Resumen de la Tesis Doctoral "Una solución a la ambigüedad semántica basada en Métodos Conexionistas para un problema de traducción automática ", que fue presentada por Vivian López Batista como Tesis para aspirar al grado de Doctor en Informática en la Universidad de Valladolid, Departamento de Informática, Facultad de Ciencias. Esta Tesis fue dirigida por el Profesor Catedrático Luis Alonso Romero y el Dr. Valentín Cardeñoso Payo y fue defendida el día 20 de julio de 1996 obteniendo la calificación APTO "CUM LAUDE"

El aumento de la incidencia de la informática en la mayoría de los campos de la actividad humana ha provocado un cambio sustancial en la relación hombre-máquina porque no sólo el experto informático es el que la emplea como herramienta de trabajo, sino que los estudiosos de otros temas deben hacer uso de ésta.

Una de las áreas de investigación más activas en IA es la relacionada con el desarrollo de procesadores del lenguaje natural que faciliten esa interrelación hombre-máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes de programación y sistemas de menús tradicionales.

Dos son las áreas de aplicación principales del lenguaje natural:

?? Aplicaciones basadas en diálogo, como interfaces hombre-máquina con diferentes propósitos: acceso a Bases de Datos, acceso a Sistemas Expertos y acceso a Sistemas Operativos.
?? Aplicaciones no basadas en diálogo, en los que se enmarcan los Sistemas de Traducción Automática y los Sistemas de tratamiento de la información textual.

La presente tesis queda enmarcada dentro de la segunda área de aplicación mencionada y se centra fundamentalmente en el estudio y revisión de los problemas y dificultades de la traducción automática y los métodos que se han desarrollado para afrontarlos, con la mirada puesta en el desarrollo de una herramienta de uso práctico.

Se ha abordado un problema básico en el que la lingüística computacional no ha llegado muy lejos en su grado de formalización y automatización: la solución de ambigüedades por el propio sistema, de manera que éste, aprovechando la mayor cantidad posible de información relevante, incorpore conocimiento lingüístico y contextual del mundo y pueda hacer una elección de significado imitando las habilidades del conocimiento humano.

Los avances más significativos en la modelización del conocimiento surgen de la investigación en computación paralela, las redes neuronales y los modelos híbridos conexionistas-simbólicos. Se cree que las funciones superiores del entendimiento del lenguaje serán mejor modeladas, si en lo posible, somos capaces de diseñar mecanismos de razonamiento similares a los que emplea el cerebro humano, que procesa una gran cantidad de información en paralelo.

Está clara la relevancia de los modelos conexionistas para el procesamiento del lenguaje natural, pues el enfoque tradicional simbólico estratificado (morfología, sintaxis y semántica), aunque es aplicable conceptual y computacionalmente, no se entiende más que como un modelo artificial del proceso de comprensión y comunicación humanos.

Es por ello que se aspira a la integración de operaciones sintácticas, semánticas y pragmáticas en modelos conexionistas híbridos que permitan combinar las ventajas de las representaciones simbólicas con las conexionistas, lo que desde luego significa un modelo superior de representación del conocimiento aún cuando, en la mayoría de los casos la implementación de las redes neuronales se haya realizado en ordenadores secuenciales.

Las arquitecturas híbridas conexionistas, gracias a sus capacidades de entrenamiento y codificación de patrones contextuales, posibilitan la adquisición automática del conocimiento lingüístico a partir de las propias regularidades de los datos y pueden ser entrenadas para procesar con alto grado de seguridad construcciones lingüísticas que no aparecen en el universo de datos de entrenamiento.

Aunque el conexionismo ha sido adoptado recientemente en el procesamiento del Lenguaje Natural y la mayoría de los trabajos en éste área se enfocan a tareas restringidas, se prevé que el empleo de estos modelos en traducción automática se incremente notablemente con el devenir del tiempo. Por otra parte, su utilización es especialmente recomendable en sistemas que hagan uso del conocimiento de los errores cometidos de cara a las tareas de postedición de la TA.

Recientemente, se ha reconocido que en los sistemas de traducción automática de aprendizaje real, la elección de significado debía realizarse automáticamente, por un mecanismo complejo de retroalimentación que aprenda constantemente de las nuevas entradas.

Partiendo de éstas ideas en esta tesis, se ha implementado un método alternativo para realizar la elección correcta de significado y resolver la ambigüedad en dependencia del contexto. El modelo utiliza los Mapas Autoorganizados de Kohonen para determinar automáticamente el significado correcto de las palabras, mediante un análisis de clases sobre los patrones presentados a la red, que puede emplearse como un clasificador entrenado automáticamente por medio de ejemplos e incorporarse de manera general a cualquier tipo de sistema de vocabulario controlado, lo cual promete mejorar la primera elección del traductor al interactuar con otros métodos de procesamiento simbólico.

Este método se ha probado en el diseño de un prototipo de ayuda a la traducción para un lenguaje restringido, con arquitectura híbrida, que utiliza una estrategia tipo Transferencia y siguiendo los cuatro principios básicos en que deben basarse los sistemas de procesamiento del lenguaje natural a nivel de teoría computacional: modularidad, integración de restricciones parciales, grado de plausibilidad y entrenamiento.

Las características más relevantes del sistema dentro de los objetivos descritos anteriormente son:

1. Análisis e interpretación de un conjunto de frases en alemán donde cada frase sufrirá un proceso de análisis léxico, morfológico, sintáctico y semántico tras el cual se obtendrá la representación semántica de ésta.
2. El conjunto de frases permitidas viene determinado por las limitaciones de la sintaxis implementada y el léxico almacenado, el cual podrá ser ampliado de forma incremental con relativa sencillez.
3. El conjunto de reglas que componen el ámbito gramatical abordado se ha definido con una Gramática Libre de Contexto, con estructura de rasgos y un mecanismo de unificación, que

puede importarse a formalismos sintácticos más recientes en lingüística computacional (LFG, GPSG, y HPSG).

4. Las reglas de la gramática poseen además procedimientos que realizan pruebas necesarias (en el análisis y la transferencia) y se ejecutan cuando se invoca una regla. Este mecanismo es muy similar al utilizado por las Redes de Transición Aumentadas (ATN).
5. Se establece un mecanismo eficiente para incorporar restricciones semánticas en el análisis sintáctico y si aún siguen existiendo ambigüedades léxicas sólo salvables a nivel contextual, se aplica el método para solucionar ambigüedades contextual que emplea el Mapa Autoorganizativo de Kohonen.
6. La implementación se realizó en Lenguaje C, bajo el Sistema Operativo UNIX y utilizando la herramienta Som-Pack, The Self-Organizing Map Program Package y el Generador de Procesamiento Compatible Yacc.

El contenido de la tesis presentada se estructura en los capítulos siguientes:

En el *Capítulo 2* se realiza una breve introducción de los conceptos básicos y los distintos formalismos en el procesamiento del lenguaje natural, que en algunos casos serán empleados en otros capítulos.

En el *Capítulo 3* se presenta el controvertido problema de la traducción automática, los principales diseños y tipos de sistemas.

En el *Capítulo 4* se hace una revisión de las principales técnicas existentes en el campo de la traducción automática, los problemas que se presentan en el análisis y cómo se aborda la transferencia y generación.

En el *Capítulo 5* se hace una evaluación de los principios básicos utilizados para eliminar la ambigüedad del significado de las palabras, junto con el estado del arte.

En el *Capítulo 6* se contraponen las ventajas y desventajas de los modelos simbólicos frente a los conexionistas para el procesamiento del lenguaje natural, la exposición de los modelos de redes neuronales artificiales que pueden tener relación con los objetivos planteados en este trabajo y el estado del arte actual de estos modelos para el procesamiento del Lenguaje Natural.

En el *Capítulo 7* presentamos una exposición original del Mapa de Kohonen para la extracción de clases semánticas y la exposición del trabajo experimental.

En el *Capítulo 8* se detalla la implementación del prototipo aplicando los métodos y modelos expuestos en capítulos anteriores.

En el último capítulo se exponen las conclusiones y líneas de investigaciones futuras, pudiéndose destacar las siguientes aportaciones:

?? Se implementa un método general para abordar la traducción automática que combina técnicas simbólicas y conexionistas.
?? Se demuestra que las Redes de Kohonen brindan buenos resultados en la tarea de solucionar la ambigüedad contextual.
?? Los Mapas Autoorganizativos pueden proporcionar una imagen global de los tipos de contextos existentes en una base de datos lingüística.

?? Se crea un método para solucionar la ambigüedad contextual donde a diferencia de otros de éste tipo, la parte de la información que se codifica manualmente es mínima.

En el conjunto de pruebas efectuadas, para un número de 1500 frases, se realizó una clasificación correcta el 87 por ciento de veces; por lo que los resultados pueden considerarse satisfactorios para este tipo de tarea.

Esta implementación se considera un modelo superior de representación del procesamiento del lenguaje natural usando redes neuronales distribuidas, donde las tareas se diseñan modularmente, se establece la comunicación entre los diferentes módulos y el conocimiento semántico y contextual emerge automáticamente de las propiedades estadísticas de los ejemplos entrenados.

Pero las redes de Kohonen tienen la dificultad de que hay que especificar a priori el número de clases de las que se desea disponer. El modelo Fuzzy ART para *cluster* adaptativo, ha demostrado ser de gran eficacia para la construcción de clases semánticas, teniendo en cuenta la mínima información necesaria para realizar su trabajo. Esta red permite obtener el grado de pertenencia de cada término a cada clase semántica, lo que puede ser utilizado para construir una red semántica de términos de la base.
Por lo que proponemos se experimente con este modelo en investigaciones futuras, pues son los únicos que ofrecen una solución sólida al problema de elasticidad-plasticidad.

Resumen de la Bibliografía utilizada:

1. Hutchins,W.J. and Somers, H. L. An Introduction to Machine Translation Academic Press. London, New York 1992
2. Donnelly Charles and Stallman Richard. The YACC-compatible Parser Generator. Bison. June 1992.
3. Allen J. Natural Language Understanding Benjamin Cummings Publishing Company, 1987.
4. McRoy Susan W. Using Multiple Knowledge Sources for Word Sense Discrimination. Artificial Intelligence Program GE Researche and Development Center,1992.
5. Noble H. Natural Language Processing. Blackwell Scientific Publications, 1990.
6. Covington Michael A. A Dependency Parser for Variable-Word-Order Languages Artificial Intelligence Programs. The University of Georgia Athens Georgia, 1993.
7. Dorffner Georg "Radical" Connectionism for Natural Language Processing Dept. of Medical Cybernetics and AI University of Vienna, Austria and Austrian Research Institute for Artificial Intelligence, 1994.
8. Kohonen T. Self-organized Formation of Topologically Correct Feature Maps. Neurocomputing, Pag. 511-512. The MIT press, Cambridge, 1989.
9. Kohonen T. The Self-organizing Map.Proceedings of the IEEE, 78(9):1464-1480, 1990.
10. Miikkulainen R. and Dyer M. Natural Language Processing With Modular PDP Networks and Distributed Lexicon.University of California, Los Angeles,1991.
11. Gallant I. Stephen A Practical Approach for Representing Context and for Performing Word Sense Disambiguation Using Neural Networks, Neural Computation 3, pp. 293-309, 1991.
12. Kohonen T. SOM-PAK. The Self-Organizing Map Program Package. Helsinki University of Technology laboratory of Computer and Information Science. Finland, 1995.
13. Wilks Y. A Preferencial, Patter - Seekeng, semántics for Natural Language inference. 53-64 Artificial Intelligence 6, 1975.