# A Method of Automatic Detection of Lexical Relationships using a Raw Corpus

Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
hjimenez@fcfm.buap.mx

**Abstract.** This work presents some results on the application of a criterion used to compare the senses of a pair of words. A measure that involves the senses of words was used to reinforce hypothesis like hyponymy relationship between the words.

## 1 Introduction

Information Retrieval Systems (IRS) use lexical relationships (hyponymy, synonymy, holonymy, etc.) in order to improve its performance. In particular, query expansion techniques help to solve this task [4, 2]. The problem is to construct these lexical resources for different domains. In [6], for two words $x$ and $y$, the relation $x$ hyperonym of $y$ is obtained, through the subsumption notion: *$x$ subsumes $y$* if $A_y \subset A_x$, where $A_w$ denotes the set of contexts (documents, sentences, etc.) that contain the word $w$. For hyperonyms in general $A_y \subset A_x$ is not satisfied, but 80% of the members of $A_y$ are in $A_x$.

We are interested in the exploration of several lexical relationships. In the present work the problem of comparing two sets is faced by computing how much a set is included in another; this is the base to define the quantity subsumption ratio. We specifically use the features of the contexts instead of the contexts. Features of a context follow the IRS idea of document representation by index terms. Now, the inclusion property of $x$ subsumes $y$ is expressed in the framework of formal concepts by duality: $B_x \subset B_y$, where $B_w$ is the set of features that each member of $A_w$ holds.

Thus, we will conceive the contexts that use a word $w$ (a subset of $A_w$) with the same meaning as instances of a concept. In the *formal concept* theory [1] we can compare two concepts using its components, namely *intent* and *extent*. The extent is the set of objects which are instances of the concept, and the intent is the set of features that are satisfied by all the objects ascribed to the concept. In this approach, a concept is more general than another if the extent of the former contains the extent of the latter or, equivalently, if the intent of the former is contained in the intent of the latter. This supports the use of the intent to compare the sets of contexts.

The next section specifies the notion of subsumption ratio. Next the experiment carried out is described, and at the end the conclusions of this work are presented.

## 2 Subsumption Ratio

Let us consider a corpus $T$ and the set of all sentences that use in $T$ the word $w$, $A_w$, and let $A_w^i \in A_w$. We may represent $A_w^i$ by its index terms or features, $B_w^i$, i.e. the most representative words of $A_w^i$ with respect to $A_w$. To determine the representative terms, $v \in B_w^i$, the discrimination value model is used ([5] presents this model). Discrimination value model defines the document frequency of $v$, $df_v$, as the number of documents that contain the term $v$. Thus, for some $i$, $v \in B_w^i$ if $df_v \in [n_w/100, n_w/10]$, where $n_w = \#A_w$. Let us denote the $k$-th sense of the word $w$ by $w_k$. $w_k$ will be represented by some subset $A(w_k) \subset A_w$, where each context, member of $A(w_k)$, uses the same sense of $w$. $A_w$ may be partitioned to group the most similar $A_w^i \in A_w$ that represent $w_k$. For this task we use $B_w^i$ instead of $A_w^i$. Thus, we may represent the concept of the $k$-th sense of $w$ with the extent $A(w_k)$ and the intent $B(w_k)$. The goal is to compare different meanings of two words, through its intents to conclude a semantic relationship between them. We define the *subsumption ratio* of $x_i$ to $y_j$ as:

$$\rho(x_i, y_j) = \frac{\#(B(x_i) \cap B(y_j))}{\#B(x_i)}$$

where $B(w_k)$ is the set of features of the $k$-th sense of the word $w$. Note that $B(x_i) \cap B(y_j)$ is the intent of a more general concept. Certainly if $B(x_i) \subset B(y_j)$ then $\rho(x_i, y_j) = 1$. Thus, we can use the ratio of features in $B(x_i)$ that are in $B(y_j)$ to indicate how much $x_i$ subsumes $y_i$. If $\rho(x_i, y_j)$ is high then $x_i$ will subsume $y_j$ with degree $\rho(x_i, y_j)$. This point is directly related with the riches of the corpus.

Besides, we need to compute $\rho(y_j, x_i)$ which will complement the knowledge about $x_i$ and $y_j$. For example, if both ratios have a high value we can strengthen the hypothesis of synonymy relationship between those word senses. From the previous remark it can be established the conditions to find some relationship between two word senses. Fixing the words $x$ and $y$, and denoting $\rho(x_i, y_j)$ by $\rho_{ij}$, we propose the following rules:

$y_j$ **synonym of** $x_i$**:** $\rho_{ij}$ is high and $\rho_{ji}$ is high
$y_j$ **hyponym of** $x_i$**:** $\rho_{ij}$ is high and $\rho_{ji}$ is low
**strongly related:** $\rho_{ij}$ is high and $\rho_{ji}$ is medium
**weakly related:** $\rho_{ij}$ is low and $\rho_{ji}$ is medium
**very low relation:** $\rho_{ij}$ is low and $\rho_{ji}$ is low

This classification has implicit two thresholds $\mu_1$, the maximum for low values of $\rho_{ij}$ and $\mu_2$, the maximum for medium values of $\rho_{ij}$.

## 3 Experiment

The corpus[1] (10Mb) is composed of 2057 articles, 61,216 sentences, a vocabulary of 136,988 signs (different words including punctuation, abbreviations and numbers) and 18,092 stems. The frequency of words was observed to choose those words that could provide sufficient contexts to the processing. Stop words were removed and the remaining words were stemmed. The test was performed on the words: *triunfo* 'triumph', *victoria* 'victory', *militar* 'military', *coronel* 'colonel', *teniente* 'lieutenant', *avión* 'plane', *aeroplano* 'airplane', *aeropuerto* 'airport', and *hijo* 'son' whose frequencies are 120, 140, 520, 251, 100, 590, 62, 166, 507, respectively. In the experiment were carried out the next steps:

1. Given a word $w$:
   (a) To obtain $A_w$.
   (b) To represent each member $A_w^k \in A_w$ by $B_w^k$, using the discriminator terms contained in $A_w$.
   (c) To partition the collection $A_w$ according to the most similar uses of $w$: $A(w_1), A(w_2), \ldots A(w_p)$, $(A_w = \cup_{q \leq p} A(w_q))$. In this task Jaccard similarity measure is used. $A(w_k)$ has the corresponding $B(w_k)$; each pair $(A(w_k), B(w_k))$ represents a sense $w_k$.
2. Given two cluster collections corresponding to $x$ and $y$, all pairs (taking only the largest groups) are produced, then $\rho(x_i, y_j)$ and $\rho(y_j, x_i)$ are applied, where $i$ and $j$ range from 1 to the number of elements in the cluster collection of $x$'s and $y$'s, respectively.
3. The highest values of $\rho(x_i, y_j)$ and $\rho(y_j, x_i)$ and its difference $(\Delta_{max})$ are used to classify the dominant relationship between $x$ and $y$.

| $x$-$y$ | $\Delta_{max}$ | $i,j$ | $\rho(x_i, y_j)$ | $\rho(y_j, x_i)$ | Class |
|---|---|---|---|---|---|
| triunfo-victoria | 0.09 | 1,1 | 0.54 | 0.45 | synonym |
| militar-teniente | 0.40 | 2,2 | 0.50 | 0.08 | hyponym |
| militar-coronel | 0.48 | 3,2 | 0.66 | 0.18 | hyponym |
| aeroplano-avión | 0.63 | 2,1 | 0.02 | 0.66 | hyponym |
| aeropuerto-avión | 0.54 | 2,2 | 0.31 | 0.85 | strongly related |
| aeropuerto-aeroplano | 0.31 | 2,2 | 0.36 | 0.05 | weakly related |
| hijo-aeropuerto | 0.07 | 4,2 | 0.18 | 0.10 | unrelated |

**Table 1.** Classification with the subsumption ratio.

In the table 1 we used the thresholds $\mu_1 = 0.2$ and $\mu_2 = 0.4$. For $x =$"militar" the number of largest clusters was six and for $y =$"teniente" was two, which gives 12 combinations. Both values $\rho(x_i, y_j)$ and $\rho(y_j, x_i)$ were calculated and plotted for each one of these combinations. The graph is shown at fig. 1-b. The first half

[1] It is a collection of selected articles from the Mexican newspaper *El Universal*, titled *80 Años Informando (1916-1996)* '80 years of information'.

of the graph corresponds to one sense of $x$ varying all senses of $y$. The similarity matrix obtained in the clustering process was used to visualize the main senses of the word *militar* (see fig. 1-a), this is an iso-analogical representation of the word senses [3].

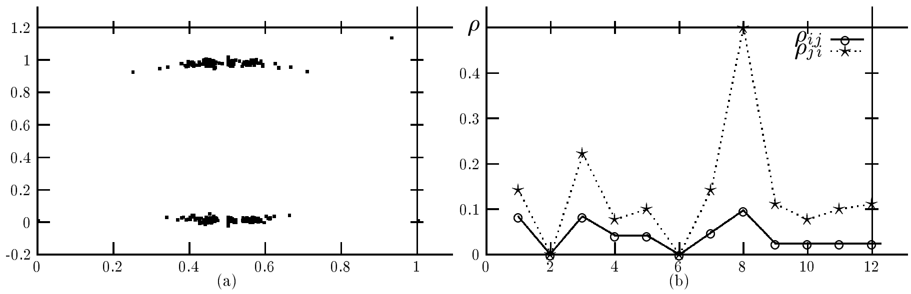


**Fig. 1.** (a) Clusters of contexts for *militar*. (b) Subsumption ratio for *teniente* and *militar*.

## 4 Conclusions

In this paper we have defined a measure based on the subsumption notion and formal concept theory in order to know the degree of relationship between two senses of words. This measure uses only a raw corpus and its advantage is that no domain knowledge is required. For this approach we have shown some examples but it is necessary an exhaustive test.

## References


1. Davey, B. & Priestley, H.: *Introduction to lattices and order*, Cambridge Mathematical Textbooks, 1990.
2. Gelbukh, A.: Lazy query enrichment: a simple method of indexing large specialized document bases, *Proc.DEXA-2000 11th Int. Conf. and Workshop on Databases and Expert Systems Applications*, 2000.
3. Lavalle-Martínez, J.: *Representación isoanalógica de objetos n-dimensionales*, M.Sc. Thesis, CINVESTAV (México), 2000.
4. Mandala, R.; Tokunaga, T. & Tanaka, H.: Combining multiple evidence from different types of thesaurus, *Proc. 22nd International Conference ACM-SIGIR*, 191-197, 1999.
5. Salton, G.; Yang, C.S. & Yu, C.T.: A theory of term importance in automatic text analysis, *Journal of American Society for Information Science*, 26(1), 33-44, 1975.
6. Sanderson, M. & Croft, W.B.: Deriving concept hierarchies from text, *Proc. 22nd International Conference ACM-SIGIR*, 206-213, 1999.