Clasificación

Graduated: Héctor Jiménez
Facultad de Ciencias de la Computación
Universidad Autónoma de Puebla
e-mail: hjimenez@cfm.buap.mx
Advisor: Guillermo Morales
Sección de Computación
Centro de Investigación y Estudios Avanzados-IPN
e-mail: gmorales@cs.cinvestav.mx

Abstract

We will assume that a term pertains to a domain or sublanguage if the term is "relevant" to the sublanguage. We introduce two methods to determine the membership degree of a term. In the first one, each term is just a word, while in the second it is a simple noun phrase. We will discuss the possibilities to apply our methods into phrase classification. In order to determine the membership degree we deal with text tagging with parts-of-speech and nominal phrases identification. We will show our approaches to solve both problems. All our methods and results are relative and have been applied to the Spanish language, and more particularly to the Mexican Spanish.

Keywords:

Phrase classification, corpus-based tagging, sublanguage of a domain, machine learning.

Resumen

La pertenencia de un término a un dominio o sublenguaje significa que el concepto asociado al término es "importante" dentro del sublenguaje. Presentamos dos métodos para determinar la pertenencia. En uno se considera al término como una palabra, y en el otro como una frase nominal simple. El estudio realizado mostró buenas perspectivas de aplicación de estos métodos en la clasificación de frases. En este reporte se presenta, además, las herramientas necesarias que fueron desarrolladas para apoyar a la solución del problema de pertenencia: el etiquetamiento de un texto con partes del discurso, y la identificación de frases nominales.

Palabras clave:

Clasificación de frases, etiquetamiento basado en corpus, sublenguaje de un dominio, aprendizaje automático.

1 Introduction

The quality of information is an important matter in our current life and its processing makes extremely necessary the use of powerful and efficient tools. The WWW, for instance, is a great information source but it does not have had an exploitation rate equivalent to its growth (Baeza-Yates, 1998).

A quite relevant component in the improvement of information quality, is the processing of non-structured information, either graphical or textual. In order to fulfill the textual information, several approaches have been developed: *text categorization, information extraction* and *summarization* (Smeaton, 1997). It is not just a matter of traditional *information retrieval*, instead semantical resources are essential in the processing. Besides *full text understanding*, some alternative successful methods have been used (Grishman, 1997).

The *context* (understood as the neighborhood of a word, or the structure that contains a phrase, or the domain of an output) plays an important role in classification and disambiguation. The main problem consists in classifying a linguistic entity into a domain. We address to the problem of such a classification for words and noun phrases, with a certain membership degree.

Our methods, obviously, depend on the available information resources concerning the NLP for Spanish in general, and for Mexican Spanish of our particular interest. We developed some tools for particular syntactical analysis as well as for tagging.

In the next two sections of this paper we present our methods for tagging and the determination of membership degrees of words and noun phrases to a given domain. In the fourth section we present an application oriented towards phrase classification.

2 Tagging

Text tagging is a procedure used quite often in NLP. In particular, it consists in the association of each word in a discourse with its own *part-of-speech* (POS). Several methods are reported in the literature. We have followed the so called *Memory-Based Learning* (MBL) (Daelemans, 1995) approach.

2.1 Memory-based Learning

MBL is a *supervised* learning method that uses a collection of instances to classify any new instance. The class assigned to any new instance X is the class of the most likely instance.

More precisely, let $\mathcal{A} = \{A_1, \ldots, A_m\}$ be a collection of attributes, for each $A \in \mathcal{A}$ let $D(A)$ be its domain and let $U = \prod_{A \in \mathcal{A}} D(A)$ be the universe of instances.

Given any set $S \subset U$ (supposed later to be of *training instances*), for any attribute $A_i \in \mathcal{A}$ and any value in its domain $a \in D(A_i)$, let $S_{A_i \leftarrow a} = \{X = (x_1, \ldots, x_m) \in S | x_i = a\}$ be the collection of instances in S that have a as their i-th value, and let $\pi_a(A) = \frac{\#(S_{A \leftarrow a})}{\#(S)}$ be the proportion of instances with the a value for the attribute A. With respect to a given collection of classes $\mathcal{C} = \{C_1, \ldots, C_n\}$, with $C \subset U$, the *entropy* of S is

$$info_{\mathcal{C}}(S) = -\sum_{j=1}^{n} \text{freq}_j(S) \cdot \log_2 \text{freq}_j(S) \qquad (1)$$

where $\text{freq}_j(S) = \frac{\#(S \cap C_j)}{\#(S)}$ is the "relative frequency of elements in S that fall in class C_j". For each attribute $A \in \mathcal{A}$ the *information gain* of S, relative to A and $\mathcal{C}$, is

$$gain_{\mathcal{C},A}(S) = info_{\mathcal{C}}(S) - \sum_{a \in D(A)} \pi_a(A) \cdot info_{\mathcal{C}}(S_{A \leftarrow a}) \qquad (2)$$

Also, let's define the following *normalized* parameters:

$$split\ info_A(S) = -\sum_{a \in D(A)} \pi_a(A) \log_2 \pi_a(A) \qquad (3)$$

$$gain\ ratio_{\mathcal{C},A}(S) = \frac{gain_{\mathcal{C},A}(S)}{split\ info_A(S)} \qquad (4)$$

(observe that $split\ info_A(S)$ does not depend on the partition $\mathcal{C}$).

Suppose fixed a set of training instances S and a current partition $\mathcal{C}$ of classes. Given two instances $X = (x_1, \ldots, x_m)$, $Y = (y_1, \ldots, y_m) \in U$ let $\Delta(X,Y) = \sum_{i=1}^{m} w_i \cdot \bar{\delta}(x_i, y_i)$, where the vector of weights $\mathbf{w} = (w_1, \ldots, w_m) \in (\mathbb{R}^+)^m$ can be chosen with several criteria as explained above and $\bar{\delta}$ is a *complementary Kroenecker delta*: $\bar{\delta}(a,b) = \begin{cases} 0 & \text{if } a = b, \\ 1 & \text{if } a \neq b. \end{cases}$ $(\Delta : (X,Y) \mapsto \Delta(X,Y)$ is a distance function realized as a *weighted average* of the "discrepancies" on attributes.) For any $X \in U$ let us denote by $\text{argmin}_{Y \in S} \Delta(X,Y)$ any element in S that minimizes the function $Y \mapsto \Delta(X,Y)$ on S: $Y_0 = \text{argmin}_{Y \in S} \Delta(X,Y) \Leftrightarrow Y_0 \in S \ \& \ \forall Y \in S : \Delta(X,Y_0) \leq \Delta(X,Y)$. Hence, given any new instance X, we will classify it in the same class as was classified $\text{argmin}_{Y \in S} \Delta(X,Y)$.

The weights vector $\mathbf{w}$ can be selected with either of the next criteria:

Information gain Let $\mathbf{w} = \left(gain_{\mathcal{C},A}(S)\right)_{A \in \mathcal{A}}$ with its components defined as in eq. (2).

Normalized gain Let $\mathbf{w} = \left(gain\ ratio_{\mathcal{C},A}(S)\right)_{A \in \mathcal{A}}$ with its components defined as in eq. (4).

Evidently, the evaluation of the weighted distance excludes ⟨*unknown features*⟩, i.e. attribute values not occurring in the corresponding domain. The technique of IGTree (Daelemans, Durieux, et. al. 1998) deals with this problem. The main advantages of this approach are the following:

- The training set is realized as an ordered decision tree, i.e. a *trie*: Each level in the *trie* is determined by an attribute in the set of instances. The root of the *trie* corresponds to the whole collection S of training instances, and any node that corresponds to a collection N may have an edge, labeled with a, towards a child that corresponds to $N_{A \leftarrow a}$, for some $a \in D(A)$. The set of attributes $\mathcal{A}$ is assumed ordered in accordance with the *gain ratio* values. This provides savings of search time and memory space.

- In the classification process, each feature of the new instance is matched with an edge in a depth-first way. The first time that there is no a matching the information corresponding to the node where the matching fails is used as default. This assumption provides the most likely class of the instances on the fail node.

Several taggers following this approach have been reported in the literature. For instance, in (Daelemans, van-den-Bosch, et. al. 1998) is reported 97.8% in accuracy when tagging a text of 89×10^3 words using 711×10^3 training instances. It is rather usual to use the tenth part of a tagged text as the training set.

Certainly, MBL is an efficient and simple method quite adequate for NLP classification tasks. In our approach, we have implemented a modification of the complementary Kroenecker delta (Jiménez, Morales, 2000) regarding that

- lacking of exact matching and ⟨unknown features⟩ is a common problem in real classification applications, and
- there is an implicit information in the training set that can be used to solve the above problem.

Indeed we may analyze also the context around an ⟨unknown feature⟩. Given any new instance $X = (x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_m)$ and an attribute index $i \in \{1, \ldots, m\}$, a *context of* x_i is a substring c_i of X around x_i: $X = X_1 * c_i * X_2$, for some possibly empty strings X_1, X_2. For any possible value y_i of the i-th attribute let $c_i(y_i)$ be the string obtained from c_i substituting x_i by y_i. For any $Y = (y_1, \ldots y_i, \ldots, y_m) \in S$ we shall estimate the probability that y_i appears in the context c_i of x_i: $p_i = Prob(c_i(y_i)|c_i)$. More precisely, let

$$\overline{\delta}'(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 - p_i & \text{if } x_i \notin D(A_i) \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

Let us proceed as in IGTree with a slight modification:

1. In case of an instance with "high" gain information whose current value is unknown with respect to the training set, then using $\overline{\delta}'$ we select the class from the closest instance,
2. else we proceed the classification as in IGTree.

2.2 Part of Speech Tagging

Regarding the frequency of most used words in written Mexican Spanish and the most common types of ambiguities we introduce the following four "super-classes". The first two are directly related to frequencies. The other two have a relevant presence in the corpus:

Frequent well-defined words. These are frequent words whose POS is invariant with respect to contexts (i.e. non-ambiguous frequent words). The tags of these words are determined by a dictionary and morphological rules. For instance articles and most prepositions fall in this class.

Frequent ambiguous words. The disambiguation of such words is realized through MBL using a training set extracted from the corpus.

Ambiguity verb/noun. These are words whose POS is either `Verb` or `Noun` or `Adjective`. As in last case, MBL is used. For instance, words as *ganas* which can be assumed as "you win" or "intentions to do something", *vista* which is either the past participle of *ver* or "sight".

Unknown words. This class is complementary to the union of the above classes. For this kind of words, we use MBL with IGTree provided of the distance $\overline{\delta}'$.

In our procedures based on MBL, we have assumed all instances consisting of the endings of neighbor words and their proper tags. We used extensively the following resources:

a. examples formed by sequences of pairs (-endings, POS) [they were extracted from a tagged discourse consisting of around 10 000 words published in the *Corpus del Español Mexicano Contemporáneo* collected by El Colegio de México (Lara, Ham-Chande, et. al. 1979)],

b. morphological rules based on the conditional probability of the occurrence of a given tag in a particular context,

c. dictionaries of high frequency words, and

d. lists of endings corresponding to conjugations of regular and irregular verbs.

We estimate the efficiency of the methods through the *precision* P and *recall* R indices: Both P and R are ratios whose numerator is the number of times that the assigned class coincides with the correct class, the denominator of P is the total number of objects to be classified while the denominator of R is the total number of objects classified previously in the super-classes introduced above. The *performance* index F involves P and R and is defined as $F = \frac{2PR}{P+R}$. Each index will be expressed as a percentual proportion.
In our tests, IGTree by itself got a performance value of 79.19%, while IGTree with $\overline{\delta}'$ got a performance of 81.5%. Particularly, when considering just the "Journalistic Genre" in the *Corpus*, which naturally corresponds to a well structured discourse scheme, the performance ratio was higher than 90% (Jiménez, Morales, 2002).

2.3 Noun Phrases Recognition

MBL was very useful in recognizing *simple noun phrases* (Jiménez, Morales, 1998). In this application the method decides whether there is the limit (or boundary) of a noun phrase between two words in the text. We proceed following the next algorithms:
Determination of an initial training set

1. Let us tag, in a manual form, the right bound (RB) of noun phrases in the corpus.

2. For each RB, both strings consisting of the five POS at the left of RB and five POS at the right of RB are considered as "positive" instances.

3. "Negative" instances are strings of length five which are not positive.

Attribute choice

1. Select the attributes that have greater information gains, according to the gains of the initial training set.

2. Substitute the current training set by the projection of the former training set onto the selected attributes.

A similar method is reported in (Veenstra, 1998) that uses a corpus in English extracted from *The Wall Street Journal*. The reported *accuracy* (i.e. when each word to be tagged is indeed tagged) is 97.2% using 203 711 examples of phrases when it was applied over a text containing 47 377 English noun phrases. In our method whenever we processed a text whose genre coincided with that of the training set we got a performance index of around 96% while in texts of different genre the index was around 88%, in spite that our training set was rather small: 1 241 phrases in Mexican Spanish (Jiménez, Morales, 1998).

3 Domain Membership

A *domain* includes the set of terms, i.e. words or phrases entailing objects, that appear in a language, but may have other elements in a non-crisp set theoretical notion. Let us precise the membership notion on domains. First of all let us introduce a convention on the domain and word representation.

Word representation. Let $\mathcal{V}$ be a vocabulary. We realize any *dictionary* as a function $D : \mathcal{V} \times \mathbb{N} \to \mathcal{M}$, $\mathcal{M} \subset \mathcal{V}^*$, that associates to each pair (w, i) the i-th *meaning* of word w. Since each word has just a finite number of meanings for all $w \in \mathcal{V}$ there is a minimal number $n_w = \text{nr\_mean}(w)$ such that $[n > n_w \Rightarrow D(w, n) = nil]$, where *nil* is the empty word in $\mathcal{V}^*$. For each word $w \in \mathcal{V}$ let $D_{-1}(w) = \{(v, i) \in \mathcal{V} \times \mathbb{N} | \exists i \in \mathbb{N} : w \text{ appears in } D(v, i)\}$ be the set of words for which there is a definition involving w. The relation $D_{-1} : \mathcal{V} \to 2^{\mathcal{V}}$ is a kind of *inverse dictionary* of D.

For any word $w \in \mathcal{V}$, the *context of w with respect to dictionary D* is

$$\mathcal{C}_D(w) = \bigcup_{(v,i) \in D_{-1}(w)} D(v, i) \qquad (6)$$

(in words: $\mathcal{C}_D(w)$ consists of any definitions which involve w).

Domain representation. Let us define procedurally this notion:

1. Let B_0 be a set of basic words in the domain, D.
2. Enrich B_0 by *lexical induction*: B.
3. Let B' the join of $\mathcal{C}_D(w)$, for each $w \in B$.

Second step is done using the notion of *mutual information* (Church, Hanks, 1990): Given a text T and two words w_1, w_2 let $Prob(w_1 w_2)$ be the probability of occurrence of string $w_1 w_2$ within the text and let $Prob(w_i)$, $i = 1, 2$, the probability of occurrence of w_i without the other word w_j. Clearly, if $Prob(w_1 w_2) > Prob(w_1), Prob(w_2)$ then a link between w_1 and w_2 should exist. Let

$$MI(w_1 w_2) = \log_2 \frac{Prob(w_1 w_2)}{Prob(w_1) Prob(w_2)} \qquad (7)$$

With respect to a threshold value $\lambda > 0$, if $MI(w_1 w_2) > \lambda$ then we will assume that the bigram $w_1 w_2$ is a *composed term*.

Let $T^{(2)}$ be the bigrams that occurs in the text T. The *lexical enrichment* procedure is the following (Gierl, Frost, 1992):

1. **Initial set of terms.** Let $L_0 = \{w_1 w_2 \in T^{(2)} | w_1, w_2 \in B_0 \ \& \ MI(w_1 w_2) > \lambda\}$ be the set of composed bigrams with components in the basic set B_0.
2. **Iterated set of terms.** For each $i > 0$, let $B_i = B_{i-1} \cup \{z | \exists w_1 w_2 \in L_{i-1} : z = w_1 \text{ or } z = w_2\}$ be the set of words appearing as components of bigrams in the current set L_{i-1} and let $L_i = \{w_1 w_2 \in T^{(2)} | w_1, w_2 \in B_i \ \& \ MI(w_1 w_2) > \lambda\}$ be the set of composed bigrams with components in the current set B_i.

In a first test, using a text related to the notion of "public health" in a public inquire surveyed between rural population in Puebla's valley in Mexico, B_0 consisted of 19 words of least rank extracted from a basic dictionary of health of 20 322 signs after suppression of the most frequent Spanish words. With $\lambda = 4$ and five iterations, the set of terms B (B_5) consisted of 102 elements, and looking at these terms we estimated an error less than 5%.

Let Dom be a domain and let w be a word. Let us denote by $pert(w, \text{Dom})$ the *membership degree* of w to domain Dom. We proceed to compare two approaches in order to calculate $pert(w, \text{Dom})$:

Using a metric. Dom and x are embedded into a metric space and we decide whether x is in the neighborhood of Dom.

Rough sets. We associate to Dom a function $\mu_{\text{Dom}} : \mathcal{V} \to [0,1]$ which is the membership function indeed.

3.1 The Metric Approach

Let us recall that $\mathcal{C}$ is the context function defined by eq. (6). For any two words $w_1, w_2 \in \mathcal{V}$ and for any set of words $J \subset \mathcal{V}$ let us define the following parameters:

Point distance.
$$\delta_{\mathcal{C}}(w_1, w_2) = \frac{\#(\mathcal{C}(w_1) \cap \mathcal{C}(w_2))}{\#(\mathcal{C}(w_1) \cup \mathcal{C}(w_2))}.$$

Minimal point-set distance.
$$\delta_{\mathcal{C}}^{min}(w_1, J) = \text{Min}\{\delta_{\mathcal{C}}(w_1, w_2) | w_2 \in J\}.$$

Average point-set distance.
$$\overline{\delta}_{\mathcal{C}}(w_1, J) = \frac{1}{\#J} \sum_{w_2 \in J} \delta_{\mathcal{C}}(w_1, w_2).$$

Normalized point-set distance.
$$\delta_{\mathcal{C}}^{norm}(w_1, J) = \delta_{\mathcal{C}}^{min}(w_1, J) \cdot \overline{\delta}_{\mathcal{C}}(w_1, J).$$

Set diameter.
$$Diam_{\mathcal{C}}^{max}(J) = \text{Max}\{\delta_{\mathcal{C}}(w_1, w_2) | w_1, w_2 \in J \ \& \ \delta_{\mathcal{C}}(w_1, w_2) \neq 1\}.$$

Average set diameter.
$$\overline{Diam_{\mathcal{C}}}(J) = \frac{1}{\#J - 1} \sum_{w \in J} \delta_{\mathcal{C}}^{norm}(w, J - \{w\}).$$

Normalized set diameter.
$$Diam_{\mathcal{C}}^{norm}(J) = Diam_{\mathcal{C}}^{max}(J) \cdot \overline{Diam_{\mathcal{C}}}(J).$$

Given a domain Dom, our classification criteria forms three groups: The words *included* in Dom, the words *close* to Dom and the words *outside* Dom:

Criterion M.1. $\delta_{\mathcal{C}}^{min}(w, \text{Dom}) \leq Diam_{\mathcal{C}}^{norm}(\text{Dom}) \Rightarrow$ w is *included* in Dom.

Criterion M.2. $\delta_{\mathcal{C}}^{min}(w, \text{Dom}) \leq Diam_{\mathcal{C}}^{max}(\text{Dom}) \Rightarrow$ w is *close* to Dom.

Criterion M.3. $\delta_{\mathcal{C}}^{min}(w, \text{Dom}) > Diam_{\mathcal{C}}^{max}(\text{Dom}) \Rightarrow$ w is *outside* Dom.

3.2 The Rough Sets Approach

Let us recall that for each word w, $\mathcal{C}(w)$ is the representation of w as defined at the beginning of this section. Two words are related with respect to the *tolerance* relation (Pawlak, 1982) $R_{\mathcal{C}}$ if their representations meet. I.e. $R_{\mathcal{C}}$ is defined as

$$\forall w_1, w_2 \in \mathcal{V}: \quad w_1 R_{\mathcal{C}} w_2 \Leftrightarrow \mathcal{C}(w_1) \cap \mathcal{C}(w_2) \neq \emptyset \quad (8)$$

$R_{\mathcal{C}}[w_1] = \{w_2 \in \mathcal{V} | w_1 R_{\mathcal{C}} w_2\}$ is the *image* of w_1 under relation $R_{\mathcal{C}}$.
For a set of words $V \subset \mathcal{V}$ contained in the vocabulary, let

$$\mu_V : \mathcal{V} \to [0,1], \quad w \mapsto \mu_V(w) = \frac{\#V \cap R_{\mathcal{C}}[w]}{\#R_{\mathcal{C}}[w]} \quad (9)$$

be the *rough set* associated to V, in other words, μ_V is the *membership* function of V.
Fixed two threshold values $\lambda_1, \lambda_2 \in [0,1]$, $\lambda_1 < \lambda_2$, we stated the following criteria

Criterion F.1. $\mu_{\text{Dom}}(w) \geq \lambda_2 \Rightarrow$ w is *included* in Dom.

Criterion F.2. $\lambda_2 > \mu_{\text{Dom}}(w) \geq \lambda_1 \Rightarrow$ w is *close* to Dom.

Criterion F.3. $\lambda_1 > \mu_{\text{Dom}}(w) \Rightarrow$ w is *outside* Dom.

3.3 Comparison of both Approaches

In our tests, we used as dictionary D (to be used for representing words) an electronic version of the *Diccionario Anaya de la Lengua Española* which comprises more than 29 000 entries. The domain Dom was the notion of "health" and the vocabulary $\mathcal{V}$ was the set of words employed by respondents in the public inquire in the Puebla valley. All words appearing in the inquire answers were put into a list of three blocks: the "included" words, the "close" words and the "outside" words. In order to calculate the performance index, we compared this list with the list of all words ordered decreasingly with respect to the rough-membership values. In summary, our results were the following:

- The performance index F, as defined in section 1.1, got a value $F = 0.53$ using the criteria **F** in the rough approach while $F = 0.35$ with the criteria **M** in the metric approach. However, a very great advantage of the metric approach was that it does not require a priori fixed threshold values. In (Haas, He, 1993) another criteria for domain membership are assumed, and there are reported indexes of the order $F \approx 0.49$.
- The classification of terms into three classes produces a cover which is not a proper partition. When we use just two disjoint classes (the "included" and the "outside" classes of terms) then with the metric approach we got $F = 0.56$.

3.4 Domain Membership for Phrases

In order to consider terms determined by phrases, not just words, we used the notion of *phrase sense* (García-Fajardo, 1995). Intuitively, we assume that the sense of a syntagma is determined by combining the properties of the words appearing in the syntagma. Formally, let us approximate this idea as follows: Let

$$\Psi : \mathcal{V} \to \mathcal{V}^*, \quad w \mapsto \Psi(w) = \left(\bigcup_{i \in \mathbb{N}} D(w,i) \right) \cup \mathcal{C}(w) \quad (10)$$

($\Psi(w)$ consists of all words appearing in definitions of w and in the context of w, with respect to dictionary D). For any two words w_1, w_2 let $\Upsilon(w_1, w_2) = \Psi(w_1) \cap \Psi(w_2)$. Then for arbitrarily long syntagma, let us define in an iterative way:

$$\xi(w_1, \ldots, w_n, w) = \begin{cases} \xi(w_1, \ldots, w_n) \cup \bigcup_{i=1}^{n} \Upsilon(w_i, w) \\ \quad \text{if } \Psi(w) \cap \xi(w_1, \ldots, w_n) \neq \emptyset, \\ \emptyset \quad \text{otherwise} \end{cases}$$

where $\xi(w_1, w_2) = \Upsilon(w_1, w_2)$. Now for any domain B, let B' be its representation as defined at the beginning of section 2. Then, for any syntagma $w_1 \cdots w_n$ let us define the membership degree of syntagma $w_1 \cdots w_n$ into the domain B as

$$\mu_B(w_1 \cdots w_n) = \frac{\#(B' \cap \xi(w_1, \ldots, w_n))}{\#\xi(w_1, \ldots, w_n)}. \quad (11)$$

A discussion on our definitions is worthy at this point: From a classical point of view, the concept associated to a syntagma is determined by the common features of the concepts associated to the words in the phrase syntagma. In case of long phrases the intersection of the components tends to be empty, whereas a simple set-theoretical union makes too wide the notions and introduce a great vagueness to the concepts.

4 Phrase Classification

We will exemplify our methods of phrase classification by the processing of the above mentioned "Public Health" inquire. Let B be this domain. Our treatment is based on the ⟨*expositive acts of speech*⟩ (Austin, 1990). Any answer to the inquire will be put on the class computed as a function of the membership degrees to the health domain, according to the words and the noun phrases contained in the answer, as well as on the number of *links* between phrases on the domain. Indeed, a *link* exists between two phrases if the two syntagma taken as a whole has sense in the domain. We recognize a sense of a text in a domain, if there exists at least $(n/2) + 1$ truly links from n total possible links among the text components. Thus one possible *illocutionary value* (⟨*refract*⟩, ⟨*circunds*⟩, ⟨*describes*⟩) shall be assigned to each such answer:

describes. Either the number of links between noun phrases having sense on the domain is greater than minimum necessary links of the phrase or the number of participating words is greater than 1.

circunds. The number of participating words is greater than one and one of the following conditions is satisfied:

1. the average membership degree of noun phrases is positive,
2. the words appearing in the noun phrases have an average class "outside" of B or
3. the words appearing in the noun phrases have an average class "close" of B.

refract. In any other case.

After processing 172 answers, we got a global accuracy index of 49.41% and 93% of successes in distinguishing far separated classes while we got just a success index of 78.2% for closer classes. Distinction of far separated classes was made by computing the number of successes when discriminating ⟨describes⟩ against ⟨refract⟩ values, whereas for near classes we computed the number of successes on discriminating ⟨describes⟩ against ⟨circunds⟩ values or on discriminating ⟨circunds⟩ against ⟨refract⟩ values.

5 Final Remarks

5.1 Solved Problems

We have constructed an automatic tagger with POS for Spanish. It is based on MBL with a modified metric which allows the treatment of unknown values in the training instances for IGTree. For unknown words, the performance index of IGTree increased from 79.19% to 81.5% with the modificated metric, nevertheless we do not have still statistically significant evidence of improvement.

With respect to noun phrases identification, we developed a system based on MBL to recognize simple noun phrases. We have begun with a lexicon of Spanish of around 10 000 words. We got higher performance indexes: 96.7% when processing texts of the same genre as the training text, and 88.9% otherwise.

Finally, we have developed two approaches to estimate membership degrees of words and noun phrases to discourse domains: one is based on rough sets while the second in a metric. In spite that the first gives better results it depends on extra criteria to select good threshold values.

5.2 Contributions

Computational system. Ours is specialized on the Spanish language. We estimate important the system since it is one of the very few tools for Spanish up to-date.

Methods. We have developed and implemented some efficient variations of already classical classification methods (MBL and IGTree).

5.3 Future Work

At present we are addressing several ulterior tasks in order to continue the development of our system:

- To develop methods and algorithms for a word tagger training able to process unrestricted text.
- To explore the advantages of new comparison functions alternative to δ'.
- To identify verbal and adjective phrases as well.
- To validate in a stronger way the performance improvement of methods by using different linguistic resources.
- To test our methods for membership degrees calculation in areas relevant to information retrieval.
- To contribute in the establishment of newer linguistic resources for the Spanish language.

References

Austin, J., *Cómo hacer cosas con palabras*, Paidos, Spain, 1990.

Baeza-Yates, Ricardo, "Searching the world wide web; challenges and partial solutions" in Coelho, Helder (Ed.), *Lecture Notes in Artificial Intelligence 1484*, Springer Verlag, Germany, 1998, pp 39-51.

Church, Kenneth & Hanks, Patrick, "Word association norms, mutual information and lexicography" in *Computational Linguistics 16*, Association for Computational Linguistics EU, 1990, pp 22-29.

Daelemans, Walter, "Memory-based lexical acquisition and processing" in *Lecture Notes in Artificial Intelligence* 898, Springer Verlag, Germany, 1995, pp 85-98.

Daelemans, Walter; van-den-Bosch, Antal; Zavrel, Jakub; Veenstra, Jorn; Buchholz, Sabine & Busser, Bertjan, "Rapid development of NLP modules with memory-based learning" in *Proc. of ELSNET in Wonderland*, 1998, pp 105-113.

Daelemans, Walter; Durieux, Gert & van-den-Bosch, Antal, "Towards inductive lexicon" in *Proc. of LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications*, Spain, 1998, http://ilk.kub.nl/.

García-Fajardo, Josefina, "Estructura conceptual y comunicación" in *Dimensión Antropológica*, Año 2, V. 3, México, 1995, pp 75-84.

Gierl, Claude & Frost, David, "Identification of domain-specific terminology by combining mutual information and lexical induction" in *European Conference on Artificial Intelligence*, 1992, pp 564-566.

Grishman, Ralph, "Information extraction, techniques and challenges" in *Lecture Notes in Artificial Intelligence* 1299, Springer Verlag, Germany, 1997, pp 10-27.

Haas, Stephanie & He, Shaoyi "Toward the automatic identification of sublanguage vocabulary" in *Information Processing and Management*, V.29 (6), 1993, pp 721-732.

Jiménez-Salazar, Héctor & Morales-Luna, Guillermo, "Noun phrases identification from a tagged text" in *Memoria TAINA98*, UNAM, México, 1998, pp 90-101.

Jiménez-Salazar, Héctor & Morales-Luna, Guillermo, "Instance metrics improvement by probabilistic support" in *Lecture Notes in Artificial Intelligence* 1793, Springer Verlag, Germany, 2000, pp 699-705.

Jiménez-Salazar, Héctor & Morales-Luna, Guillermo, "Sepe: A POS tagger for Spanish" in *Lecture Notes in Computer Science* 2276, Springer Verlag, Germany, 2002, pp 250-259.

Lara, Luis Fernando; Ham-Chande, Roberto & García-Hidalgo, Ma. Isabel, *Investigaciones lingüísticas en lexicografía*, Jornadas 89, El Colegio de México, México, 1979.

Pawlak, Z., "Rough sets" in *Int. J. Computer and Information Science*, V.11, Poland, 1982, pp 341-365.

Smeaton, Alan, "Information retrieval: still butting heads with natural language processing?" in *Lecture Notes in Artificial Intelligence* 1299, Springer Verlag, Germany, 1997, pp 115-138.

Veenstra, Jorn, "Fast NP chunking using memory-based learning techniques", *Proc. of Benelearn*, The Netherlands, 1998, pp 71-79.

Héctor Jiménez received the B.Sc. degree in physics and mathematics from Instituto Politécnico Nacional, and the M.Sc. and Ph.D. degree in electrical engineering from CINVESTAV, México. Since 1977 he has been computer science profesor at Autonomous University of Puebla. He does research and teaching in natural language processing related with information retrieval, mainly linguistics resources for Spanish language. Additional interest areas are mathematical logic and machine learning.

Guillermo Morales-Luna *received a B. Sc. in mathematics at the Mexican National Polytechnic Institute in 1977, a M. Sc. in mathematics at Mexican CINVESTAV-IPN in 1978 and a Ph. D. in the Mathematics Institute of the Polish Academy of Sciences in 1984. Since 1985 he is a researcher in his current position. His research interest are in cryptography, complexity theory and mathematical logic.*