http://dx.doi.org/10.12795/PH.1997.v11.i02.03

APLICACIÓN ESTADÍSTICA A LA TRADUCCIÓN AUTOMÁTICA

María Teresa López Soto

NLP (Natural Language Processing) aims to represent natural language using mathematical formalisms. This application is useful in areas such as MT (Machine Translation), Speech Recognition, Information Retrieval, etc. The formalization of natural langauge, however, represents serious linguistic analysis problems. One major challenge is the solution of structural ambiguity. Research in the area shows that statistical information can achieve disambiguation in many cases. Structural ambiguity occurs when the parser of a system finds two possible analysis for the same sentence. This happens when a sentential constituent can be attached to, at least, two superior nodes. A method which has proved a high degree of success in the analysis of structural ambiguous sentences is that of LA (Lexical Association). LA consists of the creation of several tables which include morphological information. These tables are the so-called "lexical tables". For the case when a PP (Prepositional Phrase) can be assigned to the VP (Verb Phrase) as a complement (OD, Direct Object), or, to an NP (Noun Phrase) as a modifier, the lexical tables create three columns: one for the nouns, one for the verbs and the third column is for the preposition associated. The parser must get information to assign the preposition to the VP or to the NP. This information is obtained by means of using a numerical value which determines the frequency with which, in a large sample corpus, a specific noun is associated with that preposition or a verb is associated with the same preposition. If the value is higher for the association between the verb and the preposition, then the analysis obtained is that the PP functions as a complement to the VP. Otherwise, the analysis of the PP is that of being a modifier to the VP.

0. Introducción.

El Procesamiento del Lenguaje Natural (PLN), pretende ofrecer una representación formalizada de la lengua que utilizamos a diario. Esta formalización encuentra luego aplicaciones diversas: al ser fácilmente procesable por ordenador, es el lenguaje formal el utilizado en reconocimiento de la voz, traducción automática, extracción y recuperación de información, etc. El medio de trabajo en que nos movemos es totalmente diferente al de una situación de comunicación natural, de lenguaje natural: la lengua que usamos para comunicarnos y que surge espontáneamente, no es en muchos aspectos comparable a un lenguaje de naturaleza formal, es decir, a un lenguaje etiquetado y estructurado a partir de reglas claras, generales y siempre derivables desde otras superiores.

En el lenguaje natural, el ser humano parte de una información que necesita comprender, asimilar e intercambiar. Lo que se pretende comunicar son significados,

conceptos; en lingüística tradicionalmente este campo de la comunicación se ha etiquetado como información semántica, estudio del significado. Por otra parte, los sonidos que emitimos, y que luego representamos con letras, son entidades físicas que se refieren a contenidos más abstractos, las palabras se organizan hasta formar oraciones que entendemos, los sonidos pertenecen al campo de estudio de la fonología o fonética; la morfología, estudia unidades menores que la oración, morfemas o palabras. Por último, la pragmática, que puede definirse como el estudio de los elementos comunicativos relacionados con el contexto.

Todos estos campos lingüísticos encuentran representación en el lenguaje formal. El parser de un sistema de PLN es simplemente el analizador sintáctico y a menudo también semántico que filtra las oraciones que son gramaticales a partir de una serie de reglas definidas. La información fonética y pragmática también es a menudo representable por medio de reglas: la información fonética se formaliza en las aplicaciones de reconocimiento y síntesis de la voz, y el análisis pragmático puede aparecer en un sistema de procesamiento.

Pero existen limitaciones. Para elaborar unas reglas gramaticales simples referidas a un lenguaje formal, el investigador suele partir de un corpus que analiza y del que extrae la información que le es relevante para el fin propuesto. La elaboración de reglas gramaticales se hace siempre a partir de enunciados reales de lenguaje formal, bien extraidos de corpus reales o de otros creados a tal fin. Es más, en la mayoría de los casos, la elaboración de un lenguaje formal se hace a partir de un corpus concreto para usar en situaciones similares a las que el corpus se refiere. Es decir, se parte de un contexto para su uso en el mismo contexto. Cuando se dan representaciones no recogidas en el corpus original, el sistema tiene que recurrir a mecanismos más productivos; cuando ni siquiera esto es posible, es preciso aumentar la capacidad del sistema creando mayor vocabulario, ampliando reglas gramaticales o incluyendo mecanismos que faciliten el análisis desde nuevos enfoques.

Los casos más conflictivos suelen ser los de ambigüedad léxica o estructural. Son casos que incluyen una doble represenación, bien léxica, bien sintáctica. en el lenguaje natural. En el lenguaje formal la ambigüedad se trata con técnicas diversas que, a menudo, no impiden la doble representación[1].

Para entender esto mejor, veamos una serie de ejemplos que se refieren a ambigüedad estructural y ambigüedad léxica. Por ambigüedad estructural entendemos la ambigüedad que se produce cuando una misma oración permite más de una representación sintáctica, lo que se deriva en más de un significado. En la oración:

(1) Vi a Juan con el telescopio

tenemos dos significados:

---

[1] El mayor problema en los casos de ambigüedad viene derivado de la doble representación que se produce en el análisis. Cuando el parser encuentra más de un camino posible de análisis, éste se detiene. Es preciso entonces hallar herramientas que sepan hallar el significado deseado.

(1a) Vi a Juan que llevaba un telescopio

(1b) Vi a Juan usando un telescopio

Otros casos:

(2) Vive en la casa al lado de la torre con ventanas que es de color rosa

(¿la casa es rosa o la torre es rosa?)

(3) Natalia me dijo que sacó la basura ayer

(¿me lo dijo ayer o sacó la basura ayer?)

(4) El jefe me despidió el día 1

(¿me dijo adiós o me he quedado sin trabajo?)

(5) Vi a Pepe corriendo

(¿iba yo corriendo o era él?)

También encontramos casos de oraciones que sólo son analizables correctamente si utilizamos los signos de puntuación, pausas o comas, por ejemplo:

(6) Los amigos que alabas a veces se lo merecen

según dónde coloquemos las comas encontramos dos significados:

(6a) Los amigos, que alabas, a veces se lo merecen

(6b) Los amigos, que alabas a veces, se lo merecen

Casos de ambigüedad léxica:

(7) Se ha roto la pantalla

(¿de la tele o de la lámpara?)

A primera vista, la ambigüedad se resuelve fácilmente en situaciones de lenguaje natural. Por ejemplo, la oración (4), la primera lectura que hacemos es la de que hemos perdido el trabajo, quizá por las connotaciones negativas que se atribuyen a la palabra "jefe" nos haría pensar que es más fácil que nos despida del trabajo que no que se despida de nosotros antes de vacaciones, por poner un caso. En otros ejemplos, como (3), existe la preferencia de asociar el adverbio "ayer" al hecho de sacar la basura y no al momento en se nos dijo. El contexto, como antes se ha dicho, también rompe la ambigüedad de (7), sobre todo si vemos a la persona con la lámpara rota en la mano.

El problema es cómo representar toda esta información en un lenguaje fórmal. Como método alternativo a la información sintáctico-semántica, en PLN una alternativa que ha obtenido resultados óptimos es la aplicación estadística. Esta herramienta es muy utilizada

porque en muchos casos sólo el contexto es capaz de evitar la ambigüedad; debido a la naturaleza espontánea del lenguaje natural, a menudo no es posible formalizar la información contextual. La estadística se observa como un método alternativo para lograr mejorar los sistemas de PLN y para casos muy concretos. A continuación se describen algunos métodos existentes que se basan en técnicas de cálculo de probabilidad para elegir uno u otro análisis.

1. Gramáticas Probabilísticas.

En este apartado vamos a presentar un ejemplo de la gramática **n-grama.**

Cualquier gramática que represente a un lenguaje natural es ambigua. Por lenguaje natural entendemos la lengua, el idioma que el ser humano utiliza para comunicarse, bien por escrito, bien en forma hablada. Una gramática que represente a un lenguaje natural puede definirse como una secuencia de reglas de re-escritura, en las que el nodo superior, el situado a la izquierda de la regla, se re-escribe en el nodo inferior, situado a la derecha. Por ejemplo:

(1: O --> SN SV)
(2: SN --> det n)
(3: SN --> pron)
(4: SV --> v)

sería un caso de gramática que serviría para analizar la oración:

(8) Nosotros compramos la leche

Una gramática **n-grama** se compone de una serie de probabilidades que se representa así:

$$P(w_n/w_1,w_2...w_{n-1})$$

que da una probabilidad en la que a w_n le sigue la cadena de palabras w_1 w_2 --- hasta w_{n-1}, para cada combinación posible de w's en el vocabulario del lenguaje. Para un vocabulario de 5.000 palabras, una gramática bigramática tendría aproximadamente 5.000 x 5.000= 25.000 de parámetros libres, y una gramática trigramática tendría aproximadamente (5.000 x 5.000) x 5.000 = 125.000.000.000.

Una gramática estocástica libre de contexto (SCFG) se compone de una serie de reglas, al lado de las que se especifica la probabilidad de elegir una producción determinada para el símbolo no terminal de la izquierda. (Se denomina producción a cada regla de re-escritura). Por ejemplo, si tenemos una CFG simple, podemos aumentarla con las probabilidades especificadas así:

O --> SN SV	[1.0]
SN --> N1	[0.4]
SN --> det N1	[0.6]
SV --> VG	[0.8]
SV --> VG SN	[0.2]

Det --> el	[0.4]
Det --> una	[0.6]
N1 --> libro	[1.0]
V --> abrir	[0.3]
V --> cerrar	[0.7]

El cálculo de probabilidad basado en **n-gramas** se obtiene de calcular las frecuencias asociadas para n-gramas y para (n-1)-gramas:

$$P(w_n/w_1\ w_2...w_{n-1}) = \frac{c(w_1...w_n/L)}{c(w_1...w_{n-1}/L)}$$

donde **c(w / L)** contabiliza el número de concurrencias para la subcadena **w** en la oración **L**. Ese decir, contabiliza el número de veces en que se asocia una determinada palabra con otra para el conjunto total de concurrencia de palabras en la oración **L**.

2. Modelos de Gramáticas Probabilísticas: Fujisaki, et al; Sharman et al; Pereira & Schabes.

En este apartado se resume brevemente las propuestas de varios autores sobre distintos modelos de gramáticas probabilísticas.

2.1 Fujisaki et al. (1989) . A partir del análisis de un corpus, se crea una gramática libre de contexto (CFG) probabilística con unas 7.550 reglas para un corpus de 4.206 oraciones. El proceso de entrenamiento del corpus consiste en asignar las probabilidades para cada regla de manera automática según sea su frecuencia de ocurrencia para todos los posibles análisis de cada oración del corpus. La probabilidad se estima usando una variante del algoritmo de Baum y Welch y el algoritmo de Viterby en conjunción con el CYK que se usa en la fase de análisis sintáctico (parsing) para seleccionar el análisis más probable una vez hecho el entrenamiento. Es decir, el modelo se reduce de manera que muchos de los parámetros (reglas) posibles definidos para el conjunto de categorías de nodos terminales y no terminales se inicializa a cero; el entrenamiento sirve entonces solamente para estimar nuevas probabilidades para un conjunto de reglas predefinidas. Fujisaki et al. sugieren que las probabilidades estables sirven para modelar limitaciones semánticas y pragmáticas. Sin embargo, esto sólo es factible si las nuevas probabilidades se corresponden con la frecuencia estimada para las reglas en los análisis correctos. De 72 oraciones examinadas para un conjunto de 84, el análisis más probable era también el análisis correcto. De los restantes, 6 eran falsos positivos y no recibían un análisis correcto, mientras que los otros 6 sí ofrecían un análisis correcto pero sin ser el más probable.

2.2. Sharman, Jelinke y Mercer (1990). Proponen una gramática que contiene 100 nodos terminales y 16 no-terminales y establecen unas probabilidades basadas en la frecuencia de las relaciones ID y LP. La frecuencia se calcula a partir del análisis manual del corpus, que consta de alrededor de un millón de palabras. Las probabilidades que resultan de la gramática ID/LP se usaron para analizar 42 oraciones de 30 o ménos palabras extraidas del mismo corpus. Además, las probabilidades léxico-sintácticas se integraron a la probabilidad inicial de las relaciones ID/LP. 18 de los análisis resultaron ser idénticos a los

obtenidos en el análisis hecho manualmente, mientras que 19 fueron sólo parecidos, lo que da un acierto de un 88%.

2.3. Pereira y Schabes (1992) usan la reestimación de Baum-Welch para inferir una gramática y asociar reglas de probabilidad desde una base de categorías que contiene 15 nodos no-terminales y 48 terminales. El corpus contenía 770 oraciones, del tipo: *"Isn't that?" o "you should"*. Se entrenó el corpus en dos modos: supervisado y semi-supervisado. La versión de análisis manual sirvió para depurar los análisis hechos en el proceso de re-estimación. En el entrenamiento semi-supervisado los análisis eran aceptados como correctos si ofrecían valores consistentes con los obtenidos manualmente. En su experimento demuestran que en modo supervisado, el entrenamiento no sólo se realiza más rápidamente, sino que también ofrece una gramática en la que el análisis más probable es compatible con el análisis conseguido de forma manual en tests de oraciones extraidas de la base de datos para un porcentaje mayor de acierto: 78%. Esto nos indica la importancia de un entrenamiento supervisado, sobre todo cuando en dicho proceso de entrenamiento se infieren, no sólo las reglas de probabilidad, sino la misma gramática.

2.4. Briscoe y Carroll. Exponen los problemas inherentes al uso de CFGs probabilísticas. En primer lugar, aunque una CFG es un modelo adecuado para representar la mayoría de las construcciones del lenguaje natural, se hace necesario que las reglas creadas en la CFG se hagan extensibles a un número muy grande de casos. El problema derivado estriba en la dificultad de desarrollar gramáticas consistentes y de llegar a manejar computacionalmente el tiempo de análisis. En segundo lugar, asociar probabilidades a reglas CF significa que se pierde la información de probabilidad para una regla cuando la aplicamos a un punto del análisis en que se produce una derivación. Esto conlleva problemas para distinguir la probabilidad de diferentes derivaciones cuando se puede aplicar la misma regla varias veces. Veámoslo con un ejemplo:

(1: O' --> O)	[1.0]
(2: O --> SN SV)	[1.0]
(3: O --> Vt SN)	[0.4]
(4: SV --> Vi)	[0.6]
(5: SN --> ProSN)	[0.4]
(6: SN --> Det N)	[0.3]
(7: SN --> SN SP)	[0.3]
(8: N --> N N)	[0.3]
(9: SP --> Prep SN)	[1.0]
(10: N --> N1)	[0.7]

es una gramática probabilística CFG en la que cada producción se asocia con una probabilidad, y las probabilidades de todas las reglas se expanden en una categoría no terminal que suma 1.

La probabilidad para un análisis concreto es la que resulta del producto de las probabilidades para cada regla que se ha usado en la derivación. Sólo se considera aceptable una probabilidad global asociada a cada produccion CF relevante. Así, el modelo de CFG probabilístico predice (incorrectamente) que dos producciones tendrán la misma

probabilidad de concurrencia. Todas estas consideraciones plantean la necesidad de usar una técnica que permitan un formalismo gramatical más adecuado que el que ofrece una CFG, así como un modelo probabilístico más dependiente del contexto.

La solución que ofrecen Briscoe y Carroll es la de usar la técnica de parsing de izquierda a derecha (LR parsing) como método para obtener una representación de estados finitos de una gramática de estados no finitos incorporando información del contexto de parsing.

El corazón de una técnica de parsing LR es un algoritmo de construcción de tabla de análisis que constituye el aspecto más complejo y con mayor gasto computacional. Un parser LR encuentra la derivación más a la derecha para una cadena y dada una CFG. La tarea fundamental se realiza en la fase de precompilación, que resulta en un mecanismo de control de análisis que le permite al parser identificar la subcadena apropiada. (Como se ve en las tablas)

3. El Método de Configuración Basado en la Concurrencia de Palabras.

El procedimiento estadístico de la concurrencia de palabras se basa en analizar la frecuencia con que dos o más términos del corpus aparecen en secuencia. De esta manera se asigna, para el caso de la ambigüedad estructural, el análisis que corresponda según sea más o menos probable.

Aparte de la exactitud en el cálculo y en el diseño de las funciones matemáticas, cualquier método de configuración basado en la estadística, ha de usar información desde un corpus para ser utilizada en cualquier corpus.

Un método que parece ha dado buenos resultados es el basado en la concurrencia de palabras. La primera exigencia es la de partir de un corpus extenso, representativo de la lengua. Eso no excluye la posibilidad de elaborar diversas tablas de valores para contextos, registros lingüísticos, etc. diferentes. Así, y usando el mismo cálculo base, con corpus diferentes, se obtendrían diversos resultados aplicables, según cada caso, a muestras lingüísticas de naturaleza distinta: contexto literario, científico, etc. Sólo que parece rentable elaborar una tabla de valores a partir de un corpus lo suficientemente representativo de la lengua en estudio.

Este método defiende el diseño de un mecanismo que permite asociar secuencias de palabras (**n-gramas**) a las relaciones que se establecen entre los núcleos de los distintos sintagmas en las construcciones sintácticas (por ejemplo, la relación que se establece entre verbo-objeto o adjetivo-nombre). La probabilidad para esta concurrencia, en la mayoría de los modelos existentes, se define en una función que estima la concurrencia de palabras: es decir, la frecuencia con que dos palabras (núcleos de sintagma) aparecen en secuencia en un corpus determinado. Cuando hablamos de dos palabras, la tabla que se deriva se llama

bigrama[2]. En los modelos de bigramas, la probabilidad $P(w_2 / w_1)$ para una palabra condicionada w_1, se calcula por la probabilidad de w_2, estimada por su frecuencia en el corpus. Este planteamiento ha encontrado defensores en varios autores, como ahora se expone; estos autores son Brown et al. (1992), Pereira, Tishby y Lee (1993), Dagan, Markus y Markovitch (1992) y Hindle y Rooth (1992).

3.1 Brown et al. (1990) sugieren un modelo de **n-grama** basado en clases en el que las palabras con distribuciones de concurrencia similares se engloban en clases morfológicas. La probabilidad de concurrencia de un par de palabras se estima según la probabilidad media de concurrencia de las dos clases correspondientes.

3.2. Pereira, Tishby y Lee (1992) proponen un esquema más sencillo, en el que, para ciertas concurrencias gramaticales, la pertenencia de una palabra a una determinada clase es probabilística. Las probabilidades de concurrencia de palabras se modelan según las probabilidades medias de concurrencias de grupos de palabras.

3.3. Dagan, Markus y Markovitch (1993) defienden una reducción a un número relativamente pequeño de grupos de palabras, previamente determinados, que pueden llegar a causar una pérdida de información importante. Los modelos basados en similitud que ellos proponen evitan tener que agrupar palabras en categorías generales. Cada palabra se considera como perteneciente a una clase específica, o bien, a un grupo de palabras con el que guarda una gran similitud (como ocurre en el enfoque "**k-nearest neighbor**" en los patrones de reconocimiento).

Usando este esquema, se puede llegar a predecir qué concurrencias son más probables. Sin embargo, no se trata de un modelo probabilístico, pues no define las estimaciones de probabilidad para concurrencias no observadas. No puede ser, por tanto, usado en un marco completo de probabilidad, como son los modelos de **lenguaje de n-grama** o las **gramáticas probabilísticas lexicalizadas.**

4. La Propuesta de Hindle y Rooth: Relaciones Léxicas.

Hindle y Rooth (1992) describen un método, con base en un corpus, para la desambiguación del SP cuando éste puede modificar al verbo de la oración principal y al OD.

La teoría de Hindle y Rooth surge como una crítica al uso de estrategias clásicas, en concreto, la "**asociación por la derecha**" y la de "**dependencia mínima**". Dichas estrategias son contradictorias entre sí y no funcionan en la práctica. Hindle y Rooth proponen un método alternativo: sugieren usar la preferencia léxica, estimada a partir de un corpus textual extenso, como método para resolver la ambigüedad estructural. Es el método de la "**asociación léxica**" (AL) con base en la técnica estadística.

---

[2] Como función con dos argumentos, un bigrama se representa normalmente como una tabla con dos entradas o "matriz bidimensional". En el contexto de este trabajo, los bigramas se usan para almacenar la información relativa a la probabilidad de concurrencia entre nombres/verbos y preposiciones.

El sistema propuesto sigue una serie de pasos para elaborar la información esadística que permita elegir el candidato "más probable" en el análisis de una oración con ambigüedad estructural. Estos pasos son los siguientes:

En primer lugar, el mecanismo elabora un análisis morfológico, que dará lugar a la asignación de las distintas categorías para cada término del léxico. Esto es importante, pues, a partir de aquí, el proceso se va a centrar en tres términos fundamentales: verbo núcleo del VP, nombre núcleo del SN y preposición incluida en el SP condicionante.

En segundo lugar, se elabora una tabla con tres entradas (**trigrama**) que recoge aquellos casos en que verbo, nombre y preposición aparecen seguidos en secuencia en el corpus: son las llamadas "tablas léxicas". De esta manera, para una oración como:

(9) The radical changes in export and customs regulations evidently are aimed at remedying an extreme shortage of consumer goods in the Soviet Union and assuaging citizens angry over the scarcity of such basic items of soap and windshield wipers

se crearía la siguiente tabla léxica:

	VERB	NOUN	PREP	SYNTAX
a.	change		in	-V
b.		regulation		
c.		aim		
d.		PRO-+	at	
e.	remedy	shortage	of	
f.		good	in	
g		DART-PNP		
h.	assuage	citizen		
i.		scarcity	of	
j.		item	as	
k.		wiper		

Estas tablas incluyen cierta información sintáctica, por ejemplo: **-V** (nombre deverbal), **PRO- +** (es una categoría vacía, aquí el OBJ que se asume del verbo en forma pasima "aimed"). **DART-PNP** quiere decir NP premodificado por artículo determinado y que, a su vez, es un nombre propio. Toda esta información sintáctica afecta a ciertos aspectos del proceso de cálculo de probabilidades.

Llegados a este punto, Hindle y Rooth se dieron cuenta de que una tabla diseñada así era imperfecta, pues el analizador sintáctico podía originar análisis incorrectos como en la oración siguiente, en que "**to**" se ha malinterpretado como preposición y no como marcador de infinitivo:

(10) The Bush administration told Congress on Tuesday it wants to [**v** preserve] [**sn** the right] [**sp** [**p to**] control entry] to the United States of anyone who was ever a Communist

Otro caso erróneo se da cuando la preposición no es tal preposición sino una conjunción subordinante:

(11) The Supreme Court today agreed to consider reinstating the murder conviction of a New York City man who confessed to [**ving** killing] [**sp** his former girlfriend] [**p** after] police illegally arrested him athis home

El sistema, sin embargo, no puede ser considerado como no adecuado. Este tipo de ambigüedad se puede eliminar si el analizador sintáctico prevé estos casos y establece las reglas necesarias en la gramática, ya sean de tipo sintáctico, semántico, etc. De hecho, esto no impide que los resultados del método de Hindle y Rooth sean óptimos.

El sistema continúa con la creación de unos bigramas, o parejas de nombre o verbo con o sin preposición asociada. Para una frecuencia (**f**) entre una preposición (**p**) y para un término léxico, verbo o nombre (**w**) se establece la siguiente función:

$$\mathbf{f(p) = f\,(w, p)}$$

Es decir, la frecuencia de aparición de una preposición es igual a la frecuencia de aparición de esa preposición con una palabra determinada, para todos los casos en que aparezca esa palabra a lo largo de un texto. La información que se recoge de las tablas léxicas las almacena el ordenador en forma de valores sucesivos de aparición de dichas secuencias. Estos valores sirven, a su vez, para calcular el número que asigna a cada secuencia un grado más o menos grande de probabilidad de aparición. Se usa un algoritmo para extraer la información sobre los tipos de dependencias posibles a partir de la tabla de concurrencias.

La función básica para determinar la combinación **SV-SP**, **SN-SP** para una **AL** con los términos **v** (verbo), **n** (nombre), y **p** (preposición) sería como sigue:

$$\mathbf{AL\,(v, n, p) = \log_2 \frac{P(p\,/\,v\,,\,n)}{P(noun\_attach\;\;p\,/\,v,\,n)}}$$

Para evitar obtener el valor 0 en el caso en el que el nombre / verbo más preposición asociada no tenga ninguna ocurrencia se puede redefinir la estimación de la probabilidad así:

$$P(p / n) = \frac{f(n, p) + \frac{f(N, p)}{f(N)}}{f(n) + 1}$$

y lo mismo para el verbo:

$$P(p / v) = \frac{f(v, p) + \frac{f(V, p)}{f(V)}}{f(v) + 1}$$

Lo más interesante de esta propuesta es el alto índice de acierto en el análisis de oraciones con ambigüedad estructural. Hindle y Rooth demuestran que, en experimentos hechos por ellos, el margen de error del método estadístico en la asignación del SP al SN o al SV oscila entre un 12% y un 15%, mientras que con el método lexicográfico se sube hasta un 20%. Esto es así porque la información que se recoge en un diccionario es información estándar y muy general que pronto ha de renovarse. Los resultados también son mejores si se comparan con los obtenidos a partir de análisis hechos por hablantes nativos. Es por todo que se puede considerar a este método como válido en la resolución de la ambigüedad estructural.

BIBLIOGRAFÍA.

Briscoe, T. y J. Carroll. 1991 *Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars*. Technical Report 224, Computer Laboratory, Cambridge University Press.

Brown, P. et al. 1990 *A Statistical Approach to Machine Translation*. Computational Linguistics **16**:79-81.

Dagan, I., S. Markus y S. Markovitch. 1993 *Contextual Word Similarity and Estimation from Sparse Data*. 30th Annual Meeting of the Association for Computational Linguistics: 164-71

Fujisaki, T. et al. 1989 *A Probabilistic Method for Sentence Disambiguation*. Actas del 1st International Workshop on Parsing Technologies, Carnegie-Mellon University:105-14

Hindle, D. y M. Rooth. 1992 *Structural Ambiguity and Lexical Relations*. Computational Linguistics **19**(1):103-20.

Pereira, F. y Y. Schabes. 1992 *Inside-Outside Re-Estimation for Partially Bracketed Corpora*. Actas del 30th Annual Meeting of the Association for Computational Linguistics:128-35.

Sharman, R. et al. 1990 *Generating a Grammar for Statistical Training*. DARPA Speech and Natural Language Workshop:267-74.