Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional
Departamento de Ingeniería Eléctrica
Sección de Computación

Grado de pertenencia a un dominio y métodos de clasificación

Tesis que presenta
Héctor Jiménez Salazar

Para Obtener el Grado de
Doctor en Ciencias

En la especialidad de
Ingeniería Eléctrica

Director de la tesis: **Dr. Guillermo Morales Luna**

México, D.F. diciembre de 2000.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología, por haber apoyado la realización de este trabajo[1].

Al Centro de Investigación y de Estudios Avanzados del IPN, por brindarme las condiciones adecuadas para llevar a cabo este proyecto.

A las autoridades de la Universidad Autónoma de Puebla, de la Facultad de Ciencias Físico Matemáticas, y en particular al Director de la Academia para el Avance de la Educación, Dr. Rafael Campos Enríquez, por haberme otorgado todo lo necesario para cumplir con las metas establecidas[2].

A los miembros del jurado, Dr. Christian Lemaitre León, Dra. Josefina García Fajardo, Dra. Ana María Martínez Enríquez, Dr. Sergio Chapa Vergara, y Dr. Guillermo Morales Luna, por sus acertados comentarios para mejorar el documento final de tesis.

En especial, a la Dra. Josefina García por sus invaluables observaciones sobre el presente trabajo.

A mi amigo y director de la presente tesis, Dr. Guillermo Morales Luna, por corregir con esmero muchos de mis planteamientos, y por la confianza que en mí siempre ha tenido.

A mis compañeros de trabajo, Javier Albores, Rodolfo Reyes, Luis Montgomery, Jesús García, Jesús Lavalle, Manuel Martín, y muchos más, por la solidaridad de siempre.

A toda mi familia, por comprender mi acrecentada ausencia.

[1] Beca CoNaCyT No. 50084.
[2] Beca ProMeP UAP-750.

A Almendra
e Itzel

Resumen

Se desarrolló un sistema de *etiquetamiento* de textos, es decir, de marcado de cada palabra con su correspondiente parte de discurso. El sistema empleó el método de *aprendizaje basado en memoria.* Dicho método fue modificado para enfrentar el problema de *rasgos desconocidos*, los cuales pueden ocurrir en el contexto de una *palabra desconocida.* Los recursos en los que se apoya el sistema son: a) ejemplos formados por una secuencia de etiquetas y una secuencia de terminaciones de palabras –las secuencias fueron extraídas de un texto formado aproximadamente por 10,000 palabras, y este texto provino del *Corpus del Español Mexicano Contemporáneo* recopilado por El Colegio de México, b) reglas morfosintácticas basadas en la probabilidad condicional de que ocurra una etiqueta determinada, dado un contexto particular, c) diccionarios de palabras de alta frecuencia, y d) listas de gramemas para la conjugación de verbos regulares e irregulares. El desempeño del sistema, para textos del género periodístico, es cercano al 93%.

También, empleando el método de aprendizaje basado en memoria, se logró identificar las *frases nominales simples* que ocurren en un texto etiquetado con partes del discurso. En este caso, el método de clasificación discierne sobre la existencia o no del límite de una frase nominal, ubicado entre dos palabras del texto. Para textos del mismo género que el de entrenamiento, el desempeño de esta identificación es de 96.7%.

Se propusieron dos métodos para determinar el *grado de pertenencia de una palabra al dominio de un sublenguaje.* La primera propuesta empleó una *función de pertenencia rugosa* de un elemento a un conjunto. La segunda propuesta usó una métrica que calcula el "diámetro" de un conjunto de palabras, y la "distancia" entre las representaciones de dos palabras. El método apoyado en una métrica mostró ventaja sobre el método alternativo, cuando no es posible establecer el umbral, usado como *corte*, que permita decidir la pertenencia rugosa de una palabra a un conjunto. Los procedimientos mencionados requirieron la *representación* del dominio de un sublenguaje y, también, la representación de las palabras involucradas. El dominio se representó, de manera aproximada, formando un conjunto con palabras básicas del sublenguaje. Este conjunto se enriqueció por *inducción léxica* usando un corpus, con alrededor de 20,000 palabras, del área de la salud. La representación de las palabras utilizó un diccionario del español (una versión, legible por un sistema de cómputo, del Diccionario Anaya de la Lengua Española, con más de 29,000 entradas). La representación de una palabra x fue conformada por la reunión de las palabras que ocurren en todas las definiciones, del diccionario, que contengan a x.

Una extensión del cálculo anterior se realizó para frases: la determinación del *grado de pertenencia de una frase a un dominio.* Esta determinación se apoya en una propuesta sobre la formación del sentido de una frase, a partir de rasgos compatibles que contengan los conceptos asociados a las palabras de la frase.

Con los métodos anteriormente descritos, se llevó a cabo la clasificación de una respuesta a una pregunta abierta de un cuestionario del área de la salud. La solución se basó en los tipos de *actos de habla expositivos.* La clase de una respuesta se determinó calculando una puntuación que considera el grado de pertenencia, al dominio de la salud, de las palabras y frases nominales contenidas en la respuesta. La distribución de 172 respuestas en tres clases obtuvo una exactitud global de 49.41%, y mayor acierto para distinguir entre clases alejadas (93%) que entre clases cercanas (78.2%).

Palabras clave: Procesamiento del lenguaje natural, etiquetamiento basado en corpus, sublenguaje de un dominio, aprendizaje automático.

Abstract

We developed a *tagging* procedure (of words in a text with their corresponding *parts-of-speech*) and it has implemented into a NLP system. The procedure has followed the memory based learning approach to deal with *unknown features* of *unknown words* contexts. The system uses a training text consisting of 10^4 terms approximately. The text is a part of the Journalistic genre provided by the *Corpus del Español Mexicano Contemporáneo*, compiled at El Colegio de México. Further, some resources were extensively employed: a) a set of morphological rules based on the conditional probabilities of a tag in a given context, b) a lexicon of high frequency words and c) a list of grammems to determine the conjugation of both regular and irregular verbs. In our study we got a performance rate near 93%. The memory based learning method was also able to recognize *simple nominal phrases* in a part-of-speech tagged text by deciding whether there is a limit of a nominal phrase between two given words in the text. Regarding nominal phrases recognition, the system reached a 96.7% performance rate, when the application text genre coincides with the training genre.

On the other hand, we characterize a sublanguage *domain* as a word "cluster" initially given as a set of basic words and later enriched by *lexical induction* with respect to a given corpus. The membership notion of words to a domain is treated in an approximate way. We developed a method to estimate the word membership value to a domain. Our experiments were realized considering a corpus, related to the domain of "public health", of around 20 000 words.

Let us say that a word y is in the *representation* of a given word x, with respect to a dictionary, if y appears in the definition of a word z that uses also x in one of its meanings. We used a Spanish dictionary, Anaya's Dictionary Edition avalaible on-line, with almost 29 000 entries to compute the representability notion. The *word membership value to a domain* can be estimated either with the aid of a threshold value (lower-bounding the distance of the word to the domain representation) or using a *rough membership function* that compares the representation of the tested word and the representations of words in the domain. Each approach has its own advantages.

In order to test our methods we proceed to estimate the *nominal phrase membership value to domains* using the sense of a nominal phrase regarding compatible features of the concepts of words appearing in that phrase.

Our final goal was the determination of the class to which pertains an answer. This answer was provided by a person to an "open" question in an inquire related to public health. Our procedures take into account the *act of speech* approach. Classification was realized according to both nominal phrases and word membership values to public health domain. The global accuracy of a classification of 172 answers to three classes was 49.41%, and were better classified answers on "far" classes (93%) than "near" classes (78.2%).

Keywords: Natural language processing, corpus based tagging, domain sublanguage, machine learning.

Introducción

El **Procesamiento del Lenguaje Natural** (PLN) es un área de investigación orientada a desarrollar sistemas de cómputo que coadyuven a la solución de problemas lingüísticos y computacionales. Dentro del área de PLN se estudia la representación y transformación de entidades relacionadas con elementos de la lengua (habla, texto o partes de éstos).

Los sistemas de PLN tratan de aligerar el procesamiento de grandes volúmenes de información, haciendo el trabajo de manera rápida y con un grado de confiabilidad aceptable para la aplicación que fueron diseñados. La problemática implícita en la construcción de tales sistemas proviene principalmente de la falta de una explicación "suficientemente" completa de cómo procesar información para que un sistema de cómputo tenga buen desempeño sobre algún aspecto de la lengua. Además, el empleo de una computadora para emprender proyectos ambiciosos en lingüística conduce a buscar métodos computacionales que exploten la información de manera más eficiente. Es por esto que el término PLN[3] corresponde a una fusión, más que al uso de la computación por la lingüística o viceversa, ya que, entre ellas, se crea una retroalimentación al resolver sus problemas[4].

Uno de los objetivos principales del PLN ha sido la construcción de *interfaces de usuario*[5], y en menor medida, aunque de gran importancia y de mayor impacto, los sistemas para realizar *gestión de información textual* como es el caso de los sistemas para la:

- *Traducción Automática* de textos de una lengua a otra,
- *Generación de Texto* asociado a información almacenada en un sistema, y
- *Revisión Ortográfica*

Una muestra de la necesidad urgente de emprender proyectos de PLN es la traducción automática. J. Abaitua [Abaitua, 99] comenta que: "En un estudio titulado *Methods for Satisfying the Needs of the Scientist and the Engineer for Scientific and Technical Communication*, Hubert Murray Jr. revela que se registran aproximadamente 20 millones de palabras por día de información técnica. Un lector capaz de leer 1.000 palabras por minuto necesitaría 45 días, a una media de 8 horas diarias, para digerir la producción de un día. Al cabo de estos 45 días, su desfase sería de 5,5 años. Una comunidad lingüística necesitaría de 2.000 esforzados traductores para poder vertir a su lengua este caudal diario de información técnica."

Los problemas que se plantean en PLN cada vez son más importantes, y no sólo es por el hecho de saber que la cultura de la humanidad está en buena medida contenida en los textos, sino además porque las necesidades actuales de procesamiento de información demandan la ayuda de nuevos mecanismos que enfrenten el

[3]Los términos *Procesamiento del Lenguaje Natural* y *Lingüística Computacional* (LC), se utilizan en la literatura hasta en un 75% en forma indistinta [Walts, Porter, et. al. 97]. Hay campos que son de mayor interés para el PLN que para la LC. El campo de interés principal del PLN es *lógica* (12%), mientras que para LC es *traducción automática* (27%).

[4]El impacto de esta fusión ha conducido a afirmar que el avance tecnológico está en relación directa con el desarrollo lingüístico [Abaitua, 99].

[5] Es sabido que el 50% del esfuerzo en crear un producto de software se dedica a la interfaz de usuario. Claro está que el lenguaje natural no es el único ingrediente que mejoraría una interfaz de usuario, pero es esencial para lograr la efectividad de la interacción con los sistemas computacionales. Sobre las interfaces en general, hay estudios que muestran que al tener presente los aspectos de la *usabilidad del software* [Slator, Anderson, et. al. 86], en la construcción de los sistemas con interfaz, se generan beneficios estimados hasta de 5,000 veces el costo del producto.

fenómeno de la “informatización” de la sociedad[6]. El problema de la informatización es causa y, a la vez, efecto de la adaptación del ser humano al uso de información copiosa. La red *WWW* tiene un crecimiento exponencial y se estima que contiene del orden de un *terabyte* de información textual [Baeza-Yates, 98]. Sin embargo, este crecimiento no está en relación proporcional a su disponibilidad “efectiva”, puesto que es requisito el procesamiento de dicha información para mejorar su aprovechamiento.

El problema de acceso a la información ha sido tratado por las *Bases de Datos* y la *Recuperación de información* (RI). La primera se dedica a responder consultas sobre información almacenada de manera *estructurada*, la segunda, tiene la misma función pero sobre información *no estructurada*, como son los textos. Bien que la recuperación de información ha mostrado éxito, ésta no mantendría su efectividad en la WWW pues el volumen de documentos es significativamente mayor al que esta área trabaja normalmente. No por ello se descarta el empleo de la RI, sino que, por el contrario, se ha enriquecido replanteando la forma de realizar las consultas, por ejemplo, mediante el uso de *perfiles* o regularidades de preferencia, lo cual permite restringir el volumen de información al que se accede. Ésta es un especialidad cercana al PLN llamada *Filtración de información* [Belkin, Croft, 92].

Para la explotación masiva de documentos, es necesario que éstos estén clasificados, tarea con un nexo claro hacia el PLN. Considerando clases muy amplias, la *Categorización de textos* se ha tomado como una vía de solución a este problema.

La importancia de estos problemas ha generado propuestas que conducen al establecimiento de más líneas de investigación [Basili, Pazienza, 97], particularmente la *Extracción de información*, que trata de concebir a un conjunto de documentos como una base de datos formada por el resultado de la extracción, según un patrón, de hechos o eventos que se encuentran en los documentos. También es subsidiaria a la clasificación de documentos la *Elaboración del resumen* de un texto.

Como podemos ver, el acceso a volúmenes extensos de texto requiere inevitablemente del PLN.

Ha habido dos formas esenciales de abordar el desarrollo de los sistemas del PLN, lo cual se manifiesta en su eficiencia. En los sistemas orientados a la *competencia* se favorece la *cobertura*, mientras que en los sistemas dirigidos hacia el *desempeño* se tienen efectos positivos en su *precisión*. Por supuesto, la adecuada combinación de estos enfoques redunda en la utilidad.

En los inicios del PLN hubo inclinación por la competencia que implicó proporcionar una solución al problema de la “comprensión de textos” (*full text understanding* [Grishman, 97]) lo cual demanda la solución de muchos problemas de índole semántica. La efectividad de este enfoque, solamente en ámbitos muy restringidos, creó condiciones para que fuera rebasado por la segunda postura, el desempeño. Ejemplo claro de este fenómeno es la RI que, frente a los problemas semánticos de los textos, prefirió asentar su desarrollo sólo en una base morfológica.

El elemento crítico para el funcionamiento de un sistema de PLN es la información lingüística “vinculada” al método que se sigue para alcanzar una meta propuesta. En el desarrollo de un sistema de PLN es, pues, necesaria la consideración de los procedimientos que transforman los elementos de la lengua, así como los *recursos de información lingüística* en los que se apoyan los primeros. Entre los recursos lingüísticos se cuentan principalmente:

- *reglas* de tipo diverso (e.g. morfológicas y sintácticas),
- *léxicos* o catálogos de vocablos con información acorde a los requisitos del sistema, y
- *corpora*, cuyo papel es el de proveer ejemplos sobre algún fenómeno de la lengua.

La importancia de construir sistemas de PLN se traslada entonces a la de crear recursos de información lingüística.

[6]La explosión informativa se ha observado desde hace más de quince años. A principios de 1990, en el mundo, cada 20 meses se duplicaba el volumen de información, había ya más de cinco millones de bases de datos y cada día crecía en más de 10^{15} bytes la información registrada en los sistemas de cómputo (contando las imágenes satelitales) [Patietsky-Shapiro, Frawley, 91].

La constitución de recursos para el PLN plantea un problema computacional pues es indispensable el procesamiento de extensas cantidades de datos. Es de interés el apoyo de métodos automáticos o semiautomáticos para llevar a cabo tal tarea. Normalmente, dentro de un dominio del lenguaje, este proceso conlleva a un primer paso completamente manual, seguido de otro semiautomático y finalmente la liberación de un método automático. Posteriormente, se trata de generalizar este método para poder aplicarlo a otros dominios del lenguaje natural, e inclusive trasladarlo a otros fenómenos del mismo. El marco de desarrollo, implícito en este enfoque empírico, pretende ahorrar el análisis exhaustivo de abundantes muestras del lenguaje y mantener el enriquecimiento de los recursos informáticos con cada experiencia de su aplicación.

En la actualidad, los *corpora* son vistos como una fuente de información básica dentro del PLN, pues a partir de ellos es posible generar otros recursos de información. Ha sido tan grande la necesidad de los recursos de información mencionados, que para explotar *corpora*, se han adherido al PLN métodos que corresponden a otras áreas de la computación, formando nuevas subáreas como es el caso de la *Minería de Textos* [Quafafou, Zytkov, 98].

Entre los pasos para construir sistemas de PLN se encuentra la creación de herramientas, tanto para emplear recursos lingüísticos como para generarlos y, por tanto, el estudio, adaptación, creación e implantación de métodos computacionales aptos para el procesamiento dentro del campo lingüístico. La consecución de recursos de información para el PLN ha consumido mucho esfuerzo. Puede afirmarse que durante la última década se afinaron los métodos para trabajar con *corpora etiquetados* y sólo de esta forma, en ámbitos particulares, se han logrado sistemas con una eficiencia razonable. Ejemplo de ello es el avance logrado en la *Desambiguación del sentido de una palabra* [Ide, Véronis, 98]; procedimiento indispensable en la elaboración de *parsers* robustos.

Los problemas de acceso a la información textual llevan a concebir la clasificación, como un tema central en el PLN.

Se planteó el problema de investigar el grado de pertenencia de una palabra a un dominio, ya que su solución coadyuvaría a los problemas de clasificación. La pertenencia de una palabra a un dominio puede expresarse como la existencia de un sentido de la palabra en ese dominio. Y un problema más general es el de averiguar en qué grado tiene sentido una frase en un dominio.

El problema de la determinación del sentido de una palabra en un dominio, está relacionado con el de la desambiguación del sentido de una palabra [Ide, Véronis, 98] siendo el primero un caso particular del segundo. También en estrecha relación con el sentido de una frase en un dominio se encuentra la estructura conceptual correspondiente a un sintagma [Fernández, Navarro, et. al. 98], [Sertula, Kourie, 98]. En nuestro caso hemos retomado la propuesta que se hace en [García-Fajardo, 95], donde se analizan las condiciones para que un sintagma adquiera un sentido.

Se desarrolló en esta tesis un método para determinar el grado de pertenencia a un domino particular, tanto para una palabra como para una frase. La propuesta para llevar a cabo esta determinación dependió en buena medida de los recursos de información disponibles pero, además, debido a las condiciones actuales del PLN para el español, en particular el español de México, y el estado del arte del PLN, hubo necesidad de explorar los métodos indispensables en la mayoría de los sistemas de PLN. Por ello, se tuvo que desarrollar algunos instrumentos básicos de cómputo que complementaran el enriquecimiento de recursos de información. Se pretende con ésto, iniciar un análisis de los métodos que mejor se adapten a la solución de problemas del PLN, y de cuáles recursos deben acompañar estos métodos para señalar una dirección en la empresa de proyectos de mayor alcance.

El análisis de los problemas planteados condujo a dos subproblemas:

A La identificación de las frases nominales de un texto (*shallow parsing*),

B La determinación del sentido de una frase en un dominio.

Además para realizar el análisis sintáctico es indispensable asignar la marca de su parte del discurso a cada palabra de un texto. Este problema de clasificación se apoyó en una técnica de aprendizaje automático.

Por su parte, ya que el sentido de una palabra o frase en un dominio vendrá dado por la pertenencia de ese sentido en el dominio, es requisito ofrecer una representación del dominio de trabajo.

Los dos subproblemas se descomponen entonces en:

A.1 La evaluación de algunos métodos de *aprendizaje automático*,

A.2 La determinación de la *parte del discurso* de las palabras de un texto, y

B.1 El *agrupamiento* de vocablos de un dominio.

Para la evaluación que establece **A.1**, se compararon tres métodos de *aprendizaje automático*. Estos métodos fueron *supervisados*, esto es, requieren de información expresa sobre la clase que se ha asignado a algunas instancias del problema, o *colección de entrenamiento*. Los métodos analizados fueron los siguientes:

1. La identificación de rasgos relevantes [Chan, Wong, 91]. Esta técnica determina la clase de un ejemplar, cuando ésta depende, significativamente, de los rasgos del ejemplar.

2. La síntesis de atributos [Quinlan, 93]. Obtiene reglas que se representan en un árbol de decisión. La asignación de la clase a una instancia se realiza aplicando la regla que empate con la instancia.

3. El aprendizaje basado en memoria [Daelemans, van-den-Bosch, et. al. 98]. Se apoya en la similitud para determinar la clase de una nueva instancia: elige la clase de la instancia que se "parezca" más a la nueva instancia.

Con los anteriores métodos, se realizó la prueba consistente en clasificar las palabras ambiguas frecuentes ("la", "que", etc.). Tomando la lista de las cien palabras más frecuentes del español de México [Lara, Ham-Chande 79] las ambiguas frecuentes fueron aquellas palabras ambiguas que ocurrieron en esta lista. El clasificador tuvo que discernir sobre la parte del discurso que le corresponde a cada una de estas palabras al ocurrir en un texto, es decir, elegir entre pronombre y artículo la clase de "la", entre conjunción y pronombre la clase de "que", etc. El experimento procedió aplicando cada uno de los anteriores métodos a un texto y observando su precisión. Además de obtener la mayor precisión, las características del método basado en memoria, como el empatamiento no exacto y el mantener sin abstracciones todas las instancias de entrenamiento –esto es, ejemplos tanto de casos generales como excepciones–, llevaron a elegir este método para realizar la clasificación.

La determinación de la parte del discurso de las palabras de un texto (**A.2**), por basarse en un método de clasificación supervisado, implicó la utilización del *Corpus del Español Mexicano Contemporáneo*. Dicho corpus se encuentra etiquetado con sus partes del discurso y con una parte de él se definió la colección de entrenamiento. En el análisis del problema de etiquetamiento de un texto con partes del discurso, fue indispensable dividir la adjudicación de partes del discurso según diferentes grupos de palabras. Se consideraron palabras: frecuentes y no ambiguas, ambiguas frecuentes, y desconocidas. En el primer caso se emplean fundamentalmente diccionarios y reglas morfosintácticas. Para el problema de la desambiguación de palabras ambiguas frecuentes se utilizó el método de aprendizaje basado en memoria apoyándose en el contexto de la palabra ambigua frecuente –contexto formado por terminaciones de las palabras vecinas que ocurren en el texto, y la parte del discurso de éstas. Antes de determinar la parte del discurso asociada a las palabras desconocidas, se distinguió entre la asignación de la parte del discurso verbo o nominal a palabras con esta ambigüedad. Por último, se llevó a cabo la asignación de la parte del discurso a palabras desconocidas. Los casos que aplicaron la clasificación basada en memoria, hicieron uso de instancias de entrenamiento formadas por terminaciones de palabras vecinas y sus partes del discurso. Así mismo se definió una función de distancia basada en la probabilidad condicional para estimar la marca de una palabra vecina, miembro de la instancia a ser clasificada.

Se consideró una forma de análisis sintáctico parcial de las frases (para **A**), pues fueron de interés las palabras contenidas en la frase misma, más que la estructura de la frase. En la implantación de un *reconocedor de frases nominales* aplicó el método basado en memoria, tomando como colección de entrenamiento instancias formadas por las partes del discurso de palabras vecinas a la frontera derecha de una frase nominal. Previamente, en forma manual, se marcó con la frontera derecha de cada frase nominal el mismo corpus empleado para la adjudicación de partes del discurso.

Previo al cálculo del sentido de una frase en un dominio (**B**) fueron probados los resultados parciales de la representación del dominio en el problema de la determinación del grado de pertenencia de una palabra al dominio (**B.1**). Lo cual exigió de la determinación de la *afinidad* [Grefenstette, 94] que pudiera existir entre dos palabras. Así, con base en los recursos disponibles, se efectuaron los siguientes pasos:

B.1.1 Determinación de un *conjunto básico de palabras* del dominio,

B.1.2 Formación de un *cúmulo de palabras* para representar el dominio, y

B.1.3 Utilización de un método para comparar representaciones de una palabra y un conjunto de palabras empleando:

- **B.1.3.1** Una *función de pertenencia rugosa.*
- **B.1.3.2** Una *función de distancia.*

La obtención del conjunto de palabras básicas (**B.1.1**) se hizo con base en el método de inducción léxica, propuesto en [Gierl, Frost, 92]. Este método procede con un conjunto de palabras inicial, extraídas manualmente de un corpus, e itera, mientras se incremente el conjunto, con las palabras que ocurran en bigramas cuya *información mutua* [Church, Hanks, 90] sea mayor que un umbral. Para generar el cúmulo, cada una de estas palabras se representa usando un diccionario.

Según el criterio de que las cosas que tienden a ocurrir en situaciones similares son descritas por palabras que tienden a ocurrir en contextos lingüísticos similares [Morris, Hirst, 91], se utilizó como representación de las palabras su *contexto* (en vista de **B.1.3**), tomado éste como un fragmento de texto en el que ocurre la palabra. De este modo, la afinidad entre dos palabras, se determina comparando sus contextos. El contexto de una palabra x es tomado como la reunión de las definiciones de cualquier otra palabra que use a x en el diccionario. Es posible, entonces, definir una función que asigne el grado de afinidad entre dos palabras y, de manera complementaria, una función que mida la "distancia entre palabras". Primeramente, se realizó un experimento usando una función de pertenencia rugosa (**B.1.3.1**) de un elemento a un conjunto, con la consecuente necesidad de hacer *cortes* para decidir si una palabra se considera incluida. Tratando de independizar el grado de pertenencia del empleo de umbrales usados en los cortes, se concibió al conjunto de palabras representativas de un dominio como un cúmulo (**B.1.2**), esto es, un conjunto para el cual la distancia entre sus elementos es menor que la distancia entre un elemento de él y cualquier otro que no pertenezca a él. Se propuso una función de distancia para medir la cercanía entre una palabra y un conjunto de palabras (**B.1.3.2**). Esta métrica se basa en el solapamiento de los contextos extraídos de un diccionario del español. Ya que se espera que palabras poco afines tengan una distancia "grande", es necesario tomar un punto de referencia para decidir cuál es el grado de pertenencia al cúmulo. Lo anterior condujo a definir el **diámetro** de un cúmulo de palabras J, y la distancia desde una palabra x hacia él. Justamente, se clasifica x como:

incluida cuando la distancia de x hacia J es menor o igual que el *diámetro normalizado* de J;

cercana cuando la distancia hacia J es menor o igual que el diámetro de J y

externa cuando la distancia de x hacia J es mayor que el diámetro de J.

Al implantar el procedimiento que determina el sentido de un sintagma (**B**) se requirió, para cada palabra que compone al sintagma, su definición y las definiciones que refirieran a la palabra a través de un diccionario. Esta aproximación permitió calcular la pertenencia rugosa del sintagma al dominio, esto es, el grado en que el sintagma tiene sentido dentro del dominio.

Con el fin de observar el funcionamiento de los métodos antes descritos, más que establecer una evaluación definitiva de éstos, se llevó a cabo una aplicación. El problema consistió en la clasificación de respuestas, de un cuestionario con preguntas abiertas, de acuerdo con el "modo de abordar el concepto" [Carrillo-Ruiz, 90]. Se tomó en consideración el valor que puede tener un *acto de habla* [Austin, 90], para proponer una solución al problema de clasificación.

En la figura 1 se muestra un diagrama de la relación que existe entre los principales métodos implantados.

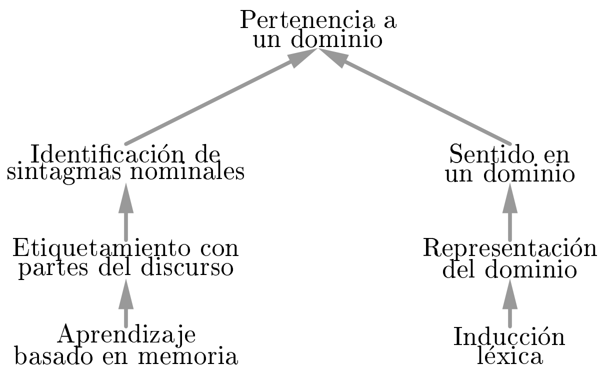

Figura 1: Relación de cooperación entre los principales métodos implantados.

La propuesta de dos funciones de distancia que ayudan a resolver problemas de clasificación, mostraron utilidad en los problemas relacionados con el PLN. La generalización de la métrica, usada en el método de aprendizaje basado en memoria, establece una pauta de mejoramiento para cualquier método de aprendizaje automático que no considere el caso de los valores desconocidos. Por su parte, la propuesta del cálculo de la distancia entre una palabra y un cúmulo se adscribe a las aplicaciones de recuperación de información. De manera especial esta métrica considera ajeno el uso de umbrales para determinar la cercanía.

Este trabajo también aporta dos herramientas y son, como se ha señalado, herramientas básicas en el desarrollo de casi cualquier sistema de PLN. El sistema etiquetador de palabras de un texto con su parte del discurso, mostró buen desempeño con el tamaño del corpus usado para su entrenamiento. En este sentido, aunque orientado a sintagmas muy simples, fue más sobresaliente el desempeño del sistema reconocedor de los límites de frases nominales. Estas contribuciones se dirigen al incremento de herramientas que complementan recursos de información lingüística. Además, respecto a las necesidades para el procesamiento del español de México, los resultados de este trabajo pueden concebirse como una plataforma para llevar a cabo proyectos ulteriores de PLN.

El presente reporte se divide en cinco capítulos y un apéndice. En el primer capítulo, *Procesamiento del lenguaje natural*, se ofrece un panorama de los campos de conocimiento, métodos y recursos que han girado en torno al PLN, se destaca entre todo la importancia del tema *Aprendizaje Automático*, que posteriormente se desarrolla en el capítulo dos. En él se expone la elección de entre tres métodos de aprendizaje automático, uno para ser empleado en problemas de clasificación. Es en el capítulo tres de *Etiquetamiento*, en el que se muestra el empleo del método de clasificación elegido para adjudicar partes del discurso a las palabras de un texto, además, se presenta un identificador de frases nominales apoyado también en el mismo método de clasificación. El capítulo cuatro aborda los problemas concernientes a la determinación de *Grado de pertenencia a un dominio*: la representación del dominio y su concepción como cúmulo, y la propuesta para considerar el sentido de una frase en un dominio particular, empleando para este propósito la obtención del grado de pertenencia de una palabra a un dominio. El capítulo cinco trata la *Aplicación a la clasificación de frases*, de las herramientas antes construidas. En el apéndice A se expone una especificación formal del sistema de etiquetamiento.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1

Procesamiento del lenguaje natural

En este capítulo haremos un brevísimo recuento del desarrollo del PLN, para acercarnos al estado del arte. En esta perspectiva, esperamos que el trabajo realizado en esta tesis pueda ser mejor ubicado. Mencionaremos las raíces del PLN, los enfoques que hacia él se han tenido en su relación múltiple con otras áreas de la computación, y el actual paradigma para resolver un buen número de problemas dentro del PLN. Finalmente, cuál es, a nuestro parecer, el área de la computación que tendrá mayor impacto en el desarrollo de sistemas de PLN [Jiménez-Salazar, 96].

Las características del lenguaje natural plantean que la construcción de los sistemas de PLN tengan un tratamiento especial dentro del conjunto de problemáticas complejas[1] de la computación. Siguiendo la caracterización que hace E. Hajičova [Hajičova, 87] , tenemos que el lenguaje natural es *universal* porque es posible expresar con él cualquier cosa que los hablantes deseemos comunicar. En este sentido, los lenguajes artificiales no pueden ser comparados con el lenguaje natural, aunque los lenguajes artificiales contengan algunos rasgos de generalidad tomados del mismo lenguaje natural. El lenguaje natural es *multifacético*. Si bien cada "sublenguaje" tiene su terminología, no se reducen en él las propiedades estructurales de la lengua que lo contiene. También, el lenguaje natural es *irregular* ya que sus expresiones no corresponden unívocamente a sus significados. Hay dos posibles fuentes que violan la relación de univocidad: una expresión puede tener más de un significado, *ambigüedad*, o bien, un sentido puede ser dado por más de una manera, *sinonimia*. El fenómeno de ambigüedad ocurre en todos los niveles del lenguaje (léxico, sintáctico y semántico) y, como veremos, representa el problema central del PLN.

A partir de la *estructuración del sistema de la lengua*[2] se han conformado en el PLN áreas de estudio alrededor de clases de problemas. Antes de exponer una visión panorámica de cómo se están tratando actualmente los problemas del PLN, presentaremos algunos antecedentes.

1.1 Antecedentes

Debido a la disponibilidad de los recursos y a los propósitos derivados de los problemas que deseaban resolverse, en una primera etapa dominó el enfoque de la construcción de los sistemas de PLN apoyado en teorías lingüísticas, esto es, el aparato conceptual de la teoría proporcionaba las reglas que eran adaptadas a los sistemas de cómputo para resolver problemas. Este enfoque contrasta con el que en la actualidad predomina que es el *enfoque empírico*, y que más adelante será explicado.

Durante los inicios de la computación, se tuvo énfasis en el *parsing*, debido a la necesidad de producir *compiladores* de los *lenguajes de programación*[3]. Aunque, en los compiladores, la semántica se reducía

[1] Problemáticas complejas son las consideradas en el programa HPC (*High Performance Computing*) [CSTNC, 92] entre las que se encuentran la *Visión por Computadora*, la *Predicción del Clima*.

[2] La caracterización que se usa para estudiar cualquier lengua: fonología, sintaxis y semántica [García-Fajardo, 96].

[3] Hubo una influencia notable de la corriente generativista, de hecho la computación retomó para sus fundamentos la je-

a la generación de código objeto[4], en alguna medida esta concepción se heredó a los sistemas de PLN [Pereira, Warren, 80] , aún cuando los sistemas profundizaran en los aspectos de representación[5], pocos sistemas tomaban con más realismo al léxico y a la semántica [Small, 87]. La raíz de la *Semántica Formal* en la *Lógica Matemática* [Frege, 73] funcionó como catalizador en el predominio de *Programación Lógica* para implantar sistemas de PLN. La atractiva propuesta de concebir un programa cómo *lógica+control* [Kowalski, 79], [Dahl, Popowich, 90], [Popowich, 94], venía a resultar ideal en el procesamiento del lenguaje pues una regla gramatical puede representarse directamente como una *cláusula de Horn* y, además, en el mismo lenguaje de programación se representa la *forma lógica* de un enunciado. Otra área subsidiaria de la programación lógica como las *Bases de Datos Deductivas*[6], y la consecuente conexión de la lógica con las *Bases de Datos* y la *Programación Distribuida y Paralela* [Ceri, Gottlob, et. al. 90] reforzaron este enfoque hasta principios de los noventa[7].

Hay que subrayar la necesidad que había por el uso de un "subconjunto" del lenguaje natural, sin pretender que a través de este sublenguaje se estableciera la forma principal de comunicación con una computadora[8], puesto que los sistemas usados como interfaces tratan sólo de facilitar el uso de algunos sistemas implantados en una computadora. Es muy probable que la justificación del interés en usar un subconjunto del lenguaje natural haya sido el mejoramiento de sistemas de *Recuperación de Información* [Zhengfu 94], que ofrecían la consulta a bases de datos a través de operaciones lógicas con palabras clave. Esto es, la forma de consulta se enriquece, por ejemplo, permitiendo el uso de frases nominales[9] en un sistema de recuperación de información [Lewis, Jones, 96], [Strzalkowski, 95], [Tzoukerman, Klavans, et. al. 97], [Zhai, 97]. Hubo también sistemas [Binot, 91] que tomaron como base la *semántica de Montague* [Montague, 77] para fortalecer el empleo de un subconjunto del lenguaje natural [Colmerauer, 82], superando muchos obstáculos que implica la deducción con una *lógica intensional* [Thayse, 89]. Sin embargo, el uso de un subconjunto del lenguaje natural tuvo problemas pues no resultaba fácil que el usuario de un sistema, con una interfaz orientada a reconocer expresiones de una parte de la lengua, se adaptara a los límites del sublenguaje definidos por el diseñador de la interfaz; ello equivale a aprender un nuevo lenguaje de programación sobre la marcha[10].

Los *Sistemas Expertos* aclararon la idea sobre la suficiencia de emplear subconjuntos de casos o elucidaciones. Un *sistema experto* puede verse como un sistema computacional que realiza inferencias tomando como base el conocimiento de un experto humano en un dominio [Cuena, Fernández, et. al. 86]. Aunque hubo contribuciones importantes, como el valioso hallazgo de molibdeno en el que participó el sistema PROSPECTOR [Duda, Hart, et. al. 76], se observó la necesidad de incorporar *conocimiento de fondo*[11], es decir, conocimiento que no es propio del dominio pero que su empleo es necesario para obtener conclusiones; por ejemplo, *reglas de generalización* [Michalski, 80]. Así pues, no bastaba para resolver un problema de un dominio

rarquía de Chomsky como una forma de clasificar los lenguajes y establecer el poder computacional de diversos mecanismos [Moll, Arbib, 87].

[4]Esta idea está implícita en la traducción dirigida por sintaxis que circunscribe una semántica composicional a las producciones gramaticales.

[5]Es relevante el trabajo realizado en el programa SHRDLU [Winograd, 76], que a principios de los años setenta mostró la necesidad de ampliar los aspectos sintácticos y semánticos. En este sistema el uso del conocimiento del mundo viene a ser su distinción.

[6] Una *base de datos relacional* con una *base de datos virtual*, que sirven para deducir nuevas relaciones.

[7]La agitación que hubo dentro del área de programación lógica entró en estabilidad en 1992 al terminar una etapa del proyecto japonés para construir la *computadora de la quinta generación* [Shapiro, Warren, 93].

[8]En 1950 A. M. Turing propuso su famosa prueba para determinar si un sistema podía ser considerado inteligente: si un usuario, al comunicarse mediante escritura a través de una terminal, no distinguía si su interlocutor es o no un sistema de cómputo, entonces se considera que ese sistema pasa la *prueba de Turing*. Ningún sistema ha pasado esta prueba. Un sustituto práctico de este reto es la conferencia MUC (*Message Understanding Conference*) a la que concurren equipos de trabajo sobre PLN para extraer información de textos. Dichas conferencias fueron iniciadas por el *Naval Ocean System Center* desde fines de los ochenta, para evaluar los sistemas de PLN [Lehnert, Sundheim, 91].

[9]Recordemos que los sistemas de consulta sólo permiten combinaciones booleanas de palabras; es fácil imaginar que "sistema" AND "distribuido" tendrá menos precisión que "sistema distribuido".

[10]No ha quedado completamente clara la necesidad de "usar lenguaje natural" –"qué tanto", "cómo" y "cuándo"– para una interfaz de un sistema de cómputo. Parece que falta tener más experiencias con los sistemas que realmente lo necesitan y, determinar así, la función del PLN dentro de una interfaz que se sirva en esencia de este recurso. En los ochenta, por ejemplo, se había previsto la dificultad de usar solamente el lenguaje natural como una forma de interacción con un sistema de cómputo [Slator, Anderson, et. al. 86], después de mayor avance en el área del PLN, se encontró "sorpresivamente" que era una forma inadecuada [Palanivel, Helander, 91].

[11] Al respecto se han tenido que enfrentar varias problemáticas como la de *ramificación del conocimiento* [McCarthy, 77] y el proceso de descubrimiento [Noé, 98].

especializado, el conocimiento subyacente en la teoría de ese dominio[12]. pues parte de este conocimiento se mezclaba con *conocimiento del mundo*[13], La reunión y sistematización de la información necesaria para resolver un tipo de problemas de un dominio o *base de conocimientos*[14] resultaron ser el "cuello de botella" en el desarrollo de muchos sistemas llamados "inteligentes", con el subsecuente impulso del área de *Aprendizaje Automático* y posteriormente la *Minería de Datos*. Era necesario construir bases de conocimiento de un dominio para lo cual se propuso disponer, además del conocimiento de un experto, de problemas resueltos en ese dominio. Es decir, explotar el cúmulo de experiencias sobre problemas resueltos. Cada problema y su solución se ve como un ejemplo que contribuye a la construcción de operaciones que permiten resolver nuevos problemas del dominio.

La generación de bases de conocimiento, coincide con la revitalización del estudio empírico del lenguaje natural o enfoque *basado en corpus* [Church, Mercer, 93]. Antes de describir esta nueva etapa en el desarrollo de sistemas de PLN es importante notar que, ampliamente hablando, el enfoque seguido en el desarrollo de sistemas, fue el establecido por el *ciclo del desarrollo del software*, esto es, el análisis del problema, la propuesta de un modelo[15], seguido de las etapas de diseño, implantación y pruebas. Para el desarrollo de un sistema computacional, la elucidación de un modelo basado sólo en una *especificación*, ha sido, durante mucho tiempo, el punto principal de arranque. El desarrollo de sistemas de PLN no fue la excepción de esta estrategia, y las descripciones sobre un fenómeno de la lengua se proporcionaban, normalmente, por el equipo de diseño [16]. Se trataba, así, de un desarrollo en el que las *pruebas* del sistema se derivaban de los *requisitos funcionales*, pero, por otra parte, la medición del desempeño del sistema se llevaba a cabo dentro de un entorno cualitativamente estrecho[17]. Fue necesario, entonces, considerar un entorno más amplio como punto de partida para el diseño y construcción de un sistema de PLN. M. Walker y J. Moore [Walker, Moore, 97] caracterizan esta etapa del PLN de la siguiente manera: "The role of empirical methods is to help researchers discover general features by analyzing specific discourse phenomena or programs that interpret or generate them. Once relevant features are identified, hypotheses about the relationship between them can be formed, and controlled studies that test the hypothesized relationships can be devised".

1.2 Enfoque basado en corpus

La mayoría de los sistemas de PLN se habían orientado a trabajar con un subconjunto del lenguaje natural y, por otro lado, el desarrollo de los mismos se apoyaba en una especificación que era producto de un modelo asentado en casos representativos de un fenómeno particular de la lengua, necesario para la aplicación desarrollada. Dominaba, en este ambiente, la necesidad de construir interfaces de usuario, pero había también proyectos de PLN aparentemente ajenos a las interfaces de usuario. Los sistemas de etiquetamiento de *partes del discurso* de un texto, que se empezaron a construir desde la década de los sesenta[18] [Franz, 96], son necesarios para el *parsing*, pues cada palabra requiere de su categoría gramatical, para poder aplicar la

[12] En PLN, hay una línea en la que se realiza la construcción de *ontologías* como un medio que suple la estructuración de conceptos alrededor de un dominio [Mahesh, Nirenburg, 95].

[13] Durante varios años, el sistema CYC [Lenat, Guha, et. al. 90] ha recolectado un número cuantioso de *axiomas* sobre el conocimiento del mundo para lograr una base de conocimientos que provea del "sentido común" a un sistema de cómputo, y, en particular, realizar PLN [Barnett, Knight, et. al. 90].

[14] *Knowledge Data Base*, o (*KDB*), llamada así, pues proviene de una extensión del concepto de bases de datos, para la que se escoge una representación de la información (*representación de conocimiento*) y que supone un mecanismo de inferencia para producir nuevo "conocimiento".

[15] Para muchos problemas "conocidos" esto se reduce a la elección de un método o combinación de éstos.

[16] Recordando que en los sistemas expertos el especialista no puede construir su base de conocimientos, lo mismo sucede con los hablantes de una lengua, quienes no pueden disponer de tal información para construir un sistema de PLN.

[17] Esto es casi equivalente a diseñar un sistema con base a ciertos datos y probarlo sólo con ellos mismos, lo cual difiere en mucho de un *sistema supervisado* en el que los datos, con los que el sistema define su comportamiento, son suficientemente representativos del entorno donde va a aplicarse. Vale la pena citar el comentario que hacen L. Guthrie, J. Pustejovsky, Y. Wilks y B. Slator [Guthrie, Pustejovsky, et. al. 96]: "In a moment of great honesty five years ago, a group of AI researchers of natural language processing admitted in public (in answer to a question by Bran Boguraev) how many words there really were in the vocabularies of their systems. Of the answers, the average was 36, a figure often taken to be a misprint when it appears, though it was all too true."

[18] Estos sistemas asignan a cada palabra que ocurre en un texto su parte del discurso que le pertenece y, por tanto, deben llevar a cabo un proceso de desambiguación.

reducción correspondiente a una regla gramatical. Un sistema pionero del interés anterior es el identificador de frases nominales de K. Church [Church, 89] con soporte estocástico, el cual presentaba ya la fisonomía de los actuales sistemas de PLN que dan servicio a otros, incluso con una precisión que fue difícil rebasar decididamente en un periodo de varios años. Tal ejemplo mostró la polarización que hubo dentro del área de PLN hacia intereses diferentes a la solución de problemas concretos.

El enfoque empírico hacia el PLN y hacia otros problemas de la computación (como los sistemas expertos y el aprendizaje automático) modifica, o más bien enriquece en sentido amplio, el ciclo de vida de un sistema de cómputo, pues añade a la componente de análisis algo más que una especificación informal y conlleva a la precisión de hipótesis, aprovechando la historia del desempeño de los sistemas anteriores[19]. El enfoque empírico ya se había ejercido en diversas tareas de la lingüística durante la primera mitad del siglo XX [McEnery, Wilson, 96], . Sin embargo, el surgimiento del programa racionalista[20] atrajo gran cantidad de proyectos para trabajar en esa línea[21]. Más adelante, las condiciones tecnológicas favorables[22] y el desempeño mostrado por los sistemas basados en corpus, invirtió la correlación anterior. La revitalización del empirismo cambió de dirección muchas de las investigaciones en lingüística[23], convirtiendo al corpus en la base esencial de múltiples proyectos [Young, Bloothooft, 97], [Marcus, Marcinkiewicz, et. al. 93], [Thomas, Short, 96], [Basili, Pazienza, et. al. 96].

El enriquecimiento de la metodología de los estudios del discurso considera los siguientes pasos [Walker, Moore, 97]: identificar características del fenómeno a estudiar; determinar con un modelo causal cuáles de estas características influyen de manera determinante en el comportamiento del fenómeno; evaluar el desempeño del modelo, considerando las características y el comportamiento observado, y generalizar las características, siempre que el modelo haga predicciones "correctas" para conjuntos de casos[24].

Para esta nueva visión del PLN, se encuentran también nuevos problemas. Es destacable el requisito de un *léxico*, para poder efectuar las tareas de casi todos los sistemas de PLN. La traducción automática, la recuperación de información, la extracción de información y muchas otras problemáticas coinciden en calificar a este recurso como indispensable. De esta forma el léxico se convierte en el núcleo de muchos sistemas de PLN. De modo muy general, el léxico debe contener un catálogo de palabras e información sobre ellas[25]. La información asociada a cada palabra varía según la aplicación, pero en todos los casos se trata de información lingüística útil para apoyar variados procesos en los sistemas de PLN [Krovetz, 93], [Basili, Pazienza, 97], [Boguraev, Pustejovsky, 96]. Se ha construido léxicos variados o diccionarios orientados a servir de apoyo a los primeros [Calzolari, 95], [Danlos, Gross, 88], [Suematsu, 94], [Rinaldo, Strutz, et. al. 94], [Brent, 93], sin ninguna homogeneización de su formato ni de su contenido, pero sólo su explotación en masa podrá lograrse con una estandarización [Calzolari, 91][26]. La forma de construir un léxico o, en particular, una base de datos léxica[27], ha variado desde el empleo de métodos manuales o semi-manuales [Beckwith, Fellbaum, et. al. 91] y más comúnmente a partir de los *diccionarios legibles por un sistema de cómputo* (o MRD de *Machine Readable*

[19] Etapa que fue posible cuando ([Patietsky-Shapiro, Frawley, 91]) las bases de datos registraron las operaciones de un sistema funcionando durante un periodo relativamente largo, después se pasó a explotar la información almacenada en este sistema con el fin de mejorar el desempeño en su dominio de aplicación.

[20] El racionalismo está orientado a desarrollar una teoría del procesamiento del lenguaje natural y, a la vez, pretende mostrar que de esa forma el ser humano lleva a cabo el procesamiento del lenguaje. Se centra, así, en el conocimiento internalizado del lenguaje [McEnery, Wilson, 96].

[21] También puede ser que haya influido un criterio "económico", pues el uso de reglas reduce el espacio de búsqueda [Church, Mercer, 93].

[22] Significa el aumento creciente de la disponibilidad de sistemas computacionales potentes a más usuarios, y además, recursos útiles para resolver problemas, como corpora legibles por sistemas de cómputo y algoritmos adecuados para su explotación.

[23] Algunos datos que precisan el impacto del empirismo lo presenta M. Walker y J. Moore [Walker, Moore, 97]: en los artículos de *Association for Computational Linguistics*, el número de trabajos sobre semántica, pragmática y discurso, con enfoque empírico fluctuaba entre 8% y 20% antes de 1993. Después de ese año ascendió a 40%. Ya en 1996 este porcentaje fue de 75%.

[24] El interés por poner en práctica esta nueva manera de resolver problemas alrededor de la lingüística impulsó el surgimiento de organizaciones para el incremento de los corpora como: Association for Computational linguistics/Data collection Initiative (*ACL/DCI*), European Corpus Initiative (*ECI*), British National Corpus (*BNC*), Consortium for Lexical Research (*CLR*), y Text Encoding Initiative (*TEI*). Se ha logrado reunir corpus hasta de cientos de millones de palabras [Thomas, Short, 96].

[25] Entre los primeros trabajos que plantean la problemática de construcción de léxicos están [Amsler, 89] y [Zernik, 91].

[26] El comité *EAGLES* (Expert Advisory Group on Language Engineering Standards) ha tenido bastante participación en este sentido [EAGLES, 93], [EAGLES, 96].

[27] O *LDB*, de *Lexical Data Base*, una base de datos puesto que debe ser legible por un sistema de cómputo y habrá que mantener un formato particular (tupla, compleja y fija, o estructurada) [Amsler, 89], [Ide, Maitre, et. al. 93].

Dictionary) [Ide, Véronis, 89], pasando por corpora con los cuales se realiza un trabajo semiautomático usando *agrupamiento* [Takenobu, Makoto, et. al. 95], [Grefenstette, 93], para establecer propiedades de sus entradas. Los léxicos, como se ha dicho, están dirigidos a ámbitos diversos y participan en los diferentes niveles de los sistemas de PLN, por ejemplo, en traducción automática para la determinación del sentidos de palabras, o bien, en el *parsing*, para ayuda al etiquetamiento de partes del discurso.

También, en este resurgimiento del empirismo, se pensó que los MRD podrían resolver otro de los "cuellos de botella" de los sistemas de PLN: la semántica[28]. Se vio [Ide, Véronis, 89] que, sin restar su utilidad, los MRD por sí solos eran insuficientes[29], pero en combinación con un corpus ayudan mucho a resolver el problema de adquisición léxica.

La etapa empírica retoma al corpus, MRD o cualquier otra información masiva sobre el lenguaje, que llamaremos **recurso lingüístico**, para constituir la fuente de información utilizada en la elaboración de un sistema de PLN. Debe señalarse, sin embargo, que un corpus es información en bruto y es necesario, a partir de él, construir una *base de conocimientos*. Anteriormente se empleaban recursos lingüísticos para guiar las decisiones de un sistema, principalmente apoyados en análisis estadísticos que conseguían [Church, Mercer, 93] "información valiosa"[30]. Los sistemas de *Reconocimiento de Voz* hicieron patente la bondad de los métodos basados en corpus, y esto impulsó su uso en otras áreas. En los reconocedores de voz se han empleado clasificadores estocásticos [Young, Hauptmann, et. al. 89] que fueron vistos como retos para otras estrategias no estocásticas [Schütze, 93], [Brill, 94]. Era necesario, entonces, obtener relaciones o clases a partir de una fuente de información. El aprendizaje automático se fue introduciendo al PLN hasta llegar a ser el punto de unión más fuerte entre la computación y la lingüística.

No es inesperado que, implícita o explícitamente, en todos los problemas de PLN antes planteados esté presente la ambigüedad [Alshawi, Carter, 94], [Azzam, 95], [Franz, 96], [Hirst, 87], [Oflazer, Tür, 97], [Palmer, Hearst, 94], puesto que tratan directamente con representaciones de elementos de la lengua los cuales portan diversas interpretaciones según su contexto. La desambiguación trata en esencia de elegir una clase, para algún elemento de la lengua entre varias posibles; lo cual se traduce en clasificación[31]. Esta es una forma de ver al aprendizaje automático como eje del PLN.

1.3 Aprendizaje automático

El *Aprendizaje Automático* (AA) refiere a la realización automática de cambios en estructuras de información o procesos de cómputo que modifican en esencia el comportamiento de un sistema basado en cualquiera de estas dos entidades: estructuras de información o procesos. Dichos cambios se representan en la misma entidad computacional que cambia. Es de interés, por tanto, para el AA estudiar la representación de entidades computacionales y, por supuesto, los métodos subyacentes a las transformaciones. La representación de la entidad computacional sujeta a transformaciones procura ampliar su rango de aplicación, particularmente, después del aprendizaje, la nueva representación será más general.

El AA tiene sus raíces en el área de *Reconocimiento de Patrones* estadístico y posteriormente sintáctico [Sun-Fu, 74], [González, Thomason, 78]. Partiendo de que el proceso de aprendizaje (en general) podía expresarse como un algoritmo, se empezó a considerar a este proceso como una búsqueda en un *espacio*

[28]Los recursos de información lingüística de tipo semántico son los que menos han sido desarrollados. Se debe, en primer lugar, por las dificultades intrínsecas de esta área de la lingüística, y por otro, a que requieren el empleo de otros recursos que han estado en construcción (e.g. *parsers*). Se ha desarrollado, a pesar de estas dificultades, por ejemplo, sistemas de desambiguación del sentido de una palabra, y sistemas para determinar el tipo de acto de habla de una frase dentro de un diálogo. Hay, por supuesto, muchos intentos de acercar las diferentes propuestas teóricas a sistemas de aplicación, como el manejo de las *Máximas de Grice* para realizar *referencias* en la generación de texto [Dale, Reiter, 95].

[29]Dentro de un diccionario, el término *genus*, que es la clase de la palabra definida en una entrada, aparece irregularmente y esto obliga nuevamente a construir un sistema de PLN para explotar un MRD; por ejemplo, fue necesario desarrollar *parsers* para las definiciones contenidas en los MRD [Boguraev, Briscoe, 89].

[30]Este concepto se refinó más en la minería de datos al definir *información interesante* [Patietsky-Shapiro, Frawley, 91] como aquélla que ayuda a describir sucintamente los hechos, tomando como punto de referencia una meta.

[31]Del libro de Z. Harris [Harris, 66] emerge esta idea con bastante claridad y van den Bosch [Bosch, Daelemans, et. al. 96] lo dice de manera explícita: "Most linguistic tasks can be described as classification tasks, i.e., given a description of an input in terms of a number of feature-values, a classification of the input is performed".

de hipótesis. Así por ejemplo R. Michalski [Michalski, 80] propone identificar un objeto a partir de reglas aprendidas mediante criterios de generalización aplicados a un conjunto de objetos de diversas clases, y el uso de conocimiento de fondo. Esta forma de abordar el aprendizaje tiene un carácter *simbólico* y *heurístico*, pero a fines de los ochenta se concibió la necesidad de integrar otros elementos como aquellos que explotan las regularidades estadísticas y el carácter algorítmico [Neri, Saitta, 97], regresando en cierta forma a los métodos originales [Minsky, Papert, 69] que de alguna manera se adscriben a la inteligencia artificial "fuerte". Tal distinción define dos tendencias en la inteligencia artificial, la "fuerte" y la "débil", que clasifican los enfoques para resolver problemas[32].

Ha habido muchos enfoques hacia el AA, en particular deben tomarse en cuenta las teorías en torno a esta área de la computación. Partiendo de la teoría de lo "aprendible"[33] se desarrolló una línea de investigación sobre lo que podía ser aprendido. Este tipo de resultados han influido crucialmente en la dirección del PLN. Por ejemplo, la *Inferencia Gramatical* trata de obtener la gramática de un lenguaje a partir de ejemplos del mismo[34]. Claramente hay una conexión entre esta área y el enfoque hacia el PLN basado en reglas. Los resultados de M. Gold [Gold, 67] han sido significativos[35], para elegir caminos diferentes al de inferencia gramatical [Lucas, 93]. No obstante que ha habido avances en la línea de inferencia gramatical aplicada al PLN [Ahonen, Mannila, et. al. 94], [Suppes, Böttner, et. al. 96], en general, hacia el AA se reforzó el enfoque empírico. Este sesgo se debe también al fenómeno de la disponibilidad de grandes bases de datos a nivel mundial que se tradujo en bautizar al AA como *Minería de datos*[36], además de la influencia del hallazgo del sistema BACON que "redescubrió" la *ley de los gases ideales* [Neri, Saitta, 97]. Con una raíz compartida a la de minería de datos se encuentra la *Programación Lógica Inductiva*, que trata justamente de inducir *cláusulas* a partir de *hechos*, [Bergadano, 93], [Dhar, Tuzhilin, 93], y que también ha sido usada en el PLN [Zelle, Mooney, 95].

Otros enfoques diferentes hacia el aprendizaje se han relacionado con características del aprendizaje humano, por ejemplo la analogía[37], el conexionismo[38] y la lógica difusa[39].

Una clasificación desde el punto de vista de la interacción de un sistema de AA con el entorno, o del grado de dependencia que tiene el sistema para obtener sus resultados, considera dos tipos de sistemas los supervisados y los no supervisados[40]. Los primeros, parten de un conjunto de *ejemplos* o *instancias* de entrenamiento, con el cual producen su nuevo comportamiento que es medido a por su desempeño frente a instancias diferentes. Los segundos [Matthews, Hearne, 91], [Cover, Hart, 67], construyen clases empleando normalmente una función que maximiza la similitud[41] de los elementos que pertenecen a la misma clase, a la vez tratan de minimizar la similitud de los elementos de diferentes clases. La relación de estos sistemas

---

[32] La postura "fuerte" toma a la *homeostasis* como criterio de equivalencia. El cerebro, por tanto, está en la misma clase que un termostato; considerando aparte la eficiencia y complejidad del sistema [Penrose, 96]. En el caso del PLN, es muy conocido el ejemplo de la "habitación china" de John Searle [Penrose, 96], razonamiento con el cual Searle confronta la postura "fuerte", argumentando que no hay "comprensión" por un sistema de cómputo aunque éste pase la prueba de Turing.

[33] L. Valiant [Valiant, 84] definió una teoría basada en el aprendizaje de fórmulas proposicionales y presenta las condiciones para poder aprender, en tiempo polinomial.

[34] Muchos trabajos se orientaron al aprendizaje de lenguajes artificiales como [Tamura, Tanaka, 73], [Kotani, 88], [Sakakibara, 90] y [Jansen-Winkeln, 85].

[35] M. Gold define el término *aprendizaje en el límite* como una forma de asegurar que un mecanismo aprenda, y demostró que no puede aprenderse un lenguaje regular solamente usando ejemplos positivos. Sin embargo, aunque hay algoritmos para aprender en el límite reglas gramaticales, su complejidad es *NP*.

[36] A diferencia del AA "clásico", la minería de datos [Patietsky-Shapiro, Frawley, 91] trabaja sobre bases de datos muy grandes, incompletas, dinámicas y con *ruido*. Este nuevo enfoque revitalizó la mirada hacia el AA ampliando su impacto en las aplicaciones.

[37] Esta operación es importante considerarla en los sistemas de AA, pues más del 60% de las operaciones que realizamos los seres humanos son de analogía. En [Hasse, 95] se ha usado este criterio para el PLN.

[38] Éste proviene de los años cuarenta y fue retomada en los ochenta. Trata de definir la estructura y función de las redes neuronales [Rumelhart, Widrow, et. al. 94]. , y se ha empleado también en el PLN [Wyard, 92].

[39] E. Rosch [Rosch, 73], desde el punto de vista de la cultura, fundamenta, con la gradación de pertenencia a una clase, la teoría de los conjuntos difusos [Morales-Luna, 88], [Dubois, Prade, 80]. En este camino puede ubicarse los trabajos con relaciones de *tolerancia* como el de los conjuntos rugosos (*rough sets*) [Iwiński, 89].

[40] También puede incluirse a los sistemas *parcialmente supervisados*, los cuales trabajan como no supervisados pero tienen etapas de retroalimentación.

[41] Puede ser expresado, también, como minimización de la *distancia* entre los elementos, o bien, como la maximización de la probabilidad de que el elemento pertenezca a la clase.

con el PLN es amplia pues ha habido necesidad, por ejemplo, de construir *tesoros* [42] [Crouch, Yang, 92], de manera que sea posible la clasificación de objetos sin tener conocimiento previo de las clases [Cover, Hart, 67], [Brill, 95]. Gran parte del trabajo en aprendizaje no supervisado es la continuación de una línea iniciada en clasificación y reconocimiento de patrones.

Los sistemas de *clasificación* [Breiman, 84] [Duda, 73] y aquellos que encuentran *similitudes* entre objetos [Rousseau, 98], [Kozima, 96], [Shapiro, 94], [Rada, Mili, et. al. 89], [Song, Johannesson, et. al. 96] pueden verse como anexos del AA porque al incorporar un clasificador a un sistema de cómputo puede distinguirse una entidad computacional que se enriquece con la clasificación. A su vez, las funciones de similitud encuentran su nicho en los clasificadores, empero, la determinación de este tipo de funciones queda sujeta a la forma particular de atacar un problema, y a menudo dependen del dominio[43]. Las funciones de similitud juegan así un papel especial en la clasificación. Esto conduce a poder plantear el problema de clasificación mediante una función complementaria, a saber, entre más similares sean dos instancias pueden ser consideradas menos "distantes". Una función de *distancia* o *métrica* puede definirse a partir de una función de similitud y viceversa. Muchos clasificadores han sido desarrollados con esta orientación . Lo importante de trasladar el problema de clasificación al trabajo con una métrica es que heredamos propiedades de los *espacios métricos*, por ejemplo, viendo las instancias como *puntos*, puede ser planteada la clasificación como la localización de la clase del punto en cuya *vecindad* se encuentre el punto a clasificar. Esta forma de abordar el problema tiene importancia, ya que se ha establecido conexión entre la semántica de la lógica de primer orden y la topología [van-Frassen, 87], el razonamiento y una representación analógica de los enunciados [Barnden, 89] [Johnson-Laird, 85] y, más recientemente, entre la representación de los conceptos y la deducción haciendo uso de los conceptos de distancia[44]. Este tipo de representaciones, es suplementaria a la forma en que comúnmente se hace dentro de los sistemas de cómputo, i.e. la *representación simbólica*, por ello conviene tomarla en cuenta dentro de la representación de conocimiento, sobre todo en los sistemas de PLN[45].

La investigación aplicada en las líneas esbozadas, debe ser ubicada en una etapa apropiada del desarrollo de los sistemas de PLN. Para lograr buen desempeño en las aplicaciones habrá que apoyarse en recursos de información lingüística, independientemente de la teoría que las sustente, i.e. en el paradigma empírico [Church, Rau, 95]. El uso de una plataformas computacionales para desarrollar sistemas de PLN permite, por tanto, precisar los planteamientos de la teoría y avanzar en la empresa de proyectos importantes dentro del PLN.

[42] Término empleado para referirse a la organización jerárquica de palabras o documentos para ser usados en la recuperación de información [Iwayama, Tokunaga, 95], [Takenobu, Makoto, et. al. 95], .

[43] Un método importante es el *Basado en Memoria*, pues en muchos casos se ha visto que su "función de similitud" no depende del dominio [Daelemans, van-den-Bosch, et. al. 98].

[44] Este es el enfoque llamado *razonamiento espacial* que se ha pretendido considerar con *adecuación psicológica*, esto es, que contenga un modelo adecuado del conocimiento y razonamiento humanos en sus aspectos relevantes [Knauff, Rauh, et. al. 97], [Tversky, Lee, 98].

[45] El debate entre quienes defendían la postura de que el ser humano representa la información de manera simbólica *vs* quienes postulaban la representación analógica, se mantuvo por varios años. En [Rivière, 86] se exponen pruebas de que son necesarias las dos formas de representación.

Capítulo 2

Aprendizaje automático

En este capítulo se presenta una revisión de tres métodos de aprendizaje automático (AA) para poder seleccionar el que más se ajuste a la solución que será dada a algunos problemas del PLN. La primera sección define los conceptos que van a ser empleados. Cada una de las tres secciones subsiguientes describe un método de AA, y, al final, se hace la elección de uno de ellos, comparando su precisión en el problema de desambiguación de palabras frecuentes con varias funciones lingüísticas en los textos. Empezaremos por motivar las nociones generales del AA.

Es general que los seres humanos asignemos un conjunto de rasgos a cada experiencia, y es común que, para resolver un problema, empecemos por clasificar sus partes. Aprendemos clasificando. La clasificación es considerada como una de las operaciones más empleadas en diversos campos de la ciencia. La relación entre aprendizaje y clasificación puede mostrarse con el área de la computación llamada *aprendizaje automático*.

El aprendizaje automático refiere a la realización automática de cambios en estructuras de información o procesos de cómputo que modifican en esencia el comportamiento de un sistema basado en esta entidad computacional, trátese de una estructura de información o de un proceso. Dichos cambios se representan en la misma entidad computacional que cambia. Es de interés, por tanto, para el aprendizaje automático estudiar la representación de entidades computacionales y, por supuesto, los métodos que subyacen a sus transformaciones. La representación de la entidad computacional sujeta a transformaciones procura ampliar su rango de aplicación, particularmente, después del aprendizaje, la nueva representación será más general. Hay dos modos de realizar el aprendizaje: supervisado, en el cual se tiene una muestra de problemas resueltos que "enseñan" al sistema cómo resolver nuevos problemas y, no supervisado, en donde deben descubrirse regularidades sin necesidad de contar con problemas ya resueltos. Nos interesa trabajar con un sistema supervisado. En este contexto, el problema de clasificación consiste en usar las regularidades "aprendidas" para resolver nuevos problemas.

Consideremos una entidad o fenómeno particular representado por sus rasgos. A dicha entidad la llamaremos *instancia*, o *ejemplar*; a cada conjunto que reuna un tipo de rasgos, *atributo*; y, consecuentemente, los *valores de un atributo* serán los *rasgos*. La clasificación supervisada se apoya en instancias previamente clasificadas. A esta colección de instancias se le designa **colección de entrenamiento**, de manera que podemos representar a la colección de entrenamiento como una reunión de *clases* de instancias.

Una forma de clasificar una instancia es adjudicándole la clase de una de las instancias de la colección de entrenamiento, aquélla que se parezca más a la instancia por clasificar. Otros métodos de clasificación tratan de ahorrar el trabajo de revisar la colección de entrenamiento completa para determinar la clase de una instancia y, de antemano, establecen reglas basadas en la colección de entrenamiento, para decidir la clase de la nueva instancia. Bastará pues, en estos casos, identificar algunos rasgos de la instancia por clasificar para determinar su clase. Presento en las siguientes secciones tres métodos dentro del campo del aprendizaje automático para analizar su empleo en el ámbito de la clasificación:

- La identificación de rasgos relevantes, para apoyar la determinación de la clase de un ejemplar.

- La síntesis de atributos que se refiere a la obtención de reglas de clasificación representadas en un árbol de decisión, y
- El razonamiento basado en memoria, cuya esencia es calcular la similitud entre instancias para determinar la clase de una nueva instancia.

En primer lugar formalizaremos algunos de los conceptos básicos de la clasificación.

2.1 Relaciones y entropía

Dada una colección de *atributos*, $\mathcal{A} = (A_p)_{p \leq P}$, y una función, Dom, que asigna a cada atributo su *dominio*, $A_p \mapsto \mathrm{Dom}(A_p) = D_p = \{x_{pj}\}_{j \leq n_p}$, definimos una *relación* R, con *esquema* $\mathcal{A}$, como un subconjunto del producto cartesiano de los dominios de los atributos de $\mathcal{A}$, $R \subset D = \prod_{p \leq P} D_p$. Veremos a R como la *colección de entrenamiento* y para cualquier *instancia* x, miembro de R, nos referimos indistintamente a su p-ésima componente x_p o *proyección* p de x, $\pi_p : R \to D_p$, $x \mapsto x_p$.

Debido a que la relación R contiene instancias clasificadas, consideraremos, para toda $x \in R$, uno de los rasgos, digamos x_q, como la clase a la cual pertenece x. Al atributo A_q lo llamaremos *atributo de decisión*. Nos interesa resolver el problema de caracterización de las instancias de una relación que tienen un determinado valor en la componente q. Es decir, cuáles son, si es que existen, los valores de otras componentes de las instancias de R que permitan asegurar que en la componente q se presenta un valor x_{qk}. De esta forma podrá establecerse una dependencia entre el valor de un atributo de una instancia por clasificar y la clase a la que pertenece.

Una *dependencia funcional* $A_p \to A_q$ entre dos atributos de una relación R es consistente con un subconjunto S de R, si $\forall x, y \in S(\pi_p(x) = \pi_p(y) \Rightarrow \pi_q(x) = \pi_q(y))$. Para un atributo particular A_p, consideramos una *partición* $\mathcal{R}_p$ de R en clases $R|_{x_{pj}} = \pi_p^{-1}(x_{pj}), j \leq n_p$, esto es: $\bigcup_{j \leq n_p} R|_{x_{pj}} = R$ y para $k \neq j$, $R|_{x_{pk}} \cap R|_{x_{pj}} = \emptyset$. Sean $\mathcal{R}_s$ y $\mathcal{R}_t$ particiones de R, decimos que $\mathcal{R}_s$ *refina* a $\mathcal{R}_t$ si $\forall R_s \in \mathcal{R}_s(\exists R_t \in \mathcal{R}_t$, tal que $R_s \subset R_t)$.

En las relaciones que exploramos no es común el cumplimiento cabal de una dependencia funcional consistente con toda la relación R, es por ello que requerimos determinar el 'grado' de dependencia entre dos atributos. Observemos por ejemplo que si $\mathcal{R}_p$ refina a $\mathcal{R}_q$ entonces $A_p \to A_q$ es consistente con R. Intuitivamente, hay mayor dependencia de A_q en A_p que de A_q en A_r si la partición que induce A_p refina a A_q en un subconjunto de R significativamente mayor que en el que lo hace A_r. Es en este sentido que nos referimos a que A_p es más *relevante* que A_r. La noción de grado de dependencia entre atributos puede precisarse de varias formas. En la sección 2.2 lo vamos a hacer con base a una prueba estadística.

Para una función f con dominio A, denotaremos con $\mathrm{argmin}_{x \in A} f(x)$ el valor que es argumento de f y cuya imagen es menor. En forma semejante $\mathrm{argmax}_{x \in A} f(x)$, es el valor que maximiza a la función.

Otro concepto muy útil es la *entropía* o *cantidad de información*. Decimos que un rasgo provee más cantidad de información que otro, si el primero tiene menos oportunidades de determinar la clase de una instancia. O bien, que entre menos relevante resulte la presencia de x_{pj}, para que cualquier instancia x con este valor pertenezca a una clase, menos cantidad de información proveerá x_{pj}. La motivación de este concepto es que los eventos muy probables tienen poca información que añadir, mientras que los eventos poco probables anuncian mucha información faltante, es decir *incertidumbre*. La *entropía* (o cantidad de información) de un estado s_i se define como

$$-\log_2 \Pr(s_i)$$

Esta medida tiene unidades de *bits*. En un sistema con estados $S = \{s_1, \ldots, s_k\}$ la cantidad de información sería:

$$H_S = -\sum_{i=1}^{k} \Pr(s_i) \log_2 \Pr(s_i). \tag{2.1}$$

Consideramos a la relación R con esquema $(A_1, \ldots, A_q)$ como un sistema cuyos posibles estados son las clases que define la partición, $\mathcal{R}_q$. Supongamos que $R_1 \subset R$, y que $R_1|_{x_{pj}}$ está contenido en un miembro de $\mathcal{R}_p$. La cantidad de información de R_1 según la repartición en clases que se hace con los valores de A_p, o simplemente respecto del atributo A_p, es el promedio de las cantidades de información que provee cada uno de los subconjuntos $R_1|_{x_{pj}}, x_{pj} \in D_p$, esto es:

$$H_{R_1(A_p)} = \sum_{x_{pj} \in D_p} H_{R_1|_{x_{pj}}} \frac{\#R_1|_{x_{pj}}}{\#R_1}. \tag{2.2}$$

La noción de cantidad de información de R_1 respecto a A_p se interpreta como una relación entre la cantidad de información de dos sistemas: la cantidad de información de R_1 (la incertidumbre de que las instancias de R_1 pertenezcan a una clase) y la cantidad de información de A_p (la incertidumbre de que las instancias de R_1 tengan un valor en su componente p). $H_{R_1(A_p)}$ es pues la cantidad de información de que dado un determinado valor de A_p una instancia de R_1 pertenezca a una clase, y es menor que la cantidad de información de R_1, ya que al conocer algo sobre R_1 no puede aumentar su incertidumbre, $H_{R_1(A_p)} \leq H_{R_1}$, más bien se dice que R_1 "pierde" información, por lo cual se define $H_{R_1} - H_{R_1(A_p)}$ como la *ganancia de información* de A_p.

La cantidad de información se empleará en dos de los algoritmos que van a exponerse. En el primero (sec. 2.3), la cantidad de información nos ayuda a optimizar la construcción de un árbol de decisión. En el segundo algoritmo (sec. 2.4), esta medida ayuda a ponderar la comparación entre los rasgos de dos instancias.

2.2 Identificación de rasgos relevantes

La idea central de este método ([Chan, Wong, 91][1]) es realizar una prueba de hipótesis que nos permita decidir si un atributo depende de otro. Para seleccionar los valores que determinan la clase de una instancia, el método parte de una colección de entrenamiento y revisa en cada clase si es significativo, estadísticamente, la ocurrencia de un rasgo. A cada rasgo, entonces, podrá corresponder una clase. Siguiendo este procedimiento para cada clase y para todos los rasgos, se encuentra una serie de parejas, (rasgo-clase): un rasgo que propone, con cierto nivel de confianza, la clase a la que pertenecería cualquier otra instancia que contenga dicho rasgo. Dada, pues, una instancia por clasificar, determinamos tantas clases como propuestas haga cada uno de sus rasgos que hayan sido seleccionados por el método. De todas estas clases elegimos aquélla que haya sido propuesta por el mayor número de rasgos.

2.2.1 Rasgos que determinan una clase

Tomemos a A_q como atributo de decisión y algún otro atributo A_p para el que deseamos identificar si depende A_q. Supongamos que $\#R = N$, y consideremos a las tablas de contingencia $O_{pq} = [o_{jk}^{pq}]_{j,k}$, donde $o_{jk}^{pq} = \#\{x | \pi_p(x) = x_{pj} \wedge \pi_q(x) = x_{qk}\}$. Recordemos que la frecuencia esperada de las instancias que tienen los valores x_{pj} y x_{qk} se calculan mediante $e_{jk}^{pq} = \sum_j o_{jk}^{pq} \times \sum_k o_{jk}^{pq}/N$. En una prueba de independencia usando χ^2 entre los atributos A_q y A_p con un nivel α, probamos que si $\sum_{j,k}(o_{jk}^{pq} - e_{jk}^{pq})^2/e_{jk}^{pq}$ es mayor al valor que determina α para χ^2 entonces se rechaza la independencia. No es posible, sin embargo, determinar, en caso de dependencia, qué clase depende de cierto valor del atributo p.

Para averiguar la caracterización de las instancias de una clase $R|_{x_{qk}}$ mediante valores del atributo A_p, comparemos $\Pr(\pi_q(x) = x_{qk} | \pi_p(x) = x_{pj})$ con $\Pr(\pi_q(x) = x_{qk})$, es decir, comparemos la probabilidad condicional de que una instancia x pertenezca a $R|_{x_{qk}}$ dado que x tiene en su p-ésima componente el valor x_{pj}, y la probabilidad de que la instancia x pertenezca a la clase $R|_{x_{qk}}$. Si

$$\Pr(\pi_q(x) = x_{qk} | \pi_p(x) = x_{pj}) > \Pr(\pi_q(x) = x_{qk}) \tag{2.3}$$

[1] Presentamos la base en que se sustenta este método, lo cual es motivado por el empleo de la *información mutua* que es de interés en la sec. 4.4.2, pero no será seguido como se presenta en [Chan, Wong, 91].

entonces podríamos reconocer que es más probable que una instancia x pertenezca a $R|_{x_{qk}}$ cuando tenga en su p-ésima componente el valor x_{pj}. En este caso elegiríamos como característico de la clase $R|_{x_{qk}}$ al valor x_{pj} de la p-ésima componente. Puesto que

$$\Pr(\pi_q(x) = x_{qk}|\pi_p(x) = x_{pj}) = \Pr(\pi_q(x) = x_{qk}, \pi_p(x) = x_{pj}) / \Pr(\pi_p(x) = x_{pj}),$$

la desigualdad 2.3 puede ser reescrita de la siguiente manera:

$$\begin{aligned} \Pr(\pi_q(x) = x_{qk}, \pi_p(x) = x_{pj}) &> \Pr(\pi_p(x) = x_{pj}) \Pr(\pi_q(x) = x_{qk}) \\ N \cdot \Pr(\pi_q(x) = x_{qk}, \pi_p(x) = x_{pj}) &> N \cdot \Pr(\pi_p(x) = x_{pj}) \cdot N \cdot \Pr(\pi_q(x) = x_{qk})/N \\ o_{jk}^{pq} &> o_{j+}^{pq} o_{+k}^{pq}/N = e_{jk}^{pq} \end{aligned} \tag{2.4}$$

donde 2.4 refiere al cálculo de la frecuencia esperada a partir de los totales marginales. Tenemos, por tanto, que la hipótesis de caracterizar una clase con algún valor de otro atributo se transforma en probar la hipótesis de que las frecuencias de los valores observados en un atributo de las instancias son mayores que las frecuencias de los valores esperados en ese mismo atributo. Complementariamente, si se tiene la igualdad entonces habría independencia entre los atributos A_q y A_p y, si se cumple además que $\Pr(\pi_q(x) = x_{qk}|\pi_p(x) = x_{pj}) < \Pr(\pi_q(x) = x_{qk})$, podría afirmarse que dado el valor x_{pj} en una instancia es menos probable que pertenezca a la clase $R|_{x_{qk}}$.

Dado que hay una correspondencia entre una distribución normal con media cero y varianza σ^2, $n(0, \sigma^2)$, con una distribución $\chi^2(1)$ podemos remplazar una prueba de independencia entre atributos por una prueba basada en la distribución normal de la variable aleatoria:

$$m_{jk} = \frac{o_{jk}^{pq} - e_{jk}^{pq}}{\sqrt{e_{jk}^{pq}}} \tag{2.5}$$

para la cual la hipótesis de independencia conduce a suponer que su media es diferente de cero. Antes de proceder con esta prueba debemos estandarizar la distribución 2.5 y, para hacerlo, acudimos al cálculo de su varianza mediante la siguiente fórmula [Chan, Wong, 91]:

$$v_{jk}^{pq} = \sqrt{\left(1 - \frac{o_{j+}^{pq}}{N}\right)\left(1 - \frac{o_{+k}^{pq}}{N}\right)}$$

por lo tanto

$$d_{jk}^{pq} = m_{jk}^{pq}/v_{jk}^{pq} \tag{2.6}$$

Así pues, d_{jk}^{pq} representa la diferencia de los valores esperados de las observaciones y las estimaciones del número de instancias que tienen los rasgos x_{pj} y x_{qk}. Según la derivación que hemos realizado, d_{jk}^{pq} tiene la densidad normal $n(\mu, \sigma^2)$, con $\mu = 0$ y $\sigma^2 = 1$,

Ahora probar que la clase k de A_q depende del valor j de A_p significa probar la hipótesis $H_0 : \mu > 0$ (vs $H_1 : \mu = 0$). En esta prueba, si $d_{jk}^{pq} > 1.96$ podemos aceptar que la clase k depende del valor x_{pj}, con 95% de confianza.

Por cada aplicación de la prueba de hipótesis es posible que encontremos el valor de un atributo del cual depende una de las clases de A_q, esto es, una pareja *(clase, rasgo indicador)*: $x_{qk} \leftarrow x_{pj}$. Al variar j, fijando un valor x_{qk}, se agregarán otros valores del atributo A_p a los rasgos indicadores que ya hemos determinado. Sucesivamente, para cada $p \in \{1, \ldots, P\} \setminus \{q\}$, encontramos todos los posibles rasgos indicadores de la clase x_{qk} que proporciona la colección de entrenamiento R: $x_{i_1 k_1} \ldots x_{i_j k_j}$.

Las parejas (clase, rasgo indicador) pueden verse como reglas que determinan que: cualquier otra instancia, x, pertenecerá a la clase que representa x_{qk}, si cualquier x contiene alguno de los valores $x_{i_1 k_1} \ldots x_{i_j k_j}$[2]. En la figura 2.1 se muestra de manera resumida los pasos anteriores, haciendo referencia a la ecuación 2.6 con la función $normaliza(p, q, j, k, R)$.

[2]Observemos que, también, podemos tener parejas que "rechacen" la asignación de una clase a una entidad, cuando $d_{jk}^{pq} < -1.96$. El método que fue implantado no usa este tipo de parejas.

Input:
The relation R*, a threshold value* s
and a fixed class k *from the* q*-th attribute*

Output:
Hypothesized dependencies pairs: $x_{qk} \leftarrow x_{pj_1}, \ldots$

Algorithm:
foreach $j \in [1, n_p]$
  println; print $R(k,q)$"←";
  foreach $p \in [1, P] \wedge p \neq q$
    let $d = normaliza(p, q, j, k, R)$
    if $d > s$: print $R(j,p)$
    else continue
end

Figura 2.1: Algoritmo para encontrar los rasgos que determinan una clase.

Observemos que aún cuando cada uno de los rasgos indicadores nos permiten, con cierto nivel de confianza, garantizar la clase de una instancia, no es necesario que todos ellos aparezcan en la instancia. Más bien, basta que uno de estos rasgos ocurra para asegurar, con cierto grado de confiabilidad, la pertenencia de la instancia a la clase que determina el rasgo. El conflicto entre las diferentes clases que pueden ser propuestas por algunos rasgos, de una instancia sin clasificar, puede ser resuelto eligiendo la clase que sea propuesta por mayor número de rasgos, considerando, en caso de empate, la clase que tenga mayor confiabilidad. Los resultados del algoritmo antes explicado se interpretan como "votaciones por una clase" [Oflazer, Tür, 97].

2.3 Síntesis de atributos

Hacer más sintéticos los patrones formados por valores de atributos que determinan la clase de una instancia significa, bajo algún criterio determinado, eliminar aquellos patrones y atributos considerados redundantes. El método que se expondrá está basado en los árboles de decisión. El uso de un árbol de decisión para clasificar instancias proviene del hecho de que con una colección de entrenamiento y eligiendo cualquier atributo, podemos *bifurcar* o *repartir* la colección original en tantas subcolecciones como valores tenga el atributo. Por cada *bifurcación* o *rama* asociada a un rasgo del atributo elegido, se tendrá una colección restringida a contener, en todas sus instancias, ese rasgo. El propósito de bifurcar una colección en varias subcolecciones es conocer si una subcolección es *homogénea*, es decir, si todos los miembros, de alguna de las subcolecciones que se obtienen, pertenecen a la misma clase. Si es así, puede establecerse que el rasgo usado para realizar esa bifurcación determina la clase de las instancias. En caso de que una subcolección sea heterogénea, ésta se compondrá de instancias de al menos dos clases diferentes, podremos entonces repartirla nuevamente eligiendo otro atributo[3]. Al final del proceso habremos conseguido una *estructura jerárquica* o *árbol* en el que cada punto de bifurcación representa una colección de instancias y cada bifurcación o rama que parte de un punto de bifurcación representa un rasgo del atributo elegido para llevar a cabo la repartición. Así pues, una bifurcación establece la *condición* de que todas las instancias de la colección a la que conduce tengan el mismo valor de un atributo. Al *recorrer el camino* que definen las bifurcaciones, empezando por el punto de bifurcación inicial, que es la colección de entrenamiento completa, pasaremos a través de las ramas, a cada vez más reducidas subcolecciones de instancias, hasta llegar a una colección homogénea. En

[3]Observemos que si dos instancias idénticas tienen la misma clase, estaremos frente a un problema de falta de información para poder discernir.

este recorrido, podremos acumular las condiciones que establecen las bifurcaciones contenidas en el camino. Tenemos así, que un recorrido desde la raíz hasta una colección homogénea proporciona un conjunto de rasgos que determinan, según la colección de entrenamiento, la pertenencia a la clase de la colección homogénea en la que termina ese recorrido. El problema de clasificación se traduce en que: cualquier otra instancia que coincida con alguna de las secuencias de rasgos, definidos por un recorrido, pertenecerá a la clase de la colección homogénea donde termina ese recorrido.

Puesto que la construcción de un árbol de decisión, como se ha expuesto, elige arbitrariamente los atributos para cada punto de bifurcación, existirán, para una colección de entrenamiento dada, varios posibles árboles de decisión. Claramente, el árbol óptimo es el que más favorece la determinación expedita de la clase de una instancia. Es decir, un árbol con la menor cantidad posible de ramas. En el método que vamos a explicar se elige, en cada punto de bifurcación, el atributo cuyos rasgos proporcionan la menor cantidad de información sobre cada una de las clases. La razón de esta elección descansa en el hecho de que con ese atributo se tienen más oportunidades de determinar la clase de las instancias de las colecciones derivadas de ese punto de bifurcación y, por tanto, es más probable que las instancias queden reunidas en subcolecciones homogéneas. Otra manera de ver lo anterior es refiriéndonos a la ganancia de información: se elige el atributo de máxima ganancia de información puesto que esperamos que, al repartir con él, las nuevas subcolecciones tengan reducida su cantidad de información.

2.3.1 Construcción de un árbol de decisión

Sea R una relación con esquema $\mathcal{A} = (A_1, \ldots, A_q)$ si $R_1, R_2 \subset R$, decimos que R_1 *precede* a R_2 en x_{pj}, $R_1 \prec_{x_{pj}} R_2$, si R_2 es el resultado de la selección de las instancias con valor x_{pj} a partir de R_1: $R_2 = R_1|_{x_{pj}}$. Llamemos *condición* a la restricción de una relación a un valor x_{pj}. Para precisar qué es un árbol de decisión, recordemos primero que un árbol W se define mediante una relación V en un conjunto de nodos E, de tal forma que existe un único nodo llamado raíz que no tiene predecesores y cuyos sucesores pueden verse como raíces de j árboles repartiendo en cada uno de ellos el resto de los nodos de E. En un árbol W, la secuencia de arcos $b_1, \ldots, b_t$, es una *trayectoria* si existe una secuencia de nodos $c_1, \ldots, c_{t+1}$ y b_h conecta a c_h con c_{h+1}, $(1 \leq h \leq t)$. Un *árbol de decisión*, W, para una relación R, se construye definiendo como raíz a R y todo nodo R_s de nivel i cumple que $\exists R_r$ de nivel $i-1$ y $x_{p_j k} \in A_{p_j}$ $(1 \leq p_j < q)$ tal que $R_r \prec_{x_{p_j k}} R_s$, además si R_s es terminal se tiene que $R_s|q_l = R_s$ para algún $q_l \in A_q$.

Si W es un árbol de decisión de R entonces cada trayectoria $b_1, \ldots, b_t$ desde la raíz hasta un nodo terminal representa una secuencia de condiciones sobre valores de atributos que son satisfechas por las instancias del nodo terminal. El problema de síntesis trata de reducir el tamaño de las trayectorias y el número de éstas. Como se ha dicho el criterio de reducción se apoya en el concepto de cantidad de información.

En un árbol de decisión W podemos ver a la clase de una instancia como un estado. Observemos, además, que si S tiene instancias de sólo una clase, entonces $H_S = 0$. Para sintetizar las instancias que contiene una relación R, debemos construir un árbol de decisión que sea óptimo, en el sentido de que sus trayectorias sean lo más cortas posibles y en menor número. Si W se construye de manera que los nodos elegidos tengan la menor cantidad de información ([Quinlan, 93]), o mayor ganancia de información, entonces estaremos eligiendo como constituyentes de esos nodos, instancias más probables y por tanto restarán menos instancias para clasificar. Al construir el árbol de esta forma es más probable que se eliminen algunas trayectorias. También, las trayectorias pueden reducirse deteniendo la bifurcación cuando $H_S = 0$.

Así pues, es necesario elegir los atributos en un orden tal que el número de trayectorias sea mínimo, y esto significa elegir cada vez el atributo que proporcione la mayor ganancia de información: $\operatorname{argmin}_{A_r \in \mathcal{A} \setminus \{A_q\}} H_{R_i(A_r)}$. Este es el criterio usado en el algoritmo de Quinlan.

Ordenemos ahora los pasos para construir el árbol de decisión óptimo W, asociado a una relación R con esquema $\mathcal{A} = (A_1, \ldots, A_q)$ y atributo de decisión A_q. Por definición se toma a R como raíz y se itera el siguiente procedimiento para definir los arcos: de acuerdo con el criterio de Quinlan se elige un atributo para realizar una partición en el nodo en curso R_i, de tal manera que los elementos de esta partición, considerados como nuevos nodos, queden pendientes hasta que sea su turno partirlos. Eventualmente, habrá subconjuntos de R que no sea necesario partir, puesto que contendrán únicamente instancias de una sola clase de $\mathcal{R}_q$ (su entropía es cero) y que, como nodos del árbol, representan a las hojas. El algoritmo mostrado en la figura 2.2 precisa estos pasos.

Input:
The relation R with scheme $\mathcal{A}$ and the decision attribute A_q.

Output:
A decision tree W given as a collection of edges and its root.

Algorithm:
$push(Queue, R)$
let $E = \emptyset$
while $Queue \neq \emptyset$
  let $R_i = \mathrm{pop}(Queue)$
  let $A_p = \mathrm{argmin}_{A_r \in \mathcal{A} \setminus \{A_q\}} H_{R_i(A_r)}$
  if $H_{R_i(A_p)} \neq 0$
    foreach $j \in [1, n_p]$
      let $E = E \cup \{(R_i, R_i|_{x_{pj}})\}$
      $\mathrm{push}(Queue, R|_{x_{pj}})$
end

Figura 2.2: Algoritmo para construir un árbol de decisión.

2.4 Razonamiento basado en memoria

Dentro del área de la Inteligencia Artificial un problema puede verse como un estado inicial y su solución como un estado meta, al cual se llega mediante una secuencia de operadores aplicados a partir del estado inicial. A fin de pasar a un nuevo estado, tendrá que aplicarse un operador y esto obliga a que sus condiciones sean satisfechas, es decir, las condiciones de un operador se cumplen si el estado en curso permite aplicar el operador. Este enfoque requiere de un mecanismo de búsqueda para encontrar la secuencia de operadores que producen el estado meta. Normalmente en esta tarea se hace uso de un empate exacto al comparar un estado con las condiciones del operador que se desea aplicar. Hay, sin embargo, criterios menos restrictivos que el empate exacto, como el siguiente [Stanfill, Waltz, 86]: Elegir el operador cuyas condiciones se acerquen más al estado en curso. Es necesario, entonces, medir la cercanía entre estados y, por tanto, definir una función de distancia.

Designada la búsqueda como el principal componente del razonamiento mecánico, un problema de clasificación puede expresarse de la siguiente manera: dada una nueva instancia, a partir de la colección de entrenamiento con instancias compuestas cada una de sus rasgos y su clase, debe elegirse la clase de la instancia de la colección cuyos rasgos se parezcan más a la nueva serie de rasgos. El método que va a exponerse recurre a la colección de entrenamiento completa y no sólo a rasgos particulares o a un subconjunto de éstos, para determinar la clase de una instancia. Este método se dice *basado en memoria* pues hace uso de una descripción concreta de todos los ejemplos, sin abstracciones.

Para clasificar una nueva instancia, es natural que se observe la colección de entrenamiento con el fin de adjudicarle a esta nueva instancia la clase de aquella instancia de la colección que más se parezca. El presente método no hace más que eso, precisando el significado de "parecido" o *similitud* entre instancias. El concepto de similitud también puede traducirse en *distancia*: a mayor similitud menor distancia y viceversa. Una forma de medir la "cercanía" entre dos instancias es contando el número de rasgos diferentes que hay entre éstas; instancias idénticas tendrán distancia cero. Esta formulación es un buen punto de partida, para definir otra métrica que enfrente situaciones comunes cuando comparamos instancias; como cuando hay atributos más importantes que otros, y, por tanto, la presencia o ausencia de los más importantes es más significativo que la ausencia o presencia de los menos importantes. La importancia de un atributo puede ser expresado por un *peso* del atributo. Este peso puede ser proporcionado por la ganancia de información del atributo, pues nos indica qué tanta información provee tal atributo. Resumiendo, dada una colección de entrenamiento S, una función que determina la distancia entre dos ejemplares X y Y, $\Delta(X,Y)$, y un nuevo ejemplar por clasificar, Z, asignamos a Z la clase de $\mathrm{argmin}_{X\in S}\Delta(X,Z)$.

Es importante señalar que el razonamiento basado en memoria hace uso intenso de los datos almacenados en la memoria sin que éstos sufran ninguna abstracción. Otra característica importante del razonamiento basado en memoria es que no depende[4] de modelos apoyados en el dominio [Stanfill, Waltz, 86].

2.4.1 Similitud entre instancias

Consideremos una colección de atributos $\mathcal{A} = (A_1, \ldots, A_m)$. Cada atributo $A \in \mathcal{A}$ tiene su dominio finito propio $\mathrm{Dom}(A)$. Sea $S \subset U$ una colección de entrenamiento y sean $C_1, \ldots, C_n \subset U$ las clases en el *universo* $U \subset \prod_{A\in\mathcal{A}} \mathrm{Dom}(A)$. Sean X y Y dos instancias representadas por la secuencia de valores de atributos $(x_1, \ldots, x_m)$ y $(y_1, \ldots, y_m)$. La distancia Δ es definida como

$$\Delta : (X,Y) \mapsto \Delta(X,Y) = \sum_{i=1}^{m} \delta(x_i, y_i) \quad \text{donde} \quad \delta(a,b) = \begin{cases} 0 & \text{si } a = b \\ 1 & \text{si } a \neq b \end{cases}$$

Para clasificar una nueva instancia, Z, es necesario obtener la clase de la instancia $\mathrm{argmin}_{X\in S}\Delta(X,Z)$.

Frecuentemente Δ, aunque es muy simple, no es la distancia más adecuada. Como alternativa, podemos asociar *pesos* a cada uno de los atributos de tal manera que, en el peor de los casos –cuando los rasgos son diferentes– sumamos pesos (y aumenta la distancia) y, en el mejor de los casos –los valores de los atributos son iguales–, no haya contribución a la distancia. La distancia se modifica así:

$$\Delta(X,Y) = \sum_{i=1}^{m} \delta(x_i, y_i) \cdot w_i. \tag{2.7}$$

Una técnica que ha tenido buenos resultados [Daelemans, van-den-Bosch, et. al. 98] asocia a cada atributo su *ganancia normalizada*. Ello se basa en el cálculo de la cantidad de información de los atributos, de manera semejante a la determinación de un árbol de decisión [Quinlan, 93] que discrimina atributos en los caminos desde la raíz. La ganancia normalizada se define con base en la ganancia de información:

Ganancia de información Para cualquier conjunto $S' \subset S \subset U$, usamos la cantidad de información necesaria para identificar una instancia de S' en una clase: $H_{S'}$, de acuerdo con la ecuación 2.1. Entonces, para cada $A \in \mathcal{A}$, podemos calcular la cantidad de información de S' respecto de A: $H_{S'(A)}$ (ec. 2.2). El valor $gain(A) = H_{S'} - H_{S'(A)}$ es la *ganancia de información* cuando S' es repartido de acuerdo con el atributo A.

Ganancia normalizada Para una colección de entrenamiento dada, S, y para cada $A \in \mathcal{A}$ sea

$$split\ info(A) = -\sum_{a\in\mathrm{Dom}(A)} \frac{\#S|_a}{\#S} \cdot \log_2\left(\frac{\#S|_a}{\#S}\right),$$

la ganancia normalizada se define por

$$gain\ ratio(A) = \frac{gain(A)}{split\ info(A)}. \tag{2.8}$$

[4]Esta dependencia se refiere a los cambios sustanciales que sufre un método al cambiar el dominio de aplicación.

La normalización en 2.8 pondera la ganancia de información con las entropías de los atributos. Los valores que calcula 2.8 serán utilizados como pesos en 2.7 para llevar a cabo nuestra aplicación.

El procedimiento del razonamiento basado en memoria, en general, consiste de reunir una colección de instancias de entrenamiento y una vez identificada la entidad a clasificar, esto es, habiendo determinado las clases posibles de la entidad, sus propiedades y los rasgos que permiten determinar su clase, tendrá que calcularse la ganancia normalizada de cada atributo, eligiendo los atributos con mayor información para definir los pesos que intervienen en la función de distancia. Como se ha dicho, para clasificar una nueva instancia z se calcula la distancia hacia cada elemento $x \in S$ en la colección de entrenamiento S, y la clase del elemento que resulte con menor distancia es elegida para clasificar a la nueva instancia z. Para las entidades lingüísticas, las instancias que formarán la colección de entrenamiento se extraen de una fuente rica de ejemplos; normalmente de un *corpus.* Los siguientes pasos, orientados a la clasificación de entidades lingüísticas, precisan la obtención de una colección de entrenamiento:

Colección de entrenamiento inicial:

- Proponer todos los atributos posibles de la entidad a clasificar y determinar las posibles clases de ésta.
- Reunir las instancias de entrenamiento compuestas por atributos contextuales y su clase. Los atributos contextuales se forman por tantos vecinos alrededor de la entidad como sean necesarios y, si se considera necesario, propiedades de los vecinos. En consecuencia, los valores de los atributos pueden ser palabras que están alrededor de la entidad a clasificar o propiedades de las palabras, estructuras gramaticales y otras entidades lingüísticas del contexto en que ocurre la entidad por clasificar. Siendo ésta una de las tareas básicas para conformar una colección de entrenamiento, ha de ser realizada por un experto (naturalmente humano) en el área. Esto obliga a que la determinación de la clase de cada instancia se haga manualmente (en contraposición a la etapa de aplicación: cuando se induce la clasificación a un texto de aplicación, partiendo del corpus de entrenamiento, se procede, en efecto, de manera algorítmica, es decir, automática).

Selección de atributos:

- Calcular y analizar la curva de ganancia normalizada para todos los atributos[5], con lo cual será posible discriminar los atributos que no aportan información para determinar la clase de la entidad.
- Proyectar la colección inicial de acuerdo con los atributos que ofrecen mayor información en una nueva colección de entrenamiento. Específicamente:
  - Una nueva instancia se forma extrayendo, de una instancia original, los rasgos con mayor peso informativo. En caso de formar dos instancias iguales pero con diferente clase decidir la clase de esa instancia tomando aquélla cuya frecuencia de ocurrencia sea mayor, procediendo de acuerdo con un criterio "voraz" (*greedy*): Se da una mayor oportunidad de asignar clases correctas a instancias ambiguas. Careciendo de un criterio determinista, se ha convenido en la aceptación de este criterio voraz basado en frecuencias.

2.5 Elección de un método

Como será visto en esta sección, el método de aprendizaje basado en memoria es el que mejor se adapta a nuestras necesidades. Con base en la precisión que tenga cada método presentado, al clasificar palabras de un texto, sustentaremos nuestra elección.

Se plantea, como problema a resolver, la clasificación de las palabras ambiguas frecuentes[6] ("la", "que", etc.). Específicamente, el clasificador tendrá que discernir sobre la parte del discurso que le corresponde a

[5]La fig. 3.4 expone un ejemplo de la realización de este paso.

[6]Puede consultarse la sec. 3.3.2 para precisar la idea de palabra ambigua frecuente, las partes del discurso empleadas, y los rasgos seleccionados para este problema.

cada una de estas palabras al ocurrir en un texto, es decir, elegir entre pronombre y artículo la clase de "la", entre conjunción y pronombre la clase de "que", etc. Para ello se tomó un texto compuesto de 7,661 signos, cada uno con su parte del discurso. De éste se extrajeron instancias para constituir la colección de entrenamiento. Las instancias se formaron por las partes del discurso de cuatro palabras vecinas a cada palabra ambigua, dos palabras contiguas a cada lado de ella. A partir de estas instancias, 654 en total, se procedió a formar: la colección de entrenamiento, para el método de aprendizaje basado en memoria; las parejas (clase, rasgo), según el método de atributos relevantes; y el árbol de decisión definido por el criterio de Quinlan.

Hemos tomado para ejemplificar los recursos antes aludidos, lo correspondiente para el etiquetamiento de la palabra "que". En este caso, las parejas que emplea el método de rasgos relevantes se muestran en la tabla 2.1. La tabla está organizada en cuatro columnas, una por cada rasgo. Cada entrada de la tabla contiene una terna formada por una pareja y el valor que le corresponde, de acuerdo con la curva normal, para proponer que la clase de la pareja depende del rasgo de la pareja. Todos los valores d_{ik}^{pq} son mayores que 1.96, pues hemos establecido 95% de confianza (conforme a lo dicho en la sec. 2.2.1). El primer renglón, correspondiente al rasgo 1, lo interpretamos como: la clase de "que" es PRON si su rasgo 1 es ART, esto es, si la marca de la $(j-2)$-ésima palabra es ART, si "que" ocupa la posición j-ésima, pues $d_{1k}^{pq} > 1.96$[7]. De manera semejante se interpretan las demás entradas.

Clase, Rasgo $1/d_{1k}^{pq}$	Clase, Rasgo $2/d_{2k}^{pq}$	Clase, Rasgo $3/d_{3k}^{pq}$	Clase, Rasgo $4/d_{4k}^{pq}$
PRON, ART /5.95	PRON, ART /7.30	PRON, VERB /6.22	PRON, VERB /5.16
PRON, CONT /2.49	PRON, COM /2.32	PRON, ADV /2.78	PRON, COM /2.49
PRON, PREP /3.54	PRON, NOM /8.35	CONJ, ADJ /4.36	CONJ, NOM /3.23
CONJ, ADV /3.78	PRON, PTO /2.02	CONJ, NP /2.49	
CONJ, COM /3.13	PRON, NP /2.02	CONJ, ART /5.21	
CONJ, CONJ /2.78	CONJ, PREP /3.70	CONJ, CONJ /2.49	
CONJ, PRON /3.49	CONJ, VERB /10.49		
CONJ, PTO /2.02	CONJ, ADV /4.64		
	CONJ, CONJ /2.62		
	CONJ, DPT /2.02		

Tabla 2.1: Parejas clase-rasgo para determinar la parte de discurso de "que"

Por el método de Quinlan, procediendo según lo explicado en la sección 2.3.1, se determinaron 90 reglas para desambiguar la palabra "que". Algunas de éstas se muestran a continuación:

ADJ NOM VERB X4 → PRON	ADJ PREP X3 NOM → CONJ	ADV NOM X3 X4 → CONJ
ADV PREP X3 VERB → CONJ	ART NOM X3 X4 → PRON	COM PREP X3 NOM → CONJ
CONJ NOM PRON X4 → PRON	CONT NOM X3 X4 → PRON	NOM NOM X3 X4 → PRON
PREP NOM PREP NOM → PRON	PREP NOM VERB ART → PRON	PREP NOM VERB NOM → CONJ
PRON GUI X3 X4 → PRON	PRON PREP X3 VERB → CONJ	VERB ADV X3 ADV → PRON
X1 ADJ ADV X4 → PRON	X1 ADJ ART X4 → CONJ	X1 ADJ COM X4 → CONJ
X1 ADJ PRON X4 → CONJ	X1 ADJ VERB VERB → PRON	X1 ADV X3 ADJ → CONJ
X1 ADV X3 VERB → CONJ	X1 ART X3 X4 → PRON	X1 COM ADJ X4 → CONJ
X1 COM PREP X4 → PRON	X1 CONJ X3 ART → CONJ	X1 CONJ X3 VERB → CONJ
X1 CONT X3 X4 → PRON	X1 NP X3 X4 → PRON	X1 PREP X3 ADJ → PRON
X1 PREP X3 PRON → PRON	X1 PRON ADV X4 → PRON	X1 PRON VERB X4 → PRON
X1 PTO X3 X4 → PRON	X1 VERB X3 X4 → CONJ	

Figura 2.3: Parte de las reglas usadas en la desambiguación de "que".

[7] Recordemos que este valor se usa cuando hay empate en la votación hecha por los rasgos.

Las reglas se interpretan de la manera siguiente: La palabra ambigua se etiqueta con la marca que indica el consecuente, si se cumple el antecedente. El antecedente de una regla está formado como lo hemos estado diciendo, por cuatro marcas. Para nuestro ejemplo: las marcas de dos palabras antes de "que", y las marcas de dos palabras después de "que". Así, la regla `ADJ NOM VERB X4` → `PRON` que ocurre en la primera entrada de la tabla 2.3, significa que la marca de "que" será `PRON`, si se cumple que: la palabra antes de "que" tiene la marca `NOM`, la marca antes de `NOM` es `ADJ`, la marca de la palabra que sigue a "que" es `VERB`, y no importa cual sea la marca posterior a `VERB` (`X4`).

El método de aprendizaje basado en memoria no modifica las instancias, trabaja directamente con ellas, según lo expuesto en la sección 2.4.1. Solamente fue calculada la ganancia de información para cada uno de los atributos. De un total de 189 instancias útiles en la asignación de la parte de discurso a la palabra "que", se muestran sólo algunas de éstas en la fig. 2.2. Según el orden, de la marca más alejada a la izquierda hacia la marca más alejada a la derecha, los valores de ganancia de información (véase la ec. 2.8) son: 0.0963056, 0.161837, 0.360231 y 0.103724. Estos valores indican que la marca de la palabra que sigue a "que" es la de mayor peso informativo, en tanto que la marca que le antecede en dos posiciones es la de menor peso informativo. Se ha dicho que estos valores serán utilizados en la ec. 2.7 como pesos (w_i). Más adelante, con estos valores, calcularemos la distancia entre dos secuencias de rasgos.

Instancia	Clase
ART NOM PRON VERB	PRON
PTO CONJ ADJ NOM	CONJ
COM ART ADV VERB	PRON
VERB VERB ART NOM	CONJ
VERB ART ART NOM	PRON
ART NOM ADV VERB	PRON
VERB VERB ART NOM	CONJ
NOM PREP PREP PRON	PRON
VERB ART PRON VERB	PRON
ADJ COM PREP NOM	PRON
PRON VERB VERB PREP	CONJ
VERB ART PRON VERB	PRON
NOM ADJ ADV PRON	PRON

Tabla 2.2: Instancias formadas por dos marcas vecinas a cada lado de la palabra "que".

Con las reglas, parejas e instancias (que hemos ejemplificado sólo para el caso de la palabra "que") se llevó a cabo el experimento de aplicar los tres métodos expuestos a un texto de prueba[8] formado por 400 palabras. Ya que eran necesarias las partes del discurso de las palabras que son vecinas a la palabra ambigua frecuente por clasificar, se tomó el texto ya marcado y se reclasificaron las palabras ambiguas frecuentes de éste.

El texto a marcar se muestra en la fig 2.4. En esta figura, la primera columna representa la posición (indicada con el signo #) de la primera palabra, que está en la siguiente columna de ese mismo renglón. El texto de la columna dos es una secuencia de 15 palabras o signos, con la secuencia de marcas adjudicadas en el siguiente renglón. Las marcas que ocurren en esta secuencia se consideran con las mismas posiciones que las palabras del texto del renglón previo.

[8]Éste es un fragmento del artículo periodístico del género editorial, titulado "El congreso y la economía", de Miguel Angel Granados Chapa, tomado del CEMC, ver sec. 3.2.3.

#	Texto/Marcas
1	. lo hizo con ramplonería agravada por la insolencia que se deriva de hablar desde
	PTO lo VERB PREP NOM VERB PREP la NOM que PRON VERB PREP VERB PREP
16	el poder , en nombre de la mayoría . mayoría que esta vez , contra
	ART NOM COM PREP NOM PREP la NOM PTO NOM que ADJ NOM COM PREP
31	su costumbre , no fue silenciosa , si bien sólo alcanzó a romper su habitual
	ADJ NOM COM ADV VERB NOM COM CONJ ADV ADV VERB PREP VERB ADJ ADJ
46	y obligado mutismo golpeando la madera de los escaños . pero todo eso es lo
	CONJ ADJ NOM ADV la NOM PREP los NOM PTO CONJ todo PRON VERB lo
61	que menos importa . más relevante debió ser la toma de posición del congreso ante
	que ADV VERB PTO ADV ADJ VERB VERB la NOM PREP NOM CONT NOM PREP
76	la carestía . formar un comité de emergencia supone que la situación es de emergencia
	la NOM PTO VERB ART NOM PREP NOM VERB que la NOM VERB PREP NOM
91	. si así es , la apreciación parlamentaria contradice el informe hecho público , el
	PTO CONJ ADV VERB COM la NOM ADJ VERB ART NOM VERB NOM COM ART
106	mismo día , por el banco de México . según este organismo , el índice
	ADJ NOM COM PREP ART NOM PREP NP PTO PREP ADJ NOM COM ART NOM
121	de precios sólo ha subido cinco por ciento de agosto a la fecha . ¿
	PREP NOM ADV VERB VERB ADJ PREP NOM PREP NOM PREP la NOM PTO AIN
136	justifica esta alza el que el congreso se sienta obligado a estudiar la situación ?
	VERB ADJ NOM ART que ART NOM PRON VERB VERB PREP VERB la NOM INT
151	¿ o los diputados y senadores , a pesar de todo , han podido encontrar
	AIN CONJ los NOM CONJ NOM COM PREP NOM PREP todo COM VERB VERB VERB
166	en la realidad datos que no corresponden a los de la estadística ? de cualquier
	PREP la NOM NOM que ADV VERB PREP los PREP la NOM INT PREP ADJ
181	modo , el caso es que el comité sólo tendrá una importancia retórica . nace
	NOM COM ART NOM VERB que ART NOM ADV VERB una NOM ADJ PTO VERB
196	sin que para ello exista fundamento legal alguno . y eso es grave , en
	PREP que PREP PRON VERB NOM ADJ ADJ PTO CONJ PRON VERB NOM COM PREP
211	un sistema constitucional cimentado en facultades expresas . es decir , el poder público sólo
	ART NOM ADJ VERB PREP NOM ADJ PTO VERB VERB COM ART NOM ADJ ADV
226	puede hacer válidamente lo que le está ordenado o permitido . de otra manera ,
	VERB VERB ADV lo que PRON VERB ADJ CONJ ADJ PTO PREP ADJ NOM COM
241	a diferencia del norteamericano , que en mucho le ha servido de modelo , el
	PREP NOM CONT NOM COM que PREP ADV PRON VERB VERB PREP NOM COM ART
256	congreso mexicano carece de facultades indagatorias , jurisdiccionales , salvo en pocas materias , ninguna
	NOM ADJ VERB PREP NOM ADJ COM ADJ COM ADV PREP ADJ NOM COM PRON
271	de las cuales tiene que ver con los precios . centralmente , ha de ocuparse
	PREP las PRON VERB que VERB PREP los NOM PTO ADV COM VERB PREP VERB
286	de legislar y de supervisar algunas actividades del poder ejecutivo . en ello debiera enfocar
	PREP VERB CONJ PREP VERB ADJ NOM CONT NOM ADJ PTO PREP PRON VERB VERB
301	su esfuerzo el congreso . primero , puede ejercer facultades como la que le atribuye
	ADJ NOM ART NOM PTO ADV COM VERB VERB NOM ADV la que PRON VERB
316	el artículo 93 de la constitución . una interpretación quizá extensiva - pero autorizada por
	ART NOM NUM PREP la NOM PTO una NOM ADV ADJ GUI CONJ VERB PREP
331	los hechos , pues el congreso ha llamado ante él , así sea de modo
	los NOM COM CONJ ART NOM VERB VERB PREP PRON COM ADV VERB PREP NOM
346	informal , aun a jefes de departamento - le permitiría interrogar al procurador del distrito
	NOM COM ADV PREP NOM PREP NOM GUI PRON VERB VERB CONT NOM CONT NOM
361	y territorios federales sobre el grado de aplicación del código penal en materia de delitos
	CONJ NOM ADJ PREP ART NOM PREP NOM CONT NOM ADJ PREP NOM PREP NOM
376	contra la economía pública . se sabría allí , con certidumbre , que jamás se
	PREP la NOM ADJ PTO PRON VERB ADV COM PREP NOM COM que ADV PRON
391	procesa a nadie por acaparamiento , tal como podría hacerse
	VERB PREP PRON PREP NOM COM PRON ADV VERB VERB

Figura 2.4: Texto a marcar con sus partes del discurso.

Se describe, a continuación, algunos etiquetamientos para "que", con cada uno de los métodos. En los ejemplos, hacemos referencia a los métodos: Quinlan, con Q; atributos relevantes, con A; y aprendizaje basado en memoria, con M. Al inicio de cada ejemplo, aparece la posición en la que ocurre "que" y sus marcas vecinas, contexto. En la fig 2.4 los ejemplos que explicaremos aparecen señalados con un recuadro.

posición 61, contexto `[VERB ART ADV VERB]`**:**

- **A** Se clasificó como `PRON`. Tres rasgos votaron por esta clase, según las parejas de la fig 2.1: el 2, 3 y 4.
- **M** La misma clase se determinó por similitud de instancias. Se empleó: `[COM ART ADV VERB]`, cuya distancia al contexto de "que" en la posición 61 fue: 0.0963056.
- **Q** La regla `X1 ART X3 X4` → `PRON` empató con el contexto y asignó, también, la misma marca.

posición 140, contexto `[NOM ART ART NOM]`**:**

- **A** La clase determinada por dos rasgos contra uno fue `CONJ`. A favor 3 y 4, en contra el 2.
- **Q** El método basado en árbol de decisión, aplicó la misma regla que en la posición 61, clasificándola como `PRON`.
- **M** La instancia `[VERB ART ART NOM]` cuya clase es `PRON`, fue la más cercana.

posición 230, contexto `[ADV ART PRON VERB]`**:**

- **A** Tenemos que los rasgos relevantes que determinaron esta clase fueron dos entre el 1, 2 y 4. La clase propuesta fue `PRON`.
- **M** La instancia `[VERB ART PRON VERB]` con clase `PRON` se acercó a una distancia 0.0963056.
- **Q** Para este contexto no hubo empate con ninguna regla de "que".

posición 388, contexto `[NOM COM ADV PRON]`**:**

- **A** Los rasgos 2 y 3 propusieron la clase: `PRON`.
- **M** Con la instancia `[NOM ADJ ADV PRON]` se clasificó `PRON`.
- **Q** Tampoco hubo regla que pudiera ser aplicada a este contexto.

En total, en el texto de la fig. 2.4 hay 13 ocurrencias de la palabra "que". El método de Quinlan tuvo 8 aciertos, el método de atributos relevantes 9 aciertos, y el método de aprendizaje basado en memoria, 10 aciertos.

Los aciertos, en general (considerando todas las palabras ambiguas frecuentes), se muestran en la siguiente tabla 2.3.

Método	Aciertos (%)
Síntesis de atributos	75.0
Rasgos relevantes	90.0
Similitud entre instancias	92.5

Tabla 2.3: Precisión de los métodos al etiquetar palabras ambiguas frecuentes.

Para que el método de Quinlan pueda clasificar correctamente un ejemplar, deberá contar con una regla sin variables en su antecedente que empate completamente con el contexto de la palabra que va a ser marcada, o bien, deberá alimentarse con suficientes ejemplos para empatar solamente con algunos rasgos contenidos en el contexto de la palabra a etiquetar, y los demás rasgos tuvo que haberlos generalizado el método. En otras palabras, para este método, las fallas en la clasificación de nuevos ejemplos, se pueden interpretar como la necesidad de más ejemplos para el método, de tal forma que realice una adecuada generalización. En el texto de la fig 2.4 hubo cinco casos en los que no pudo ser aplicada una regla (dos de ellos ocurren en las posiciones ejemplificadas: 230 y 388.).

En contraste con el método anterior, el método de identificación de atributos relevantes reduce la clasificación de una instancia, a la existencia de los atributos más relevantes en la instancia por clasificar. El no comprometer la ocurrencia de rasgos no relevantes de una instancia para concluir su pertenencia a una clase, permite incrementar la flexibilidad para clasificar diversas instancias.

La mayor ventaja del método basado en memoria es que conserva, en las instancias de entrenamiento, las regularidades y las excepciones. La flexibilidad de comparar instancias que no sean exactamente iguales, en todos sus rasgos, ofrece también una ventaja en situaciones que ocurren con frecuencia en el procesamiento del lenguaje natural [Daelemans, van-den-Bosch, et. al. 98]. Este tipo de empatamiento lo tiene, en cierta medida, el método de síntesis, pero solamente en los atributos que se generalizaron.

Aunque la búsqueda exhaustiva del vecino más cercano en el método basado en memoria es costosa, respecto al uso de reglas, el trabajo puede aligerarse representando la colección de entrenamiento en un árbol *trie*, mismo que permite efectuar podas en el cálculo de la distancia. Esta implantación del método basado en memoria se conoce como *IGTree* [Bosch, Daelemans, et. al. 96].

Con la elección hecha en esta sección estamos en posibilidades de llevar a cabo el desarrollo de otros sistemas [Daelemans, van-den-Bosch, et. al. 98] para el PLN, como el análisis sintáctico parcial.

Pero más aún, la aplicación de los métodos de aprendizaje automático abarcan un espectro amplio de problemas en el campo del PLN, ya que la clasificación es una forma de desambiguación.

En el siguiente capítulo se presenta el sistema que, en general, asigna partes del discurso a las palabras de un texto y, además, muestra una aplicación al análisis sintáctico parcial. Ambas aplicaciones con el apoyo del método de aprendizaje basado en memoria.

Capítulo 3

Etiquetamiento

En este capítulo presentamos algunas técnicas para asignar marcas o etiquetas a los componentes de un texto. Posteriormente, se describen dos etiquetadores desarrollados en este trabajo, uno para asignar categorías gramaticales, o partes del discurso, a las palabras de un texto, y otro para adjudicar una marca, entre dos palabras consecutivas de un texto, marca que representa el límite de una frase nominal (análisis sintáctico parcial).

3.1 Etiquetamiento basado en corpus

Nos interesa *etiquetar* a cada uno de los componentes de un texto con cierta información lingüística, esto es, asignarle a cada palabra del texto una marca, categoría o etiqueta. Para este fin, resulta conveniente apoyarse en otros textos que tengan explícita tal información. Un *corpus* es un texto representativo para analizar un determinado fenómeno lingüístico. Cuando un corpus contiene información lingüística explícita, se le llama *corpus anotado*, marcado o etiquetado. El **corpus de trabajo** o de entrenamiento es un corpus particular, al cual hará referencia nuestra aplicación.

Partes del discurso y tipos de palabras

La función que desempeña una palabra en un discurso se llama *parte del discurso*, categoría o simplemente marca. Hay varias maneras de clasificar las palabras de acuerdo con su función y, en consecuencia, diversos conjuntos de partes del discurso. La cardinalidad de los diferentes conjuntos de categorías oscila entre una y varias decenas. Nos referiremos al conjunto de marcas en uso como *Partes del Discurso* (PD).

El problema de etiquetamiento

El problema de etiquetamiento de un texto con partes del discurso se enuncia, en general, de la siguiente manera:

> *Dado un texto cualquiera, asignar a cada palabra, o signo que lo compone, la parte del discurso que le corresponde.*

Consideraremos el problema del etiquetamiento restringido a un género, es decir, etiquetar solamente textos de un determinado género. Concebimos un **etiquetador**, como un sistema automático que resuelve el problema antes mencionado. El etiquetador del que hablamos es, específicamente, un sistema computacional que toma como entrada un texto, del género para el cual se ha orientado el etiquetador y obtiene, como resultado, la asignación de su parte del discurso a cada uno de los componentes del texto de entrada.

Algunos sistemas computacionales que realizan de manera automática esta asignación dependen, no sólo del algoritmo empleado, sino además, en forma crucial, de una fuente de información. Es común utilizar un corpus etiquetado a partir del cual se extrae información lingüística, como diccionarios, terminaciones, etc, que son útiles para llevar a cabo la tarea de adjudicación de partes del discurso. A este tipo de sistemas se les conoce como sistemas supervisados, en contraste con los no supervisados que carecen de esta fuente de información.

Las siguientes subsecciones se ocupan de precisar algunos términos, útiles en la descripción de los etiquetadores que veremos posteriormente.

3.1.1 n-gramas, frecuencia y terminaciones

Al conjunto $\mathcal{V}$ de cadenas formadas por elementos de un alfabeto $\mathfrak{A}$, que tienen sentido para una comunidad lingüística, lo llamaremos *vocabulario*. Un *texto* es una secuencia de cadenas de un vocabulario, y que tiene aceptación en una comunidad lingüística.

Consideremos un texto $T = [v_i]_{i \leq n_T}$ y una palabra $v \in \mathcal{V}$, $v = a_1 a_2 \dots a_p$, $a_i \in \mathfrak{A}$. La *frecuencia* de v en T, el número de ocurrencias de v en T, se denota con $\mathrm{fr}(v)$. El *rango* de v es la posición que ocupa v en la lista de palabras ordenadas y agrupadas descendentemente según su frecuencia de ocurrencia en T. El conjunto de n-**gramas** de T es

$$T_{\langle n \rangle} = \{x_1 x_2 \dots x_n | \exists j \leq n_T - n + 1, x_1 = v_j, x_2 = v_{j+1}, \dots x_n = v_{j+n-1}\}.$$

Un *contexto* de tamaño n para v, es un n-grama de T que contiene a v. Denotamos al conjunto de n-gramas que contienen a v en la k-ésima posición con:

$$T_{\langle k-1:n-k \rangle}(v) = \{X = [x_1, \dots, x_n] \in T_{\langle n \rangle} | x_k = v\}.$$

La *probabilidad condicional* de que ocurra la palabra v en la secuencia $w_1 w_2 \dots w_{k-1} v\, w_{k+1} w_{k+2} \dots w_n$ es:

$$\Pr(w_1 w_2 \dots w_{k-1} v\, w_{k+1} w_{k+2} \dots w_n | w_1 w_2 \dots w_{k-1} w\, w_{k+1} w_{k+2} \dots w_n),$$

con $w_1 w_2 \dots w_{k-1} w\, w_{k+1} w_{k+2} \dots w_n \in T_{\langle k-1:n-k \rangle}(w)$, para cualquier w de T. La probabilidad anterior se aproxima por:

$$\frac{\mathrm{fr}(w_1 w_2 \dots w_{k-1} v\, w_{k+1} w_{k+2} \dots w_n)}{\mathrm{fr}(w_1 w_2 \dots w_{k-1} w w_{k+1} w_{k+2} \dots w_n)}.$$

Si T_{ejem} =[`ejemplo de un texto de prueba`], entonces $T_{\mathrm{ejem}\langle 2 \rangle}$ ={`ejemplo de`, `de un`, `un texto`, `texto de`, `de prueba`}, y $\Pr(\texttt{de un} | \texttt{de } w) = \mathrm{fr}(\texttt{de un}) / \mathrm{fr}(\texttt{de } w) = 0.5$.

Si $v = a_1 \dots a_p \in V$, la *terminación* de longitud menor o igual a k de v es la subsecuencia de v:

$$v_{\_k} = \begin{cases} a_{p-k+1} a_{p-k+2} \dots a_p & \text{si } k < p \\ v & \text{si } k \geq p \end{cases}$$

Trabajaremos también con n-gramas de terminaciones: $T_{\langle n \rangle \_k}$. Así, $T_{\mathrm{ejem}\langle 1 \rangle \_3}$ ={`plo`, `de`, `un`, `xto`, `eba`} son todas las terminaciones de tamaño 3 de las palabras de T_{ejem}.

Formalmente, veremos a un etiquetador como una función que asigna la parte del discurso a cada palabra de T, esto es, la función: $\mathcal{E}^* : \mathcal{V}^* \to PD^*$. Así, si $T = [v_i]_{i \leq n_T}$ y $\mathcal{E}^*(T) = [m_i]_{i \leq n_T}$, entonces m_k es la marca de v_k, $k \leq n_T$.

3.1.2 Medición de la eficiencia

En contraste con la definición de $\mathcal{E}^*$, hay que precisar que no siempre un etiquetador adjudica una marca a todas las palabras del texto al que se aplica, ni lo hace siempre de manera correcta. Consideremos un etiquetador $\mathcal{E}^*$ que se aplica a un texto $T = [v_i]_{i \leq n_T}$, $T^* = \mathcal{E}^*(T) = [m_i]_{i \leq n_T}$. Sea $\tilde{m}$ una marca especial que representa una marca vacía, es decir, si m_k es un elemento de la secuencia T^* y $m_k = \tilde{m}$, entonces la palabra v_k de la secuencia T no pudo ser marcada por $\mathcal{E}^*$. Podemos entonces hablar del número de marcas asignadas por $\mathcal{E}^*$, $n_{T^*} = \#\{k|m_k$ ocurre en $T^* \wedge m_k \neq \tilde{m}\}$. Supongamos que se conoce el etiquetamiento correcto de T, $T' = [e_i]_{i \leq n_T}$. Para medir la eficiencia de $\mathcal{E}^*$ definimos los siguientes índices: La *precisión* del etiquetador $\mathcal{E}^*$ está dada por:

$$P = \frac{\#\{k|m_k = e_k\}}{n_{T^*}}, \tag{3.1}$$

esto es, la cantidad de marcas asignadas correctamente por $\mathcal{E}^*$, en proporción al número de marcas puestas por $\mathcal{E}^*$. La *evocación* es la cantidad de marcas asignadas correctamente en proporción al número de marcas preasignadas:

$$R = \frac{\#\{k|m_k = e_k\}}{n_T}. \tag{3.2}$$

Observemos que $R \leq P$, cuando $n_{T^*} \leq n_T$.

Es útil manejar estas dos cantidades en una sola que represente el desempeño de un etiquetador. Para este fin se emplea una proporción del número de aciertos respecto a la cantidad de marcas asignadas y preasignadas, la cual se calcula como:

$$F = \frac{2 \cdot P \cdot R}{P + R}. \tag{3.3}$$

Si $n_{T^*} \leq n_T$, entonces $n_{T^*} \leq (n_{T^*} + n_T)/2 \leq n_T$ y $R \leq F \leq P$. Cuando $R = 1$, identificaremos a P con la *exactitud*.

En la siguiente sección presentaremos algunos marcadores que servirán de base para ubicar el etiquetador desarrollado en el presente trabajo.

3.2 Técnicas para adjudicar partes del discurso

Antes de describir con detalle algunos marcadores que muestran diferentes enfoques, haremos referencia a los orígenes de los marcadores. Esto se hará con una descripción sucinta de los etiquetadores TAGGIT y CLAWS[1].

El sistema TAGGIT fue creado para etiquetar el corpus Brown[2]. TAGGIT realiza dos pasos: primero, identifica todas las etiquetas para cada palabra y, segundo, elige del conjunto de etiquetas de una palabra, la etiqueta más adecuada. El primer paso se realiza buscando cada palabra dentro de un diccionario, llamado diccionario de "excepciones" que incluye la mayoría de las palabras de clase *cerrada* (preposiciones, artículos, etc.). Posteriormente se procesan las contracciones, los símbolos especiales y palabras con letra mayúscula. Al final de este paso se verifica si la palabra contiene algún sufijo y con base en ello le asigna su etiqueta. Si aún no ha sido posible asignarle una etiqueta a la palabra, entonces se marca la palabra de manera múltiple con "sustantivo, adjetivo y verbo"[3]. El segundo paso se basa en reglas heurísticas, construidas manualmente, que para determinar la marca de una palabra consideran un contexto formado por dos palabras atrás y dos adelante.

CLAWS[4], se ha usado para etiquetar el corpus LOB (Lancaster-Oslo-Bergen). Su diccionario fue extraído del corpus Brown. Las reglas de CLAWS contemplan sólo un vecino alrededor de una palabra y están dispuestas

[1]En [Franz, 96] se describen estos sistemas. TAGGIT fue desarrollado a principios de 1970 por B. Green & G. Rubin, CLAWS se reportó por R. Garside & G. Leech en 1987.

[2]El corpus Brown fue un proyecto de recolección de recursos de información para ser empleados en sistemas de cómputo. En 1982, fue reportado este trabajo por Francis & Kučera Consiste de textos escritos del inglés americano, que reunen un millón de palabras [Franz, 96].

[3]Entrecomillado que ocurre en [Franz, 96]. N o se menciona el procesamiento de adverbios como una clase más que, por omisión, se incluya dentro de la marca múltiple asignada a las palabras que no pudieron marcarse.

[4]La principal aplicación de este etiquetador ha sido el etiquetamiento del *Britsh National Corpus* (BNC), un corpus de 100 millones de palabras[Garside, 96].

en una matriz, *matriz de colocaciones*. Esta matriz contiene las probabilidades de que una marca esté sucedida por otra. La entrada ij contiene la probabilidad de ocurrencia del bigrama $m_i m_j$:

$$\Pr(m_i, m_j) \approx \frac{\mathrm{fr}(m_i, m_j)}{\mathrm{fr}(m_i)\mathrm{fr}(m_j)}$$

donde fr denota la frecuencia de aparición de su argumento en el corpus Brown. Las probabilidades se usan para estimar la *máxima verosimilitud* de un etiquetamiento, de acuerdo con el siguiente criterio: Sea $\bar{M} = \bar{m}_1\bar{m}_2 \ldots \bar{m}_n$ una secuencia de marcas múltiples, la cual se desambigua adecuadamente con una de las secuencias $M_1, M_2, \ldots M_p$. La secuencia $M_q = m_{q1} \ldots m_{qn}$ que logra esta desambiguación se le llama la de *máxima verosimilitud*. El valor que determina la máxima verosimilitud de cada una de las secuencias desambiguadas, se calcula con base a las probabilidades que contiene la matriz de colocaciones:

$$val(M_i) = \prod_{j=1}^{n-1} \Pr(m_{i_j}, m_{i_{j+1}}).$$

Este método integró, además, la identificación de *lexicalizaciones* y consideraciones morfológicas para mejorar su desempeño.

En las siguientes subsecciones se revisarán dos métodos recientes, que han dado pauta a diversas aplicaciones en el PLN.

3.2.1 El modelo oculto de Markov

El *modelo oculto de Markov* trata con un conjunto de estados $M = \{m_1, m_2, \ldots, m_j\}$, una matriz de transiciones entre estados, $\mathbf{M}$, y una secuencia de observaciones $V = [v_1, v_2, \ldots v_n]$. Para el problema de etiquetamiento, los estados representan etiquetas y las observaciones las palabras que conforman al texto. Ya que sólo se observa la secuencia V y no M, el modelo se llama *oculto*.

De manera intuitiva [Poritz, 88] puede pensarse que se tiene $j+1$ tarros $T_0, T_1, \ldots, T_j$. Para i entre 1 y j el tarro T_i, por un lado, corresponde a la etiqueta m_i y, por otro, "contiene" etiquetas. Podemos pensar que al extraer una etiqueta m_k de T_i se selecciona la etiqueta siguiente a m_k en el texto a etiquetarse. Al tarro T_0 lo llamaremos *inicial*. Pensemos también que hay j urnas $U_1, \ldots, U_j$, donde la urna U_i contendrá las palabras que han de etiquetarse con m_i.

Inicialmente del tarro T_0 se extrae una primera etiqueta m_{i_1}. En cada paso del proceso, si se tiene la etiqueta m_{i_ι}, de la urna U_{i_ι} se extrae la palabra v_{i_ι} y del tarro T_{i_ι} se extrae la etiqueta $m_{i_{\iota+1}}$. Con esto la palabra v_{i_ι} habrá quedado marcada con la etiqueta m_{i_ι}. Y se repite el proceso.

La forma de aplicar el modelo oculto de Markov al etiquetamiento de textos se transforma en: maximizar la probabilidad de asignación de una secuencia de etiquetas a un texto Abreviando la aplicación del modelo oculto de Markov, diremos que: primero deben estimarse los parámetros del modelo y, después, llevar a cabo la asignación de etiquetas.

Sea $V = [v_1 v_2 \ldots v_n]$ un texto y $M = [m_1 m_2 \ldots m_n]$ una secuencia de etiquetas. Deseamos encontrar una secuencia M_{max} de etiquetas que maximice la probabilidad de asignación de etiquetas a V:

$$M_{max} = \mathrm{argmax}_M \ \Pr(M|V) \tag{3.4}$$

De acuerdo con la fórmula de Bayes, 3.4 se escribe como:

$$M_{max} = \mathrm{argmax}_M \ \frac{\Pr(M)\Pr(V|M)}{\Pr(V)}$$

luego se puede hacer

$$M_{max} = \mathrm{argmax}_M \ \Pr(M)\Pr(V|M), \tag{3.5}$$

ya que $\Pr(V)$ no influye en la elección del argumento que hace máxima la probabilidad. Finalmente expresamos 3.5 como:

$$M_{max} = \text{argmax}_M \ \Pr(m_1, m_2, \ldots m_n) \Pr(v_1, v_2, \ldots v_n | m_1, m_2, \ldots m_n)$$

Notemos que el término $\Pr(m_1, m_2, \ldots m_n) \Pr(v_1, v_2, \ldots v_n | m_1, m_2, \ldots m_n)$ contiene una cantidad excesiva de parámetros por estimar. Haciendo uso de la definición de probabilidad conjunta, tendríamos:

$$\begin{aligned} \Pr(m_1, m_2, \ldots m_n) \Pr(v_1, v_2, \ldots v_n | m_1, m_2, \ldots m_n) &= \\ \Pr(m_1, v_1, m_2, v_2, \ldots m_n, v_n) &= \\ \Pr(m_1) \Pr(v_1|m_1) \Pr(m_2|m_1, v_1) \Pr(v_2|m_1, v_1, m_2) \ldots & \end{aligned}$$

Se recurre entonces a una simplificación cuya utilidad ha sido comprobada, llamada, la *suposición de independencia*: la *i-ésima* etiqueta no depende de toda la historia anterior sino solamente de las dos etiquetas anteriores a ella. Esta suposición nos permite hacer la aproximación:

$$\Pr(m_i | m_1, \ldots m_{i-1}) \approx \Pr(m_i | m_{i-1}, m_{i-2})$$

y por tanto $\Pr(M)$ puede expresarse por:

$$\begin{aligned} \Pr(M) &= \Pr(m_1, m_2, \ldots m_n) \\ &= \Pr(m_1) \Pr(m_2|m_1), \ldots \Pr(m_n|m_1, \ldots m_{n-1}) \\ &\approx \prod_{i=1}^{n} \Pr(m_i | m_{i-1}, m_{i-2}) \end{aligned}$$

donde $\Pr(m_1|m_0 m_{-1}) = \Pr(m_1)$ y $\Pr(m_2|m_1 m_0) = \Pr(m_2|m_1)$. De la misma manera se supone que: la ocurrencia de una palabra es independiente de la ocurrencia de las palabras anteriores, mas no de la etiqueta de ella:

$$\begin{aligned} \Pr(V|M) &= \Pr(v_1, v_2, \ldots v_n | m_1, m_2, \ldots m_n) \\ &= \Pr(v_1|m_1) \Pr(v_2|v_1, m_1, m_2) \ldots \Pr(v_n|v_1, m_1, \ldots v_{n-1}, m_{n-1}) \\ &\approx \prod_{i=1}^{n} \Pr(v_i|m_i) \end{aligned}$$

La función final a maximizar es:

$$M_{max} = \prod_{i=1}^{n} \Pr(v_i|m_i) \Pr(m_i|m_{i-1}, m_{i-2})$$

Todo esto es suponiendo correcto el modelo. La estimación de las probabilidades implicadas puede realizarse de muchas formas. Mencionaremos aquí sólo el caso de la estimación de los parámetros del modelo mediante interpolación lineal [Merialdo, 91], y usando un corpus ya etiquetado. Dadas tres etiquetas consecutivas m_1, m_2 y m_3 la probabilidad $\Pr(m_3|m_1, m_2)$ se aproxima por la frecuencia de ocurrencia de ese trigrama, $\text{fr}(m_3|m_1, m_2)$, y se interpola con una distribución uniforme de la asignación de etiquetas:

$$\Pr(m_3|m_1, m_2) \approx \lambda \text{fr}(m_3|m_1, m_2) + (1 - \lambda)/N_m$$

donde $0 \leq \lambda \leq 1$ y N_m es el número de etiquetas posibles. Ya que entre mayor sea el corpus, mayor la confiabilidad hacia la aproximación de la probabilidad por las frecuencias, es natural establecer una relación inversa entre λ y el tamaño del corpus. Y esto lo apreciamos en la fórmula de interpolación; cuando el corpus es "pequeño" habremos de apoyar la estimación con la distribución uniforme. Similarmente la probabilidad de un vocablo dada una etiqueta se aproxima por la frecuencia observada en un corpus etiquetado. También esta probabilidad se suaviza con una distribución basada en la selección de una de las palabras que conforman al corpus:

$$\Pr(v|m) \approx \lambda \text{fr}(v|m) + (1 - \lambda)/N_p$$

aquí N_p es el número de palabras en el corpus. Debemos señalar que la suavización de las distribuciones observadas tiene un peso importante en el algoritmo de etiquetamiento, ya que de otra forma los valores de etiquetas ausentes pueden sesgar sensiblemente la asignación de probabilidades y por tanto la determinación de las marcas. Las probabilidades con valor cero, por tanto, se suplantan por valores de omisión, calculados por las interpolaciones expuestas.

3.2.2 Marcador basado en transformaciones

E. Brill [Brill, 94] propuso la estrategia de *etiquetamiento basado en transformaciones*, estrategia que se ha extendido a otras aplicaciones [Samuel, Carberry, et. al. 98], [Brill, Resnik, 94]. Su algoritmo en forma supervisada hace necesario contar con un texto marcado que sirve como "corrector", además de un diccionario que asigna una marca[5] a cada una de las palabras del texto por etiquetar.

El método basado en transformaciones genera reglas que sucesivamente se ajustan para reducir el error del etiquetamiento. Sucintamente, los pasos para asignar partes del discurso a un texto son:

1. Realizar un etiquetamiento inicial a través del diccionario.
2. Comparar con el texto corrector, a fin de generar reglas de transformación de las etiquetas asignadas.
3. Aplicar al texto marcado las reglas encontradas para reducir el error.
4. Iterar en tanto el error decrezca.

En el diccionario se trata de disponer del mayor número de palabras con su etiqueta más probable.

A cada palabra w del texto a marcar se le asigna la marca **b**, si tenía la marca **a** en el contexto [**c a d**]. La marca **b** (transformación) es tal que para w ocurre más veces en el contexto [**c b d**] dentro del texto corrector.

Para precisar las ideas anteriores representemos con T_c el texto corrector y con T_m el texto a marcar, sigamos denotando con letras **sólidas** etiquetas constantes, y con w una palabra de T_m, cualquiera y fija. Sean X, Y y Z variables con dominio en las etiquetas, y el predicado $A(T, w, X, Y, Z)$ = "w tiene la marca Y en el contexto $[X\ Y\ Z]$ dentro de T", y la puntuación de una transformación $\mathbf{e}$ de w en T_m, $\text{puntua}(A, w, \mathbf{a}, \mathbf{e})$, como "el número de veces que $A(T_c, w, X, \mathbf{e}, Z)$ es verdadero, si $A(T_m, w, X, \mathbf{a}, Z)$ es verdadero".

Para cada w de T_m con marca **a** se le asigna la marca:

$$\mathbf{b} = \text{argmax}_{\mathbf{e}}\ \text{puntua}(A, w, \mathbf{a}, \mathbf{e})$$

Otros esquemas de reglas pueden conseguirse variando los contextos en los antecedentes considerados.

El método anterior puede adaptarse al aprendizaje de transformaciones en forma no supervisada, cuya diferencia es que no emplea un texto marcado, pero sí un diccionario que asigna todas las marcas posibles a cada una de las palabras que ocurren en el texto a etiquetar. Las reglas tienen la forma: *Cambiar la marca* χ *de una palabra* w *por* $\mathbf{d}$*,* $(\mathbf{d} \in \chi)$*, en el contexto* C=$[w_{-n} \ldots w_{-1} w w_1 \ldots w_n]$. El proceso de generación de transformaciones se realiza considerando que para cada $\mathbf{c} \in \chi$ $(\mathbf{c} \neq \mathbf{d})$ se ha calculado su puntuación:

$$\frac{\text{fr}(\mathbf{c})}{\text{fr}(\mathbf{d})} * incontext(\mathbf{c}, C), \tag{3.6}$$

[5]Este diccionario se conforma con las marcas de mayor frecuencia que ocurren en el texto corrector.

donde $incontext(\mathbf{c}, C)$ representa el número de ocurrencias de la marca $\mathbf{c}$ en el contexto C. El método elige la marca que maximice 3.6:

$$\mathbf{d} = \underset{\mathbf{e}\in\chi,\mathbf{e}\neq\mathbf{c}}{\operatorname{argmax}} \frac{\mathrm{fr}(\mathbf{c})}{\mathrm{fr}(\mathbf{e})} * incontext(\mathbf{e}, C)$$

La precisión del etiquetador basado en transformaciones es uno por ciento mejor que el método estocástico [Brill, 95]. Por otro lado, este marcador se considera el más rápido –diez veces la velocidad del marcador estocástico– y con ventajas de ser aplicado a variados propósitos dentro del procesamiento del lenguaje natural [Samuel, Carberry, et. al. 98].

* * *

Se resumen las anteriores subsecciones con la tabla 3.1 que compara los marcadores presentados, donde la precisión se calcula como lo indica la ec. 3.1.

Etiquetador	Tamaño del diccionario	No. de Reglas	Precisión (%)	Tamaño del corpus (1,000×palabras))
TAGGIT	3450	3300	77.0	1000
CLAWS	7700	-	94.0	-
Markov	50000	10000	96.7	1000
Brill	5000	267	97.2	600

Tabla 3.1: Tabla comparativa de los marcadores de partes del discurso.

Como puede observarse, el etiquetador basado en transformaciones muestra ventajas por tener mejor desempeño al usar un corpus menor al que refiere el marcador estocástico. El método basado en transformaciones usa el diccionario para iniciar la asignación de conjuntos de marcas a cada una de las palabras del texto por etiquetar, lo cual supone que no hay palabras desconocidas. Esta última consideración le crea al método de transformaciones un grado particular de debilidad. Al lado de este marcador está el etiquetador basado en el modelo oculto de Markov que ha contribuido notablemente a la solución de los problemas de reconocimiento de voz. Sin embargo, parece necesario el empleo de un corpus grande para poder alcanzar un buen desempeño, así como la incorporación de más elementos lingüísticos, ajenos al método, para usarlo en lenguas diferentes al inglés.

La exposición de las técnicas de etiquetamiento no pretende ser completa[6], sólo trata de mostrar algunos procedimientos seguidos en la adjudicación de la parte del discurso a las palabras de un texto. Excepto TAGGIT y CLAWS, en los etiquetadores presentados hacen mayor énfasis en el contexto de ocurrencia de las palabras, predomina la información sintáctica sobre la morfológica, lo cual es esperado por orientarse a una lengua, como el inglés. Para la lengua española, es necesario dar un peso mayor al aspecto morfológico, si se desean buenos resultados en la asignación de partes del discurso. La siguiente subsección presenta un marcador construido específicamente para el español mexicano, que pone de manifiesto la anterior observación.

3.2.3 Técnica basada en tablas de precedencia

La importancia de este marcador[7] radica en los resultados que se obtuvieron para el *Corpus del Español Mexicano Contemporáneo* (CEMC), base para la construcción del *Diccionario del Español de México* (DEM) de El Colegio de México. Además, la estrategia que sigue constituye una guía frente al tratamiento de las particularidades del español usado en México. En la tabla 3.3 se muestra la composición del CEMC[8],

[6] Otros etiquetadores son los reportados en: [Samuelsson, Voutilainen, 97] y [Schütze, 93].

[7] Hasta donde se conoce el marcador descrito en esta subsección es el único marcador desarrollado en México para el español y construido antes de 1995 ("La formalización del analizador gramatical del DEM", Ma. Isabel García Hidalgo, [Lara, Ham-Chande, et. al. 79] pp 85-155).

[8] Tomada de [Lara, Ham-Chande, et. al. 79].

en ella se exhiben la clases de lenguas muestreadas (columna 1), su porcentaje en el CEMC y los géneros considerados, así como el porcentaje de cada género participante (en la columna final). Las partes del discurso que emplea este sistema aparecen abreviadas, a lo largo de este trabajo, como lo indica la tabla 3.2, con una cadena en fuente `diferente` a la usual.

Símbolo	Nombre	Símbolo	Nombre
`AAD`	*A. Admiración*	`PTC`	*Punto y coma*
`ADJ`	*Adjetivo*	`GUI`	*Guión*
`ADM`	*C. Admiración*	`INT`	*C. Interrogación*
`ADV`	*Adverbio*	`NOM`	*Sustantivo*
`AIN`	*A. Interrogación*	`NP`	*Nombre propio*
`ART`	*Artículo*	`NUM`	*Número*
`COM`	*Coma*	`PREP`	*Preposición*
`CONJ`	*Conjunción*	`PRON`	*Pronombre*
`CONT`	*Contracción*	`PTO`	*Punto*
`DPT`	*Dos puntos*	`VERB`	*Forma Verbal*
`SUS`	*Puntos susp.*		

Tabla 3.2: Lista de las partes del discurso consideradas.

El conocimiento sobre la lengua permite construir tablas de precedencia y postcedencia, para definir funciones útiles en la determinación de la parte del discurso de un vocablo que aparece en cierto contexto. Además, en este sistema, se consideraron otras regularidades, como el uso de información morfológica y listas de palabras frecuentes.

Dado un vocabulario $\mathcal{V}$ y suponiendo que PD es el conjunto de partes del discurso, haremos uso de los siguientes términos: Las *categorías funcionales* de $v \in \mathcal{V}$, son todas las partes del discurso con las que puede ocurrir v en cualquier texto. Sea $\Phi \subset \mathcal{V} \times \mathrm{PD}$ la relación que formaliza este concepto:

$$\Phi(v,e) \ \Leftrightarrow\ [v \text{ ocurre con la parte del discurso } e.]$$

Llamaremos a $\mathcal{V}_{\min} \subset \mathcal{V}$ el conjunto de *vocablos básicos*. Para los vocablos básicos se emplean las *categorías de vocablos básicos*: $\Phi_{\min} \subset \mathcal{V}_{\min} \times \mathrm{PD}_{\min}$, con $\mathrm{PD}_{\min} = \{e | \exists v \in \mathcal{V}_{\min}, \Phi_{\min}(v,e)\}$. Denotemos con $\phi(v) = \{e | \Phi_{\min}(v,e)\}$ al conjunto de categorías posibles de v.

Recursos usados para etiquetar un texto

Los recursos que emplea este sistema son básicamente dos: terminaciones que permiten identificar la categoría de un vocablo haciendo uso de un diccionario de excepciones y, un conjunto de tablas de precedencia para adjudicar una marca de acuerdo con el contexto donde aparezca.

El marcador dispone también de información para asignar una categoría a ciertos vocablos a partir de su terminación. Con la función morfo representamos la categoría de un vocablo v dependiendo del sufijo máximo de v. En los casos indefinidos o ambiguos asignamos $\between$.

Las tablas de precedencia en que se basa este etiquetador son de los siguientes tipos:

- *Precedencia simple.* Suponemos dada una transformación $\mathrm{F} : \mathrm{PD}_{\min} \times \mathcal{V}_{\min} \to \mathrm{PD}_{\min}$ que decide la categoría de un vocablo a partir de la categoría de su vecino izquierdo en el texto procesado. Para los casos en los que no se cuenta con suficiente información F da el valor *ambiguo*: $\mathrm{F}(e,v) = \between$.
- *Precedencia doble.* Si para un elemento de $T_{\langle 3 \rangle}$, $v_1 v_2 v_3$, se tiene que $v_1, v_2 \in \mathcal{V}_{\min}$ y sus categorías son diferentes a $\between$, es posible, para algunos casos, asignar una categoría a v_3. Suponemos dada una

Input:
- $T = [v_i]_{i=1}^{n_T}$: *A text.*
- *F,G,H: Mappings that codify context-sensitive information on words.*
- *morfo: Mapping codifying morphological information*

Output:
The tagging $[m_i]_{i=1}^{n_T}$ *corresponding to* T

Algorithm:
$j := 2;$
while $j < n_T$ **do**
    case
        $\alpha(v_j) \wedge \phi(v_j) = \{e_k\}: \;\; m_j := e_k;$
        $\alpha(v_j) \wedge \beta(\psi(v_{j-1}), v_j): \;\; m_j := \mathrm{F}(\psi(v_{j-1}), v_j);$
        $\alpha(v_j) \wedge \neg\beta(\psi(v_{j-1}), v_j) \wedge \eta(v_j, v_{j+1}): \;\; m_j := \mathrm{H}(v_j, v_{j+1});$
        $\neg\alpha(v_j) \wedge \theta(v_j): \;\; m_j := \mathrm{morfo}(v_j);$
        else $m_j := \between$
    end ;
    if $\gamma(\psi(v_{j-1}), \psi(v_j), v_{j+1})$ **then**
        $m_{j+1} := \mathrm{G}(\psi(v_{j-1}), \psi(v_j), v_{j+1});$
        $j++;$
    $j++$
end

Figura 3.1: Algoritmo de etiquetamiento basado en tablas de precedencia.

transformación $\mathrm{G} : \mathrm{PD}_{\min} \times \mathrm{PD}_{\min} \times \mathcal{V} \to \mathrm{PD}$ que asigna una categoría a un vocablo considerando las categorías de sus dos vecinos previos.

- *Postcedencia simple.* Supongamos que para $v_1v_2v_3 \in T_{\langle 3\rangle}$ se tiene $\#\phi(v_1) > 1$, i.e. v_1 puede asumir varias categorías, y por tanto $\mathrm{F}(\between, v_2)$ no podría calcularse. Si, por otro lado, $\mathrm{morfo}(v_2) \neq \between$, el proceso de elegir algún elemento de $\phi(v_1)$ y asignarlo a v_1 queda facultado por $\mathrm{morfo}(v_2)$, en este caso podríamos hacer $\mathrm{H}(v_1, v_2) = e \in \phi(v_1)$. H es pues una función de postcedencia que obtiene la categoría de v_1 a partir de la categoría de v_2, habiendo fijado esta última a partir de la morfología de v_1.

Etiquetamiento de vocablos de un texto T

Supongamos que se tiene disponible la información concerniente a $\Phi, \Phi_{\min}, \mathrm{F}, \mathrm{G}, \mathrm{H}$ y morfo, descritos anteriormente. Sean $\alpha, \beta, \gamma, \eta$, y θ los siguientes predicados:

$$\begin{aligned}
\alpha(v) &\doteq [v \in \mathcal{V}_{\min}] \\
\beta(e, v) &\doteq [\mathrm{F}(e, v) \neq \between] \\
\gamma(e, e', v) &\doteq [\mathrm{G}(e, e', v) \neq \between] \\
\eta(u, v) &\doteq [\mathrm{H}(u, v) \neq \between] \\
\theta(v) &\doteq [\mathrm{morfo}(v) \neq \between]
\end{aligned}$$

Representemos con $\psi(v)$ la categoría del vocablo v previamente analizado. Un algoritmo para etiquetar un texto se presenta en la Figura 3.1.

Muestra	% absoluto	Género	% relativo
Lengua Culta	66.8		100.00
		Literatura	22.45
		Periodismo	26.34
		Ciencia	26.94
		Técnica	15.26
		Discurso Político	02.69
		Religión	01.79
		Habla de la Cd. de México	04.49
Lengua Sub-Culta	11.7		100.00
		Literatura Popular	53.00
		Conversaciones grabadas	47.00
Lengua No-estándar	21.5		100.00
		Textos Regionales	60.46
		Documentos de antropólogos	15.34
		Jergas	13.95
		Conversaciones grabadas	10.25

Tabla 3.3: Composición del Corpus del Español Mexicano Contemporáneo.

El analizador del DEM es de mediados de los setenta [Lara, Ham-Chande, et. al. 79] y el único antecedente del etiquetamiento de textos del español de México con partes del discurso. Dicho analizador tuvo entre sus fines el de ayudar al etiquetamiento del CEMC. Más recientemente, para el español se han aplicado otros algoritmos como el del modelo oculto de Markov, usando aprendizaje basado en memoria [Daelemans, van-den-Bosch, et. al. 98] y árboles de decisión [Màrquez, Rodríquez, 98]. El corpus objeto de la aplicación del método de tablas de precedencia fue producto de recolectar aproximadamente 1000 textos con 2000 palabras cada uno, haciendo un total de dos millones de palabras. Estos requisitos fueron fijados por el análisis derivado de los objetivos del proyecto: costo de la adquisición de dicha información, y la obtención de alrededor de 30,000 vocablos.

3.3 Etiquetador basado en instancias

En las anteriores secciones se han presentado marcadores que siguen estrategias muy diferentes. El etiquetador que va a ser expuesto en esta sección se inspira en gran medida en el marcador basado en tablas de precedencia, por seguir algunos de sus pasos, y porque retoma los resultados que de él se obtuvieron para usarlos como corpus de entrenamiento. Asimismo se hace uso de la técnica de clasificación elegida en el capítulo anterior, por parecer muy prometedora en el campo del procesamiento del lenguaje natural ([Daelemans, 95], [Daelemans, Durieux, et. al. 98], [Bosch, Daelemans, et. al. 96], [Zavrel, Daelemans, et. al. 97], [Veenstra, 98]).

3.3.1 Elementos de diseño del sistema

Recordemos que la función que desempeña una palabra en un discurso se llama parte del discurso. Por otro lado, independientemente de la parte del discurso de una palabra, distinguiremos dos tipos de palabras: Las **palabras definidas** son aquéllas que tienen sólo una función en cualquier contexto que ocurran y, por tanto, siempre tendrán la misma parte del discurso. Las **palabras ambiguas** son palabras para las cuales existen al menos dos contextos en los que la palabra ambigua tiene diferente parte del discurso. Es natural pensar que cada uno de estos tipos de palabras reciba un tratamiento diferente en el proceso de etiquetamiento. El hecho de apoyarnos en un corpus, para etiquetar cualquier otro texto, nos lleva a considerar el uso de una lista de palabras definidas con sus marcas, de esta manera podremos adjudicarle a algunas palabras, de otro texto, su parte del discurso. Aunque este recurso es bastante útil debe reconocerse su limitación, pues no habrá corpus suficientemente grande para garantizar que toda palabra definida sea etiquetada. La anterior observación nos presenta un nuevo tipo de palabra que encontraremos en los textos, y que será llamada **palabra desconocida**. Las palabras desconocidas pueden ser ambiguas o definidas, pero de ellas no tenemos información para poderlas incluir en alguno de estos conjuntos. El etiquetador tendrá que abocarse a decidir la marca por asignar a las palabras conocidas (definidas o ambiguas), y a las palabras desconocidas. Un primer acercamiento a los pasos que deberá seguir el etiquetador son:

1.	Etiquetar las palabras definidas.
2.	Desambiguar las palabras ambiguas.
3.	Determinar la parte del discurso de las palabras desconocidas.

Figura 3.2: Pasos globales del método de etiquetamiento.

Es necesario, como primer paso del etiquetamiento, precisar cómo se adjudicará su parte del discurso a las palabras definidas. Se ha propuesto la construcción de una lista de palabras definidas de acuerdo con el corpus de trabajo. Ya que el corpus de trabajo contiene las marcas de cada una de sus palabras, lo manejaremos como dos secuencias: el texto $T_C = [v_i]_{i \leq n_{T_C}}$ y sus marcas $T_C^* = [m_i]_{i \leq n_{T_C}}$. Tenemos, entonces, que el primer paso de etiquetamiento se puede realizar con base en la función $\mathcal{E}^1 : DEF \to \text{PD}$, donde

$$DEF = \{v_j \in T_C | \forall k \neq j (v_j = v_k \Rightarrow m_j = m_k)\}.$$

Sea T el texto a marcar, justamente si x ocurre en T y $x \in DEF$, $x \mapsto \mathcal{E}^1(x) = m_j$, si $x = v_j \in DEF$.

Como segunda medida, dentro del primer paso, podemos reunir algún otro tipo de información lingüística. En particular, a partir de catálogos que nos indiquen con seguridad la adjudicación de una marca, a una palabra de cualquier otro texto, pueden definirse reglas morfosintácticas. En este caso caen las palabras para las que siendo aún desconocidas, puede ser identificada su parte del discurso. Estas palabras pueden llamarse *parcialmente conocidas*, puesto que no están completamente registradas en un inventario. Un ejemplo de esto es: adjudicar a las palabras con terminación "ción" la parte del discurso sustantivo.

Podemos acudir, también, a una lista de parejas (palabra, marca), en la que la marca de cada pareja sea la de mayor frecuencia con que ocurre la palabra en el corpus de trabajo. Es interesante saber que el etiquetamiento con esta lista, extraída de nuestro corpus de trabajo, arroja una precisión de 0.9528, sin embargo su evocación es de 0.7097, lo cual hace un desempeño, medido por el número F, de 0.8134. Otro tipo de asignación de marcas es la que se basa en la probabilidad condicional de ciertos n-gramas. La eficiencia de esta asignación está comprometida seriamente con el tamaño del corpus, y en nuestras condiciones conviene apoyarse en la siguiente observación. Fijado un corpus, la cantidad de n-gramas aumenta si se consideran terminaciones en lugar de palabras, de esta manera con la ayuda del cálculo de la probabilidad condicional se elige la marca más probable. Por ejemplo, una palabra v puede etiquetarse con la marca m, cuando a v preceden inmediatamente dos palabras con terminaciones t_{-2} y t_{-1}, y además la probabilidad $\Pr(t_{-2}t_{-1}m|t_{-2}t_{-1})$ alcanza su máximo valor con m. Según el corpus de trabajo, si t_{-2} es "." y t_{-1} es "la" entonces, con una probabilidad de 0.8333, la marca de la palabra siguiente a la de terminación t_{-1} es sustantivo.

Los últimos pasos de etiquetamiento se dedican a la desambiguación de palabras y a la determinación de su parte del discurso de las palabras desconocidas. Estos casos tendrán que apoyarse con más amplitud en el corpus de trabajo y particularmente en el método de clasificación. Es de interés determinar una lista de las palabras ambiguas (al menos las más frecuentes) pues esta identificación permite obtener directamente los contextos que servirán para desambiguarlas. Basta extraer, del corpus de trabajo, los contextos para cada una de las palabras de este inventario y aplicar el método de clasificación usando estos contextos o n-*gramas*.

Por último, el caso de las palabras desconocidas puede abordarse por exclusión, es decir, cualquier palabra que no haya podido marcarse con los pasos previos caerá en esta categoría. En primer lugar, se marca cada palabra del corpus de trabajo aplicando los anteriores pasos de etiquetamiento. Cada palabra que no haya sido marcada se considerará desconocida y, por tanto, puede extraerse el n-grama donde ocurre. Es posible, entonces, alimentar el método de clasificación con este conjunto de n-gramas.

Observemos, por otro lado, que entre más casos abarque nuestro corpus de trabajo se traducirá en mayor desempeño. Puesto que el corpus de trabajo es pequeño tenemos que sacar provecho de otras características, tanto de las estrategias para marcar, como de los métodos para clasificar. En nuestro caso haremos uso de algunos diccionarios que pueden ser extraídos del corpus de trabajo, además seguiremos dos estrategias para aplicar el método de clasificación:

- La desambiguación de palabras basada en el método de clasificación deberá aplicarse a conjuntos lo más específicos posibles, sin detrimento de la cantidad de casos.

- Deberá tratarse de adecuar el método de clasificación para mejorar la eficiencia del etiquetamiento.

3.3.2 Requisitos de información y especificaciones informales

El corpus[9] utilizado en la presente aplicación está compuesto de 9,224 elementos y su vocabulario es de 2,644 palabras, incluyendo signos de puntuación e inflexiones. Las palabras se distribuyen en 21 clases de partes del discurso incluidas en la siguiente tabla:

Rango	P. del D.	%	Rango	P. del D.	%
1	NOM	20.50	12	CONT	1.79
2	PREP	13.37	13	NUM	0.67
3	VERB	12.70	14	GUI	0.53
4	ART	10.70	15	DPT	0.29
5	ADJ	10.41	16	PTC	0.27
6	COM	6.07	17	INT	0.23
7	ADV	5.37	18	AIN	0.20
8	CONJ	5.31	19	SUS	0.09
9	PRON	5.23	20	ADM	0.05
10	PTO	3.65	20	AAD	0.05
11	NP	2.42			

Tabla 3.4: Porcentaje de ocurrencia de las partes del discurso en el corpus de trabajo.

De manera importante aparece la coma con 6.07% y también, dentro de los rangos de palabras, ocurre el punto con 3.65%. Siguen los demás signos de puntuación (1.74% en total), precedidos en rango por la marca numérica, esta última con 0.67%.

Como puede observarse las palabras pertenecientes a las clases representadas por `NOM`, `VERB`, `ADJ`, y `ADV` tienen casi un 50% de ocurrencia y contienen varios casos de ambigüedad (`NOM/VERB`, `ADJ/ADV`, etc.). Bien que muchas palabras de estos conjuntos pueden identificarse por características morfológicas, consideraremos en especial la desambiguación de `NOM/ADJ/VERB`. Los verbos derivan nominales que a su vez juegan el papel de sustantivos o adjetivos, por ejemplo "dibujo", que es una forma verbal y sustantivo. Se espera que un tratamiento especial para este tipo de ambigüedad mejore el rendimiento del etiquetamiento.

Aparentemente, éstos son los casos más difíciles, sin embargo, veamos, por ejemplo, las diez palabras más frecuentes del español mexicano[10]:

[9]Este corpus es una parte del *Corpus del Español Mexicano Contemporáneo*, compuesto de artículos periodísticos del género editorial.
[10]Tomado de [Lara, Ham-Chande, et. al. 79].

Vocablo	*la*	*el*	*de*	*y*	*que*	*en*	*a*	*se*	*no*	*ser*
P. del D.	ART	ART	PREP	CONJ	PRON	PREP	PREP	PRON	ADV	VERB
Rango	1	2	3	4	5	6	7	8	9	10

Tabla 3.5: Rango de las palabras más frecuentes del español mexicano.

Los vocablos "la", "que" y algunas formas del verbo "ser" son ambiguos. Específicamente, en el CEMC, "la" en su rol de pronombre tiene rango 46. Desempeñando el rol de conjunción, "que" tiene rango 28. En el corpus de trabajo, las formas ambiguas de "la", incluyendo "las", tienen un 5.31% de ocurrencia y rango dos. Las formas ambiguas de "que" ocurren 2.98% con rango 5. Por último, está "los" de rango diez. Excepto las anteriores, las demás palabras de rango menor a diez, son definidas, entre las cuales se encuentran los signos punto y coma. En suma, en el corpus de trabajo se tienen 32.07% ocurrencias de palabras con más de una marca y 2.02 partes del discurso promedio por palabra con más de una marca[11]. Ya hemos hecho la observación sobre la necesidad de dedicar una etapa del etiquetamiento a este caso de desambiguación. Pasemos ahora a describir los recursos en los cuales se basa cada paso de etiquetamiento, según lo expresado en la figura 3.2.

Recursos del sistema

Un etiquetador basado en un corpus no debe despreciar la información antedicha. Más aún, el mismo corpus ofrece un recurso potencial de marcas, extrayendo de él una lista que permita decidir la marca que puede ser asignada a cada palabra de otro texto. Desde que se inició el trabajo de construcción de etiquetadores esta lista ha sido muy útil y es identificada como un *diccionario* del sistema de etiquetamiento.

Diccionarios

También resulta conveniente usar diccionarios para un conjunto amplio de palabras y no sólo las que ocurren en el corpus, de esta forma se reforzará el vocabulario del etiquetador desarrollado. Los verbos son candidatos para aplicar el recurso del diccionario, ya que hay listas disponibles de verbos y de sus terminaciones para cada caso de conjugación.

Los diccionarios de un sistema de etiquetamiento contienen, pues, las palabras conocidas. Por otro lado, y como es de esperarse, habrá siempre un porcentaje importante de palabras desconocidas, para las cuales no es posible el recurso del diccionario. De éstas hablaremos más adelante.

Para construir una lista de palabras frecuentes, se tomó como base la lista de las cien palabras más frecuentes del español mexicano [Lara, Ham-Chande, et. al. 79]. Esta lista de cien palabras fue repartida en dos: una lista de vocablos definidos y otra de vocablos ambiguos. La primera (DICCD) con el fin de usarla como diccionario y asignar partes del discurso de manera directa y, la segunda (DICCA), para preparar un tratamiento especial dirigido a las palabras ambiguas frecuentes. Los verbos que ocurrieron entre las palabras más frecuentes fueron extendidos a todas sus conjugaciones (DICCV) y, de estas conjugaciones, las formas verbales ambiguas fueron trasladadas a DICCA.

De acuerdo con la tabla 3.4, no hay mucho que hacer con los conjuntos cerrados de palabras invariables que en su mayoría son definidas como el caso de PREP (donde se exceptúan algunas como "para", "sobre", etc.). Aunque con un rango bajo, observemos la categoría de nombre propio. Este tipo de palabras no es fácil de reconocer debido a que se confunden con las palabras que inician las oraciones. Como los nombres propios forman un conjunto abierto de palabras, el problema de la determinación de su parte del discurso conviene manejarse como el de una palabra desconocida. Sin embargo, la baja frecuencia de los nombres propios y el uso de un corpus pequeño produce n-gramas insuficientes. Se optó por emplear una lista de

[11] Obsérvese que en el corpus de trabajo puede haber otras palabras que son ambiguas pero ocurren sólo con una marca.

nombres propios (NOMP) extraída de textos periodísticos adicionales[12]. De un total de 222,867 palabras se extrajeron todas las que iniciaran con mayúscula y se descartaron manualmente las que no fueran nombres propios. El resultado fue una lista con 3,337 nombres propios.

Agregando las marcas de los signos de puntuación, cuya frecuencia es 11.20%, resumimos en la siguiente tabla los elementos que hemos descrito y su ocurrencia en el corpus de trabajo[13]:

Nombre del dicc.	Tamaño	Tipo de palabra	% ocurrencia
DICCD	184	definida frecuente	28.38
DICCA	163	ambigua frecuente	17.59
DICCS	11	puntuación	11.20
DICCV	835	forma verbal frecuente	5.20
NOMP	3,337	nombre propio	2.08
Total	*4,530*		*64.47*

Tabla 3.6: Diccionarios conformados por palabras conocidas y su ocurrencia en el corpus de trabajo.

La tabla 3.6 nos lleva a concluir[14] que el uso de diccionarios de palabras definidas, i.e. DICCD, DICCS, DICCV y NOMP, garantiza marcar aproximadamente la mitad de palabras contenidas en un texto cualquiera de género semejante al del corpus de trabajo. Por su parte, el diccionario de palabras ambiguas, DICCA, será útil para llevar a cabo un tratamiento especial orientado a este tipo de palabras.

Reglas morfosintácticas

Un conjunto de palabras con frecuencia de ocurrencia alta y bastantes casos de ambigüedad, es el de los nominales. Si bien el enfoque que hemos tratado de seguir está basado en corpus, no conviene encerrarse en este camino y omitir información lingüística muy conocida que se tiene acerca del español. Hay otras características de las palabras que permiten aligerar el trabajo de la determinación de su parte del discurso. Estoy refiriéndome a las terminaciones. En el libro de José Moreno de Alba [Moreno, 86] se documentan las terminaciones de nominales y es posible extraer una lista de terminaciones que asigne unívocamente la categoría de sustantivo. Sea TERD la lista de terminaciones para sustantivos derivados. La validación de esta lista de terminaciones se hace verificando que toda palabra que tenga una terminación de TERD sea sustantivo, ya que no todas las palabras que tienen una terminación de TERD son sustantivos. Para $t \in$ TERD, se comprueba entonces que:

$$\Pr(v_{i\_k} = t, m_i = \texttt{NOM} | v_{i\_k} = t) = 1,$$

donde k es la longitud de t. Claramente habrá terminaciones que no satisfagan la igualdad anterior, por ejemplo: "acia", "adora", "alla", etc. Pero también, para ciertos casos, algunas palabras que violan la igualdad (palabras que no son sustantivos) son detectados en su totalidad en el paso anterior, manteniendo así la regla morfosintáctica correspondiente. Éste es el caso de "acia" donde $\Pr(v_{i\_4} = \text{acia}, m_i = \texttt{NOM} | v_{i\_4} = \text{acia}) = 0.66$, lo cual se debe a la palabra "hacia". Las terminaciones elegidas para definir sustantivos fueron: "acia", "ad", "amento", "amiento", "ancia", "anda", "ato", "encia", "icia" "idumbre", "ón", "tad", "tura" y "ud".

Uno de los fines de esta primera etapa de etiquetamiento fue la de marcar la mayor parte de palabras, para ello, como se ha señalado, se puso énfasis en las palabras frecuentes. Para determinar si algún n-grama era útil en este propósito, se constató que su probabilidad condicional fuera uno. Un ejemplo de este tipo

[12]Esta colección de textos corresponde a 127 artículos tomados al azar, de la revista *Proceso* del año 1998.

[13]A partir de las palabras extraídas de la lista de las cien más frecuentes, los diccionarios DICCD, DICCA y DICCV fueron enriquecidos con palabras bastante comunes y rango promedio de 46.71.

[14]Debido a que el concepto de palabra frecuente que estamos manejando se basa en las cien palabras más frecuentes, la tabla 3.6 es una expresión del resultado que presenta P. Guiraud [Guiraud, 54]: las palabras con rango menor o igual que 100 cubren el 60% del léxico.

de regla es asignar a cualquier elemento del conjunto $loas = \{la, las, lo, los\}$ la marca ART, siempre que se satisfaga que la palabra previa sea preposición. En suma, se validaron en el corpus de trabajo las siguientes reglas:

$$\Pr(m_{i-1} = \texttt{PREP}, v_i \in loas, m_i = \texttt{ART} | m_{i-1} = \texttt{PREP}, v_i \in loas) = 1$$
$$\Pr(m_{i-1} = \texttt{CONT}, m_i = \texttt{NOM}, v_{i+1} \in ddel | m_{i-1} = \texttt{CONT}, v_{i+1} \in ddel) = 1$$
$$\Pr(v_{i-1} \in \{el, al\}, m_i = \texttt{NOM}, v_{i+1} \in ddel | v_{i-1} \in \{el, al\}, v_{i+1} \in ddel) = 1$$
$$\Pr(m_{i-2} = \texttt{PREP}, m_{i-1} = \texttt{ART}, m_i = \texttt{NOM}, v_{i+1} \in ddel | m_{i-2} = \texttt{PREP}, m_{i-1} = \texttt{ART}, v_{i+1} \in ddel) = 1$$

donde $ddel = \{de, del\}$.

Un recurso más, fue una lista de más de 7,000 verbos extraídos del sistema COES[15]. Se conformaron dos listas de verbos, una de verbos regulares y otra de verbos irregulares, además, para cada tipo de verbo, se compiló una tabla de conjugación. El número de reglas de conjugación, agrupadas por terminación, cubre 1,370 casos. Designemos con el conjunto VERBO a la lista de verbos del español en forma infinitiva. Sea $\mathcal{M}$ el conjunto de gramemas verbales. La tabla que ayuda a identificar las conjugaciones válidas puede representarse como la función:

$$\text{inflex} : \mathcal{M} \rightarrow 2^{\mathcal{M}}$$

Esta función asocia una terminación verbal a otras, y su funcionamiento es el siguiente: Sea w una palabra, tal que $w = xy$, donde $y \in \mathcal{M}$. Supongamos que aplicamos inflex a y y obtenemos $\text{inflex}(y) = [y_i]_{i \leq ny}$, entonces para que w sea una conjugación válida, debe verificarse, para alguna $k \leq ny$, que $xy_k \in$ VERBO. Por ejemplo, "anduviesen" se separa[16] en "and" y "uviesen", por su parte $\text{inflex}(uviesen) = [ener, ar]$ y claramente con $k = 2$, resulta válida la conjugación.

La identificación de verbos se refuerza con la consideración de clíticos. El funcionamiento es semejante al caso anterior: se divide la palabra en dos partes, $w = xy$, y si y es un clítico basta agregar a x una terminación de infinitivo r y verificar que xr sea miembro de VERBO. La verificación de conjugaciones verbales para clíticos utiliza una expresión regular que representa clíticos en reversa. En el caso de los verbos regulares se utiliza la siguiente expresión:

```
^s?[oa]l(son|e[mts])r[\'a\'e\'{\i}]|
^s?[oae]lr[aei]|
^s?([oae]l|[oa]l(e[mts]|son))odn\'a|
^s?([oae]l|[oa]l(e[mts]|son))odn\'ei|
^s?([oae]l|[oa]l(e[mts]|son))odn\'e
```

En todo el proceso anterior se evitó marcar palabras ambiguas. Para lograr este propósito sólo se etiquetaron las palabras que cumplieran con algunas de las condiciones que se han expuesto y no pertenecieran a DICCA, ni tampoco tuvieran una terminación común con los sustantivos o adjetivos[17]. Algunas de las terminaciones comunes para verbos y nominales son: "ías", "ado", "ido", "ía", "to", "so", "es", "as", "o", "a", "era", "ijo", "iño", "ite" y "uelo". En total se manejan 36 terminaciones comunes; conjunto que denominaremos TERC.

Con la identificación de verbos según se expuso, aumenta el número de palabras marcadas en 5%, es decir, después de usar los diccionarios del sistema y las conjugaciones verbales, se alcanza un total del 56% de palabras marcadas, con un error promedio del 2%.

Contextos

Quedan pendientes por procesar las palabras ambiguas frecuentes y las palabras desconocidas. Observemos que todavía hay información en algunas de las palabras revisadas por la etapa anterior, pero que no fueron

[15] Este es un sistema de ayuda ortográfica para el español, orientado al trabajo de edición en ambientes UNIX [Rodríguez, Carretero, 97].

[16] Debe aclararse que el proceso de separación da preferencia a los sufijos más grandes.

[17] Para esto se acudió nuevamente al inventario que ofrece J. Moreno de Alba [Moreno, 86].

marcadas. Hablamos de las palabras que satisfacen la condición de verbo y, a la vez, tienen terminación de nominal. Podemos entonces establecer las siguientes etapas finales del etiquetamiento:

- Desambiguar palabras ambiguas frecuentes.
- Desambiguar entre verbos y nominales.
- Determinar la etiqueta de las palabras desconocidas.

Palabras ambiguas frecuentes.
La desambiguación de palabras ambiguas frecuentes se realizó aplicando el método de aprendizaje basado en memoria. Recordemos que es necesario tomar en cuenta el n-grama en el que ocurra la palabra ambigua para decidir su clase. A estas alturas del etiquetamiento la palabra ambigua puede tener más de dos vecinos inmediatos, pues también podemos emplear la marca, tanto de su vecino izquierdo como del derecho, si es el caso de que hubieran adquirido una marca en alguno de los pasos de etiquetamiento anterior. Así pues, utilizaremos las palabras vecinas y sus marcas como n-grama para desambiguar una palabra ambigua frecuente.

Un factor importante, que debe tenerse presente, es que nuestro corpus de trabajo no puede proveer suficientes n-gramas formados por palabras. Una alternativa para aumentar el número de n-gramas es conformarlo por las terminaciones de las palabras vecinas. Observemos la curva de ganancia de las instancias formadas por 38 terminaciones de longitud ocho. Cada terminación corresponde a una de las 19 palabras vecinas contiguas, que se encuentran a cada lado de una palabra ambigua frecuente:

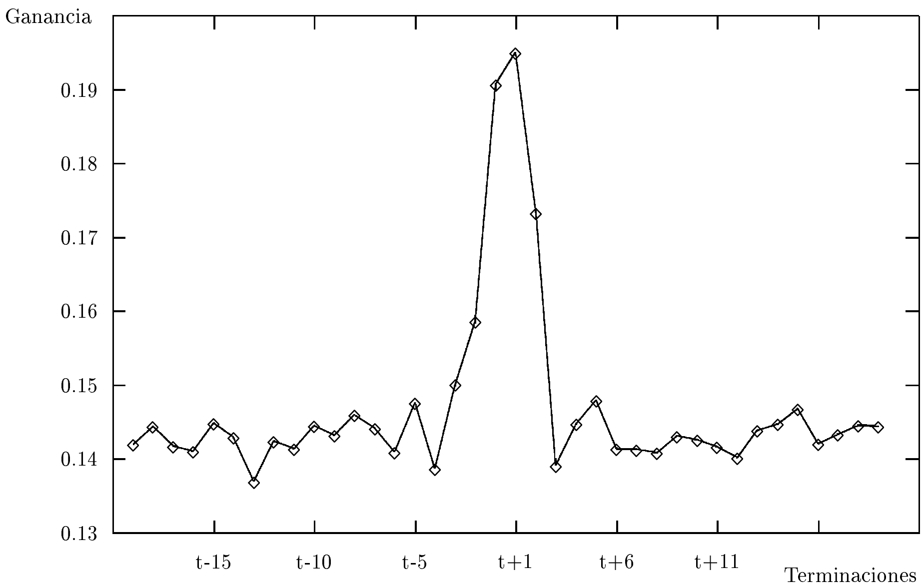

Figura 3.3: Curva de ganancia para 19 terminaciones a cada lado de una palabra ambigua frecuente.

La curva muestra, claramente, que son cuatro los vecinos de una palabra ambigua frecuente, (t-2, t-1, t+1 y t+2, en la fig. 3.3) los dos más cercanos a cada uno de sus lados, los que proporcionan mayor información para determinar la parte del discurso de este tipo de palabra. Observemos, por otro lado, la curva de ganancia de las marcas que rodean a una palabra ambigua frecuente:

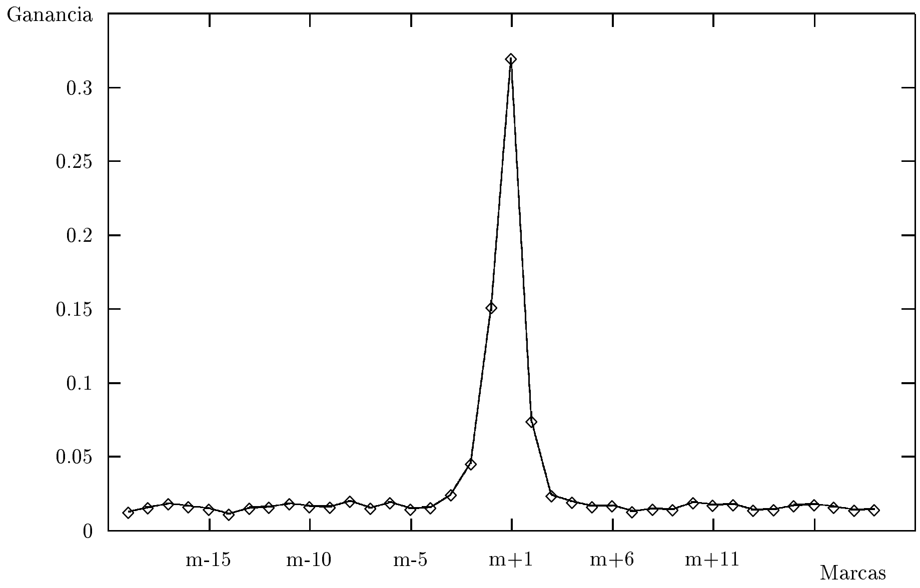

Figura 3.4: Curva de ganancia para 19 marcas a cada lado de una palabra ambigua frecuente.

Comparando con la ganancia de las terminaciones elegidas para desambiguar (fig. 3.3), cuya mínima ganancia es 0.1586 bits, tenemos que sólo son dos las marcas que contribuyen con mayor información: una marca a la izquierda y otra a la derecha (m-1 y m+1 en la fig. 3.4), con ganancias 0.1514 bits y 0.3198 bits, respectivamente. En síntesis, el contexto para desambiguar una palabra ambigua frecuente estará constituido por:

- Las terminaciones de dos palabras vecinas inmediatas a cada lado de la palabra por desambiguar.
- Una marca contigua a cada lado de la palabra por desambiguar.

Verbos y nominales.
La siguiente etapa es resolver la ambigüedad de asignación de la etiqueta verbo o nominal a una palabra. En el momento que se etiquetaron los verbos por medio de las tablas de conjugación, se detectaron las palabras con la ambigüedad verbo/nominal, haciendo uso de TERC. Para poder contar con información adicional a estas palabras, se les asignó la marca `VERP`, de tal manera que se pudiera construir el contexto que ayudara a desambiguarlas. El contexto de una palabra con ambigüedad verbo/nominal se forma siguiendo los pasos:

1. Márquese el corpus de trabajo con los pasos previos a la desambiguación de palabras ambiguas frecuentes.
2. Extráigase los n-gramas formados por marcas y terminaciones de palabras vecinas alrededor de la marca `VERP`.

Al analizar las curvas de ganancia de estas instancias se concluyó, como en el caso de las palabras ambiguas, que: los atributos contextuales que proporcionan mayor información son las terminaciones de cuatro palabras, dos a cada lado de un verbo probable, y una marca a cada lado del mismo verbo probable, i.e t-2, t-1, t+1, t+2, m-1 y m+1.

Palabras desconocidas.
Por último, se asignarán etiquetas a las palabras desconocidas, que son casi el 20% del corpus de trabajo.

Se esperaba que estas palabras fueran principalmente nominales. Al extraer los n-gramas alrededor de las palabras no marcadas hasta el paso anterior -e identificadas con la marca `IND-`, se determinaron las categorías de las palabras desconocidas: sustantivo, adjetivo, verbo, adverbio, preposición y nombre propio. Para ellas se generaron también los contextos que permitieran decidir su clase. De igual manera, como en los dos casos anteriores, se observó las curvas de ganancia para elegir ahora: dos marcas y dos terminaciones inmediatamente antes de la palabra desconocida, como contexto para ayudar a etiquetar este tipo de palabra.

Debemos agregar que, en este caso, el algoritmo se comportó más sensible a la ausencia de marcas. Esto es, al no haber marca de un vecino considerado como parte del n-grama de una palabra desconocida, este valor se contabiliza como diferente en todas las instancias con las que se mide su distancia. A la vez que la distancia se desplaza por una constante hacia arriba, otros atributos influyen de manera importante en la determinación de la clase elegida. Por ello es que se propuso estimar la presencia de una marca dentro del n-grama de una palabra desconocida. La estimación antedicha se hace mediante la probabilidad condicional de que esta marca ocurriera dado que se tuvieran las terminaciones de dos palabras previas a ella. En la fig. 3.5 se muestra claramente por qué se tomaron las dos terminaciones previas (t-2 y t-1), para estimar una marca vecina a una palabra desconocida:

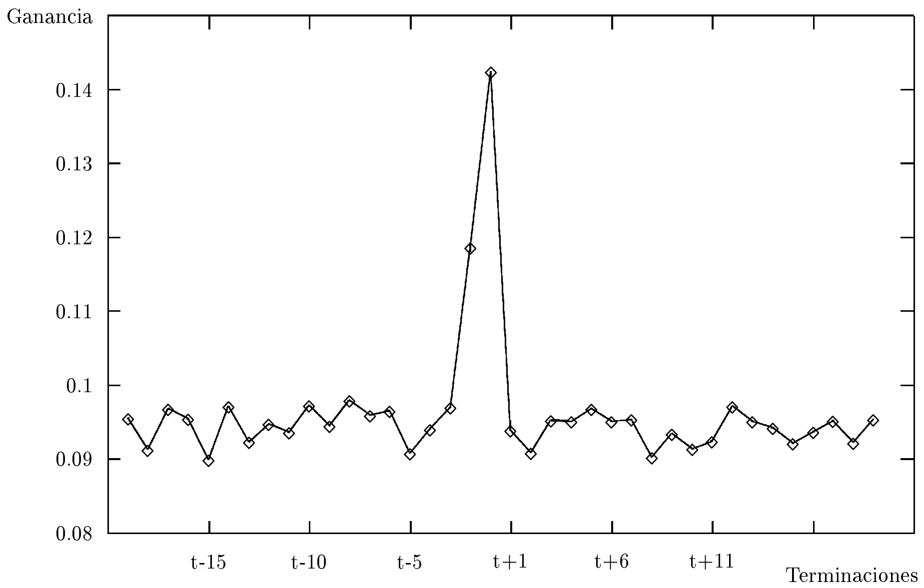

Figura 3.5: Curva de ganancia para 19 terminaciones a cada lado de una palabra desconocida.

Un ejemplo de esta estimación es el siguiente. La pareja "en el" ocurre 14 veces y, en el corpus de trabajo, hay cuatro posibles marcas después de ella.

t_{-2}	t_{-1}	$marca$	# ocurrencias	$\Pr(t_{-2}t_{-1}marca \mid t_{-2}t_{-1})$
en	el	NOM	9	0.642857
en	el	NUM	1	0.071428
en	el	ADJ	3	0.214286
en	el	ADV	1	0.071428

Tabla 3.7: Probabilidad condicional de encontrar una determinada marca después de dos terminaciones previas inmediatas.

Con este tipo de tablas, para cada trigrama, podemos pesar nuevamente la función de distancia. Sea X el n-grama de una palabra desconocida y Y una instancia. La distancia entre los componentes de X y Y está definida por:

$$\delta'(x_i, y_i) = \begin{cases} 0 & \text{si } x_i = y_i \\ p_i & \text{si } x_i = \texttt{IND} \\ 1 & \text{en cualquier otro caso} \end{cases}$$

donde $p_i = 1 - \Pr(t_{i-2}t_{i-1}y_i|t_{i-2}t_{i-1})$. Usemos la tabla 3.7 en un ejemplo: Si una instancia contiene como atributos las marcas `ART` y `NOM`, correspondientes a la marca izquierda y derecha, pero en el n-grama de una palabra desconocida no tiene marca la palabra inmediata derecha, entonces la contribución a la distancia por parte de este atributo es 0.357143 y no uno como podría obtenerse sin el uso de p_i.

3.3.3 Pasos de etiquetamiento y ejemplos

Hemos visto que la estrategia de etiquetamiento planteada en esta sección puede ofrecer buenos resultados. Estamos entonces en posición de clarificar los pasos seguidos para llevar a cabo el etiquetamiento de las palabras de un texto. En el apéndice A se encuentra una especificación formal que complementa la presente descripción.

El refinamiento practicado a los pasos de la fig. 3.2, conducen a contar en total siete pasos. Refiriéndonos a ellos por el recurso que emplean son:

1. Diccionarios
2. Reglas morfológicas
3. Inflexiones verbales regulares
4. Inflexiones verbales irregulares
5. Instancias para ambiguas frecuentes
6. Instancias para verbo/nominal
7. Instancias para palabras desconocidas

Consideremos un texto $T = [v_i]_{i \leq n_T}$ a ser etiquetado. En cada paso de etiquetamiento se obtendrá un etiquetamiento parcial de T. $T_j^* = [m_{j,i}]_{i \leq n_T}$ representa el etiquetamiento obtenido en el paso j. En el paso 7 obtendremos T^*.

Paso 1. Para cada elemento v_i del texto T:

1. Si $v_i \in$ DICCD, entonces $m_{1,i}$ tendrá la marca que indica DICCD.
2. Si $v_i \in$ DICCV, entonces $m_{1,i} = \texttt{VERB}$.
3. Si $v_i \in$ NOMP, entonces $m_{1,i} = \texttt{NP}$.
4. Si $v_i \in$ DICCS, entonces $m_{1,i}$ tendrá la marca que indica DICCS.
5. Si alguna terminación k de v_i cumple $v_{i\_k} \in$ TERD, entonces $m_{1,i} = \texttt{NOM}$.
6. Si v_i no obtuvo marca por los pasos anteriores, lo indicamos con la marca `IND`.

Paso 2. Consideramos los contextos de las palabras de T y los correspondientes de T_1^*.

Para cada elemento v_i no marcado del texto T:

1. Si $[v_i m_{1,i-1}] \in \{\texttt{la}, \texttt{las}, \texttt{lo}, \texttt{los}\}\texttt{PREP}$, $m_{2,i} = \texttt{ART}$.
2. Si $[v_{i+1} m_{1,i-1}] \in \{\texttt{de}, \texttt{del}\}\texttt{CONT}$, $m_{2,i} = \texttt{NOM}$.
3. Si $[v_{i-1} v_{i+1}] \in \{\texttt{el}, \texttt{al}\}\{\texttt{de}, \texttt{del}\}$, $m_{2,i} = \texttt{NOM}$.

4. Si $[v_{i+1} m_{1,i-1} m_{1,i+1}] \in \{\texttt{de}, \texttt{del}\}\texttt{PREPART}$, $m_{2,i} = \texttt{NOM}$.
5. En caso de no haber asignado etiqueta a v_i actualícese T_2^* con T_1^*: $m_{2,i} = m_{1,i}$.

Paso 3. Consideramos los gramemas de inflexión verbal regular, $\mathcal{M}_1$, y la función que asigna a cada uno de éstos las terminaciones infinitivas posibles, que al combinarse con el lexema producen un verbo en infinitivo: $\text{inflex}_1 : \mathcal{M}_1 \to 2_1^{\mathcal{M}}$. En este proceso también se identificarán palabras con ambigüedad verbo/nominal.

Para cada elemento v_i no marcado del texto T:

1. Si existe k tal que $v_{i\_k} \in$ TERC, asignamos a $m_{3,i}$ la marca `VERP`.
2. Si existe k tal que $v_{i\_k} \in \mathcal{M}_1$, $v_i = wv_{i\_k}$, y para alguna $t \in \text{inflex}_1(v_{i\_k})$, se tiene que $wt \in$ VERBO, entonces $m_{3,i} = \texttt{VERB}$.
3. En cualquier otro caso, $m_{3,i} = m_{2,i}$.

Paso 4. De igual manera que en el paso anterior debemos contar con lo correspondiente para verbos irregulares: $\mathcal{M}_2$ e $\text{inflex}_2 : \mathcal{M}_2 \to 2_2^{\mathcal{M}}$.

Para cada elemento v_i no marcado del texto T:

1. Si existe k tal que $v_{i\_k} \in$ TERC, asignamos a $m_{4,i}$ la marca `VERP`.
2. Si existe k tal que $v_{i\_k} \in \mathcal{M}_2$, $v_i = wv_{i\_k}$, y para alguna $t \in \text{inflex}_2(v_{i\_k})$, se tiene que $wt \in$ VERBO, entonces $m_{4,i} = \texttt{VERB}$.
3. En cualquier otro caso, $m_{4,i} = m_{3,i}$.

Paso 5. Haremos uso de una colección de entrenamiento que permita etiquetar palabras ambiguas frecuentes, según DICCA. Dicha colección está compuesta por contextos de palabras ambiguas frecuentes, extraídas del corpus (CEMC), cuyos componentes son las terminaciones de longitud seis de cuatro palabras alrededor de cada palabra ambigua frecuente, dos a la derecha y dos a la izquierda, además las marcas de las palabras vecinas inmediatas a la palabra ambigua frecuente. Sea INST5 la colección antedicha. Representemos con X_{v_i} el contexto de v_i (necesario para calcular la distancia entre X_{v_i} y una instancia de INST5, según la fórmula 2.7): $[v_{i-2\_6} v_{i-1\_6} v_{i+1\_6} v_{i+2\_6} m_{4,i-1} m_{4,i+1}]$.

Para cada elemento v_i no marcado del texto T:

1. Si $v_i \in$ DICCA, entonces asignar la clase de la instancia Y que minimiza la distancia entre el contexto de v_i y Y: $\text{clase}_{\text{INST5}}(\text{argmin}_{Y \in \text{INST5}} \Delta(X_{v_i}, Y))$.
2. En cualquier otro caso, $m_{5,i} = m_{4,i}$.

Paso 6. De igual manera que en el paso anterior haremos uso de una colección de entrenamiento, ahora para palabras con la ambigüedad verbo/nominal, INST6, con las mismas características que INST5. Es decir, requerimos ahora contextos alrededor de palabras con ambigüedad nominal/verbo, también de la forma: $[v_{i-2\_6} v_{i-1\_6} v_{i+1\_6} v_{i+2\_6} m_{5,i-1} m_{5,i+1}]$. La variante principal respecto al paso anterior es que se emplea el método IGTree: si el contexto de una palabra no coincide con alguno de la colección, entonces se toma la clase de la instancia que "más empata", comparando los rasgos en orden según su ganancia normalizada, lo cual representaremos con $\text{clas}_{\text{INST6}}(X_{v_i})$.

Para cada elemento v_i marcado como `VERP` del texto T:

1. Si $X_{v_i} \in$ INST6, entonces $m_{6,i} = \text{clase}_{\text{INST6}}(X_{v_i})$.
2. Si $X_{v_i} \notin$ INST6, entonces $m_{6,i} = \text{clas}_{\text{INST6}}(X_{v_i})$.
3. En cualquier otro caso, $m_{6,i} = m_{5,i}$.

Paso 7. En este paso se etiquetan todas las palabras restantes, que hemos llamado desconocidas. La colección INST7 está orientada a obtener información sobre este tipo de palabras. Ahora los contextos tienen la forma: $[v_{i-2} v_{i-1} m_{6,i-2} v_{6,i-1}]$. Este paso es una variante más al método de aprendizaje basado en memoria. En esencia sigue la idea del paso anterior, pero además verifica si el contexto de la palabra por marcar tiene entre sus rasgos más importantes (mayor ganancia normalizada) alguno desconocido por la colección de entrenamiento y, en este caso se realiza la búsqueda de la instancia más cercana al contexto de la palabra desconocida utilizando la métrica Δ', substituyendo a δ por δ' en la ec 2.7. Cuando los rasgos desconocidos no son muy importantes, lleva a cabo la asignación de acuerdo con el método IGTree.

Para cada palabra no marcada de T:

1. Si $X_{v_i} \in \mathrm{INST7}$, entonces $m_{7,i} = \mathrm{clase}_{\mathrm{INST7}}(X_{v_i})$
2. Si $X_{v_i} \notin \mathrm{INST7}$, y ninguno de sus dos rasgos más importantes es desconocido, $m_{7,i} = \mathrm{clas}_{\mathrm{INST7}}(X_{v_i})$.
3. Si $X_{v_i} \notin \mathrm{INST7}$, y alguno de sus dos rasgos más importantes son desconocidos, entonces asignar la clase de la instancia Y que minimiza la distancia entre el contexto de v_i y Y: $m_{7,i} = \mathrm{clase}_{\mathrm{INST7}}(\mathrm{argmin}_{Y \in \mathrm{INST7}} \Delta'(X_{v_i}, Y))$.
4. En cualquier otro caso, $m_{7,i} = m_{6,i}$.

Se presenta a continuación un texto de ejemplo[18], para mostrar el funcionamiento del etiquetamiento.

y el beneficiario directo de todo ello será , sin duda , el PRI , según reconoció el secretario de hacienda ante analistas , académicos e inversionistas en el consejo de las américas de Nueva York , horas antes de participar en la reunión de los organismos financieros internacionales . dijo , sin ambages , que si la economía sigue bien y se mantiene la disciplina , el partido revolucionario institucional tendrá buenas posibilidades no sólo en la contienda por la presidencia , sino en la elección del gobierno capitalino y , aun , en recuperar la mayoría absoluta en la cámara de diputados . ... en efecto , en su examen anual de la economía mexicana , que hizo público el jueves 29 , la OCDE - el club de los 29 países más ricos del mundo , al que México ingresó en 1994 - admite que el desempeño económico del país fue positivo en los últimos tres años , pero señala una serie de ineficiencias en la política económica - inestabilidad presupuestal , dependencia del petróleo y deficiente sistema tributario - que impiden sacar al país del subdesarrollo y contrarrestar los niveles de pobreza extrema .

Figura 3.6: Texto de ejemplo (**A**).

El paso 1 adjudicó al texto de la fig.3.6 103 etiquetas. La tabla 3.8 contiene una muestra de su comportamiento. Su contenido, por orden de columnas, es 1) la frecuencia del patrón o diccionario que habilitó la etiqueta, 2) la posición de la palabra etiquetada, 3) la parte del discurso asignada a la palabra, 4) la palabra, y 5) el patrón o diccionario utilizado. En esta tabla se muestran sólo algunos casos, los más frecuentes al aplicar el paso 1. al texto de 3.6. Para identificar el número de palabra indicado puede verse la fig. 3.7, en la cual la primera columna contiene la posición de la primera palabra de ese renglón.

fr	#Palabra	PD	Palabra	Patrón o diccionario
57	95	`PREP`	en	DICCD
25	105	`PTO`	.	DICCS
8	44	`NOM`	reunión	TERD
5	8	`VERB`	será	DICCD
5	142	`NP`	México	NOMP
3	145	`NUM`	1994	/[[:digit:]]+(.[[:digit:]]*)?%?/

Tabla 3.8: Muestra del comportamiento del paso 1.

[18]Este texto es un fragmento del artículo periodístico contenido en la revista **PROCESO** de Carlos Acosta Córdova y Guillermo Correa, del dos de mayo de 1999.

La tabla 3.9 presenta una estructura igual a la 3.8, excepto que en la columna 5 no aparece el recurso patrón o diccionario, sino el antecedente de la regla morfosintáctica aplicada para efectuar el etiquetamiento. En el primer ejemplo de esta tabla tenemos el etiquetamiento de “la” cuyo contexto, observando la fig. 3.7, coincide con el antecedente de la regla: “en la”, ya que “en” fue marcada en el paso previo con PREP.

fr	#Palabra	PD	Palabra	Antecedente de regla
11	86	ART	la	$m_{i-1} = \mathtt{PREP}\ \ t_i \in loas$
6	33	NOM	américas	$m_{i-1} = \mathtt{CONT}\ \ t_{i+1} \in ddel$

Tabla 3.9: Muestra del comportamiento del paso 2.

Tanto en la tabla 3.10 como en la 3.11 se muestran ejemplos de etiquetamiento, ahora usando el recurso de la inflexión verbal. En la columna 5 aparece el gramema que permitió ejercer una combinación con el lexema de la palabra y conseguir un verbo en infinitivo. En la columna 5 de la tabla 3.10 aparece una marca adicional que especifica el tipo de gramema.

fr	#Palabra	PD	Palabra	Gramema
4	66	VERP	disciplina	2gram:a
3	96	VERB	recuperar	inf
3	190	VERP	subdesarrollo	2gram:o
2	165	VERB	serie	1pgram:e
1	69	VERP	partido	2gram:odi
1	194	VERP	niveles	2gram:se
1	143	VERB	ingresó	1pgram:ó

Tabla 3.10: Muestra del comportamiento del paso 3.

fr	#Palabra	PD	Palabra	Gramema
1	79	VERB	contienda	ienda
1	64	VERB	mantiene	iene
1	60	VERB	sigue	igue
1	21	VERB	hacienda	ienda
1	186	VERB	sacar	INF
1	185	VERB	impiden	iden
1	17	VERB	reconoció	ió

Tabla 3.11: Muestra del comportamiento del paso 4.

La estructura de la tabla 3.12 es diferente a las anteriores. En primer lugar, la frecuencia se refiere a la distancia y no a la instancia. De la columna 5 a la 8 tenemos parte de la instancia que minimizó la distancia hacia el contexto de la palabra por etiquetar, éstas son las terminaciones de dos palabras que rodean a otra (también ambigua frecuente, pero del corpus de entrenamiento). Siguen, en las columnas 9 y 10, las marcas de las palabras que rodean a la palabra ambigua frecuente en la instancia. En la columna final aparece la distancia entre el contexto de la palabra por etiquetar y la instancia mostrada. Veamos que la palabra “gobierno” ocurre en la posición 89 y su contexto formado por las terminaciones de las palabras que la rodean: “ección”, “del”, “talino” e “y”, además la marca de “del”, que es CONT y la marca de “capitalino”, aún sin determinar en este paso, IND. Resumiendo, el contexto de “gobierno”, siguiendo el orden que presenta la tabla 3.12, es: $X =$[“ección”, “del”, “talino”,“y”, CONT, IND]. El contexto X se compara con todas las instancias de la colección de entrenamiento. La más cercana resultó: $Y =$[“rmidad”, “del”, “licano”, “,”, CONT, ADJ]. La ganancia que se obtuvo para los rasgos, de la colección que estamos usando fue $W = [0.1660,\ 0.2000,\ 0.2084,\ 0.1758,\ 0.1514,\ 0.3193]$, por tanto, $\Delta(X,Y) = w_1 + w_3 + w_4 + w_6 = 0.1660 + 0.2084 + 0.1758 + 0.3193 = 0.8695$.

fr	#Palabra	PD	Palabra	v_{i-2_6}	v_{i-1_6}	v_{i+1_6}	v_{i+2_6}	m_{i-1}	m_{i+1}	Distancia
5	89	NOM	gobierno	rmidad	del	licano	,	CONT	ADJ	0.869667
5	46	ART	los	umento	de	recios	del	PREP	NOM	0.869667
5	184	CONJ	que	ndante	-	México	no	GUI	NP	0.869667
5	132	ART	los	umento	de	recios	del	PREP	NOM	0.869667
5	126	ART	la	uiente	,	cusión	lativa	COM	NOM	0.869667
2	97	ART	la	nombre	de	ayoría	.	PREP	NOM	1.01273
2	86	ART	la	golpe	en	noche	del	PREP	NOM	0.374478
2	84	CONJ	sino	ricano	,	en	los	COM	PREP	0.341839
2	81	ART	la	,	por	dencia	ritual	PREP	NOM	0.341839
2	6	ADJ	todo	.	en	ello	,	PREP	PRON	0.54192
2	58	ART	la	ngreso	y	onomía	.	CONJ	NOM	0.861307
2	56	CONJ	que	,	lvidan	si	no	VERB	CONJ	0.693341
2	32	ART	las	el	de	ciones	de	PREP	NOM	0.374478
2	169	ART	la	egunda	sobre	lítica	gética	PREP	NOM	0.861307
2	148	CONJ	que	se	nfiere	el	estado	VERB	ART	0.54192
2	141	CONJ	que	ndante	-	México	no	GUI	NP	0.693341
2	135	ADJ	más	tres	veces	ricos	.	NOM	NOM	1.01273
1	78	PRON	la	;	hasta	lsaron	nsando	PREP	VERB	0.750361
1	65	ART	la	se	ienten	paz	,	VERB	NOM	0.727909
1	61	ADV	bien	las	siones	o	mal	NOM	CONJ	0.901782
1	43	ART	la	icipar	en	rencia	de	PREP	NOM	0.208441
1	193	ART	los	y	a	jefes	de	PREP	NOM	0.87933
1	164	ART	una	clara	de	serie	de	PREP	NOM	0.836926
1	157	ART	los	que	en	óximos	tres	PREP	ADV	0.693865
1	119	PRON	que	isoras	,	son	amente	COM	VERB	0.55028
1	115	ART	la	efensa	de	onomía	opular	PREP	NOM	0.661226
1	101	ART	la	rchivo	de	cámara	de	PREP	NOM	0.366118

Tabla 3.12: Muestra del comportamiento del paso 5.

La siguiente tabla (3.13) reporta ejemplos del paso que emplea el método IGTree. Es por ello que en su columna final aparecen sólo algunos de los rasgos de la instancia, que empataron con el contexto de la palabra con ambigüedad verbo/nominal. En la cabeza de la columna 5 aparece, según la ganancia normalizada, el orden de los rasgos. Nótese que estos prefijos son cortos respecto al tamaño de la instancia (6 valores).

Como se ha dicho, el paso 6 no realiza el cálculo de la distancia, solamente lleva a cabo una comparación de los rasgos de la palabra por etiquetar con los rasgos de las instancias. Los rasgos que "más empataron", de entre todas las instancias, determinan marcas de palabras, que ocurren en el corpus de entrenamiento, con el contexto formado por esos rasgos. La marca más frecuente de las palabras en ese contexto es la que se asignó. La palabra "partido" tiene claramente el contexto [`ART` "el"] (ver fig. 3.7) y éste es también parte de varias instancias de la colección. Así, en el corpus, la marca más frecuente de todas las palabras con el contexto [`ART` "el"] fue `NOM`.

fr	#Palabra	PD	Palabra	Prefijo: m_{i-1} v_{i-1_6} v_{i+1_6}
2	69	NOM	partido	ART el
2	150	NOM	desempeño	ART el
1	66	NOM	disciplina	ART la ,
1	197	VERB	extrema	IND
1	194	NOM	niveles	ART los de
1	190	NOM	subdesarrollo	CONT del y
1	163	VERB	señala	CONJ pero
1	155	VERB	positivo	VERB fue
1	11	NOM	duda	PREP sin ,

Tabla 3.13: Muestra del comportamiento del paso 6.

Ya que el último paso combina el método IGTree con la función de distancia δ', la columna 5 de la tabla 3.14 muestra mayor heterogeneidad. Los casos que tuvieron valores desconocidos después del segundo rasgo se resolvieron como en el paso anterior, y se reportan en la tabla 3.14, de igual forma que en la tabla 3.13. Los contextos de palabras desconocidas que contuvieron rasgos desconocidos muy importantes (rasgo 1 o 2, según el orden de la ganancia normalizada) se reportan como en la tabla 3.12, esto es, la instancia más cercana y la distancia hacia el contexto de la palabra por etiquetar.

fr	#Palabra	PD	Palabra	Prefijo: m_{i-1} v_{i-1_6} v_{i-2_6} m_{i-2} /Distancia
4	59	NOM	economía	ART la
3	93	ADV	aun	COM , y CONJ
3	178	NOM	petróleo	CONT del dencia NOM
2	3	NOM	beneficiario	ART el y CONJ
2	196	NOM	pobreza	PREP de
2	158	NOM	últimos	ART los en PREP
1	90	ADJ	capitalino	NOM bierno del CONT
1	73	NP	buenas	PTO . cional ADJ /0.265352
1	71	ADJ	institucional	ADJ onario
1	70	ADJ	revolucionario	NOM artido
1	4	ADJ	directo	NOM sábado el ART /0.163999
1	54	NOM	ambages	PREP sin , COM
1	49	VERB	internacionales	ADJ ctivas uerzas NOM /0.341505
1	48	ADJ	financieros	NOM ientos los ART /0.163218
1	27	ADJ	inversionistas	CONJ e
1	23	NOM	analistas	PREP en niendo VERB /0.341505
1	182	ADJ	tributario	ADJ ctivas uerzas NOM /0.341505
1	181	ADJ	sistema	NOM gas y CONJ /0.135299
1	174	ADJ	presupuestal	NOM ilidad
1	171	ADJ	económica	NOM lítica la ART
1	99	ADJ	absoluta	NOM vida la ART /0.180399

Tabla 3.14: Muestra del comportamiento del paso 7.

El etiquetamiento final se muestra en la fig.3.7. La tabla 3.8 resume el comportamiento de cada paso del etiquetamiento del texto. Podemos observar, en la tabla 3.8 que el conjunto TERC tiene carencias. En los pasos de etiquetamiento de verbos se clasifican como tales, palabras que son nominales pues tienen una terminación común a alguna inflexión verbal que no aparece en TERC. La mayor dificultad observada, en el ejemplo que estamos revisando, es el etiquetamiento de las palabras con ambigüedad verbo/nominal, dificultad que puede suponerse debida a la escasa colección de entrenamiento usada: 514 instancias en contraste con 1623 instancias para las palabras ambiguas y 1803 para las palabras desconocidas.

#	Texto/Marcas
1	y el beneficiario directo de todo ello será , sin
	CONJ ART NOM ADJ PREP ADJ PRON VERB COM PREP
11	duda , el PRI , según reconoció el secretario de
	NOM COM ART NP COM PREP VERB ART NOM PREP
21	hacienda ante analistas , académicos e inversionistas en el consejo
	VERB PREP NOM COM NOM CONJ ADJ PREP ART NOM
31	de las américas de Nueva York , horas antes de
	PREP ART NOM PREP NP NP COM NOM ADV PREP
41	participar en la reunión de los organismos financieros internacionales .
	VERB PREP ART NOM PREP ART NOM ADJ VERB PTO
51	dijo , sin ambages , que si la economía sigue
	VERB COM PREP NOM COM CONJ CONJ ART NOM VERB
61	bien y se mantiene la disciplina , el partido revolucionario
	ADV CONJ PRON VERB ART NOM COM ART NOM ADJ
71	institucional tendrá buenas posibilidades no sólo en la contienda por
	ADJ VERB NP NOM ADV ADV PREP PRON VERB PREP
81	la presidencia , sino en la elección del gobierno capitalino
	ART NOM COM CONJ PREP ART NOM CONT NOM ADJ
91	y , aun , en recuperar la mayoría absoluta en
	CONJ COM NOM COM PREP VERB ART NOM ADJ PREP
101	la cámara de diputados . / en efecto , en
	ART NOM PREP NOM PTO SUS PREP NOM COM PREP
111	su examen anual de la economía mexicana , que hizo
	ADJ NOM ADJ PREP ART NOM ADJ COM PRON VERB
121	público el jueves 29 , la OCDE - el club
	NOM ART NOM NUM COM ART NP GUI ART NOM
131	de los 29 países más ricos del mundo , al
	PREP ART NUM NOM ADJ NOM CONT NOM COM CONT
141	que México ingresó en 1994 - admite que el desempeño
	CONJ NP VERB PREP NUM GUI VERB CONJ ART NOM
151	económico del país fue positivo en los últimos tres años
	ADJ CONT NOM VERB VERB PREP ART NOM ADJ NOM
161	, pero señala una serie de ineficiencias en la política
	COM CONJ VERB ART VERB PREP NOM PREP ART NOM
171	económica - inestabilidad presupuestal , dependencia del petróleo y deficiente
	ADJ GUI NOM ADJ COM NOM CONT NOM CONJ NOM
181	sistema tributario - que impiden sacar al país del subdesarrollo
	ADJ ADJ GUI CONJ VERB VERB CONT NOM CONT NOM
191	y contrarrestar los niveles de pobreza extrema .
	CONJ VERB ART NOM PREP NOM VERB PTO

Figura 3.7: Texto **A** con sus partes del discurso.

Paso	Marcas adjudicadas	Marcas correctas	%
Diccionarios	103	102	99.02
R. morfológicas	17	17	100.00
V. regulares	6	5	83.33
V. irregulares	7	5	71.42
Ambiguas frec.	16	14	87.50
Verbo/nominal	9	7	77.77
Desconocidas	40	34	85.00
Total	**198**	**184**	**92.92**

Figura 3.8: Porcentaje de exactitud en cada paso del etiquetamiento para el texto **A**.

3.3.4 Desempeño

El desempeño final del etiquetador desarrollado se calculó realizando diez experimentos, numerados del cero al nueve, y en cada experimento se hicieron nueve pruebas de etiquetamiento de un texto. Para este fin se dividió el corpus de entrenamiento en diez secciones. En el experimento i se tomó como texto a etiquetar la sección i. La primera prueba del experimento i tomó como corpus de entrenamiento a la sección $(i+1)\bmod 9$. La j-ésima prueba consideró como corpus de entrenamiento al resultante de agregar la sección $(j+1)\bmod 9$ al corpus de la anterior prueba. Tomando cada una de las primeras diez pruebas de cada experimento se obtuvo el máximo, mínimo y promedio de la primera prueba. Lo mismo se hizo para las 8 pruebas restantes de los diez experimentos.

El desempeño del etiquetador se muestra en la siguiente gráfica de barras de error. Esta gráfica expone lo antes explicado sobre el crecimiento del corpus de entrenamiento, con los valores del eje horizontal. La exactitud mínima, máxima y promedio aparecen indicadas con marcas sobre cada barra vertical.

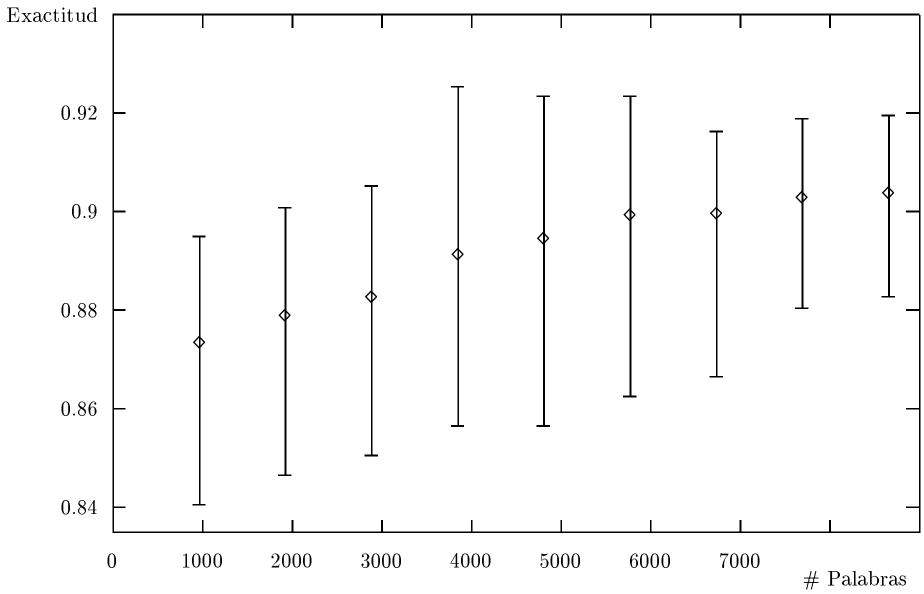

Figura 3.9: Desempeño final del etiquetador.

La gráfica de exactitud del método propuesto, usando δ' se compara con el método IGTree (indicada con δ) en la figura 3.10. Como en la gráfica anterior, también se muestra el crecimiento del corpus en el eje horizontal.

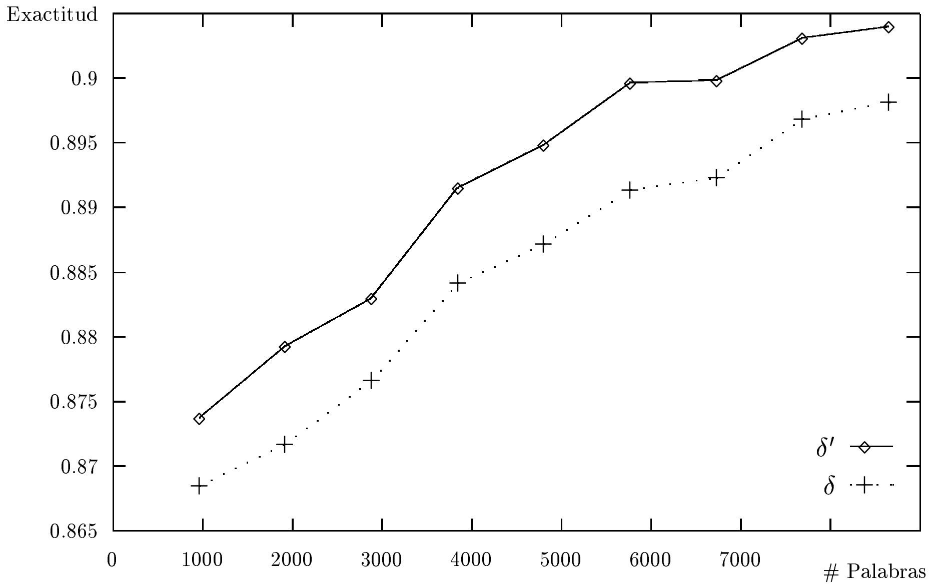

Figura 3.10: Desempeño promedio del etiquetador usando δ y δ'.

3.4 Análisis sintáctico parcial

El problema de *parsing*, en especial para frases nominales, ha sido múltiplemente atacado haciendo uso de una gramática. Lo cual exige un esfuerzo de análisis lingüístico combinado con herramientas que relajen las particularidades o excepciones no contempladas [Voutilainen, Padró, 97], [Chen, Chen, 94]. El objetivo en computación está centrado en la automatización de procesos, y trata de proveer métodos que aligeren los procesos manuales. Particularmente, en las estrategias que se basan en una gramática, el intento de inferir reglas a partir de ejemplos del lenguaje, es poco *escalable* puesto que el esfuerzo invertido en obtener una gramática no es proporcional al desempeño, cuando se aumenta el número de ejemplos; conclusión que se desprende de la complejidad del problema [Lucas, 93]. También se han construido muchos *parsers* con base a métodos cuantitativos y simbólicos [Church, 89], [Grefenstette, 96], para la lengua inglesa. Para los propósitos del presente trabajo es suficiente un método de *análisis sintáctico parcial*, que significa identificar solamente algunos sintagmas.

Esta sección presenta otra aplicación de etiquetamiento. En contraste con los analizadores sintácticos "completos", ha habido tendencia a construir analizadores parciales que sólo identifican algunos sintagmas, e inclusive no proponen estructura de éstos. El que aquí se presenta identifica únicamente los límites de las frases nominales simples, etiquetando para este propósito la frontera derecha de la frase nominal.

Uno de los requisitos para clasificar una frase es el que corresponde a la formación adecuada de los componentes. Como ya se ha planteado, al inicio del capítulo, se espera que la respuesta sea una frase nominal. No se pretende realizar un análisis exhaustivo, pues además de que llevaría una elaboración larga, para el presente propósito es suficiente saber si la respuesta tiene la "forma" de una frase nominal. Nos concentraremos entonces, en el problema de identificar las frases nominales simples[19].

El enfoque seguido para este análisis se apoya en la consideración de que los diferentes tipos de frases se distinguen no sólo por su forma sino además las fronteras que delimitan a las frases son características propias de ellas. Identificaremos a la frase nominal simple mediante el reconocimiento de una secuencia de

[19]También se exploró la identificación de las proposiciones sustantivas y adjetivas, con el mismo método que el empleado en las frases nominales simples, usando como corpus de trabajo parte del material que se presenta en dos de los trabajos realizados sobre el estudio del habla en la Ciudad de México [Arjona-Iglesias, 86], [Arjona-Iglesias, Luna-Traill, 91], [Sámano, 83]. La precisión de la identificación usando los ejemplos antedichos (alrededor de 300 proposiciones) fue de 70%.

palabras que corresponda a la definición que de ésta se hace gramaticalmente, además que, por ocurrir en un contexto determinado, sus fronteras pueden distinguirse de las fronteras de otras formas gramaticales. Se procedió, con el empleo de un corpus, a llevar a cabo el etiquetamiento de los límites de cada frase nominal y la elección de instancias para reunir una colección de entrenamiento que permitiera discernir si cualquier otra secuencia de palabras contenía alguno de los límites posibles de una frase. Es decir, se utilizó el método basado en memoria para clasificar una secuencia de palabras como portadora de una frontera de la frase de interés.

3.4.1 Identificación de frases nominales

Una frase nominal simple puede verse como una secuencia de palabras compuesta por al menos un sustantivo y algunos modificadores que extienden o precisan su significado [Alonso, Enriquez-Ureña, 44]. En la frase nominal se distingue un sustantivo o pronombre que es el núcleo de la frase. El inicio de una frase nominal puede determinarse con facilidad pues puede empezar con un determinante o un sustantivo. No es el caso de la terminación de la frase, en la que hay diferentes maneras de finalizar y continuar con el texto. Consideremos, entonces, para identificar una frase nominal, la frontera derecha de ella. Se eligió, para este propósito, el uso del aprendizaje basado en memoria y se determinó la frontera derecha de una frase nominal, dado un texto etiquetado con sus partes del discurso. Previamente a la aplicación de este método, hubo que formar dos clases de secuencias de palabras: las que contuvieran una frontera derecha de frase nominal en una posición determinada y las que no la contuvieran. Dichas secuencias fueron extraídas de una parte del CEMC. Específicamente, se procedió de la forma que se indica:

1. Se marcaron manualmente las fronteras derechas de frases nominales que ocurrieran en el corpus.

2. Las instancias de la clase de frontera derecha de frase nominal se formaron tomando cinco partes del discurso a la izquierda de ella y cinco partes del discurso a la derecha de la misma. Los decagramas de partes del discurso que no contuvieran una frontera derecha de frase nominal en el centro formaron la clase complementaria.

3. Analizando la curva de ganancia para la colección de entrenamiento compuesta de las dos clases de decagramas, se eligieron los atributos más importantes.

4. La colección de entrenamiento final fue el resultado de proyectar los rasgos correspondientes a los atributos más importantes. En este proceso se controló la ocurrencia de ambigüedades haciendo uso de la frecuencia de los decagramas: Si sucede que una instancia, resultado de la proyección, coincide con otra de clase diferente se mantiene, para esa instancia, la clase de las instancia original que tenga mayor frecuencia.

La figura 3.11 muestra que, de acuerdo con nuestro corpus de trabajo, son seis partes del discurso, tres a cada lado de la frontera derecha de las frases nominales, las que pueden tomarse como rasgos importantes para determinar cuándo una secuencia de partes del discurso contiene una frontera derecha en su centro.

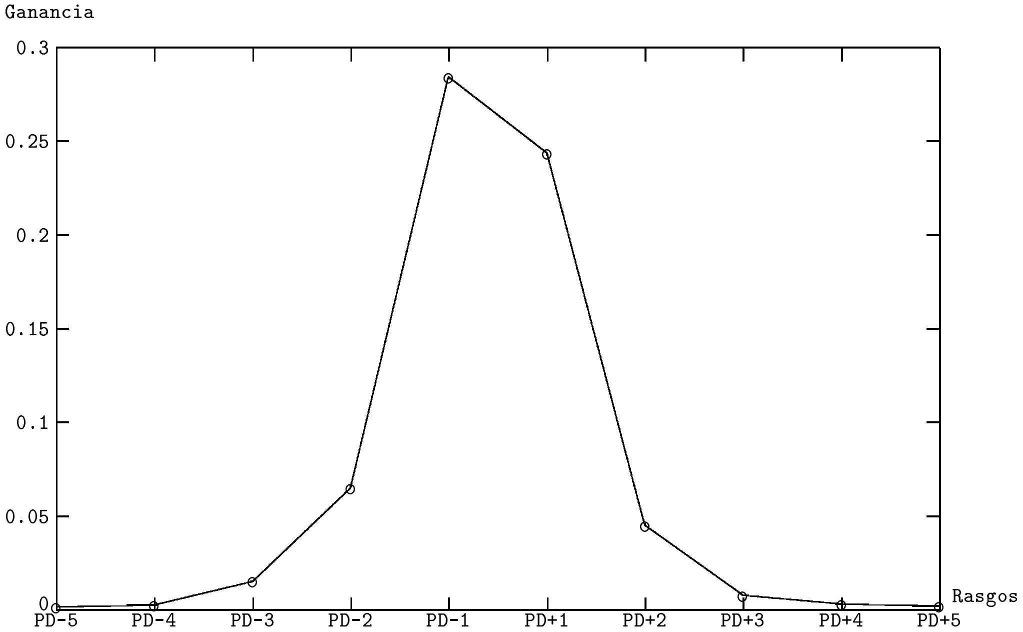

Figura 3.11: Curva de ganancia para decagramas con centro en una frontera derecha de una frase nominal.

3.4.2 Desempeño

Se realizó el experimento de identificación de frases nominales en dos textos, uno del mismo género que el del corpus de trabajo y otro de género diferente. La tabla 3.15 muestra los resultados de esta aplicación del reconocedor de frases nominales. El descenso en el desempeño para texto de género diferente es una muestra de que un reconocedor para texto irrestricto necesita de abundantes muestras para conformar subcolección de entrenamiento, y de esta manera alcance un desempeño al menos,no bajo,como el obtenido para texto del mismo género que también muestra la tabla 3.15.

También se puede observar el desempeño del método, aumentado el tamaño de la colección de entrenamiento sucesivamente. Se procedió dividiendo el corpus de trabajo en diez partes. Cada parte se usó como texto de prueba y el resto como fuente para constituir colecciones de entrenamiento crecientes. En las figuras 3.12 y 3.13 se puede comparar el alcance del método, dentro y fuera del género de trabajo. Observamos una relación de complementariedad entre la evocación y la precisión. El aumento de evocación obedece a que hay instancias que identifican mayor número de marcas, aunque no siempre sean correctas. Dichas instancias provienen de textos con marcas para muchos casos, que pueden no coincidir con los del textos de prueba. Esto reduce la precisión.

Experimento	P	R	F
mismo genero	0.967	0.968	0.967
diferente genero	0.869	0.920	0.889

Tabla 3.15: Precisión y evocación en la identificación de la frontera derecha de frases nominales.

La fig. 3.14 muestra las frases nominales encontradas en el texto de prueba A. En esta prueba encontramos errores debidos al etiquetador de partes del discurso, como el caso del etiquetamiento de la palabra “buenas” que, al ser desconocida, fue etiquetada por el paso correspondiente como `NP`. Esto, a la vez, causó que el identificador de frases nominales adjudicara una frontera inválida, excluyendo a “posibilidades” de la frase nominal en proceso de identificación.

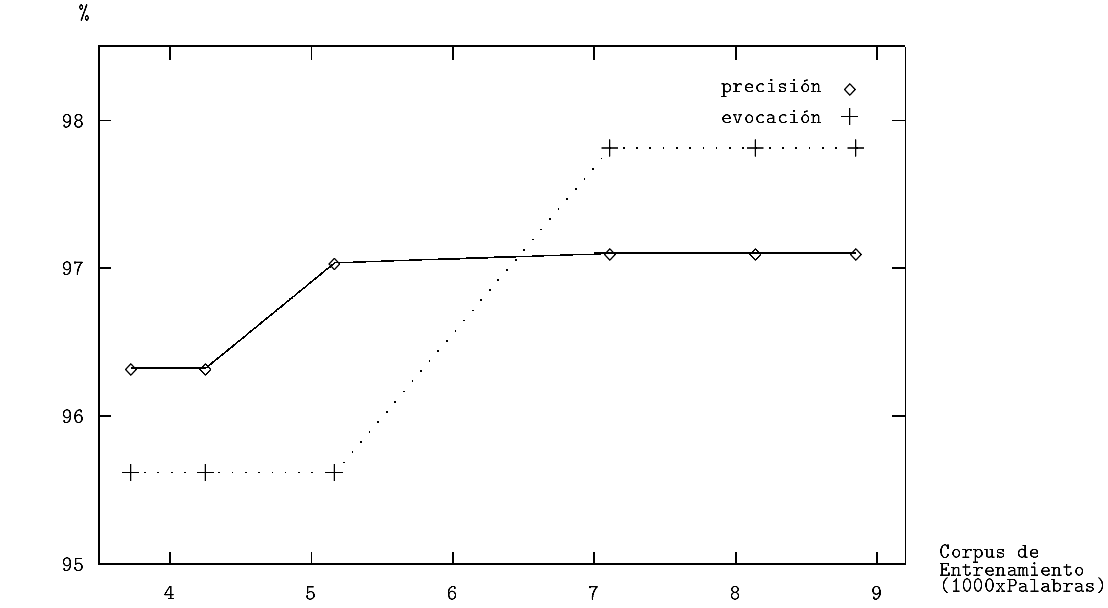

Figura 3.12: Precisión y evocación de la identificación del límite derecho de frase nominal para un texto del mismo género.

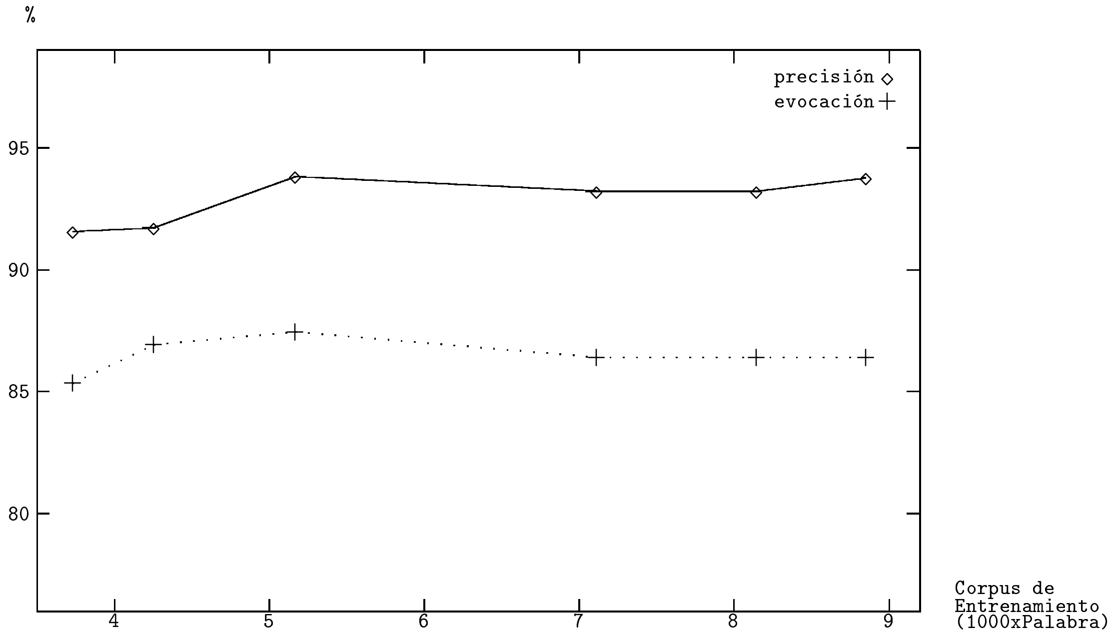

Figura 3.13: Precisión y evocación de la identificación del límite derecho de una frase nominal para un texto de género diferente.

Distancia	Frase nominal
0.00216972	duda
0.0201272	el PRI
0.00216972	el secretario
0.00671217	analistas
0.00671217	inversionistas en el consejo de las américas de Nueva York
0.013415	la reunión de los organismos financieros
0.00216972	ambages
0.00671217	la economía
0	la disciplina
0.0155847	el partido revolucionario institucional
0.0775838	buenas
0	la presidencia
0.00671217	la elección del gobierno capitalino
0.00671217	aun
0.013415	la mayoría absoluta en la cámara de diputados
0.00454245	efecto
0	su examen anual de la economía mexicana
0.00671217	público
0.00671217	
0.023039	la OCDE
0.00671217	países más ricos del mundo
0.0201272	México
0.00671217	el desempeño económico del país
0.00454245	los últimos tres años
0.00216972	ineficiencias en la política económica
0.0179574	inestabilidad presupuestal
0	dependencia del petróleo
0.0155847	deficiente sistema tributario
0.013415	al país del subdesarrollo
0.0155847	los niveles de pobreza

Figura 3.14: Distancia más corta a una instancia de clase frontera derecha de frase nominal y la frase obtenida.

La identificación de frases nominales se ha realizado de diferentes maneras, obteniendo un desempeño favorable para muchas aplicaciones. Aunque no podemos comparar los resultados aquí obtenidos con los de otras lenguas, sí lo podemos ubicar en una posición favorable. Una solución muy parecida fue desarrollada por J. Veenstra quien usó parte del corpus *Wall Street Journal* [Veenstra, 98].

* * *

Hemos observado que al conjuntar el método basado en memoria con ideas básicas de la lengua española obtuvimos un sistema de etiquetamiento de partes del discurso con un buen desempeño respecto al tamaño de su colección de entrenamiento.

También hemos observado que el último paso requirió de mayor precisión para obtener un desempeño adecuado. La atención a este problema se hizo modificando la métrica empleada en el método basado en memoria, así, se tomó en cuenta el problema de valores desconocidos en las instancias de entrenamiento. La modificación mejoró el desempeño del método en una fracción que no puede aún ser considerada estadísticamente significativa [Jiménez-Salazar, Morales-Luna, 2000].

W. Daelemans et.al. [Daelemans, van-den-Bosch, et. al. 98] aplican el método IGTree al etiquetamiento con partes del discurso a textos del español, usando un corpus de entrenamiento de 711,000 palabras, y obtienen una precisión de 97.8%. Por otro lado, Màrquez y Rodríguez, usan un método basado en árboles de decisión y los resultados son comparables al método IGTree [Màrquez, Rodríquez, 98]. Además, muestran el desempeño de su etiquetador para palabras desconocidas, y comparan sus resultados con los del método IGTree usando un corpus creciente que va desde 2,000 hasta 50,000 palabras. Obtienen así que: para un corpus de entrenamiento de 10,000 palabras, la precisión del método IGTree es de 79.18% y para el método basado en árbol de decisión 83.30%.

El método IGTree con la modificación explicada, que estima la marca ausente dentro del proceso de palabras desconocidas, arroja una precisión mayor al 81%. No es posible, después de todo, comparar los resultados obtenidos con los que presenta el método de Màrquez y Rodríguez, por varias razones. Primero, no se tiene el mismo corpus. Segundo, los conjuntos de partes del discurso utilizados en cada caso son diferentes y, tercero, no se tiene programado el método de árboles de decisión que ellos usan para realizar pruebas con el corpus disponible.

La identificación de frases nominales también se apoyó en el método basado en memoria y alcanzó un desempeño bastante alto, respecto a la colección de entrenamiento usada [Jiménez-Salazar, Morales-Luna, 98].

Aunque la exactitud obtenida por el etiquetador puede calificarse como no baja, es necesario aumentar el corpus que proporciona la colección de entrenamiento, pues con ello se logrará no solamente un mejor desempeño, sino, además, ofrecer etiquetamiento a texto de cualquier género, *texto irrestricto*.

Los sistemas desarrollados en este capítulo son de aplicación amplia dentro del PLN. Justamente, nuestro interés en abordar el problema del sentido de una frase en un dominio requiere de ellos.

Capítulo 4

Grado de pertenencia a un dominio

En este capítulo, expondremos primero una solución al problema de determinar cuándo una palabra puede ser concebida como parte del vocabulario de un sublenguaje. Puesto que los léxicos son conjuntos no completamente determinados (*conjuntos abiertos*), es necesario plantear el problema anterior como el cálculo del grado de pertenencia de una palabra a un dominio. También presentaremos una estructura conceptual que permitirá el análisis sobre la elección de una representación para las palabras, así como la representación del dominio mismo. Finalmente, haremos una propuesta para calcular el grado de pertenencia de una frase a un dominio. Antes motivaremos, muy brevemente, el propósito de resolver los anteriores problemas.

En las aplicaciones de uso frecuente del PLN, los procedimientos basan, comúnmente, la solución del problema de similitud –entre documentos, y entre una consulta y una clase de documentos- en un análisis morfológico, etiquetamiento de texto, y análisis sintáctico. Además, estas aplicaciones llegan a emplear algunos elementos semánticos como los roles que desempeñan los componentes de las frases que se analizan, aunque todavía no se ha extendido el uso de recursos semánticos "potentes" como las jerarquías semánticas.

Los sistemas de recuperación de información, continúan expresando interés por el empleo de palabras normalizadas que ocurren con cierta frecuencia en un texto para la representación del mismo: los *índices*. La consulta de un usuario, se atiende, a menudo, con la localización de los documentos cuya representación maximicen la similitud con la representación de la consulta. Herederos de estas técnicas son los sistemas de filtración, categorización de textos, elaboración de resúmenes, y extracción de información [Guthrie, Pustejovsky, et. al. 96].

En última instancia, la similitud se apoya en las palabras que componen al documento pero difícilmente descansan en representaciones conceptualmente ricas, pues se ha visto que los sistemas, que incluyen representaciones semánticas elaboradas, resultan costosos al momento de su ejecución en línea [Smeaton, 97].

Los sistemas de recuperación de información han carecido de un tratamiento regular de las frases nominales [Strzalkowski, 95]. Es importante, entonces, explorar alternativas que enriquezcan la representación de los términos y, por tanto, también analizar el proceso de similitud que está involucrado.

4.1 Pertenencia de una palabra a un dominio

Una solución para un problema parecido al de pertenencia de una palabra a un dominio la presentaron S. Haas y S. He [Haas, He, 93], mediante la determinación del vocabulario de un dominio. Es indispensable para esta propuesta contar con un diccionario (MRD) que contenga el campo que clasifica cada entrada del diccionario. Con un sentido más empírico se encuentra el trabajo reportado en [Grefenstette, 93] que emplea corpora no menores a un millón de palabras para construir *tesoros* o jerarquías del léxico de un dominio, y también una *función de similitud* entre palabras[1]. El tesoro de un dominio es útil para establecer la posible

[1]Estaremos refiriéndonos a palabras similares o con algún tipo de relación [Zgusta, 71]. G. Grefenstette [Grefenstette, 94] se refiere a las palabras *afines*.

relación entre palabras y el dominio. Por ejemplo, en [Morris, Hirst, 91] se emplea el tesoro *Roget* para definir *cadenas léxicas* que ayudan a establecer la estructura de un discurso. Muy cercana a la construcción y uso de tesoros se encuentra la línea que investiga la *representación de conceptos*. Entre los diferentes enfoques para la representación de conceptos destaca la *teoría de conceptos formales* [Davey, Priestley, 90]. Ésta se distingue por su base teórica, los nexos que se extienden hacia otras áreas como la lógica matemática, y la utilidad que ha mostrado, por ejemplo en la recuperación de información. Aunque hay diversos enfoques para resolver el problema de *similitud entre palabras* [Kozima, 96] , es necesario señalar que los MRD tienen utilidad en este problema [Amsler, 89], [Krovetz, 89], [Krovetz, 93], llegando a su máxima aplicación en la *desambiguación del sentido de una palabra* [Agirre, Rigau, 96], [Wilks, Stevenson, 97], [Ide, Véronis, 98]. El problema de determinar el grado de pertenencia de una palabra a un dominio se abordó aquí tomando ideas de la desambiguación del sentido de una palabra.

4.1.1 Desambiguación del sentido de una palabra

El problema consiste en asociarle a una palabra, de un texto, un sentido o significado, que debe ser elegido de entre varios sentidos potenciales de la palabra. Este proceso involucra necesariamente dos subproblemas [Ide, Véronis, 98]:

1. La determinación de todos los sentidos para cada palabra considerada en el texto y
2. Un mecanismo para asignar a cada ocurrencia de las palabras consideradas el sentido adecuado.

Ambos subproblemas hacen uso extensivo de recursos lingüísticos. El primer subproblema se ha apoyado en: (1) listas de sentidos para cada palabra o (2) grupos de rasgos, categorías o palabras asociadas. Y el segundo subproblema ha hecho uso del contexto y fuentes de información externas al texto de la aplicación. El *contexto de una palabra* a ser desambiguada, incluye información del texto en el que ocurre la palabra, identificando palabras vecinas, estructuras sintácticas, y cualquier otra información contenida en el texto, además de información extralingüística, referente, por ejemplo, a la situación. Por otro lado, las *fuentes de información externa* se refieren a diccionarios, enciclopedias y, en general, cualquier recurso que presente el uso de las palabras con sus sentidos. Los enfoques para resolver este problema han sido variados pero puede decirse que, en todos los casos, se trata de identificar el sentido de una palabra "empatando" su contexto con una instancia de la fuente de información y, de esta forma se le asigna el sentido correspondiente. Una de las soluciones para la desambiguación de sentido [Ide, Véronis, 98] propone el uso de un diccionario en donde cada palabra aparece con sus diferentes sentidos representados como "firmas". Una *firma* es el conjunto de palabras que componen a la definición en una de sus acepciones. La desambiguación se lleva a cabo eligiendo aquél sentido cuya firma intersectada con los vecinos, de la palabra por desambiguar, tenga el mayor número de elementos. Este método logra una desambiguación correcta en un nivel alrededor del 60%. Posteriormente, en esta misma línea de trabajo, se propuso el uso de una métrica entre palabras para conocer su grado de relación. La métrica considerada está basada en el producto interno de dos vectores. Cada uno de los componentes del vector $\overrightarrow{W}$ representa una palabra del vocabulario. La representación de una palabra a desambiguar, w, es también un vector $\overrightarrow{W}$ definido con base al contexto en el que ocurre w. Si el contexto de w contiene las palabras $p_1, p_2 \dots p_k$, y éstas corresponden a las componentes $i_1, i_2 \dots i_k$ de $\overrightarrow{W}$, entonces $\overrightarrow{W}$ tendrá cero en todas sus componentes, excepto en las posiciones $i_j, j \leq k$, donde tendrá el valor $\mathrm{fr}(p_j)$. De esta manera es que podemos referirnos, por ejemplo, al *ángulo* que forman dos palabras. Para la desambiguación de w, en un contexto que contenga a $p_1, p_2 \dots p_k$, primero se contempla la formación del vector $\overrightarrow{W}$, en segundo lugar, cada una de las m firmas de w, en el diccionario, constituye un contexto y, por tanto, un vector $\overrightarrow{W}_i, i \leq m$. El sentido de w se elige de acuerdo con la firma cuyo vector maximice el producto $\overrightarrow{W} \cdot \overrightarrow{W_i}$. Este método ha logrado una eficiencia del 90% en la desambiguación de palabras homógrafas.

Al tratar con un dominio lo estaremos haciendo con un universo de objetos y relaciones entre estos objetos. A los objetos del dominio podemos referirnos mediante los elementos de un vocabulario. Una forma de representar un dominio es a través de la representación de cada uno de los términos del vocabulario con los sentidos de esos términos.

Claramente toda palabra que pertenezca al vocabulario pertenecerá al dominio, pero siempre habrá algunas palabras que no ocurran en el vocabulario y sí pertenezcan al dominio, o tengan algún tipo de relación con él. El problema de pertenencia de una palabra a un dominio trata de averiguar si una palabra, aún cuando no pertenezca al vocabulario, pertenezca al dominio en algún grado.

Este problema se traduce en conocer si la palabra en cuestión tiene algún sentido en el dominio.

El dominio podemos entonces verlo como un conjunto de palabras, representada cada una por el sentido que tiene en el dominio. Hay pues una relación entre el problema de la desambiguación del sentido de una palabra y el sentido de una palabra en un dominio. En el primer caso se busca el sentido de una palabra, en el segundo, se trata de conocer si la palabra tiene un sentido específico.

En nuestro problema, además de encontrar el sentido de una palabra en un dominio, la solución requiere de la determinación del "grado" de pertenencia de la palabra al dominio, por supuesto, a partir del sentido que se ha dado a la palabra en el dominio.

El punto de vista anteriormente expuesto plantea la necesidad de:

- Un conjunto de palabras que representen al dominio de trabajo,
- El empleo de un diccionario que proporcione información sobre los sentidos de las palabras y,
- Un mecanismo que permita determinar si los sentidos de una palabra mantienen un grado de relación "suficiente" con el sentido que tienen las palabras del dominio de trabajo, para validarla como perteneciente al dominio.

El último punto se ha replanteado de la siguiente forma. Consideremos al conjunto de palabras que representan al dominio como un cúmulo para el cual podemos calcular la distancia entre cada una de las representaciones de sus miembros; esta distancia tiene una correspondencia con la métrica empleada en el problema de desambiguación antes descrito. Es posible, entonces, determinar el diámetro del cúmulo, y usarlo como un valor de referencia en el momento de comparar la proximidad de otra palabra hacia el cúmulo, para saber si es parte de él o qué tan cerca está de él. Antes de abordar esta forma de determinación, veremos una alternativa que ocupa la pertenencia rugosa de un elemento a un conjunto, en el marco de una estructura conceptual.

4.2 Contextos y relaciones

Dado un diccionario es posible atribuirle un contexto a una palabra, reuniendo todos los "usos" que se tengan de ella en el diccionario. Un *uso* de una palabra, en un diccionario, es la definición de otra palabra en la cual se emplea la primera. Un *diccionario* está formado por *entradas*, y cada entrada tiene dos partes: el *vocablo*, que es la palabra a definir y la *definición*. El *diccionario inverso* de un diccionario D, se construye definiendo el vocablo de cada entrada k como aquél que ocurre en la definición de un vocablo que está en D, y la definición de k se compondrá por todas las palabras que usan, en D, al vocablo de k. El *contexto* que usaremos para una palabra x se compondrá de las definiciones, en D, de las palabras que ocurren en la definición de x en el diccionario inverso.

Sea $\mathcal{V}$ el vocabulario de una lengua, las *expresiones*, E, con base en $\mathcal{V}$ es el conjunto de secuencias formadas con elementos de $\mathcal{V}$ y que tienen sentido para una comunidad lingüística, $E \subset \mathcal{V}^+$. Representemos una expresión de E con $e = (v_1, v_2, \ldots, v_k)$. Un *diccionario* D es una función parcial que asigna a una pareja, formada por una palabra y un número natural, una expresión, i.e. $D : \mathcal{V} \times \mathbb{N} \to E$. A los componentes de la terna $(x, y, z) \in \mathcal{V} \times \mathbb{N} \times E$, tal que $(x, y) \mapsto z = D(x, y)$ los llamaremos *vocablo*, *acepción* y *definición*, respectivamente. Sea D un diccionario y x un vocablo, la función parcial I_D, *diccionario inverso*, asigna a ciertas palabras el conjunto de referencias hacia x que ocurren en D. Esto es, $I_D : \mathcal{V} \to 2^{\mathcal{V} \times \mathbb{N}}$. Mediante nacep$(x)$, haremos referencia al número de acepciones de un vocablo x, que es el máximo $i \in \mathbb{N}$ para el cual la función $D(x, i)$ está definida.

Es importante explotar las relaciones que hay entre los elementos de un conjunto, pues de esta manera se encuentran nuevas relaciones útiles en diferentes situaciones. De interés particular, en nuestro caso, son las relaciones de *equivalencia*, *orden*, y *tolerancia*. Cada una de estas relaciones tiene características especiales, y facilitan el trabajo con las propiedades del conjunto de objetos que se maneja. Las relaciones de orden se identifican en el PLN con las relaciones de hiponimia e hiperonimia. Las relaciones equivalencia con la sinonimia y las de tolerancia con la afinidad u otros tipos de relaciones semánticas entre palabras.

Sea $x \in \mathcal{V}$, tal que $I_D(x)$ está definida. El valor de la función $D' : \mathcal{V}^2 \to \mathcal{V}^+$, representa el *uso* de una palabra x por otra y: $(x,y) \mapsto D'(x,y) = \{e | x \in e = D(y,i), \text{ para algún } i \leq \text{nacep}(y)\}$. El *contexto* de $x \in \mathcal{V}$ dentro de D, está formado por todos los usos de x en D y se define como:

$$\mathcal{C}(x) = \bigcup_{y \in I_D(x)} D'(x,y). \tag{4.1}$$

Sea P un conjunto ordenado y S un subconjunto de P, decimos que un elemento $x \in P$ es *cota superior* de S si $y \leq x, \forall y \in S$, similarmente, un elemento $x \in P$ es *cota inferior* de S si $x \leq y, \forall y \in S$. El conjunto de todas las cotas superiores de $S \subset P$ lo denotamos con S^u, esto es, $S^u = \{x \in P | y \leq x, \forall y \in S\}$. Análogamente, $S^\ell = \{x \in P | x \leq y, \forall y \in S\}$, es el conjunto de cotas inferiores. Para $S \subset P$, también definimos, si existe, la *cota mínima superior*, o $\mathbf{sup}(S)$, como la cota inferior de S^u y la *cota máxima inferior* o $\mathbf{inf}(S)$, como la cota superior de S^ℓ. Particularmente, para dos elementos $x, y \in P$, escribimos $\mathbf{sup}(x,y)$ mediante $x \vee y$, e $\mathbf{inf}(x,y)$ como $x \wedge y$.

Si P es un conjunto ordenado en el que $x \wedge y$ y $x \vee y$ existen para toda pareja $x, y \in P$, llamamos a P *retícula* (o *lattice*). Una retícula P es *completa* si para cualquier subconjunto $S \subset P$, existen $\bigwedge S$ y $\bigvee S$. Consideramos, así a una retícula como una estructura con un *orden*, $\langle P, \leq \rangle$ o una estructura con dos operaciones, *conjunción* y *disyunción*: $\langle P, \wedge, \vee \rangle$.

Una relación de *tolerancia* R en un conjunto G satisface reflexividad y simetría, esto es, para todo elemento $x \in G$, xRx, y para toda pareja $x, y \in G$ si xRy entonces yRx. Hay una exigencia más para las relaciones de *equivalencia* que es la transitividad: para toda terna $x, y, z \in G$ si xRy y yRz entonces xRz.

4.3 Estructura conceptual

Una teoría muy elegante que ha sido empleada en el ámbito computacional para representar conceptos ha sido la *Teoría de Conceptos Formales* debida a Wille[2]. Con base a esta teoría se han construido, por ejemplo sistemas de recuperación de información [Carpineto, Romano, 94], [Pedersen, 93], y de ayuda a la traducción automática [Sertula, Kourie, 98]. En términos generales, se trata de establecer relaciones de generalización y particularización entre conceptos. A cada objeto relacionado con un concepto se asocia un conjunto de rasgos. Esta representación permite conocer, por ejemplo, si un concepto es más general que otro: cuando los objetos que describen al segundo están incluidos en la descripción del primero. Notemos que si dos conceptos, A_1 y A_2, se generalizan por otro, A_3, entonces su grado de *afinidad* puede ponerse en función de la "altura" del subárbol con raíz en A_3 y terminales A_1 y A_2. Resulta, entonces, un buen punto de partida la exploración de estas ideas para abordar la representación de palabras.

Enseguida haremos una presentación de la teoría de conceptos formales basada en B. Davey y H. Priestley [Davey, Priestley, 90], posteriormente, daremos las ideas en las que se basa Para finalizar, propongo el empleo de una *función de pertenencia rugosa* y para conocer de esta forma nos aproximaremos a la solución de la pertenencia de una palabra a un dominio.

[2]En el libro de B. Davey y H. Priestley [Davey, Priestley, 90] se hace referencia al trabajo inicial de Wille: Rival, I.: Ordered sets, *NATO ASI*, (Ed.), V. 83, Reidel Dordretch, 1982.

4.3.1 Teoría de conceptos formales

Sea G un conjunto de objetos y M un conjunto de n rasgos ambos finitos, y sea ψ la correspondencia que asocia a cada objeto $g \in G$ un conjunto de rasgos, $\{m_1 \dots m_k\} \subset M$. Representemos con $\Psi(g)$ el conjunto $\{m \in M | \psi(g) = m\}$. A la terna $\langle G, M, \psi \rangle$ la llamamos *contexto*. Un *concepto* en el contexto $\langle G, M, \psi \rangle$, es una pareja (A, B), compuesta del *extento* A, y del *intento* B, tal que $A \subset G$, $B \subset M$ y, además,

$$\bigcap_{a \in A} \Psi(a) = B. \tag{4.2}$$

Para dos conceptos, (A_1, B_1) y (A_2, B_2), definidos en $\langle G, M, \psi \rangle$, decimos que (A_1, B_1) es *más particular* que (A_2, B_2), $(A_1, B_1) \leq (A_2, B_2)$, o bien, que (A_2, B_2) es *más general* que (A_1, B_1), $(A_2, B_2) \geq (A_1, B_1)$, si se cumple que $A_1 \subset A_2$. Notemos también que si $A_1 \subset A_2$ entonces $B_2 \subset B_1$, ya que A_1 está contenido en A_2 se tiene que $\bigcap_{a \in A_2} \Psi(a) \subset \bigcap_{a \in A_1} \Psi(a)$. En otras palabras, $(A_1, B_1) \leq (A_2, B_2)$ sii $B_2 \subset B_1$. El conjunto de todos los conceptos definidos en un contexto $\langle G, M, \psi \rangle$, se denota $\mathfrak{B}\langle G, M, \leq \rangle$.

Conviene reconocer la relación R_Ψ, que Ψ induce en el conjunto $G \times M$, como el establecimiento de $gR_\Psi m$ sii $m \in \Psi(g)$. De esta forma tendremos que para un concepto (A, B), el conjunto asociado a un extento A: $\{m \in M | gR_\Psi m, \forall g \in A\}$ es igual a B. Denotamos por A' al anterior conjunto y de manera semejante expresamos una operación sobre los subconjuntos de M: $B' = \{g \in G | gR_\Psi m, \forall m \in B\}$ para lo cual se satisface $B' = A$, siempre que $(A, B) \in \mathfrak{B}$. En general, para cualquier $A \subset G$ tendremos que $A \subset A''$, puesto que A' es el conjunto de rasgos comunes a todos los objetos de A, y A'' incluirá a todos los objetos de G con rasgos A', así A'' incluye, en particular, a los objetos de A. De igual forma puede verse que $B'' \subset B, \forall B \subset M$. Observemos, además, que se cumplen las siguientes igualdades para un contexto $\langle G, M, \psi \rangle$: Si $A \subset G$ y $B \subset M$, tenemos que $A = A'''$, $B = B'''$, y si $\{A_j\}_{j \in J}$ y $\{B_j\}_{j \in J}$ son dos familias de subconjuntos de G y M entonces $(\bigcap_{j \in J} A_j)' = \bigcup_{j \in J} A'_j$ y $(\bigcap_{j \in J} B_j)' = \bigcup_{j \in J} B'_j$. Lo anterior conduce a enunciar el siguiente teorema:

Teorema 1 *Sea $\langle G, M, \psi \rangle$ un contexto, entonces $\mathfrak{B}\langle G, M, \leq \rangle$ es una retícula completa en la cual las operaciones* **sup** *y* **inf** *están dadas por:*

$$\bigvee_{j \in J} (A_j, B_j) = ((\bigcup_{j \in J} A_j)'', \bigcap_{j \in J} B_j),$$

$$\bigwedge_{j \in J} (A_j, B_j) = (\bigcap_{j \in J} A_j, (\bigcup_{j \in J} B_j)'').$$

Podemos construir una retícula, por ejemplo, a partir de un diccionario: tomando a los objetos como el vocablo de las entradas y los rasgos como las palabras que constituyen la definición. Sin embargo, la condición $\bigcap \Psi_{a \in A}(a) = B$, es restrictiva pues dado un concepto (A, B), es posible que exista un objeto que comparta rasgos con varios objetos de A pero estos rasgos no pertenezcan a B.

La objeción anterior se atiende en cierta medida por una representación menos restringida. Optaremos por una representación más simple que la antes presentada, pero suficiente para responder al problema de la determinación de la pertenencia de una palabra a un dominio.

4.3.2 Pertenencia rugosa de una palabra a un dominio

La construcción de una retícula de conceptos exige la determinación de una correspondencia. Desafortunadamente, no todas las combinaciones de rasgos pueden verse como intentos de algún concepto. Conviene analizar la aproximación a un concepto partiendo de un conjunto de rasgos. Enseguida sólo mencionaremos las bases en las que se apoya la propuesta de los *conceptos rugosos*, de modo suficiente para hacer uso de ellos en nuestra aplicación.

Sea G un conjunto de objetos y M un conjunto de rasgos. Retomamos la correspondencia ψ y la función Ψ presentadas en la sección anterior. Definamos ahora la relación $R_\Psi \subset G \times G$, como:

$$xR_\Psi y \text{ sii } \Psi(x) \cap \Psi(y) \neq \emptyset. \tag{4.3}$$

Si la correspondencia ψ está definida en todas partes, entonces tendremos claramente que R_Ψ es reflexiva y simétrica. En el caso especial de que ψ sea una función entonces R_Ψ es una *relación de equivalencia.* Consideramos mayor aplicabilidad al caso en que ψ sea una correspondencia definida en todas partes y, por tanto, R_Ψ será una relación de tolerancia.

Cualquier conjunto $C \subset G$ es una *preclase* inducida por la relación R_Ψ en G, sii toda pareja $x, y \in C$ es tolerante, i.e. $xR_\Psi y$. La *preclase maximal*, K, que contiene a la preclase C, es tal que cualquier otra preclase $D \supset C$ entonces $K \supset D$. En tanto no haya confusión llamaremos simplemente *clase* a una preclase maximal. Una familia de conjuntos $\{S_j\}_{j\in J}$ es una *cobertura* de G sii $G = \bigcup_{j\in J} S_j$. Como G es finito, en forma simple y constructiva podemos completar cualquier preclase C en una clase: si existe un elemento $x \notin C$ tal que $xR_\Psi y, \forall y \in C$, entonces podemos agotar éstos convirtiendo C en $C \cup \{x\}$, de manera que siempre una preclase estará incluida en una clase. Con las ideas anteriores puede demostrarse que la familia de clases $\{C_j\}_{j\in J}$ formada con base en una relación de tolerancia es una cobertura de G.

Sea R una relación de tolerancia en G, para $x \in G$ definamos $C_R(x) = \{y \in G | xRy\}$ y sea $V \subset G$. Los subconjuntos:

$$\begin{aligned} \underline{R}V &= \{x \in G | C_R(x) \subset V\} \text{ y} \\ \overline{R}V &= \{x \in G | C_R(x) \cap V \neq \emptyset\}, \end{aligned}$$

son llamados la *aproximación inferior* y la *aproximación superior* de V. La *aproximación rugosa* de $V \subset G$ es la pareja $(\underline{R}V, \overline{R}V)$.

Z. Pawlak [Pawlak, 82] definió el concepto de *conjunto rugoso* como la aproximación rugosa de un conjunto, tomando a R como relación de equivalencia. Posteriormente, se amplió el concepto de conjunto rugoso imponiendo menos restricciones a ψ.

Es importante mencionar que ha habido numerosas aplicaciones de conjuntos rugosos, en muchas áreas ligadas a la representación de información o conocimientos [Ziarko, 91], y [Polkowski, Skowron, 98] Particularmente, en [Iwiński, 89] se hace referencia a los conceptos formales como *conceptos exactos* en la terminología de conceptos rugosos (cuando un concepto (A, B), $B' \neq A$). Experimentaremos entonces con los conjuntos rugosos en el ámbito de la estructura conceptual, particularmente con la noción de *función de pertenencia rugosa* de un elemento a un conjunto.

Dado un conjunto $V \subset \mathcal{V}$ y una relación de tolerancia R, en concordancia con 4.3, y tomando a $\Psi(x)$ como $\mathcal{C}(x)$ (ec. 4.1) puede definirse una función, $\mu_V : \mathcal{V} \to [0, 1]$, que mida el grado de pertenencia de un elemento u al conjunto V, del siguiente modo:

$$\mu_V(u) = \frac{\#(C_R(u) \cap V)}{\# C_R(u)}, \tag{4.4}$$

donde $C_R(u) = \{x \in G | \Psi(u) \cap \Psi(x) \neq \emptyset\}$. Para evaluar la utilidad de la ec. 4.4 es necesario contar con una representación del dominio, esto es el conjunto V, lo cual es tema de la sección 4.4.

4.4 Representación del dominio

La forma en que se pretende determinar si una palabra pertenece a un dominio considera una representación aproximada del dominio. y por tanto La representación que se propondrá usa un conjunto básico de palabras del dominio de trabajo, además de acompañarse de mecanismos que permitan decidir si una palabra es miembro de este dominio en cierto grado. Con el fin de realizar pruebas en una aplicación del área de la salud, el conjunto básico de palabras hubo de conformarse extrayendo palabras de fuentes de información, tales como textos del área de la salud. Se trata de folletos informativos sobre salud publicados por compañías farmacéuticas y la Secretaría de Salud.

Para construir una aproximación del dominio se eligió manualmente un conjunto inicial de palabras tomadas de los folletos sobre salud. El paso siguiente enriquece este conjunto para obtener la aproximación del dominio. Este conjunto es esencial pues con él se decidirá el grado de pertenencia de una palabra al dominio de trabajo. La construcción antedicha se presenta en el resto de esta sección.

4.4.1 Inducción léxica

En los dominios científico y técnico se hace referencia a conceptos mediante términos formados por varias palabras. La *terminología* de un dominio es vista como la reunión de palabras o frases que refieren a los objetos de estudio en el dominio en cuestión. Nuestra meta es, pues, la identificación de este tipo de términos para constituir la terminología del dominio de trabajo. Con la técnica que se presenta en esta sección, se pretende extraer la terminología de un determinado dominio, a partir de una fuente textual.

La *inducción léxica* consiste en determinar los miembros de un conjunto, L, que aproxima a la terminología, a partir de un corpus. El primer paso, del procedimiento de inducción léxica, es formar un conjunto básico inicial, B_0, que se extrae de forma manual del corpus. Se toman, normalmente, las palabras que ocurren en los encabezados de cualquier nivel del corpus. El segundo paso es medir la *información mutua* de los bigramas que ocurren en el corpus. Aquellos bigramas que tengan una información mutua alta y que alguno de sus componentes esté en B_0, se incluyen como miembros del conjunto L_0. En el tercer paso, se enriquece el conjunto base B_{i+1} con todas las palabras que ocurren en los términos de L_i y se repite el proceso: mientras se cuente con un conjunto base nuevo, a partir de él se incrementa L_{i+1} con bigramas que tengan información mutua alta y que alguno de sus componentes ocurra en B_{i+1}.

4.4.2 Información mutua

La *información mutua* de bigramas ha sido una medida para extraer información a partir de textos [Church, Hanks, 90]. Con esta medida se trata de averiguar si los elementos que conforman un patrón que ocurre en un texto mantienen una asociación significativa o si la ocurrencia del patrón con esos componentes es sólo incidental. Analizaremos el caso de patrones de dos palabras. Dado un texto, T, y dos palabras de él, x y y, para saber si existe una verdadera asociación entre x y y, comparemos la probabilidad de observar a x y y contiguas en el texto, $\Pr(x, y)$, con la probabilidad de observarlas independientes, $\Pr(x)$ y $\Pr(y)$. Esperamos que $\Pr(x, y)$ sea significativamente mayor que $\Pr(x)\Pr(y)$ si es el caso que la ocurrencia del bigrama xy no es casual. Mientras que si $\Pr(x, y)$ es mucho más pequeña que $\Pr(x)\Pr(y)$, podemos decir que su ocurrencia contigua es incidental. Particularmente, no es posible afirmar algo sobre xy, cuando $\Pr(x, y) \approx \Pr(x)\Pr(y)$. Estas ideas conducen a definir la *información mutua* entre x y y como:

$$\mathrm{IM}(xy) = \log_2 \frac{\Pr(x, y)}{\Pr(x)\Pr(y)}$$

Cuando $\mathrm{IM}(xy) > \lambda$, $\lambda > 0$, podemos afirmar que el bigrama xy representa un término compuesto, mientras que cuando $\mathrm{IM}(xy) < \lambda$, xy no representa un término compuesto.

4.4.3 Método de inducción

Ya se ha mencionado que el procedimiento para encontrar los nuevos términos del conjunto L es partir de B y realizar un enriquecimiento con base a la información mutua de los bigramas del corpus. Formalmente, sea T un texto, sus bigramas $T_{\langle 2\rangle}$ y B_0 el conjunto base inicial. El conjunto inicial de términos, L_0, es:

$$L_0 = \{xy \in T_{\langle 2\rangle} | \mathrm{IM}(xy) > \lambda \wedge (x \in B_0 \vee y \in B_0)\},$$

recursivamente, B_i, para $i > 0$, se obtiene como

$$B_i = B_{i-1} \cup \{z | \exists xy \in L_{i-1} \wedge (z = x \vee z = y)\}$$

y los L_i se determinan de igual manera que L_0, usando ahora B_i:

$$L_i = \{xy \in T_{\langle 2 \rangle} | \mathrm{IM}(xy) > \lambda \wedge (x \in B_i \vee y \in B_i)\}.$$

A los conjuntos finales L_i y B_i los llamaremos simplemente L y B.

C. Gierl y D. Frost [Gierl, Frost, 92] presentan sus resultados basados en este método usando como umbral $\lambda = 7$. Con un corpus de 153,065 palabras constituyeron un conjunto base de 2,643 palabras y obtuvieron 562 términos al inicio. Al final de diez ciclos los términos encontrados fueron 2,721.

4.4.4 Representación aproximada del dominio

El corpus, de esta aplicación, se formó por un diccionario de salud[3] y el glosario de un libro de higiene de nivel medio superior[4]. El glosario que resultó de esta fusión es un texto, al que llamaremos V, compuesto de 2,130 entradas en las se que definen términos de salud. El total de signos es de 20,322, y su vocabulario, incluyendo inflexiones y puntuación, consta de 4,597 elementos.

Es importante señalar que se centró el interés en obtener un conjunto base confiable, B, más que en el conjunto aproximado de la terminología, L, pues nuestro interés es determinar si una palabra pertenece en algún grado al dominio de la salud, es decir, si tiene sentido su uso en este dominio.

Para crear B_0 se eligieron, de V, las palabras con rango menor a 40, eliminando las de mayor frecuencia en cualquier texto del español de México[5]. La tabla 4.1 expone las palabras seleccionadas, e incluye datos sobre el rango y la frecuencia de las palabras para dar una idea del uso que se hace de éstas en el corpus considerado.

Palabra	Rango	Frecuencia
inflamación	20	117
enfermedad	21	91
sustancia	22	88
piel	23	76
sangre	24	71
órgano	26	51
cuerpo	29	44
medicamento	30	41
tejido	31	39
líquido	35	34
organismo	37	29
corazón	37	29
estado	38	28
anormal	38	28

Figura 4.1: Palabras que conformaron el conjunto inicial (B_0) que representa al dominio de la salud.

Descartando los verbos de mayor frecuencia en general, se aumentaron los verbos que ocurrieran con frecuencia alta en V: producir, faltar, tener, aumentar y sentir; para éstos sus rangos van del 37 al 60. Estas 19

[3]Este es un diccionario experimental que se ofrece en Internet para realizar traducciones entre varias lenguas.

[4]Se eligió el libro de C. E. Turner [Turner, 87] , en él se emplea un lenguaje poco técnico para describir fenómenos de la salud y la enfermedad.

[5]Estas son las 100 palabras más frecuentes del español mexicano.

palabras constituyen el conjunto base inicial, B_0. Es necesario, contar con los bigramas, $V_{\langle 2\rangle}$, para determinar su información mutua y así llevar a acabo el proceso de enriquecimiento. $V_{\langle 2\rangle}$ fue restringido a contener palabras con parte de oración adjetivo o sustantivo, pues son el tipo de palabra que pueden referir a los objetos del universo de la salud-enfermedad. Al ejecutar el proceso de inducción con diferentes valores de λ se obtuvieron los resultados que se muestran en la tabla 4.2 y las figuras 4.3 y 4.4. Mediante una inspección directa se determinó que el conjunto $B_{(4)}$ (B con $\lambda = 4$) tiene un error del 5%. Sin embargo, el conjunto $B_{(3)}$, que incluye 1003 palabras, tiene un error de más del 20%, y muchas palabras no se incluyen en el diccionario. La reducción de λ significa la reducción de la exigencia de una legítima asociación de términos y, por tanto, la posible ocurrencia de palabras en el conjunto B que no tienen una relación segura con el dominio. Por supuesto, esto conduce a un aumento de palabras espurias en el conjunto. Elegimos, entonces, el conjunto $B_{(4)}$ para realizar muestra aplicación.

λ	7	6	5	4
# Palabras	19	19	37	102
# Ciclos	1	1	2	5

Figura 4.2: Resultado del enriquecimiento por inducción léxica.

Hasta el momento hemos obtenido un conjunto base enriquecido por el método de inducción léxica. Para resolver el problema de determinar si una palabra refiere al dominio de la salud-enfermedad, no basta contar con un conjunto base grande, pues, además de que este conjunto es abierto, habrá muchas palabras que tengan sentido en ese dominio y no aparezcan en el conjunto base.

acidosis agalactia anafilaxis anormal aumento avance consumo corazón cuerpo depleción enclavamiento enfermedad estable estado falta hidrocefalia hiperlipidemia impotencia incarceración inflamación inmadurez lentitud líquido medicamento morbilidad organismo piel producir protrusión proyección sangre sentir sustancia tejido tiene trombocitosis órgano

Figura 4.3: Conjunto base, con $\lambda = 5$.

aberrante acción acidosis aerofagia agalactia almidón amniótico anafilaxis analéptico animal anormal antagónica antispasmódico antitusivo asimilación aspiración aumento auscultación avance cantidad claro confeccionado congénito conjuntivo consumo contráctil conversión corazón cuerpo curvatura deglución depleción desorganizado disacárido enclavamiento enfermedad epitelial espasmolítico específica estable estado este exagerado expectorante expresa falta fermentación fibroso fisicoquímico fisiológica folicular hidrocefalia hiperlipidemia hipnótica huésped impotencia incarceración inconsciente inflamación inhalación inmadurez intermedia labios lateral lentitud lordosis líquido lúteo mayor mayores medicamento metabolización micción microscópico minoso morbilidad morboso mínima oblongo organismo perforación piel producir protrusión proyección refracción sangre sano sentir seroalbuminoso sustancia tejido tiene tranquilizante transformación trombocitosis tuberculostático vermífugo vigilia órgano

Figura 4.4: Conjunto base, con $\lambda = 4$.

Como hemos mencionado, daremos respuesta a la pregunta de en qué grado pertenece una palabra a un dominio. Tenemos ya condiciones para llevar a cabo una prueba con la ec. 4.4. Antes de ello, tratemos de caracterizar al dominio para tener otro punto de referencia hacia el comportamiento de la ec. 4.4.

4.5 El dominio como cúmulo

Juzgamos necesario que las palabras afines semánticamente compartan sus sentidos. De tal manera que, representando las palabras por sus sentidos, un dominio se caracterizará porque las palabras que pertenezcan a él tendrán representaciones más similares entre ellas respecto a otras que no pertenecen al dominio. Esto es, la similitud entre la representación de dos palabras tomadas del dominio será mayor que la similitud entre las representaciones de dos palabras, una tomada del dominio y otra que no lo sea. Ésta es la hipótesis en la que descansa la propuesta de manejar al dominio como un cúmulo. Viendo a la similitud como el complemento de la distancia, las ideas anteriores pueden expresarse en forma abreviada diciendo que el dominio es un cúmulo, pues éste es un conjunto que reune elementos cercanos. La ventaja de trabajar con este enfoque es que podemos decidir el grado de pertenencia de un elemento considerando una medida del cúmulo, y otra entre el elemento y el cúmulo. En la subsección que sigue se analiza la forma de realizar estas medidas.

4.5.1 Distancia y diámetro

Para medir la distancia entre dos palabras es necesario contar con la representación de cada una de ellas, y que tomaremos como el contexto donde ocurre la palabra. Necesitamos, también, definir una distancia entre las representaciones de las palabras. La distancia que se empleó aquí está basada en la *similitud semántica* [Song, Johannesson, et. al. 96] [6]. Esta última se calcula como el número de elementos que tiene la intersección de las representaciones de las palabras dividido por el número de elementos que tiene la unión de las representaciones, de tal forma que para representaciones idénticas la similitud vale uno y para representaciones que no tienen nada en común la similitud es cero. Es fácil ver que el complemento a uno de la similitud puede funcionar como una distancia. Con estas ideas podemos referirnos al **diámetro de un cúmulo** de palabras como la máxima distancia entre parejas de este conjunto. El diámetro nos servirá como punto de referencia para valorar la distancia desde cualquier otro elemento hacia el cúmulo, y así, conocer la distancia hacia él. El **diámetro medio** de un cúmulo puede definirse como el promedio de las distancias entre cada elemento y todos los demás. El diámetro medio caracteriza a los miembros del conjunto, esto es: los elementos parecidos o los que son similares, serán aquellos para los que la distancia entre ellos es menor que cierto valor. Se espera, entonces, que si otra palabra pertenece a este conjunto su distancia hacia las demás palabras no sea mayor al diámetro. Debe, también, tenerse presente que si una palabra pertenece al conjunto, entonces su distancia hacia el conjunto debe ser cero. Precisemos esta idea analizando algunas situaciones que pueden ocurrir con la distancia.

Indiscutiblemente, la distancia máxima tomada sin condición, de todas las parejas del conjunto, no representa adecuadamente a su diámetro, pues deberá evitarse la participación de palabras espurias, incluidas por ruido de alguno de los métodos, o bien, por ocurrencia casual de un contexto ajeno para alguna palabra admitida. Ambas causas podrían conducir a una distancia máxima igual a uno. La máxima distancia, restringida a ser diferente de uno, puede ser más representativa. Esta medida no evita del todo los casos antes citados, pero resulta más conveniente si es ponderada con un valor que reuna todas las distancias entre los miembros del conjunto. Justamente, el diámetro medio es una forma de conjuntar en un factor todas las distancias del cúmulo. Así, el **diámetro normalizado** se define como la distancia máxima restringida a distancias diferentes de uno multiplicada por el diámetro medio.

Algo semejante ocurre con la **distancia entre una palabra y un cúmulo**. La distancia, tomada como el mínimo de las distancias entre la palabra y cada elemento del cúmulo, perdería representación cuando la palabra en cuestión se encuentra muy cerca de un elemento del conjunto, y a su vez, este elemento está muy alejado de todos los demás. La distancia resulta pequeña cuando en realidad no lo es. Hay que agregar, también, en este caso, un factor que pondere esta medida. Luego, la distancia entre una palabra y un cúmulo puede definirse como la distancia mínima, entre la palabra y los elementos del cúmulo, multiplicada por el promedio de las distancias de la palabra hacia cada elemento del cúmulo.

La ponderación del diámetro, visto como la distancia máxima entre los elementos de un conjunto, y la distancia de una instancia hacia un conjunto, tomada como la distancia mínima, consigue una normalización

[6] En [Rousseau, 98] y [Thomas, Short, 96] se analizan y ejemplifican otros usos de este tipo de funciones.

para hacer comparable la distancia hacia un conjunto con su diámetro. Por último, podemos definir tres grados de cercanía de una palabra x hacia un cúmulo J, mediante los *valores lingüísticos* "incluida", "cercana" y "externa":

Incluida. Cuando la distancia de x hacia J es menor o igual que el diámetro normalizado de J.

Cercana. Cuando la distancia a J es menor o igual que el diámetro de J.

Externa. Cuando la distancia de x a J es mayor que el diámetro de J.

Formalmente tendremos las siguientes definiciones.

Sean dos palabras $x, y \in \mathcal{V}$ con contexto $\mathcal{C}(x)$ y $\mathcal{C}(y)$[7], la **distancia entre** x **y** y es:

$$\delta_{\mathcal{C}}(x,y) = 1 - \frac{\#(\mathcal{C}(x) \cap \mathcal{C}(y))}{\#(\mathcal{C}(x) \cup \mathcal{C}(y))}$$

Sea J un conjunto de palabras, $\#J = n$ y x cualquier otra palabra,

$$\underset{min}{\Delta_J}(x) = \min\{d | d = \delta_{\mathcal{C}}(x,y), y \in J\}, \tag{4.5}$$

es la **distancia desde** x **hacia** J. La **distancia promedio desde** x **hacia** J se calcula como:

$$\overline{\Delta}_J(x) = \frac{1}{n} \sum_{y \in J} \delta_{\mathcal{C}}(x,y). \tag{4.6}$$

La **distancia normalizada entre** x **y** J combina 4.5 y 4.6, del siguiente modo:

$$\Delta_J(x) = \underset{min}{\Delta_J}(x) \cdot \overline{\Delta}_J(x). \tag{4.7}$$

Para calcular el diámetro normalizado de un cúmulo J haremos uso de la distancia normalizada, pero antes debemos definir el **diámetro** de J:

$$\phi_{max}(J) = \max\{d | d = \delta_{\mathcal{C}}(x,y) \neq 1, x, y \in J\}, \tag{4.8}$$

y el **diámetro medio** de J lo calculamos con la expresión:

$$\overline{\phi}(J) = \frac{1}{n-1} \sum_{x \in J} \Delta_{J \setminus \{x\}}(x). \tag{4.9}$$

Con 4.8 y 4.9 obtenemos el **diámetro normalizado** de J:

$$\phi(J) = \phi_{max}(J) \cdot \overline{\phi}(J). \tag{4.10}$$

Observemos que al efectuar el cálculo indicado por la ec.4.8 puede realizarse, apoyados en la memoria, el cálculo del sumando de 4.9. El cálculo indicado por 4.10 es costoso ($O(n^3)$), aunque el diámetro se calculará sólo una vez. De igual manera la ec. 4.7 puede optimizarse y calcular la distancia normalizada con complejidad en tiempo $O(n)$. La ventaja de este enfoque es su independencia sobre el uso de umbrales para poder decidir la pertenencia de un elemento al conjunto.

En la sección siguiente haremos un experimento para comparar la ec. 4.7 con la ec. 4.4.

[7] Ha de decirse que, para favorecer la similitud entre palabras, el contexto $\mathcal{C}$, según la ec. 4.1, se *lematiza* previamente: cada nominal que compone la representación es reescrito en singular y masculino.

4.5.2 Comparación entre pertenencia rugosa y distancia normalizada

En esta sección se presentan los resultados obtenidos al aplicar los dos métodos para decidir el grado de pertenencia de una palabra a un dominio. En ambos experimentos se contó con un diccionario[8] que incluye 29,002 entradas y un total de 61,540 acepciones. Con este diccionario se obtuvo uno nuevo, eliminando de las definiciones todas las palabras de mayor frecuencia del español de México y, eliminando también, palabras muy frecuentes en el diccionario como "dícese". Además, se dio un formato adecuado para localizar con facilidad las acepciones de cada entrada. A esta nueva versión la podemos considerar como D. Se formó el diccionario inverso como se mencionó en 4.2, y con él fue posible determinar el contexto de cada palabra del cúmulo $B_{(4)}$, y establecer su representación: $B'_{(4)}$. La fórmula 4.1 se emplea para determinar el contexto de una palabra x. Esta fórmula utiliza cada una de las definiciones de las palabras que contiene una entrada del diccionario inverso. Sin embargo, no todos estos usos deben ser considerados, puesto que si la palabra para la que se desea encontrar su contexto tiene varias acepciones, entonces sólo algunos usos estarán relacionados con el sentido que se está dando a x. Para cada palabra del conjunto base se obtuvo su definición inversa y fueron descartados los miembros de esta definición que no tuvieran el sentido esperado (en el dominio salud). Por tanto, el sentido correcto, elegido para cada palabra, es lo que da el carácter de dominio al conjunto de palabras.

Experimento con μ.

Se aplicó la fórmula 4.4, identificando a V con el conjunto $B_{(4)}$, y usando el diccionario inverso I_D como la función Ψ. Se calculó el grado de pertenencia de una lista de 162 palabras extraídas de un folleto de salud comunitaria, lista que llamaremos E. Previamente se clasificaron las palabras de E en tres grupos mutuamente excluyentes:

E_1 (Incluidas) Palabras referentes a la salud.

E_2 (Cercana) Palabras de terminología médica.

E_3 (Externa) Palabras de tipo general.

El nombre de la clase trata de referirse al dominio que estamos considerando. Para determinar el desempeño del método se procedió, en primer lugar, a ordenar descendentemente, según el grado de pertenencia de cada elemento de E al conjunto $B_{(4)}$: E'.

Para calcular la precisión se determinaron "intervalos" en E' asociados a cada clase E_i, $[E_i]$: Para la clase E_1 su intervalo inicia con la primera palabra de E' y abarca hasta la última palabra en E' que pertenece a E_1. El intervalo de E_2 contiene desde el primer elemento de E_2 que ocurre en E' hasta el último que pertenece a E_2. El intervalo $[E_3]$ cubre desde la primera palabra de E_3 hasta el final de E'. Los intervalos definidos determinan implícitamente cortes en el grado de pertenencia.

El cálculo de la precisión para una clase E_i fue realizado con base a la fórmula:

$$P(E_i) = \frac{\#E_i}{[E_i]}$$

Debe notarse que no hablamos de evocación, ya que $[E_i]$ contiene todos los elementos de E_i, es decir, la evocación es 1. Con la fórmula anterior se calculó la precisión y el número F para cada E_i y los resultados se muestran en la tabla 4.1. Estos valores muestran un desempeño bajo, y sólo nos servirán de referencia en cálculos posteriores. Debido a que este es un problema de tipo semántico, es necesario disponer de mayores recursos que los que se emplean en problemas sintácticos. De esta forma, esperaríamos que el desempeño mejorara ampliando el cúmulo $B_{(4)}$, y usando contextos más ricos. Lo que puede resultar una objeción es la necesidad de definir umbrales para efectuar "cortes" en la lista E'.

[8]Se trata de la mayor parte del *Diccionario Anaya* obtenido a través de Internet (`http://www.anaya.es/dict`).

En el siguiente apartado se hará una comparación, de estos resultados, con el método basado en una métrica, el cual también permite decidir la relación entre una palabra y un dominio.

E_i	$P(E_i)$	$F(E_i)$
incluida	0.349315	0.517766
cercana	0.358974	0.528302
externa	0.387324	0.558376
Promedio	*0.365204*	*0.534815*

Tabla 4.1: Desempeño de la clasificación de palabras con μ_B.

Experimento con Δ_B

Después de representar el conjunto base $B_{(4)}$ se determinó con 4.10, el diámetro normalizado: $\phi(B_{(4)}) = 0.675958$.

El experimento consistió en la aplicación de la fórmula 4.7 usando $B_{(4)}$, a cada una de las palabras de E. El desempeño fue calculado según los criterios que enseguida se exponen:

1. Los conjuntos de palabras determinadas por su cercanía a $B_{(4)}$, se revisaron manualmente y los resultados para esta clasificación aparecen en la tabla 4.2.
2. Las palabras que no estaban definidas en $I_D(x)$ fueron calificadas por el algoritmo como externas (en la figura 4.7 ocurren con la marca "-" en su distancia).

En las figuras 4.5, 4.6 y 4.7 se muestran las clases que determinó la métrica Δ_B. Los casos incorrectos se deben a varios factores. Uno es el "ruido" que un diccionario introduce en la representación, ya que en las expresiones que aparecen en el diccionario siguen un patrón particular. La representación de la palabra "ejemplo", incluida en el conjunto B (fig. 4.5) contiene 2221 signos, el número de palabras de la representación de B consta de más de 13,800 palabras. "ejemplo" es una palabra que se emplea en muchos contextos y en particular en el dominio de la salud, como la definición

> **albuminoide** Cualquiera de las proteínas pertenecientes a un gran grupo, como el colágeno o la queratina, que son constituyentes principales de las células de los animales y de los vegetales. Se encuentran, por ejemplo, en los huesos y en los tendones.

Es más probable que ocurran intersecciones con la representación de B y, por tanto, reduzcan la distancia. Las palabras frecuentes en el diccionario sesgan el cálculo de la distancia. Por otro lado, las palabras poco frecuentes en el diccionario conducen a valores de Δ_B altos, como "crónico" que aparece en la fig 4.7.

Notemos que, si comparamos los resultados de la pertenencia rugosa, i.e. la precisión y número F mostrados en la tabla 4.2, con los de la tabla 4.1, resultan más precisos los de la la función μ_B. Los errores promedio son 46% para μ_B y 55% para Δ_B. Hay que tomar en cuenta, también, que al aplicar directamente la ec. 4.4 habría que definir de algún modo los intervalos, por ejemplo, usando umbrales que permitan discernir si la palabra está en la clase incluida, cercana o externa. Lo cual sí pudimos hacer sin umbrales en este experimento gracias al uso del diámetro.

Clase	R	P	F
incluida	0.351648	0.666667	0.460432
cercana	0.20000	0.327586	0.248366
externa	0.345455	0.339286	0.342342
Promedio	*0.299034*	*0.444513*	*0.35038*

Tabla 4.2: Desempeño de la clasificación de palabras con Δ_B.

El resultado de las evaluaciones es muy discutible, pues debe señalarse que los grupos conformados (E_i), no tienen porque ser considerados como clases en el sentido estricto, y es posible que éstos se sobrepongan para cubrir a B (como preclases maximales).

El caso de construir un vocabulario de un dominio para decidir si una palabra pertenece a él cambiaría el criterio de evaluación, ya que podemos decidir si la palabra pertenece al vocabulario cuando está incluida o es cercana, lo cual es útil en los problemas de recuperación de información. El problema que se conoce más parecido al de pertenencia a un dominio se encuentra en [Haas, He, 93] donde obtienen una eficiencia (F) de 0.4932. La tabla 4.3 se muestra la evaluación considerando la pertenencia "dura" al dominio. Aunque tenemos en la tabla 4.3 mayor eficiencia que la publicada en [Haas, He, 93], no podemos establecer aún ninguna conclusión definitiva sobre el método propuesto, ya que carecemos de compatibilidad entre los recursos que empleamos y los citados en la referencia.

Clase	R	P	F
perteneciente	0.714286	0.613208	0.659898
ajena	0.422535	0.535714	0.472441
Promedio	*0.56841*	*0.574461*	*0.56617*

Tabla 4.3: Desempeño de la pertenencia de palabras a un dominio (Δ_B).

x	$\Delta_B(x)$	x	$\Delta_B(x)$	x	$\Delta_B(x)$
ajustar	0.517249	enfermo	0.502862	pérdida	0.675871
alimento	0.57475	filtrar	0.497076	recibir	0.5407
aparecer	0.51577	función	0.665456	recuperación	0.504262
bloqueo	0.509716	general	0.528816	resfriado	0.650491
causa	0.631685	glándula	0.666411	resultar	0.513292
causar	0.561414	infección	0.474453	sangre	0
crónica	0.622813	inflamación	0.0182395	sano	0.0419209
cuerpo	0	ingerir	0.508324	sistema	0.486348
célula	0.670498	largo	0.416814	suceder	0.535352
daño	0.575341	llamar	0.535381	sustancia	0
derecho	0.641574	mantener	0.6191	sólido	0.633784
diabetes	0.66713	medicamento	0.00510438	súbito	0.511371
digestivo	0.655756	medicina	0.577064	tipo	0.6409
ejemplo	0.651139	normal	0.643803	trasplante	0.495153
eliminar	0.477008	proceso	0.591986	tratar	0.563655
enfermedad	0.00107075	producto	0.627916	tumor	0.673277

Figura 4.5: Palabras incluidas en el conjunto B de acuerdo con Δ_B.

x	$\Delta_B(x)$	x	$\Delta_B(x)$	x	$\Delta_B(x)$
abuso	0.740916	llegar	0.764341	quemadura	0.690664
accidente	0.816186	lupus	0.859262	realizar	0.736837
agua	0.68492	manejar	0.752178	riñón	0.769118
agudo	0.741893	mayoría	0.747515	sanguíneo	0.772718
balance	0.861496	menudo	0.701921	ser	0.93125
caso	0.688631	metabolismo	0.683594	subyacente	0.692728
complejo	0.900568	mineral	0.701326	síntoma	0.697778
deshacer	0.832564	morir	0.743239	tener	0.9496
dieta	0.85942	médico	0.724274	torrente	0.984386
diálisis	0.965051	necesario	0.764487	trabajar	0.736256
dos	0.961296	necesitar	0.945125	trabajo	0.744712
equilibrio	0.721762	nivel	0.802086	tratamiento	0.754501
fluir	0.834327	orina	0.704366	traumatismo	0.785952
forma	0.827676	paciente	0.721546	urea	0.86681
gripe	0.786993	partícula	0.797449	urinario	0.693022
hacer	0.742295	pesar	0.704972	venenoso	0.765525
hipertensión	0.986328	potasio	0.801336	vez	0.744712
hombre	0.961396	prostático	0.939042	vivir	0.797249
izquierdo	0.707039	proteína	0.833788		
lado	0.733634	provenir	0.969312		

Figura 4.6: Palabras cercanas al conjunto B de acuerdo con Δ_B.

x	$\Delta_B(x)$	x	$\Delta_B(x)$	x	$\Delta_B(x)$
acumular	1	exterior	1	pobre	1
alargamiento	-	extra	1	poder	1
alteración	-	falla	1	poliquístico	-
arterial	-	fragmentar	1	prescribir	1
calcio	1	fósforo	1	renal	1
cloruro	1	glomérulos	-	saber	1
compensar	1	hereditario	1	seria	-
com un	-	importante	1	severa	-
condición	-	ingesta	-	significa	-
conforme	1	mayor	1	sodio	1
crónico	1	mejor	1	temprano	0.999558
deber	0.993071	nefrona	1	tiempo	1
desarrollar	1	originar	1	tracto	-
desecho	1	parte	0.999847	uno	0.996017
dura	-	partir	1	usar	1
electrolito	-	pasar	0.994664	utilizar	1
empeorar	1	pequeõ	1	viajar	1
equivocado	-	permanente	1	zona	1
existir	1	persona	1		

Figura 4.7: Palabras externas al conjunto B de acuerdo con Δ_B.

Es importante mencionar que en los experimentos (μ y Δ_B) no se recurre a más información que un diccionario sin clasificación sobre sus entradas. Aunque la desventaja que prevalece es el paso manual en la selección de los sentidos de los elementos del conjunto base (B).

4.6 El sentido de una frase

Por ahora sólo podemos decidir en qué grado una palabra pertenece a un dominio. Ya hemos dicho que es importante dar respuesta a la misma pregunta para una frase. De acuerdo con estructura conceptual presentada, resulta muy restrictivo calcular el sentido de una frase usando la fórmula 4.2. En esta sección se hará el cálculo del sentido también composicionalmente, pero sin imponer las condiciones de la ec.4.2.

El problema de la determinación del sentido de un sintagma ha tenido algunas propuestas como las que hacen G. Hirst [Hirst, 87], J. Binot [Binot, 91] y J. García [García-Fajardo, 95]. Los dos primeros casos siguen la semántica de R. Montague [Montague, 77] y emplean recursos lingüísticos importantes. Aunque estos sistemas tienen realización práctica, van más allá de las necesidades y posibilidades nuestras. El tercer caso es una propuesta que analiza la formación del sentido de un sintagma en el marco de un modelo del *universo del discurso*. Este último enfoque es el que se adoptó para ofrecer una solución al problema del grado del sentido de una frase en un dominio.

4.6.1 Una representación más apropiada de los conceptos

Haciendo uso de un diccionario, las operaciones en una retícula corresponden esencialmente a relaciones de hiponimia e hiperonimia, que son muy útiles en el PLN como lo ha mostrado *WordNet* [Beckwith, Fellbaum, et. al. 91] en diversas aplicaciones [Hasse, 95], [Agirre, Rigau, 96]. Sin embargo, hay necesidad de establecer otro tipo de relaciones como: *sinonimia*, *antonimia* y *holonimia*, entre otras [Lyons, 77].

La determinación de relaciones como las mencionadas tendrán que apoyarse en un análisis sobre la representación de los conceptos. Estas necesidades demandan que la representación de conceptos, como se ha indicado –con retículas o conceptos rugosos– sea enriquecida.

Ahora procederemos a revisar la propuesta que hace J. García-Fajardo [García-Fajardo, 95] sobre representación de conceptos. En primer lugar se considera el *universo del discurso* como el universo en el que se contextualiza una interpretación. Éste abarca tres componentes:

ID Información gramatical y conceptual.

RP Representación perceptual de la situación del habla.

REP Representación de estados, hechos y situaciones particulares de experiencias previas al acto lingüístico.

De acuerdo con los elementos de la estructura sintagmática que pretendemos interpretar, se toma información de alguno o algunos de los componentes del universo del discurso. En esencia, la representación de conceptos implícita que se propone es que para combinar dos conceptos, dentro de una situación, sea suficiente encontrar rasgos *compatibles*, y así crear el sentido que provee la estructura sintagmática donde ocurren los términos correspondientes a esos conceptos. Algo semejante sucede al momento que interactúan lingüísticamente dos hablantes, ya sea que se encuentren *propiedades compartidas* (rasgos compatibles) en sus emisiones, o bien, se crée un conflicto que conduce a una reestructuración conceptual o a un desacuerdo. Se considera, entonces, que cada término t de la lengua corresponde a un concepto $\mathcal{K}$ formado por la intersección de, por lo menos, dos propiedades A_i, A_j, que dan sentido a t. Sucesivamente, se consigue un nuevo sentido al combinar las propiedades de otro concepto $\mathcal{K}'$ con $\mathcal{K}$, si es el caso de que algunas de las propiedades de $\mathcal{K}'$ sean comunes con las de $\mathcal{K}$, justamente agregando las propiedades comunes que forman a $\mathcal{K}$ y a $\mathcal{K}'$. En general, si denotamos con Ψ la función que "recupera" las propiedades de un concepto a partir del universo del discurso, el sentido que se obtiene por la combinación de una secuencia de conceptos $\mathcal{K}_1 \ldots \mathcal{K}_i$ $(i > 2)$ puede expresarse con la función:

$$\xi(\mathcal{K}_1 \ldots \mathcal{K}_i) = \begin{cases} \xi(\mathcal{K}_1 \ldots \mathcal{K}_{i-1}) \cup (\Psi(\mathcal{K}_i) \bigcap_{j<i} \Psi(\mathcal{K}_j)) & \text{si } \Psi(\mathcal{K}_i) \cap \xi(\mathcal{K}_1 \ldots \mathcal{K}_{i-1}) \neq \emptyset \\ \emptyset & \text{en caso contrario.} \end{cases} \tag{4.11}$$

donde $\xi(\mathcal{K}_1, \mathcal{K}_2) = \Psi(\mathcal{K}_1) \cap \Psi(\mathcal{K}_2)$.

Deseamos aprovechar la anterior propuesta en la aproximación del sentido de una frase nominal simple, limitando con ésto muchos casos en los que tendríamos que precisar la combinación subyacente en 4.11. Por ejemplo, “las señoras que estaban cansadas fueron a sentarse” no tiene el mismo sentido que “las señoras, que estaban cansadas, fueron a sentarse”[9]. Esto es, cada estructura sintagmática contribuirá de manera particular en la forma de hacer la combinación de las representaciones[10]. Más aún, una palabra aislada puede tener muchos sentidos y es necesario referirnos a un contexto extralingüístico para elegir el sentido más adecuado. Podemos considerar una *situación* a través de un conjunto de conceptos “en uso” o “activos”, tomados de **ID**, **RP** y **REP**, que permitan definir el sentido de un término. Esto significa que no es suficiente una correspondencia ψ como ha sido empleada anteriormente, sino que tendremos que referirnos a una situación s, por ejemplo, con la correspondencia $\psi_s : G \to M$. Por otro lado, de acuerdo con la dinámica de los diálogos y al proceso en general de la comunicación, tenemos que hacer uso de estados o secuencias de situaciones [Filgueiras, Damas, et. al. 90] de manera que dada una situación y un nuevo concepto obtengamos una nueva situación.

4.6.2 Grado de pertenencia de una frase a un dominio

El cálculo del sentido de una frase según la ec. 4.11 se comporta menos restrictivo que lo propuesto por la ec. 4.2, ya que 4.11 no exige que el sentido se forme únicamente por los rasgos comunes a cada concepto involucrado en la frase, como lo hace 4.2. Más bien en 4.11 se enriquece el sentido de la frase agregando los sentidos que se comparten entre al menos dos conceptos contenidos en el sintagma. En fácil comprobar que, para los términos que participan en la formación del sentido de un sintagma, $x_1, \ldots, x_n$:

$$\bigcap_{i \leq n} \Psi(x_i) \subset \xi(x_1, \ldots, x_n).$$

Tomemos a $\Psi : \mathcal{V} \to \mathcal{V}^*$ tal que $x \mapsto \Psi(x) = \cup_j D(x,j) \cup \mathcal{C}(x)$, como una aproximación a la función que accede al universo del discurso, tal como se explicó en 4.6.1. Sea el B dominio de trabajo y B' su representación según lo explicado en la sec. 4.5.2. El conjunto que se obtiene a través de la ec. 4.11 puede ser cuantificado respecto al dominio empleando la función de pertenencia rugosa (ec. 4.4). Si los términos que dan sentido a un sintagma son $x_1, \ldots, x_n$, tendríamos que el sentido del sintagma tiene un **grado de pertenencia en el dominio** B:

$$\mu_B(x_1, \ldots, x_n) = \frac{\#(\xi(x_1, \ldots, x_n) \cap B')}{\#\xi(x_1, \ldots, x_n)}. \tag{4.12}$$

En esta fórmula estamos considerando una relación de tolerancia debida a la condición

$$\Psi(\mathcal{K}_i) \cap \xi(\mathcal{K}_1 \ldots \mathcal{K}_{i-1}) \neq \emptyset$$

.

* * *

Una forma de probar las anteriores propuestas se presenta en el siguiente capítulo, aplicando las fórmulas del grado de pertenencia a un dominio, en un problema de clasificación de frases.

[9]Ejemplo de A. Bello tomado de [Alonso, Enriquez-Ureña, 44].
[10]La idea es modificar 4.11 *a là Montague*.

Capítulo 5

Aplicación a la clasificación de frases

Uno de los problemas centrales en el procesamiento del lenguaje natural ha sido elegir la interpretación más adecuada de un enunciado. Para este problema se han propuesto algunas soluciones [Hirst, 87], [Binot, 91], dependiendo de los recursos con que se cuenta y el fin para el que la solución va a ser empleada. En este capítulo se ofrece una solución al problema de clasificación de una respuesta dada a una pregunta específica. En las secciones subsecuentes se fundamenta el método empleado en la clasificación, y se exponen los resultados obtenidos de su aplicación.

Son dos las preguntas a las que se responde con las expresiones que queremos clasificar:

¿Qué es para usted la salud?

¿Qué es para usted la enfermedad?

Estas preguntas son parte de un cuestionario que fue aplicado a varios grupos de pacientes. Fue necesario realizar la clasificación antedicha para llevar a cabo una exploración en la población sobre el concepto "salud" y el concepto "enfermedad".

Las respuestas se clasificaron de acuerdo con los siguientes valores, o "modos de abordar el concepto" [Carrillo-Ruiz, 90]

- *Refracta: la respuesta no se centra en el concepto a definir.*
- *Circunda: menciona cosas relacionadas con el concepto sin llegar a definirlo.*
- *Sustituye: define el concepto por analogía.*
- *Diferencia: en la respuesta aparecen elementos que no pertenecen al concepto, pero que contrastan con él.*
- *Describe: define el concepto mencionando sus cualidades y propiedades.*

La clase de cada respuesta puede verse como un tipo de acto de habla, ya que se desea determinar cómo se expresa dicha respuesta. Las clases que aquí consideraremos constituyen el conjunto VAL={Refracta, Circunda, Diferencia y Describe}[1].

Tenemos, pues, que resolver el siguiente problema de clasificación:

[1]De acuerdo con la forma en que se clasificaron las respuestas, el valor "describe" es equivalente al de "diferencia", en ambos se hace una definición, la diferencia entre ellos es que en el segundo se emplea la negación o *términos de polaridad negativa* [Bosque, 80] para realizar la definición. Por otro lado, son pocas las respuestas clasificadas con el valor "sustituye", y muestran una descripción metafórica del concepto ("`es algo feo`"). El problema de analizar un enunciado metafórico está relacionado con el de isomorfismo entre digráficas, y emplea recursos que no están a nuestra disposición (http:// www.compapp.dcu.ie/ ˜tonyv/ metaphor.html). Es decir, esta determinación es difícil hacerla y resolvería pocos casos de nuestro problema, motivo por el cual fue descartada.

Dada una respuesta a una pregunta abierta del cuestionario, clasificarla en concordancia con VAL.

Se cuenta con 172 respuestas a las preguntas antes mencionadas. Ya que las preguntas refieren a los sustantivos "salud" o "enfermedad", se espera que las respuestas sean frases nominales tal vez precedida por el verbo "ser"[2]. En una muestra de 26 respuestas tomadas al azar, se observaron frases nominales simples y proposiciones subordinadas o no a frases nominales simples en la proporción que muestra la tabla 5.1. La tabla da una idea de las respuestas que van a ser procesadas.

Tipo de frase	#	%
Frases sustantivas simples	7	26
Proposiciones sustantivas	16	61
Proposiciones adjetivas	3	11

Tabla 5.1: Tipos de frases en las respuestas.

Algunos ejemplos de respuestas y su clase se muestran a continuación:

(Diferencia) `no tener ninguna enfermedad tanto física como mental.`

(Refracta) `es que está uno bien.`

(Circunda) `no tener deficiencias en absoluto.`

Las variantes para determinar el tipo de acto de habla de un enunciado son diversas. Así, por ejemplo, M. Nagata y Y. Morimoto [Nagata, Morimoto, 94] utilizan el *contexto local* en un diálogo para determinar, en una pareja de enunciados, el tipo de acto de habla del segundo de ellos, conociendo el tipo de acto de habla del primero. La solución que ellos proponen se basa en el uso de una función de similitud y una colección de diálogos formados por pares de enunciados, y etiquetados con su tipo de acto de habla. Muchos trabajos han abordado el problema de la interpretación tratando de proveer una teoría que relacione las *intenciones* y *creencias* con la *acción* y su efecto en el *estado mental* de un interlocutor, como en el caso de C. Perrault [Perrault, 89]. Ha habido pocos sistemas que usen plenamente una teoría para realizar interpretaciones. Cuando se recurre a las teorías es necesario realizar varias restricciones para enfrentar el problema de interpretación. T. Kumamoto [Kumamoto, Ito, 92] reconoce la *intención comunicativa* de los enunciados que da el usuario de un sistema de ayuda dentro de un sistema de cómputo. En su modelo considera que el intercambio comunicativo se realiza en una relación "maestro-esclavo" y esto permite reducir los estados mentales, además, de acuerdo con su análisis clasifica y elige las características que le permiten determinar la intención comunicativa.

Para la presente aplicación se desarrolló un método híbrido, tratando de aprovechar los recursos lingüísticos disponibles. Las respuestas a la pregunta de salud-enfermedad están restringidas, como ya se vio, desde el punto de vista sintáctico. Semánticamente, habrá que tomar en cuenta el *contexto* de la salud y la enfermedad para poder llevar a cabo una interpretación. Es requisito para la solución del problema de clasificación de frases, determinar algunas partes de la respuesta que permitan encontrar su clase, esto es, identificar algunos sintagmas en los que habrá que observarse, más que su estructura, los sentidos que habiliten la asignación de una clase a la respuesta. Como fue señalado, se hará la identificación de las frases nominales de la respuesta, ya que en ellas se presentan potencialmente los atributos que permiten tomar decisiones sobre el tipo de acto de habla.

El análisis que apoya la solución que aquí se propone está basado en la teoría de *actos de habla* [Austin, 90] y considera que el tipo de acto de habla de una respuesta puede determinarse a partir de un conjunto de rasgos de ella. En las siguiente sección se fundamenta esta propuesta.

[2]De acuerdo con las características de este tipo de interacción lingüística. [Lara, 97] profundiza estos aspectos.

5.1 Actos de habla y clasificación

De acuerdo con J. Austin [Austin, 90], las expresiones son *realizativas* porque al decir algo se realiza una acción (adicional a la de emitir sonidos). Algunos verbos hacen explícito el realizativo (e.g. "prometer"). Estos verbos tienen ciertas características de tiempo y modo ("permita"). Pero también hay *expresiones primarias* ("estaré allí"), que contienen realizativos implícitos ("prometo ir"), además de otros elementos como frases adverbiales que aumentan o debilitan la fuerza del realizativo. En vista del carácter realizativo, una expresión puede tomar un valor *afortunado* o *desafortunado*. El valor de fortuna de una expresión está condicionado por la estructura contextual en la que se realiza. La estructura considera el procedimiento convencional, los agentes y circunstancias, la forma de ejecutar el procedimiento, y el comportamiento de los agentes ante la situación. Con esto, es posible sistematizar los realizativos en diferentes *clases de infortunios*, por ejemplo:

- Desaciertos (acto intentado pero nulo)
  - mala aplicación del procedimiento
  - acto no autorizado
  - mala ejecución
- Abusos
  - acto insincero

Además de determinar el valor de la expresión, la estructura contextual ayuda a interpretar las expresiones. Si una condición no se cumple para que una expresión sea considerada afortunada, la misma expresión, en las mismas condiciones puede ser afortunada en tanto que el tipo de acto de habla que se realiza sea otro. Esto surge por las ambivalencias de las expresiones. Por ejemplo, al decir "lo apruebo" en circunstancias que permiten considerar a esta expresión como afortunada, la misma expresión, sin posibilidad de afirmar que estoy convencido respecto de la aprobación, sería un acto insincero. Pero la ambivalencia de "apruebo" como una forma convencional de decir "concuerdo", no requiere la condición de convencimiento y, por tanto, sería afortunada viéndola como acto de concordar.

Austin presenta una subclasificación de los realizativos en: judicativos, comportativos, compromisorios, ejercitativos y expositivos. Las respuestas a una pregunta pueden ser consideradas como descripciones o aproximaciones a descripciones, por lo que caen en la última subclase; específicamente están ligadas con algún expositivo, "defino", "creo", "opino", presente de manera implícita en una expresión primaria.

Ya hemos señalado que las diferentes formas de calificar el "modo de abordar el concepto", en las respuestas a preguntas abiertas, es un tipo de acto de habla, habría, por tanto, que analizar las condiciones en las cuales se tiene uno u otro valor.

Los clases de VAL se orientan a reconocer los elementos que están presentes en diferentes niveles de análisis del enunciado, tales como [Perrault, 89]:

A-1 los conocimientos que el emisor tiene sobre el dominio,

A-2 la forma en cómo expresa sus respuestas en general y

A-3 las creencias y deseos que el emisor tiene en el memento que emite la respuesta.

Una manera muy simple que podemos proponer para analizar la clase de una emisión es:

- Si en la respuesta ocurren palabras del dominio entonces podemos afirmar que A-1 no es nulo.
- Si A-1 no es nulo pueden ocurrir en la respuesta diversas formas correspondientes a A-2 y A-3, por ejemplo

- Palabras aisladas del dominio.
- Al menos una frase con elementos del dominio.
- Frases del dominio relacionadas.

Tenemos en síntesis que:

- La respuesta a una pregunta puede verse como un acto de habla de tipo expositivo (opino, creo, defino ...).
- Un expositivo debe clasificarse en valores que consideren las clases "circunda", "refracta", y "describe".
- A partir de las condiciones que cumpla la respuesta es posible asignar un tipo de expositivo:
  1. acto nulo → refracta
  2. acto incompleto → circunda
  3. acto afortunado → describe

Un *acto intentado* o *mal ejecutado* es un desacierto y se debe a que los agentes o circunstancias no son los adecuados, o bien, no se realizó completamente o pertinentemente el procedimiento. Éste es el caso del valor "circunda". Por ello, podemos decir que la descripción que se hace del concepto es incompleta, es decir, en la respuesta ocurren algunos elementos del dominio, pero no llegan a establecerse relaciones válidas entre estos elementos. Así mismo, el valor "refracta" designa a aquellas respuestas que no tienen elemento alguno del dominio (acto nulo). Por último, los únicos valores que pueden ser considerados como actos afortunados son "diferencia" y "describe", éstos tendrán que cumplir con lo establecido en A-1, A-2 y A-3. En tanto que, al tener una realización con valor desafortunado, podemos reconsiderar el tipo de acto de la realización para tomarlo como un acto afortunado.

La clasificación de una respuesta requiere de: el etiquetamiento de las respuestas con su parte del discurso, la identificación de las frases nominales de la respuesta, y determinar si una palabra o frase de la respuesta pertenece al dominio de trabajo. Veremos, entonces, en la siguiente sección la aplicación de las herramientas desarrolladas en los capítulos anteriores.

5.2 Procedimiento de clasificación

Se presenta en esta sección los resultados de la evaluación de respuestas, según lo propuesto en la sección 5.1.

El procedimiento para clasificar las respuestas considera que se tiene el valor "refracta" cuando no hay elementos del dominio, el valor "circunda" cuando hay palabras aisladas del dominio y, por último, habrá que discernir con otros parámetros si, en caso de existir al menos una frase del dominio, se trata del valor "circunda" o "describe". Podemos emplear el grado de pertenencia de una frase para determinar cuántas relaciones entre dos frases hay y, de acuerdo con ello, asignar una clase a la respuesta. Ya que, dados n vértices, con $n-1$ aristas podemos construir una gráfica conectada, sería aceptable que más de la mitad de las frases estén relacionadas –esto es, al menos $(n-1)/2+1 = (n+1)/2$ relaciones entre frases– para clasificar la respuesta como "describe". De no cumplirse lo anterior, la determinación de la clase estará sujeta a la pertenencia de las frases al dominio, dependerá así del promedio de los grados de pertenencia de las frases al dominio.

Los criterios anteriores se derivan de la propuesta hecha en la sec. 5.1 y no consideran información alguna de las respuestas a evaluar. Al aplicar el procedimiento implícito en estos criterios se encontraron situaciones inesperadas, como el caso de respuestas de una sola palabra que pertenece al dominio, para la que no es posible calcular el grado de pertenencia de acuerdo con las ecs. 4.11 y 4.4. Por tanto, se introdujo una ponderación de los parámetros iniciales, de tal forma que fuera posible la atención de casos como el citado. Se expone en la siguiente figura la determinación de la clase de una respuesta:

describe En dos casos:

- D.1 Si el número de nexos entre frases nominales que tienen sentido en el dominio es mayor que el número mínimo de nexos necesarios, o bien
- D.2 Si el número de nexos entre frases nominales que tienen sentido en el dominio es menor o igual que el número mínimo de nexos necesarios y mayor o igual que cero, y el número de palabras participantes es mayor que uno.

circunda En dos casos:

- C.1 Si el grado de pertenencia promedio de las frases no es cero, la distancia promedio de las palabras que forman las frases nominales hacia B es mayor que el diámetro normalizado de B (externa), y el número de palabras participantes es mayor que uno, o bien
- C.2 Si la distancia promedio de las palabras que forman las frases nominales hacia B es menor que el diámetro de B (cercana), y el número de palabras participantes es mayor que uno.

refracta En los demás casos.

Figura 5.1: Clasificación de respuestas

Los pasos previos que se llevan a cabo para clasificar una respuesta son:

- A Tomar cada una de las respuestas sin la clasificación asignada manualmente.
- B Asignar su parte del discurso a cada una de las palabras
- C Extraer las frases nominales y para cada una de ellas:
  - C.1 Eliminar las palabras de las clases cerradas.
  - C.2 Contabilizar el número de palabras que participan en la clasificación.
  - C.3 Tomar cada una de las palabras restantes en la frase y medir su distancia hacia el cúmulo B.
  - C.4 Calcular el promedio de las distancias al cúmulo.
  - C.5 Determinar el grado de pertenencia de la frase al dominio.
  - C.6 Acumular el grado de pertenencia de la frase para calcular el promedio.
  - C.7 Obtener el número de frases anteriores a la frase en curso con las que forma un sentido en el dominio y, si es el caso, incrementar el número de nexos entre frases.

La respuesta se califica considerando el promedio de las distancias de las palabras al dominio, el promedio de los grados de pertenencia de las frases, el número de nexos entre las frases, y el número de palabras que participaron (no eliminadas), según se indica en 5.1.

Por ejemplo, la respuesta:
`sería el conjunto de factores que determinan el completo bienestar del organismo .`
es etiquetada con su parte del discurso como:
`VERB ART NOM PREP NOM PRON VERB ART NOM ADJ CONT NOM PTO`
de ésta se extraen las frases nominales:
`<el conjunto de factores> <el completo bienestar del organismo>`
de cada frase nominal se retienen los nominales y se cuantifica su distancia promedio hacia B y el grado de pertenencia de la frase:

`<conjunto factores 0 5.072464e-02>`
`<completo bienestar organismo 0 2.752294e-02>`
Como último paso se calcula el promedio de las distancias, el promedio de los grados de pertenencia de las frases, el número de nexos entre las frases, el mínimo número de nexos, y la cantidad de palabras participantes:
`0.000000 3.912379E-02 1.000000 1.000000 5.`

Según el procedimiento explicado, en este ejemplo, tenemos que la clase de la respuesta es "describe".

Al aplicar el procedimiento antedicho a 172 respuestas se obtuvo una precisión de 49.41%, contando globalmente los aciertos de la clasificación. La tabla 5.2 detalla el desempeño del algoritmo practicado.

Clase	Refracta	Circunda	Describe
Precisión	0.4545	0.3396	0.6981
Evocación	0.6818	0.3103	0.5285
Índice F	0.5454	0.3242	0.6015

Tabla 5.2: Desempeño de la clasificación de respuestas.

Los resultados expresan menor dificultad en clasificar respuestas con elementos dentro del dominio de la salud y mayor dificultad cuando se emplean menos términos del dominio (véase el índice F para la clase Describe en la tabla 5.2). La dificultad que los mismos evaluadores humanos tenemos en la distinción de la clase "circunda" de las demás, se refleja en la elección de un umbral para decidir si una expresión se ubica en esta clase. Por ejemplo, para las 172 respuestas se encontró que en 93% de los casos se hizo la distinción correcta entre las clases "alejadas" ("refracta" y "describe"), mientras que la distinción entre clases "colindantes" fue, en promedio, de 78.2%. Lo anterior se detalla en la tabla 5.3. Esta tabla se calculó con base en la cantidad de fracasos, por ejemplo, repuestas clasificadas como "refracta" (`r`) siendo "describe"(`d`).

Clases	alejadas		colindantes		colindantes	
propuesta	r	d	c	d	r	c
observada	d	r	d	c	c	r
%	6.39	0.58	12.79	8.72	14.53	7.55

Tabla 5.3: Porcentaje de fracasos en la clasificación de respuestas.

Por inspección directa se determinó que más del 20% de los errores de la clasificación provienen de errores de etiquetamiento de las palabras con su parte del discurso, así como de la determinación de las frases nominales contenidas en las respuestas. En parte, este error se había previsto al evaluar el etiquetador con texto de género diferente. Lo mismo ocurre con el identificador de frases nominales, y sus errores se presentan acumulados con los de etiquetamiento.

Los resultados muestran en qué medida los recursos empleados pueden ayudar a resolver problemas semejantes al que fue explorado. La representación semántica que estamos empleando es vasta y se apoya en un diccionario. Resulta necesario emplear más recursos, para establecer una evaluación determinante sobre los métodos propuestos. Además, con las herramientas construidas, hay buenas condiciones de experimentar con corpora grandes. Ya que las representaciones que toman en cuenta la estructura sintáctica son mejores –cuando se trabaja dentro de un género particular– que las que no lo hacen [Grefenstette, 96], es necesario involucrar otros elementos de los sintagmas como los verbos, lo que impondría, por ejemplo, el requisito de bases de datos exprofeso. Lo anterior también implica ampliar el tratamiento morfológico a derivaciones, de tal manera que puedan compararse con mayor fidelidad las representaciones de las palabras.

Capítulo 6

Conclusiones

En este capítulo resumimos los resultados, presentamos los aportes y, además, señalamos algunas tareas futuras relativas a este trabajo.

6.1 Resumen de resultados

Presentaremos una discusión sólo de los principales métodos expuestos.

6.1.1 Etiquetamiento con partes del discurso

Se diseñó e implantó un etiquetador de partes del discurso para textos del español. En la estrategia seguida, se observó que el paso correspondiente al etiquetamiento de las palabras desconocidas, requirió de mayor precisión para obtener un desempeño adecuado. Esto problema fue atendido modificando la métrica empleada por el método de aprendizaje basado en memoria. La nueva métrica toma en cuenta el problema de valores desconocidos en las instancias de entrenamiento. La modificación mejoró el desempeño del método en una fracción que no puede aún ser considerada estadísticamente significativa [Jiménez-Salazar, Morales-Luna, 2000]. En suma, para palabras desconocidas se obtuvo un desempeño de 79.19% con el método IGTree puro y 81.5% con el método IGTree modificado.

6.1.2 Identificación de frases nominales

Se diseño e implantó un sistema para identificar las palabras que están contenidas en un sintagma nominal simple. El método empleado para este fin es, también, el de aprendizaje basado en memoria, al cual se alimentó con instancias extraídas de un texto de alrededor de 10,000 palabras. Respecto al corpus empleado los resultados de desempeño son altos [Jiménez-Salazar, Morales-Luna, 98]: 96.7% para texto del mismo género y 88.9% para textos de género diferente.

6.1.3 Pertenencia a un dominio

Se propuso un método para determinar en qué grado una palabra pertenece a un dominio. Este método requirió tanto de la representación del dominio como de una aproximación de su vocabulario. Se probaron los anteriores procedimientos en el caso de la clasificación de palabras para una aproximación del dominio de la salud. Aunque no se cuenta aún con una evaluación de su desempeño, los resultados encontrados presentan una pauta para enfrentar este tipo de problema.

En el caso de la determinación del sentido de una frase, se siguió una propuesta que enriquece la representación de los términos. La aplicación de este método, al problema de clasificación de respuestas a preguntas abiertas de un cuestionario, mostró que es indispensable, para mejorar los resultados, el empleo de mayores recursos de información lingüística, e.g. un corpus grande para incrementar la precisión de las herramientas morfosintácticas, y más recursos morfológicos, como reglas de derivación que faciliten la comparación de representaciones de palabras y frases.

6.2 Contribuciones

El aporte de este trabajo es en dos direcciones:

Sistemas Ante la escasez de herramientas para PLN, dentro de la lengua española, consideramos que los sistemas desarrollados resultan de mucha utilidad, puesto que son requisito de casi cualquier sistema de PLN.

- Se desarrolló un etiquetador de palabras de un texto con su parte del discurso, cuya utilidad fue mostrada al emplearse en la identificación de frases nominales.
- El reconocedor de frases nominales tuvo un desempeño alto respecto del corpus de entrenamiento utilizado.

Métodos Los métodos propuestos, aunque tienen calidad exploratoria, resultan de interés pues centran su objetivo en problemas básicos de la clasificación dentro del PLN: representación y transformación de entidades lingüísticas.

- La métrica que modifica al método IGTree, al ser usada como una alternativa en el tratamiento de valores desconocidos, tiene repercusiones en un espectro amplio de métodos de aprendizaje (i.e. supervisados y basados en similitud, como los apoyados en reglas).
- Puede suponerse que hay pocos trabajos que traten el problema de pertenencia de una palabra a un dominio ([Losee, Haas, 94] [Haas, He, 93]), por lo que las propuestas realizadas en la presente tesis enfrentan esta carencia.
- La propuesta del grado de pertenencia de una frase a un dominio tiene perspectivas muy amplias por impactar muchos campos de trabajo del PLN, como en los de gestión de textos: categorización, resúmenes, y extracción de información.

6.3 Trabajo futuro

1. El etiquetador de un texto con partes del discurso debe ampliar su corpus de trabajo al menos hasta 50,000 palabras, con el fin de obtener una colección de entrenamiento más grande y, en consecuencia, mejorar el desempeño, así como enfrentar el etiquetamiento de texto irrestricto. El uso de la métrica δ' sugiere explorar otros recursos que encierra el corpus, con el fin de aumentar la eficiencia, e.g. alternativas al uso de la probabilidad de n-gramas. También es posible mejorar el desempeño con relativa facilidad teniendo mayor apoyo en recursos morfológicos. El uso reglas de inflexión proporcionaría una forma de extender el etiquetamiento a palabras desconocidas que son inflexiones de palabras conocidas.
2. La identificación de proposiciones sustantivas, adjetivas y adverbiales, es el siguiente paso para el identificador de sintagmas. Se requiere, luego, de un corpus exprofeso para este desarrollo. Este trabajo podría apoyarse en bases de datos existentes respecto a valencias de verbos, o relaciones sintácticas verbo-sujeto, verbo-objeto u otras [Pool-Westgaard, 90].
3. Los resultados de la representación de un dominio, el sentido de una palabra en un dominio, y el sentido de una frase en un dominio, los cuales se apoyan en el uso de un diccionario, constituyen un punto de partida para investigar más en torno a la representación conceptual. Conviene experimentar con

corpora grandes, substituyendo al diccionario. De esta forma podríamos aclarar el desempeño de estos métodos.

4. Hay todavía muchas pruebas que deben realizarse con las propuestas sobre la determinación del grado de pertenencia de una palabra y de una frase. La recuperación de información y los sistemas de gestión, son casos donde sería valioso aplicar estas ideas. Pero también, es muy importante, llevar a cabo aplicaciones orientadas a la consecución de recursos lingüísticos, como la obtención de relaciones de hiponimia.

Referencias

[Abaitua, 99] Abaitua, Joseba: http://orion.deusto.es/ abaitua, 1999.

[Agirre, Rigau, 96] Agirre, Eneko & Rigau, German: Word sense disambiguation using conceptual density, *Proc. of the 16th Int. Conf. on Computational Linguistics* 22, pp 1-7, 1996.

[Ahonen, Mannila, et. al. 94] Ahonen, Helena; Mannila, Heikki & Nikunen, Erja: Forming grammars for structured documents: an application of grammatical inference, *Lecture Notes in Artificial Intelligence*, 862, Springer Verlag, pp 153-67, 1994.

[Alonso, Enriquez-Ureña, 44] Alonso, Amado & Enríquez-Ureña, Pedro: *Gramática castellana*, 2o. curso, Losada, Buenos Aires, 1944.

[Alshawi, Carter, 94] Alshawi, Hyan & Carter, David: Training and scaling preference functions for disambiguation, *Computational Linguistics*, 20, pp 635-48, 1994.

[Amsler, 89] Amsler, Robert: Reserch toward the development of a lexical knowledge base for natural language processing, *Association for Computing Machinery, Special Interest Group on Information Retrieval*, pp 242-9, 1989.

[Arjona-Iglesias, Luna-Traill, 91] Arjona-Iglesias, Marina & Luna-Traill, Elizabeth: *El infinitivo hablado en la ciudad de México*, UNAM, 1991.

[Arjona-Iglesias, 86] Arjona-Iglesias, Marina: *Estudios sintácticos sobre habla popular mexicana*, UNAM, 1986.

[Austin, 90] Austin, J.: *Cómo hacer cosas con palabras*, Paidos, Barcelona, 1990.

[Azzam, 95] Azzam, Saliha: Anaphors, PPs and disambiguation process for concepts analysis, *International Joint Conference on Artificial Intelligence*, pp 1354-9, 1995.

[Baeza-Yates, 98] Baeza-Yates, Ricardo: Searching the world wide web; challenges and partial solutions, Progress in Artificial intelligence, Coelho, Helder (Ed.), IBERAMIA98, *Lecture Notes in Artificial Intelligence*, 1484, Springer Verlag, pp 39-51, 1998.

[Barnden, 89] Barnden, John: Simulation of composit, a supra-connectionist architecture for commonsense reasoning, *Proc. of 2nd. Symposium of the Frontiers of Massively Parallel Computation*, pp 311-5, 1989.

[Barnett, Knight, et. al. 90] Barnett, Jim; Knight, Kevin; Mani, Inderjeet & Rich, Elaine: Knowledge and natural language processing, *Communications of the Association for Computing Machinery*, V.33 (8), pp 50-71, 1990.

[Basili, Pazienza, 97] Basili, Roberto & Pazienza, Ma. Teresa: Lexical acquisition for information extraction, *Lecture Notes in Artificial Intelligence*, 1299, Springer Verlag, pp 44-72, 1997.

[Basili, Pazienza, et. al. 96] Basili, Roberto; Pazienza, Maria Teresa & Velardi, Paola: An empirical approach to natural language processing. *Artificial Intelligence*, V. 85, pp 59-99, 1996.

[Beckwith, Fellbaum, et. al. 91] Beckwith, Richard; Fellbaum, Christiane; Gross, Derek & Miller, George: WordeNet: a lexical database organized on psycholinguistic principles, en *Lexical acquisition: exploting online resources to build a lexicon*, Zernik, Uri (Ed.), Lawrence Erlabum Association, pp 211-32, 1991.

[Belkin, Croft, 92] Belkin, Nicholas & Croft, Bruce: Information filtering and information retrieval: two sides of the same coin?, *Communications of the Association for Computing Machinery*, V.35 (12), pp 29-38, 1992.

[Bergadano, 93] Bergadano, F.: Inductive database relations, IEEE *Transactions on Knowledge and Data Engineering*, V.5, pp 969-72, 1993.

[Binot, 91] Binot, Jean-Louis: Natural language processing and logic, en *From natural language to logic for expert systems*, Thayse, A. (Ed.), John Wiley, pp 49-116, 1991.

[Boguraev, Briscoe, 89] Boguraev, K. & Briscoe, T.: *Computational lexicography for natural language processing*. Longman, London, 1989.

[Boguraev, Pustejovsky, 96] Boguraev, Branimir & Pustejovsky, James: *Corpus processing for lexical acquisition*, The MIT Press, 1996.

[Bosque, 80] Bosque, Ignacio: *Sobre la negación*, Ed. Cátedra, Madrid, 1980.

[Breiman, 84] Breiman, Leo: *Classification and regression trees*, Wadsworth Inc. 1984.

[Brent, 93] Brent, Michael: From grammar to lexicon: unsupervised learning of lexical syntax, *Computational Linguistics* 19, pp 243-62, 1993.

[Brill, 94] Brill, Eric: Some advances in transformational-based part of speech tagging, *Proc. of 12th, National Conf. on Artificial Intelligence*, AAAI, pp 722-7, 1994.

[Brill, Resnik, 94] Brill, Eric & Resnik, P.: A rule-based approach to prepositional phrase attachment disambiguation, In *Proc. of COLING-94*, http://www.cs.jhu.edu/b̃rill/papers.html, 1994.

[Brill, 95] Brill, Eric: Unsupervised learning of disambiguation rules for part of speech tagging, *Proc. of 3rd. Workshop on very large corpora*, http://www.cs.jhu.edu/b̃rill/papers.html, 1995.

[Calzolari, 95] Calzolari, Nicoletta: Acquiring and representing semantic information in a lexical knowledge base, *Lexical semantics and knowledge representation*, Pustejovsky, J. & Bergler, S. (Eds.), Springer Verlag, pp 235-43, 1995.

[Calzolari, 91] Calzolari, Nicoletta: Lexical databases and textual corpora: perspectives of integration for lexical knowledge bases, *Lexical acquisition: exploting online resources to build a lexicon*, Zernik, Uri (Ed.), Lawrence Erlabum Association, pp 191-204, 1991.

[Carpineto, Romano, 94] Carpineto, Claudio & Romano, Giovanni: Ulysses: a lattice-based multiple interaction strategy retrieval interface, *Lecture Notes in Computer Science*, 1015, Springer Verlag, pp 91-104, 1994.

[Carrillo-Ruiz, 90] Carrillo-Ruiz, Maya: Una herramienta que ayuda a obtener conclusiones, Tesis de Licenciatura en Computación, FCFM-BUAP, 1990.

[Ceri, Gottlob, et. al. 90] Ceri, S.; Gottlob, G. & Tanca, L.: *Logic programming and databases*, Springer Verlag, 1990.

[Chan, Wong, 91] Chan, K. & Wong, A.: A statistical technique for extracting classificatory knowledge from databases, Knowledge discovery in databases, Patietsky-Shapiro, G. & Frawley, W. (Eds.), *American Association of Artificial Intelligence*, AAAI Press, pp 107-24, 1991.

[Chen, Chen, 94] Chen, Kuang-hua & Chen, Hsin-Hsi: Extracting noun phrases from large scale texts: a hybrid approach and its automatic evaluation, http://lib-www.lanl.gov/cmp-lg/9405034, 1994.

[Church, 89] Church, Kenneth W.: A stochastic parts programs and noun phrase parser for unrestricted text, *Int. Conf. on Acoustics, Speech and Signal Processing*, V.2, pp 695-8, 1989.

[Church, Hanks, 90] Church, Kenneth & Hanks, Patrick: Word association norms, mutual information and lexicography, *Computational Linguistics*, 16 , pp 22-9, 1990.

[Church, Mercer, 93] Church, Kenneth & Mercer, Robert: Introduction to special issue on computational linguistics using large corpora, *Computational Linguistics* 19, pp 1-24, 1993.

[Church, Rau, 95] Church, Kenneth & Rau, Lisa: Commercial applications of natural language processing, *Communications of the Association for Computing Machinery*, V.38 (11), pp 71-9, 1995.

[Colmerauer, 82] Colmerauer, Alain: An interesting subset of natural language, en *Logic Programming*, Clark, K.L. & Tärnlund, S.A. (Eds.), Academic Press, pp 45-66, 1982.

[CSTNC, 92] CSTNC, Computer Science and Technology of the National Council: Computing the future, *Communications of the Association for Computing Machinery*, V.35 (11), pp 30-40, 1992.

[Cover, Hart, 67] Cover, T. & Hart, P.: Nearest neighbour pattern classification, IEEE, *Transactions on Information Theory*, V.13 (1), pp 21-7, 1967.

[Crouch, Yang, 92] Crouch, Carolyn; Yang, Bokyoung: Experiments in automatic statistical thesaurus construction, Workshop, *Association for Computing Machinery, Special Interest Group on Information Retrieval*, pp 77-88, 1992.

[Cuena, Fernández, et. al. 86] Cuena, José; Fernández, Gregorio; López-de-Mántaras, Ramón & Verdejo, Ma. Felisa: *Inteligencia Artificial: Sistemas Expertos*, Alianza Editorial, Madrid, 1986.

[Daelemans, 95] Daelemans, Walter: Memory-based lexical acquisition and processing, *Lecture Notes in Artificial Intelligence*, 898, Springer Verlag, pp 85-98, 1995.

[Daelemans, Durieux, et. al. 98] Daelemans, Walter; Durieux, Gert & van-den-Bosch, Antal: Towards inductive lexicon, *Proc. of LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications*, Granada, http://ilk.kub.nl/, 1998.

[Daelemans, van-den-Bosch, et. al. 98] Daelemans, Walter; van-den-Bosch, Antal; Zavrel, Jakub; Veenstra, Jorn; Buchholz, Sabine & Busser, Bertjan: Rapid development of NLP modules with memory-based learning, *Proc. of ELSNET in Wonderland*, pp 105-113, 1998.

[Dahl, Popowich, 90] Dahl, Verónica & Popowich, Fred: Parsing and generation with static discontinuous grammars, *New Generation Computing*, N.8, pp 245-74, 1990.

[Dale, Reiter, 95] Dale, Robert & Reiter, Ehud: Computational interpretations of the Gricean maxims in the generation of referring expressions, *Cognitive Science*, V.18, pp 233-63, 1995.

[Danlos, Gross, 88] Danlos, L. & Gross, M.: Building electronic dictionaries for natural language processing, *Programming in future generation computers II*, Fuchi, K. & Kott, L. (Eds.), North Holland, pp 123-31, 1988.

[Davey, Priestley, 90] Davey, B. & Priestley, H.: *Introduction to lattices and order*, Cambridge Mathematical Textbooks, 1990.

[Dhar, Tuzhilin, 93] Dhar, Vasant & Tuzhilin, Alexander: Abstract-driven pattern discovery in databases, IEEE, *Transactions on Knowledge and Data Engineering*, V.5 (6), pp 926-38, 1993.

[Dubois, Prade, 80] Dubois, Didier & Prade, Henri: *Fuzzy sets and systems, theory and applications*, Academic Press, 1980.

[Duda, 73] Duda, R.: *Pattern Classification and scene analysis*, Wiley Interscience, 1973.

[Duda, Hart, et. al. 76] Duda, R.; Hart, P. & Nilsson, N.: Subjective bayesian methods for rule-based inference systems. *Proc. AFIPS*, V.45, pp 1075-1082, 1976.

[EAGLES, 93] EAGLES: Lexicon architecture, http://coral.lili.uni.bielefeld.de/~gibbon/EAGLES, 1993.

[EAGLES, 96] EAGLES: Recommendations for the morphosyntactic annotation of corpora, http://coral.lili.uni.bielefeld.de/~gibbon/EAGLES, 1996.

[Fernández, Navarro, et. al. 98] Fernández, Baltasar; Navarro, Antonio; Cigarran, Juan & Fernández, Alfredo: Integration of formal concept analysis in a knowledge-based assistant, *Lecture Notes in Artificial Intelligence*, 1415, pp 112-23, 1998.

[Filgueiras, Damas, et. al. 90] Filgueiras, M.; Damas, L.; Moreira, N. & Tomás, A.: Natural language processing, (Eds.), *Lecture Notes in Artificial Intelligence*, 476, Springer Verlag, 1990.

[Franz, 96] Franz, Alexander: *Automatic ambiguity resolution in natural language processing, an empirical approach*, Springer Verlag, 1996.

[Frege, 73] Frege, Gottlob: Sobre sentido y referencia, en *Estudios sobre semántica*, Frege, G. (Ed.), Ariel, 1973.

[García-Fajardo, 95] García-Fajardo, Josefina: Estructura conceptual y comunicación. *Dimensión Antropológica*, Año 2, V. 3, p75-84, México, 1995.

[García-Fajardo, 96] García-Fajardo, Josefina: *De los sonidos a los sentidos, introducción al lenguaje*, Trillas, México, 1996.

[Garside, 96] Garside, Roger: The robust tagging of unrestricted text: the BNC experience, *Using corpora for language research*, Thomas, Jenny & Short, Mick (Eds.), England, Longman, pp 167-80, 1996.

[Gierl, Frost, 92] Gierl, Claude & Frost, David: Identification of domain-specific terminology by combining mutual information and lexical induction, *European Conference on Artificial Intelligence*, pp 564-66, 1992.

[Gold, 67] Gold, Mark: Language identification in the limit, *Information and Control*, V.10, pp 447-474, 1967.

[González, Thomason, 78] González, Rafael & Thomason, Michael: *Syntactic pattern recognition*, Addison-Wesley, 1978.

[Grefenstette, 93] Grefenstette, Gregory: Automatic thesaurus generation from raw text using knowledge-poor techniques, *Making sense of words, 9th. Annual Conference of the UW Centre of the new OED and text Research*, http://www.xerox.fr/sys/sitemap.html, 1993.

[Grefenstette, 94] Grefenstette, Gregory: Corpus-derived first, second and third-order word affinities, *6th. Euralex International Congress, Amsterdam*, pp 1-10, 1994.

[Grefenstette, 96] Grefenstette, Gregory: Light parsing as finite-state filtering, *Workshop on Extended finite state models of language, ECAI96*, Budapest, http://www.xerox.fr/sys/sitemap.html, 1996.

[Grefenstette, 96] Grefenstette, Gregory: Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches, *Corpus processing for lexical acquisition*, Boguraev & Pustejovsky (Eds.), MIT Press, pp 205-15, 1996.

[Grishman, 97] Grishman, Ralph: Information extraction, techniques and challenges, *Lecture Notes in Artificial Intelligence*, 1299, Springer Verlag, pp 10-27, 1997.

[Guiraud, 54] Guiraud, Pierre: *Les caractères statistiques du vocabulaire*, Mouton, 1954.

[Guthrie, Pustejovsky, et. al. 96] Guthrie, Louise; Pustejovsky, James; Wilks, Yorick & Slator, Brian: The role of lexicon in natural language processing, *Communications of the Association for Computing Machinery*, V.39 (1), pp 63-72, 1996.

[Haas, He, 93] Haas, Stephanie & He, Shaoyi: Toward the automatic identification of sublanguage vocabulary, *Information Processing and Management*, V.29 (6), pp 721-32, 1993.

[Hajičova, 87] Hajičova, Eva: Procesamiento de lenguaje natural, 1er. curso internacional de sistemas expertos, CINVESTAV, Martínez-Enríquez, Ana Ma. (Ed.), México, pp 113-26, 1987.

[Harris, 66] Harris, Zellig: *Structural linguistics*, Phoenix Books, 1966.

[Hasse, 95] Hasse, Kenneth: Analogy in the large, *International Joint Conference on Artificial Intelligence*, pp 1375-80, 1995.

[Hirst, 87] Hirst, Graeme: *Semantic interpretation and the resolution of ambiguity*, Cambridge University Press, 1987.

[Ide, Véronis, 98] Ide, Nancy & Véronis, Jean: Introduction to the special issue on word sense disambiguation: the state of the art, *Computational Linguistics* 24, pp 1-40, 1998.

[Ide, Véronis, 89] Ide, Nancy & Véronis, Jean: Knowledge extraction from machine-readable dictionaries: an evaluation, *Lecture Notes in Artificial Intelligence*, 898, Springer Verlag, pp 19-34, 1995.

[Ide, Maitre, et. al. 93] Ide, Nancy; Le Maitre, Jacques & Véronis, Jean: Outline of a model for lexical databases, *Information Processing and Management*, V.29 (2), pp 159-86, 1993.

[Iwayama, Tokunaga, 95] Iwayama, Makoto & Tokunaga, Takenobu: Hierarchical bayesian clustering for automatic text classification, *International Joint Conference on Artificial Intelligence*, pp 1322-7, 1995.

[Iwiński, 89] Iwiński, Tadeusz: Rough sets and knowledge representation, 2o. curso internacional de sistemas expertos, CINVESTAV, Morales-Luna, Guillermo (Ed.), México, pp 121-30, 1989.

[Jansen-Winkeln, 85] Jansen-Winkeln, Roman: LEGAS: inductive learning of grammatical structures, *Advances in artificial intelligence*, Tim O'Shea (Ed.), North Holland, pp 169-81, 1985.

[Jiménez-Salazar, 96] Jiménez-Salazar, Héctor: Conocimiento de fondo para procesamiento de lenguaje natural, *Análisis Semánticos*, García-Fajardo, Josefina, (Ed.), El Colegio de México, pp 235-45, 1996.

[Jiménez-Salazar, Morales-Luna, 98] Jiménez-Salazar, Héctor & Morales-Luna, Guillermo: Noun phrases identification from a tagged text, *Memoria TAINA98*, México, pp 90-101, 1998.

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

[Johnson-Laird, 85] Johnson-Laird, P.: *Mental models, towards a cognitive science of language inference and consciousness*, Harvard University Press, Cambridge, 1985.

[Knauff, Rauh, et. al. 97] Knauff, Markus; Rauh, Reinhold & Renz, Jochen: A cognitive assesment of topological spatial relations: results from an empirical investigation. en *Lecture Notes in Computer Science*, 1329, Hirtle S.C & Frank A.U. (Eds.), pp 193-206, 1997.

[Kotani, 88] Kotani, Yushiyuki: Grammatical inference on the basis of linear environments of symbol strings, *Proc. 6th. Int. Joint Conf. on Artificial Intelligence*, pp 487-9, 1988.

[Kowalski, 79] Kowalski R.: *Logic for problem solving*, North Holland, 1979.

[Kozima, 96] Kozima, Hideki: Text segmentation based on similarity between words, http://lib-www.lanl.gov/cmp-lg/9601005, 1996.

[Krovetz, 89] Krovetz, Robert: Word sense disambiguation using machine-readable dictionaries, *Proc. 12th. Annual Int. Association for Computing Machinery, Special Interest Group on Information Retrieval, Conf. on IR*, pp 127-36, 1989.

[Krovetz, 93] Krovetz, Robert: Viewing morphology as an inference process, *Association for Computing Machinery, Special Interest Group on Information Retrieval*, pp 191-202, 1993.

[Kumamoto, Ito, 92] Kumamoto, Tadahiko & Ito, Akira: Recognizing user communicative intention in a consultant system with a natural language interface, IEEE, *Int. Workshop on Robot and Human Communication*, pp 311-6, 1992.

[Lara, Ham-Chande, et. al. 79] Lara, Luis Fernando; Ham-Chande, Roberto & García-Hidalgo, Ma. Isabel: *Investigaciones lingüísticas en lexicografía*, Jornadas 89, El Colegio de México, 1979.

[Lara, 97] Lara, Luis Fernando: *Teoría del diccionario monolingüe*, El Colegio de México, México, 1997.

[Lehnert, Sundheim, 91] Lehnert, Wendy & Sundheim, Beth: A performance evaluation of text-analysis technologies, *American Association of Artificial Intelligence Magazine*, pp 81-94, 1991.

[Lenat, Guha, et. al. 90] Lenat, Douglas; Guha, Ramanathan; Pitmann, Karen; Pratt, Dexter & Shepherd, Mary: CYC: toward programs with common sense. *Communications of the Association for Computing Machinery*, V.33 (8), pp 30-49, 1990.

[Lewis, Jones, 96] Lewis, David & Sparck Jones, Karen: Natural language processing for information retrieval, *Communications of the Association for Computing Machinery*, V.37, pp 93-100, 1996.

[Losee, Haas, 94] Losee, Robert & Haas, Stephanie: Sublanguage terms: dictionaries, usage and automatic classification, http://lib-www.lanl.gov/cmp-lg/9411001, 1994.

[Lucas, 93] Lucas, M: New directions in grammatical inference, *Colloquium on Grammatical Inference, Theory, Applications and Alternatives*, pp 1-7, 1993.

[Lyons, 77] Lyons, J.: *Semantics*, Cambridge University Press, 1977.

[Mahesh, Nirenburg, 95] Mahesh, Kovi & Nirenburg, Sergei: Semantic classification for practical natural language processing, *Classification Research Workshop*, pp 1-10, 1995.

[Marcus, Marcinkiewicz, et. al. 93] Marcus, Mitchell; Marcinkiewicz, Mary & Santorini, Beatrice: Building a large annotated corpus of English: The Penn Treebank, *Computational Linguistics* 19, pp 313-30, 1993.

[Màrquez, Rodríquez, 98] Màrquez, Lluís & Rodríquez, Horacio: Part-of-speech tagging using decision trees, *Lecture Notes in Artificial Intelligence*, 1398, pp 25-33, 1998.

[Matthews, Hearne, 91] Matthews, Geoffrey & Hearne, James: Clustering without a metric, IEEE, *Transactions on Pattern Analysis and Machine Intelligence*, V.13 (2), pp 175-84, 1991.

[McCarthy, 77] McCarthy, J.: Epistemological problems of artificial intelligence, *International Joint Conference on Artificial Intelligence*, pp 1038-44, 1977.

[McEnery, Wilson, 96] McEnery, T. & Wilson, A.: Corpus linguistics, Edinburgh University Press, pp 809-12, 1996.

[Merialdo, 91] Merialdo, Bernard: Tagging text with a probabilistic model, *Int. Conf. on Acoustics, Speech and Signal Processing*, V.2, pp 809-12, 1991.

[Michalski, 80] Michalski, R: Pattern recognition as a rule-guided inductive inference. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 2, pp 387-99, 1980.

[Minsky, Papert, 69] Minsky, M & Papert, S: *Perceptrons*, MIT Press, Cambridge, 1969.

[Moll, Arbib, 87] Moll, Robert; Arbib, Michael & Kloury A.: *An introduction to formal language theory*, Springer Verlag, 1987.

[Montague, 77] Montague, Richard: *Ensayos de filosofía formal*, Alianza Universidad, Madrid, 1977.

[Morales-Luna, 88] Morales-Luna, Guillermo: Lógica difusa, *Reporte técnico*, CINVESTAV, serie verde, No. 17, 1988.

[Moreno, 86] Moreno de Alba, Jose G.: *Morfología derivativa nominal en el español de México*, UNAM, 1986.

[Morris, Hirst, 91] Morris, Jane & Hirst, Graeme: Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Computational Linguistics* 17, pp 21-48, 1991.

[Nagata, Morimoto, 94] Nagata, Masaaki & Morimoto, Ysuyoshi: First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance, *Speech Communication*, V.15, 193-203, 1994.

[Neri, Saitta, 97] Neri, Filippo & Saitta, Lorenza: Machine learning for information extraction. *Lecture Notes in Artificial Intelligence*, 1299, pp 181-91, 1997.

[Noé, 98] Noé, Keiichi: Philosophical aspects of scientific discovery, a historical survey. *Lecture Notes in Artificial Intelligence*, 1394, pp 1-11, 1998.

[Oflazer, Tür, 97] Oflazer, Kemal & Tür, Gökhan: Morphological disambiguation by voting constraints, http://lib-www.lanl.gov/cmp-lg/9704011, 1997.

[Sámano, 83] Palacios de Sámano, Margarita: *Sintaxis de los relativos en el habla culta de la Ciudad de México*, UNAM, México, 1983.

[Palanivel, Helander, 91] Palanivel, Thiagarajan & Helander, Martin: Human-factors issues in dialog design, *Advances in Computer*, V.33, Academic Press, pp 115-55, 1991.

[Palmer, Hearst, 94] Palmer, David & Hearst, Marti: Adaptive sentence boundary disambiguation, http://lib-www.lanl.gov/cmp-lg/9411022, 1994.

[Patietsky-Shapiro, Frawley, 91] Patietsky-Shapiro, G. & Frawley, W.: Knowledge discovery in databases, (Eds.), *American Association of Artificial Intelligence*, AAAI Press, 1991.

[Pawlak, 82] Pawlak, Z.: Rough sets, *Int. J. Computer and Information Science*, V.11, pp 341-65, 1982.

[Pedersen, 93] Pedersen, Gert: A browser for bibliographic information retrieval based on a application of lattice theory, *SIGIR forum*, pp 270-9, 1993.

[Penrose, 96] Penrose, Roger: *La mente nueva del emperador*, Fondo de Cultura Económica, México, 1996.

[Pereira, Warren, 80] Pereira, F. & Warren, D.: Definite clause grammar for language analysis - a survey approach and a comparison with ATN, *Artificial Intelligence*, V.13, pp 231-78, 1980.

[Perrault, 89] Perrault, C.R.: Speech acts in multimodal dialog, *The structure of multimodal dialog*, Taylor, M.; Neel, F. & Bouwhuis, G. (Eds.), North Holland, pp 33-45, 1989.

[Polkowski, Skowron, 98] Polkowski, Leech & Skowron, Andrzej: Rough sets and current trends in computing, (Eds.) *Lecture Notes in Artificial Intelligence*, 1424, Springer Verlag, 1998.

[Pool-Westgaard, 90] Pool-Westgaard, Marianna: *Papeles temáticos y relaciones sintácticas: un estudio de argumentos verbales del español*, Tesis doctoral, El Colegio de México, 1990.

[Popowich, 94] Popowich, Fred: Improving the efficiency of a generation algorithm for shake and bake machine translation using head-driven phrase structure grammar, *Tech. Rep.* CMPT-TR 94-07, Simon Fraser University, pp 1-9, 1994.

[Poritz, 88] Poritz, Alan: Hidden Markov models: a guided tour, *Int. Conf. on Acoustics, Speech and Signal Processing*, V.1, pp 7-13, 1988.

[Quafafou, Zytkov, 98] Quafafou, Mohamed & Zytkow, Jan: Principles on knowledge discovery in databases and datamining, *Lecture Notes in Artificial Intelligence*, 1510, Springer Verlag, 1998.

[Quinlan, 93] Quinlan, Ross: *C4.5: programs for machine learning*, Morgan Kaufmann, California, 1993.

[Rada, Mili, et. al. 89] Rada, Roy; Mili, Hafedh; Bicknell, Ellen & Blettner, María: Development and application of a metric on semantic nets, IEEE, *Transactions on Systems, Man and Cybernetics*, V.19 (1), pp 17-30, 1989.

[Rinaldo, Strutz, et. al. 94] Rinaldo, Frank & Strutz, Robert & Evens, Martha: Developing a lexicon for automating knowledge acquisition, *Second Int. Conf. on Expert Systems for Development*, pp 74-8, 1994.

[Rivière, 86] Rivière, Angel: *Razonamiento y representación*, Siglo XXI de España, 1986.

[Rosch, 73] Rosch, E.: On the internal structure of perceptual and semantic categories. *Cognitive development and the acquisition of language*, Moore, T. (Ed.), Academic Press, Nueva York, 1973.

[Rodríguez, Carretero, 97] Rodríguez, Santiago & Carretero, Jesús: Building a Spanish speller, http://www.datsi.fi.upm.es, 1997.

[Rousseau, 98] Rousseau, Ronald: Jaccard similarity leads to the Marczewiski-Steinhaus topoplogy for information retrieval, *Information Processing and Management*, V.34, pp 87-94, 1998.

[Rudnicky, Hauptmann, et. al. 94] Rudnicky, Alexander; Hauptmann, Alexander & Lee, Kai-Fu: Survey of Current speech technology. *Communications of the Association for Computing Machinery*, V.37 (3), pp 52-7, 1994.

[Rumelhart, Widrow, et. al. 94] Rumelhart, David; Widrow, Bernard & Lehr, Michael: The basic ideas in neural networks, *Communications of the Association for Computing Machinery*, V.37 (3), pp 87-92, 1994.

[Sakakibara, 90] Sakakibara, Yasubumi: Learning context free grammars from structural data in polynomial time, *Theoretical Computer Science*, V.76, North Holland, pp 233-42, 1990.

[Samuel, Carberry, et. al. 98] Samuel, Ken; Carberry, Sandra & Vijay-Shanker, K.: Dialogue act tagging with transformation-based learning, http://lib-www.lanl.gov/cmp-lg/9806006, 1998.

[Samuelsson, Voutilainen, 97] Samuelsson, Christer & Voutilainen, Atro: Comparing a linguistic and stochastic tagger, http://lib-www.lanl.gov/cmp-lg/9706005, 1997.

[Schütze, 93] Schütze, Hinrich: Distributed syntactic representation with an application to part-of-speech tagging, IEEE *Int. Conf. on Neural Networks*, V.3, pp 1504-9, 1993.

[Sertula, Kourie, 98] Sertula, Lebelo & Kourie, Derrick: Sentence analysis using a concept lattice, *Lecture Notes in Artificial Intelligence*, 1529, Springer Verlag, pp 225-35, 1998.

[Shapiro, Warren, 93] Shapiro, Ehud & Warren, David: The fifth generation computer: personal perspectives, *Communications of the Association for Computational Machinery*, V.36 (3), pp 46-8, 1993.

[Shapiro, 94] Shapiro, Linda G.: Relational matching, *Handbook of pattern recognition and image processing: computer vision*. Academic Press, pp 475-96, 1994.

[Slator, Anderson, et. al. 86] Slator, Brian M.; Anderson, Mattew P. & Conley, Walt: Pigmalion at the interface, *Communications of the Association for Computing Machinery*, V.29 (7), pp 599-604, 1986.

[Small, 87] Small, Steven.: A distributed word-based approach to parsing, en *Natural language processing systems*, Bolc, Leonard (Ed.), pp 161-200, Springer Verlag, 1987.

[Smeaton, 97] Smeaton, Alan: Information retrieval: still butting heads with natural language processing?, *Lecture Notes in Artificial Intelligence*, No. 1299, Springer Verlag, pp 115-38, 1997.

[Song, Johannesson, et. al. 96] Song, William; Johannesson, Paul & Bubenko, Janis: Semantic similarity relations and computation in schema integration, *Data and Knowledge and Engineering* V.19, pp 65-97, 1996.

[Stanfill, Waltz, 86] Stanfill, C. & Waltz, D.: Towards memory-based reasoning. In *Communications of the Association for Computational Machinery*, V.29 (12), pp 1213-1228, 1986.

[Strzalkowski, 95] Strzalkowski, Tomek: Natural language information retrieval, *Information Processing and Management*, V.31 (3), 1995, Elsevier, pp 397-417, 1995.

[Suematsu, 94] Suematsu, Hiroshi: Current status of the EDR electronic dictionary project and its evaluation, research, *New Generation Computing*, V.12, pp 311-6, 1994.

[Sun-Fu, 74] Sun-Fu, King: *Syntactic pattern recognition*, Academic Press, 1974.

[Suppes, Böttner, et. al. 96] Suppes, Patrick; Böttner, Michael & Lang, Lin: Machine learning comprehension grammars, *Computational Linguistics* 22, pp 329-50, 1996.

[Takenobu, Makoto, et. al. 95] Takenobu, Tokunaga; Makoto, Iwayama & Hazumi, Tanaka: Automatic thesaurus construction based on grammatical relations, *International Joint Conference on Artificial Intelligence*, pp 1308-13, 1995.

[Tamura, Tanaka, 73] Tamura, Shinishi & Tanaka, Kokichi: Learning of fuzzy formal language, IEEE, *Transactions on Systems, Man and Cybernetics*, pp 98-102, 1973.

[Thayse, 89] Thayse, André: *From modal logic to deductive databases*, (Ed.), John Wiley and Sons, 1989.

[Thomas, Short, 96] Thomas, Jenny & Short, Mick: *Using corpora for language research*, Longman, 1996.

[Turner, 87] Turner, C.E.: *Higiene del individuo y la comunidad*, La Prensa Médica Mexicana, México, 1987.

[Tversky, Lee, 98] Tversky, Barbara & Lee, Paul: How space structures language, *Lecture Notes in Artificial Intelligence*, 1404, Freska C.; Habel C. & Wender K. (Eds.), Springer Verlag, pp 157-75, 1998.

[Tzoukerman, Klavans, et. al. 97] Tzoukerman, Evelyne; Klavans, Judith L. & Jacquemin, Christian: Effective use of natural language processing techniques for automatic conflation of multi-word terms: the role of derivational morphology, part of speech tagging, and shallow parsing, Proc. of 20th. Annual Int. *Association for Computing Machinery, Special Interest Group on Information Retrieval*, Conf. on Information Retrieval, pp 148-55, 1997.

[Valiant, 84] Valiant, L.: A theory of the learneable, *Communications of the Association for Computing Machinery*, V.27, pp 1134-42, 1984.

[Veenstra, 98] Veenstra, Jorn: Fast NP chunking using memory-based learning techniques, *Proc. of Benelearn*, The Netherlands, pp 71-79, 1998.

[Voutilainen, Padró, 97] Voutilainen, Atro & Padró, Lluís: Developing a hybrid NP parser, http://lib-www.lanl.gov/cmp-lg/9704009, 1997.

[Walker, Moore, 97] Walker, Marilyn & Moore, Johanna: Empirical studies in discourse, *Computational Linguistics* 23, pp 1-12, 1997.

[Walts, Porter, et. al. 97] Walts, R.; Porter, A.; Cunningham, S & Zhu, S.: TOAS: intelligent mining; analysis of NLP and computational linguistics, Foundation of intelligent systems, Raś, Zbigniew & Skowron, Andrzej (Eds.), *Lecture Notes in Artificial Intelligence*, 1325, pp 322-34, 1997.

[Wilks, Stevenson, 97] Wilks, Yorick & Stevenson, Mark: Sense tagging: semantic tagging with a lexicon, *Proc. of the SIGLEX Workshop*, http://lib-www.lanl.gov/cmp-lg/9705016, 1997.

[Winograd, 76] Winograd, Terry: *Understanding natural language*, Academic Press, 1976.

[Wyard, 92] Wyard, J.P.: Connectionist natural language processing: an introduction, *Neural networks for vision, speech and natural language*, Linggard, R.; Myers, D. & Nightingale, C. (Eds.), Chapman & Hall, pp 193-288, 1992.

[Young, Hauptmann, et. al. 89] Young, Sherly; Hauptmann, Alexander; Ward, Wayne; Smith, Edward & Werner, Philip: High level knowledge sources in usable speech recognition systems. *Communications of the Association for Computing Machinery*, V.32, pp 183-94, 1989.

[Young, Bloothooft, 97] Young, S. & Bloothooft, G.: *Corpus-based methods in language and speech processing*, Academic Press, 1997.

[Zavrel, Daelemans, et. al. 97] Zavrel, Jakub; Daelemans, Walter; Veenstra, Jorn: Resolving PP-attachment ambiguities with MBL, *CoNLL*, http://ilk.kub.nl/, 1997.

[Zelle, Mooney, 95] Zelle, John & Mooney, Raymond: Comparative results on using inductive logic programming for corpus-based parser construction, *Lecture Notes in Artificial Intelligence*, 1040, Springer Verlag, pp 355-69, 1995.

[Zernik, 91] Zernik, Uri: Introduction, *Lexical acquisition: exploting online resources to build a lexicon*, Zernik, Uri (Ed.), Lawrence Erlabum Association, 1991.

[Zgusta, 71] Zgusta, Ladislav: *Manual of lexicography*, Mouton, 1971.

[Zhai, 97] Zhai, Chengxiang: Fast stochastic parsing of noun phrases for document indexing, http://lib-www.lanl.gov/cmp-lg/9702009, 1997.

[Zhengfu 94] Zhengfu, Geoffrey: The semantic vector space model, a text representation and searching techniques, *Proc. of the 27-th annual Hawaii Int. Conf. on Systems Sciences*, pp 928-37, 1994.

[Ziarko, 91] Ziarko, W.: The discovery, analysis, and representation of data dependencies in databases, *Knowledge discovery in databases*, Patietsky-Shapiro, G. & Frawley, W. (Eds.), *American Association of Artificial Intelligence*, AAAI Press, pp 195-209, 1991.

[van-Frassen, 87] van-Frassen, Bas: *Semántica formal y lógica*, UNAM, México, 1987.

[Bosch, Daelemans, et. al. 96] van-den Bosch, Antal; Daelemans, Walter; Weijters, Ton: Morphological analysis as classification: an inductive-learning approach, http://lib-www.lanl.gov/cmp-lg/9607021, 1996.

Apéndice A

Especificación del sistema de etiquetamiento

Para formalizar el método de etiquetamiento desarrollado tendremos que precisar las nociones expuestas en el capítulo 3.

Sea $\mathcal{V}$ el vocabulario y T un texto, i.e., una secuencia formada por elementos de $\mathcal{V}$, $T = [v_i]_{i \leq n_T}$. Llamaremos a $\mathcal{V}_T \subset \mathcal{V}$ el vocabulario de T, $\mathcal{V}_T = \{v | v \text{ ocurre en } T\}$.

Trataremos con subsecuencias especiales de T. Para un texto T y $p, q \in [0, n_T - 1](p + q + 1 \leq n_T)$, definimos la *secuencia p:q alrededor* del i-ésimo elemento de T: $[i]_{T,\langle p:q\rangle} = v_{i-p}, \ldots, v_i, \ldots, v_{i+q}$, y la *secuencia p::q alrededor* del i-ésimo elemento de T: $[i]_{T,\langle p::q\rangle} v_{i-p}, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{i+q}$. Refiriéndonos a terminaciones de elementos de T, tenemos la *secuencia p:q* de *terminación k* alrededor de i en T: $[i]_{T,\langle p:q\rangle\_k} = v_{i-p\_k}, \ldots, v_{i\_k}, \ldots, v_{i+q\_k}$. Análogamente, tenemos la correspondiente secuencia p::q de terminación k alrededor de i en T: $[i]_{T,\langle p::q\rangle\_k}$.

Es claro que el i-éismo elemento de T es $[i]_{T,\langle 0:0\rangle}$. Además, T puede verse como una subsecuencia de sí mismo: $T = [1]_{T,\langle 0:n_T-1\rangle} = [n_T]_{T,\langle n_T-1:0\rangle}$, o bien como una concatenación de subsecuencias: $[1]_{T,\langle 0:0\rangle} \cdots [n_T]_{T,\langle 0:0\rangle}$. A diferencia de las secuencias p:q, las secuencias p::q no ocurren en T como tal, pues de éstas nos interesa más bien el patrón que representan.

Sea T un texto y PD el conjunto de etiquetas. Un *etiquetamiento* de T con PD es una secuencia $T^* = [m_i]_{i \leq n_{T^*}}$ tal que $n_T = n_{T^*}$, $\mathcal{V}_{T^*} \subset \mathrm{PD}$, y a cada elemento $[i]_{T,\langle 0:0\rangle}$ de T le corresponde el elemento $[i]_{T^*,\langle 0:0\rangle}$ de la secuencia T^*.

Sea T un texto etiquetado. La *secuencia combinada* de (T, T^*) con parámetros $p : q, i, r : s$ y k, es una concatenación de dos secuencias: $X_T(p : q, i, r : s, k) = [i]_{T,\langle p:q\rangle\_k}[i]_{T^*,\langle r:s\rangle}$. De manera semejante se definen las secuencias combinadas: $X_T(p :: q, i, r : s, k)$, $X_T(p : q, i, r :: s, k)$, y $X_T(p :: q, i, r :: s, k)$.

La función $\mathrm{index}_T : \mathcal{V}_T \to 2^{[1,n_T]}$ determina todas las *posiciones* en las que ocurre v en T, $\mathrm{index}_T(v) = \{i | v \text{ ocurre en la posición } i \text{ de } T\}$.

Si T es un texto etiquetado podemos construir la *colección de entrenamiento* para el predicado $\mathcal{H}$ con los parámetros $p, q, r, s, k \in \mathbb{N}$:

$$\mathrm{INST}_{T,\mathcal{H}}(p, q, r, s, k) = \bigcup_{\substack{\mathcal{H}(v) \\ i \in \mathrm{index}(v)}} X_T(p :: q, i, r : s, k).$$

Representaremos con $\mathring{\pi}_k(X)$ la secuencia que se obtiene por la *eliminación de la componente* k de la secuencia X.

Observemos que si $X_T(p :: q, i, r : s, k) \in \mathrm{INST}_{T,\mathcal{H}}(p, q, r, s, k)$, entonces la marca de v_i en T es $\pi_{p+q+r+1}(X)$.

Suponiendo formada una colección de entrenamiento para T y X un elemento de la colección, llamaremos $\mathrm{CL_{INST}}(X)$ a $\pi_{p+q+r+1}(X)$, y $\mathrm{CL'_{INST}}(X)$ a su complemento respecto a la instancia original, $\mathring{\pi}_{p+q+r+1}(X)$.

Una instancia $Y \in \mathrm{INST}$ es un *prefijo* de longitud l de la secuencia X si $\pi_i Y = \pi_i X (1 \leq i \leq l)$, y $\pi_{l+1}Y \neq \pi_{l+1}X$. Sea $\mathrm{INST}_l(X)$ el *conjunto de prefijos* de longitud l de la secuencia X, y $l_0(X) = \underset{l}{\mathrm{argmax}}\, \mathrm{INST}_l(X)$, la *máxima longitud de los prefijos* de X. Si $l_0(X) < |X|$ diremos que X tiene un *valor desconocido* en la componente $l_0 + 1$. De no tener X valores desconocidos, entonces existe $Y \in \mathrm{INST}$ que satisface $X = \mathrm{CL'_{INST}}(Y)$ pues $\Delta(\mathrm{CL'_{INST}}(Y), X) = 0$. Si X tiene un valor desconocido en la componente $l_0(X)+1$ podemos calcular la clase de X a partir de sus prefijos de longitud l_0, eligiendo de entre ellos la clase más frecuente:

$$\mathrm{CL0_{INST}}(X) = \underset{m \in \mathrm{PD}}{\mathrm{argmax}}\, \#\{Y | Y \in \mathrm{INST}_{l_0}(X), \mathrm{CL_{INST}}(Y) = m\}$$

Con el fin de que los prefijos contengan máxima información, ordenaremos sus rasgos de acuerdo con la ganancia normalizada. Representemos el hecho de ordenar descendentemente la secuencia $x_1, \ldots, x_n$ con $\mathrm{per}([1, \ldots, n], [x_1, \ldots, x_n]) = ([j_1, \ldots, j_n], [x_{j_1} \ldots, x_{j_n}])$, donde $j_1, \ldots, j_n$ es la permutación de los índices originales $(1, \ldots, n)$, y $x_{j_1} \geq x_{j_2} \geq \ldots \geq x_{j_n}$. La función que ordena una secuencia $y_1, \ldots, y_n$ según la permutación $[i_1, \ldots, i_n]$, será indicada con aper:

$$\mathrm{aper}_{[i_1, \ldots, i_n]}(y_1, \ldots, y_n) = [y_{i_1}, \ldots, y_{i_n}].$$

El ordenamiento de una colección de entrenamiento INST, según la ganancia normalizada de sus atributos es:

$$\mathrm{INSTO} = \bigcup_{Y \in \mathrm{INST}} \mathrm{aper}_{[i_1, \ldots, i_m]}(Y),$$

donde $i_1, \ldots, i_m$ cumple $\mathrm{per}([1, \ldots, m], [w_1, \ldots, w_m]) = ([i_1, \ldots, i_m], [w_{i_1}, \ldots, w_{i_m}])$, y w_i es el valor de la ganancia normalizada para el atributo i de la colección, en concordancia con la ec. 2.8.

Sea DICCD la función que asigna a las palabras definidas frecuentes ($\mathcal{V}_1 \subset \mathcal{V}$) su parte del discurso, $\mathrm{DICCD} : \mathcal{V}_1 \rightarrow \mathrm{PD}$, y DICCS, la función que asigna a los signos de puntuación sus etiquetas, i.e.: $\mathrm{DICCS} : \mathcal{V}_2 \rightarrow \mathrm{PD}$, $\mathcal{V}_2 \subset \mathcal{V}$. Además sean $\mathrm{DICCV}, \mathrm{TERD}, \mathrm{DICCA}, \mathrm{NOMP}, \mathrm{VERBO}_r, \mathrm{VERBO}_i \subset \mathcal{V}$, los diccionarios de formas verbales frecuentes, terminaciones que determinan sustantivos, palabras ambiguas frecuentes, nombres propios, verbos regulares en infinitivo y verbos irregulares en infinitivo, respectivamente. Distinguimos, además, un subconjunto de $\mathcal{M}_1 \cup \mathcal{M}_2$, cuyos miembros pertenecen a terminaciones de nominales: TERC.

$X_T(p : q, i, r : s, k)$ es una secuencia particular alrededor del i-ésimo elemento de (T, T^*). Indicaremos con $X_T(p : q, -, r : s, k)$ cualquier secuencia posible alrededor de un elemento de T, con parámetros $p : q$, $r : s$ y k. Si una secuencia combinada particular $X_T(p : q, i, r :: s, k)$ satisface que

$$\Pr(X_T(p : q, -, r : s, k) = X_T(p : q, i, r : s, k) | X_T(p : q, -, r :: s, k) = X_T(p : q, i, r :: s, k)) = 1,$$

podemos afirmar que según el texto etiquetado T, la clase de cualquier vocablo $v = [i]_{T,\langle 0:0 \rangle}$ es $\mathrm{CL_{INST}}(X_T(p : q, i, r : s, k))$, siempre que ocurra en la secuencia $X_T(p : q, -, r :: s, k)$.

Identificaremos a $\mathcal{M}_1$ y a $\mathcal{M}_2$ con los gramemas de inflexión regular, y los gramemas de inflexión irregular. La función que asigna a cada gramema un conjunto de gramemas para verbos infinitivos son: $\mathrm{inflex}_1 : \mathcal{M}_1 \rightarrow 2^{\mathcal{M}_1}$, e $\mathrm{inflex}_2 : \mathcal{M}_2 \rightarrow 2^{\mathcal{M}_2}$.

Se concibe el proceso de etiquetamiento de un texto T como una manera de construir la secuencia T^*. Este proceso será descrito en varios pasos. En el paso de etiquetamiento j consideramos el etiquetamiento parcial T^*_j para T. Además, asociados al paso de etiquetamiento j se establecen los predicados: $\mathcal{A}_{j,1}, \ldots, \mathcal{A}_{j,n_j}$. Designemos con $\mathcal{B}_j(i)$ al predicado $([i]_{T^*_{j-1},\langle 0:0 \rangle} = \mathtt{IND})$.

Además de los recursos de información antes descritos, hemos de precisar que es necesario un texto etiquetado que proporcione las colecciones de entrenamiento y al que llamaremos corpus de trabajo. El corpus de trabajo empleado fue una parte del Corpus del Español Mexicano Contemporáneo del Diccionario del Español de

México, de aquí en adelante T_C. La ejecución de los pasos de etiquetamiento requiere de tres colecciones de entrenamiento:

Ambigua frecuente

$$\mathrm{INST5} = \mathrm{INST}_{T_C, v\in\mathrm{DICCA}}(2,2,1,1,6)$$

Verbo/nominal

$$\mathrm{INST6} = \mathrm{INSTO}_{T_C, v\_k\in\mathrm{TERC}}(2,2,1,1,6)$$

Desconocida [1]

$$\mathrm{INST7} = \mathrm{INSTO}_{T_C, v\in\mathrm{DESC}}(2,0,2,0,6)$$

Asociamos a los pasos de etiquetamiento indicados los predicados que enseguida se exponen:

Paso 1

$$\begin{aligned}
\mathcal{A}_{1,1}(v,i) &= (v_i \in \mathcal{V}_1)\\
\mathcal{A}_{1,2}(v,i) &= (v_i \in \mathrm{DICCV})\\
\mathcal{A}_{1,3}(v,i) &= (v_i \in \mathrm{NOMP})\\
\mathcal{A}_{1,4}(v,i) &= (v_i \in \mathcal{V}_2)\\
\mathcal{A}_{1,5}(v,i) &= (v_{i_k} \in \mathrm{TERD}).
\end{aligned}$$

Paso 2

$$\begin{aligned}
\mathcal{A}_{2,1}(v,i) &= (X_{T_1}(0:0,i,1::0,\infty) \in \{\mathtt{la},\mathtt{las},\mathtt{lo},\mathtt{los}\}\mathtt{PREP}),\\
\mathcal{A}_{2,2}(v,i) &= (X_{T_1}(0::1,i,1::0,\infty) \in \{\mathtt{de},\mathtt{del}\}\mathtt{CONT}),\\
\mathcal{A}_{2,3}(v,i) &= (X_{T_1}(1::1,i,0::0,\infty) \in \{\mathtt{el},\mathtt{al}\}\{\mathtt{de},\mathtt{del}\}),\\
\mathcal{A}_{2,4}(v,i) &= (X_{T_1}(0::1,i,2::0,\infty) \in \{\mathtt{de},\mathtt{del}\}\mathtt{PREP}\ \mathtt{ART}).
\end{aligned}$$

Paso 3

$$\begin{aligned}
\mathcal{A}_{3,1}(v,i) &= (v_{i_k} \in \mathcal{M}_1) \wedge (\exists t \in \mathrm{inflex}_1(v_{i_k})) \wedge (v_{i_k}t \in \mathrm{VERBO}_1),\\
\mathcal{A}_{3,2}(v,i) &= (v_{i_k} \in \mathrm{TERC}).
\end{aligned}$$

Paso 4

$$\begin{aligned}
\mathcal{A}_{4,1}(v,i) &= (v_{i_k} \in \mathcal{M}_2) \wedge (\exists t \in \mathrm{inflex}_2(v_{i_k})) \wedge (v_{i_k}t \in \mathrm{VERBO}_2),\\
\mathcal{A}_{4,2}(v,i) &= (v_{i_k} \in \mathrm{TERC}).
\end{aligned}$$

Paso 5

$$\mathcal{A}_{5,2}(v,i) = (v_i \in \mathrm{DICCA}).$$

Paso 6

$$\begin{aligned}
\mathcal{A}_{6,1}(v,i) &= (v_i = \mathtt{VERP}),\\
\mathcal{A}_{6,2}(Y,i) &= ((\exists Y \in \mathrm{INST6})(X_{T_5}(2::2,i,1::1,6) = \mathrm{CL}'_{\mathrm{INST6}}(Y))).
\end{aligned}$$

[1]El conjunto DESC se constituye por todas las palabras de T_C que no pueden ser etiquetadas en los pasos 1 a 6.

Paso 7

$$\begin{aligned}\mathcal{A}_{7,1}(Y,i) &= ((\exists Y \in \mathrm{INST7})(X_{T_6}(2::0,i,2::0,6) = \mathrm{CL}'_{\mathrm{INST7}}(Y))),\\ \mathcal{A}_{7,2}(i) &= (l_0(X_{T_6}(2::0,i,2::0,6)) \leq 2).\end{aligned}$$

Con lo anterior podemos definir la especificación formal de cada uno de los pasos de etiquetamiento.

Dado un texto T, las marcas de T obtenidas por el etiquetador basado en instancias es el texto T_7^* que se obtiene siguiendo:

Paso 1

$$\begin{aligned}&(\forall v \in \mathcal{V}_T)(\forall i \in \mathrm{index}_T(v))(\\ (\mathcal{A}_{1,1}(v,i) &\rightarrow [i]_{T_1^*,\langle 0:0\rangle} = \mathrm{DICCD}(v)) \vee\\ (\mathcal{A}_{1,2}(v,i) &\rightarrow [i]_{T_1^*,\langle 0:0\rangle} = \mathtt{VERB}) \vee\\ (\mathcal{A}_{1,3}(v,i) &\rightarrow [i]_{T_1^*,\langle 0:0\rangle} = \mathtt{NP}) \vee\\ (\mathcal{A}_{1,4}(v,i) &\rightarrow [i]_{T_1^*,\langle 0:0\rangle} = \mathrm{DICCS}(v)) \vee\\ (\mathcal{A}_{1,5}(v,i) &\rightarrow [i]_{T_1^*,\langle 0:0\rangle} = \mathtt{NOM}) \vee\\ (\sim (\mathcal{A}_{1,1}(v,i) \vee \mathcal{A}_{1,2}(v,i) \vee \mathcal{A}_{1,3}(v,i) \vee \mathcal{A}_{1,4}(v,i) \vee \mathcal{A}_{1,5}(v,i)) &\rightarrow [i]_{T_1^*,\langle 0:0\rangle} = \mathtt{IND})).\end{aligned}$$

Paso 2

$$\begin{aligned}&(\forall v \in \mathcal{V}_T)(\forall i \in \mathrm{index}_T(v))(\\ (\mathcal{A}_{2,1}(v,i) \wedge \mathcal{B}_2(i) &\rightarrow [i]_{T_2^*,\langle 0:0\rangle} = \mathtt{ART}) \vee\\ ((\mathcal{A}_{2,2}(v,i) \vee \mathcal{A}_{2,3}(v,i) \vee \mathcal{A}_{2,4}(v,i)) \wedge \mathcal{B}_2(i) &\rightarrow [i]_{T_2^*,\langle 0:0\rangle} = \mathtt{NOM}) \vee\\ (\sim (\mathcal{A}_{2,1}(v,i) \vee \mathcal{A}_{2,2}(v,i) \vee \mathcal{A}_{2,3}(v,i) \vee \mathcal{A}_{2,4}(v,i)) &\rightarrow [i]_{T_2^*,\langle 0:0\rangle} = [i]_{T_1^*,\langle 0:0\rangle})).\end{aligned}$$

Paso 3

$$\begin{aligned}&(\forall v \in \mathcal{V}_T)(\forall i \in \mathrm{index}_T(v))(\\ (\mathcal{A}_{3,1}(v,i) \wedge (\sim \mathcal{A}_{3,2}(v,i)) \wedge \mathcal{B}_3(i) &\rightarrow [i]_{T_3^*,\langle 0:0\rangle} = \mathtt{VERB}) \vee\\ (\mathcal{A}_{3,2}(v,i) \wedge \mathcal{B}_3(i) &\rightarrow [i]_{T_3^*,\langle 0:0\rangle} = \mathtt{VERP}) \vee\\ (\sim (\mathcal{A}_{3,1}(v,i) \vee \mathcal{A}_{3,2}(v,i)) &\rightarrow [i]_{T_3^*,\langle 0:0\rangle} = [i]_{T_2^*,\langle 0:0\rangle})).\end{aligned}$$

Paso 4

$$\begin{aligned}&(\forall v \in \mathcal{V}_T)(\forall i \in \mathrm{index}_T(v))(\\ (\mathcal{A}_{4,1}(v,i) \wedge (\sim \mathcal{A}_{4,2}(v,i)) \wedge \mathcal{B}_4(i) &\rightarrow [i]_{T_4^*,\langle 0:0\rangle} = \mathtt{VERB}) \vee\\ (\mathcal{A}_{4,2}(v,i) \wedge \mathcal{B}_4(i) &\rightarrow [i]_{T_4^*,\langle 0:0\rangle} = \mathtt{VERP}) \vee\\ (\sim (\mathcal{A}_{4,1}(v,i) \vee \mathcal{A}_{4,2}(v,i)) &\rightarrow [i]_{T_4^*,\langle 0:0\rangle} = [i]_{T_3^*,\langle 0:0\rangle})).\end{aligned}$$

Paso 5

$$\begin{aligned}&(\forall v \in \mathcal{V}_T)(\forall i \in \mathrm{index}_T(v))(\\ (\mathcal{A}_{5,1}(v,i) \wedge \mathcal{B}_5(i) &\rightarrow [i]_{T_5^*,\langle 0:0\rangle} = \mathrm{CL}_{\mathrm{INST5}}(\operatorname*{argmin}_{Y \in \mathrm{INST5}} \Delta(\mathrm{CL}'_{\mathrm{INST5}}(Y), X_{T_4}(2::2,i,1::1,6)))) \vee\\ (\sim \mathcal{A}_{5,1}(v,i) &\rightarrow [i]_{T_5^*,\langle 0:0\rangle} = [i]_{T_4^*,\langle 0:0\rangle})).\end{aligned}$$

Paso 6

$$\begin{aligned}&(\forall v \in \mathcal{V}_T)(\forall i \in \mathrm{index}_T(v))(\\ (\mathcal{A}_{6,1}(v,i) \wedge \mathcal{A}_{6,2}(Y,i) &\rightarrow [i]_{T_6^*,\langle 0:0\rangle} = \mathrm{CL}_{\mathrm{INST6}}(Y)) \vee\\ (\mathcal{A}_{6,1}(v,i) \wedge (\sim \mathcal{A}_{6,2}(Y,i)) &\rightarrow [i]_{T_6^*,\langle 0:0\rangle} = \mathrm{CL0}_{\mathrm{INST6}}(X_{T_5}(2::2,i,1::1,6))) \vee\\ (\sim \mathcal{A}_{6,1}(v,i) &\rightarrow [i]_{T_6^*,\langle 0:0\rangle} = [i]_{T_5^*,\langle 0:0\rangle})).\end{aligned}$$

Paso 7

$$
\begin{array}{rcl}
(\forall v \in \mathcal{V}_T)(\forall i \in \mathrm{index}_T(v))(& & \\
(\mathcal{A}_{7,1}(Y,i) \wedge \mathcal{B}_7(v,i) & \rightarrow & [i]_{T_7^*,\langle 0:0\rangle} = \mathrm{CL}_{\mathrm{INST7}}(Y)) \vee \\
(\sim \mathcal{A}_{7,1}(Y,i) \wedge \mathcal{A}_{7,2}(i) \wedge \mathcal{B}_7(v,i) & \rightarrow & [i]_{T_7^*,\langle 0:0\rangle} = \mathrm{CL}_{\mathrm{INST7}}(\operatorname*{argmin}_{Y \in \mathrm{INST7}} \Delta'(\mathrm{CL}'_{\mathrm{INST7}}(Y), X_{T_6}(2::0,i,2::0,6)))) \vee \\
(\sim \mathcal{A}_{7,1}(Y,i) \wedge \sim \mathcal{A}_{7,2}(i) \wedge \mathcal{B}_7(v,i) & \rightarrow & [i]_{T_7^*,\langle 0:0\rangle} = \mathrm{CL0}_{\mathrm{INST7}}(X_{T_6}(2::0,i,2::0,6))) \vee \\
(\sim \mathcal{B}_7(i) & \rightarrow & [i]_{T_7^*,\langle 0:0\rangle} = [i]_{T_6^*,\langle 0:0\rangle})).
\end{array}
$$

Índice Temático

Índice de Autores