Notas sobre el enfoque empírico en el procesamiento del lenguaje

Héctor Jiménez Salazar
Fac. de Ciencias Físico Matemáticas, UAP
hjimenez@fcfm.buap.mx

septiembre, 2000

1 Procesamiento del lenguaje natural

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[1] 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[2].

Las características del lenguaje natural plantean que la construcción de los sistemas de PLN tengan un tratamiento especial dentro del conjunto de problemáticas complejas[3] 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.

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

[1] 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%).

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

[3] 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*.

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 fenómeno de la "informatización" de la sociedad[4]. 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

[4]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].

- *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.

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.

En la siguiente sección haremos un brevísimo recuento del desarrollo 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].

2 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*[5]. Aunque, en los compiladores, la semántica se reducía a la generación de código objeto[6], 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[7], pocos

[5]Hubo una influencia notable de la corriente generativista, de hecho la computación retomó para sus fundamentos la jerarquía de Chomsky como una forma de clasificar los lenguajes y establecer el poder computacional de diversos mecanismos [Moll, Arbib, 87].

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

[7]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

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*[8], 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[9].

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[10], 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[11] 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[12].

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*[13], 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 especializado, el conocimiento subyacente en la teoría de ese dominio[14]. pues parte de este conocimiento se mezclaba con *conocimiento del mundo*[15], La reunión y sistematización de la información necesaria para

distinción.

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

[9]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].

[10]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].

[11]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".

[12]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].

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

[14]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].

[15]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].

resolver un tipo de problemas de un dominio o *base de conocimientos*[16] 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[17], 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 [18]. 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[19]. 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".

3 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[20] [Franz, 96], son necesarios para el *parsing*, pues cada palabra requiere de su categoría gramatical, para poder aplicar la 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 exper-

[16] *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".

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

[18] 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.

[19] 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."

[20] 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.

tos 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[21]. 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[22] atrajo gran cantidad de proyectos para trabajar en esa línea[23]. Más adelante, las condiciones tecnológicas favorables[24] 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[25], 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[26].

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[27]. 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][28]. La forma de construir un léxico o, en particular, una base de datos léxica[29], 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 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

[21] 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.

[22] 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].

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

[24] 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.

[25] 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%.

[26] 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].

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

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

[29] 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].

de botella" de los sistemas de PLN: la semántica[30]. Se vio [Ide, Véronis, 89] que, sin restar su utilidad, los MRD por sí solos eran insuficientes[31], 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"[32]. 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[33]. Esta es una forma de ver al aprendizaje automático como eje del PLN.

4 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 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

[30] 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].

[31] 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].

[32] 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.

[33] 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".

distinción define dos tendencias en la inteligencia artificial, la "fuerte" y la "débil", que clasifican los enfoques para resolver problemas[34].

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"[35] 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[36]. 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[37], 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*[38], 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[39], el conexionismo[40] y la lógica difusa[41].

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[42]. 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[43] 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 con el PLN es amplia pues ha habido necesidad, por ejemplo, de construir *tesoros* [44] [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

[34]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.

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

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

[37]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*.

[38]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.

[39]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.

[40]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].

[41]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].

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

[43]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.

[44]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], .

[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[45]. 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[46]. 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[47].

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.

Referencias

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

[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.

[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.

[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.

[45] 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].

[46] 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].

[47] 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.

[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.

[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, 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/~brill/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.

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

[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, 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, 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/g̃ibbon/EAGLES, 1993.

[EAGLES, 96] EAGLES: Recommendations for the morphosyntactic annotation of corpora, http://coral.lili.uni.bielefeld.de/g̃ibbon/EAGLES, 1996.

[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, 96] García-Fajardo, Josefina: *De los sonidos a los sentidos, introducción al lenguaje*, Trillas, México, 1996.

[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.

[Grishman, 97] Grishman, Ralph: Information extraction, techniques and challenges, *Lecture Notes in Artificial Intelligence*, 1299, Springer Verlag, pp 10-27, 1997.

[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.

[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.

[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, 93] Krovetz, Robert: Viewing morphology as an inference process, *Association for Computing Machinery, Special Interest Group on Information Retrieval*, pp 191-202, 1993.

[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.

[Lucas, 93] Lucas, M: New directions in grammatical inference, *Colloquium on Grammatical Inference, Theory, Applications and Alternatives*, pp 1-7, 1993.

[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.

[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.

[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.

[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.

[Palanivel, Helander, 91] Palanivel, Thiagarajan & Helander, Martin: Human-factors issues in dialog design, *Advances in Computer*, V.33, Academic Press, pp 115-55, 1991.

[Patietsky-Shapiro, Frawley, 91] Patietsky-Shapiro, G. & Frawley, W.: Knowledge discovery in databases, (Eds.), *American Association of Artificial Intelligence*, AAAI Press, 1991.

[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.

[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.

[Quafafou, Zytkov, 98] Quafafou, Mohamed & Zytkow, Jan: Principles on knowledge discovery in databases and datamining, *Lecture Notes in Artificial Intelligence*, 1510, Springer Verlag, 1998.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

Índice Temático

Índice de Autores