Algoritmo LVQ aplicado a tareas de procesamiento del lenguaje natural

María Teresa Martín Valdivia
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, s/n. Edif.. A3
maite@ujaen.es

Resumen: Tesis doctoral en Informática realizada por María Teresa Martín Valdivia bajo la dirección de los doctores L. Alfonso Ureña López (Univ. de Jaén) y Francisco Triguero Ruiz (Univ. de Málaga). El acto de defensa de tesis tuvo lugar el 7 de mayo de 2004 ante el tribunal formado por los doctores Manual Palomar Sanz (Univ. de Alicante), Amparo Ruiz Sepúlveda (Univ. de Málaga), Emilio Sanchís Arnal (Univ. Politécnica de Valencia), Horacio Rodríguez Hontoria (Univ. Politécnica de Barcelona), Manuel de Buenaga Rodríguez (Univ. Europea). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: Categorización de texto, Desambiguación léxica, Redes neuronales, Modelo de Kohonen, Algoritmo de aprendizaje por cuantificación vectorial, Recuperación de información, Reconocimiento de multipalabras, Fusión de documentos en recuperación de información multilingüe

Abstract: PhD Thesis in Computer Science written by María Teresa Martín Valdivia under the supervision of Dr. L. Alfonso Ureña López (Univ. of Jaén) and Dr. Francisco Triguero Ruiz (Univ. of Málaga). The author was examined in May 6th 2004 by the commitee formed by Dr. Manual Palomar Sanz (Univ. of Alicante), Dr. Amparo Ruiz Sepúlveda (Univ. of Málaga), Dr. Emilio Sanchís Arnal (Univ. Politécnica of Valencia), Dr. Horacio Rodríguez Hontoria (Univ. Politécnica de Barcelona), Dr. Manuel de Buenaga Rodríguez (Univ. Europea). The grade obtained was *Sobresaliente Cum Laude*.
Keywords: Text Categorization, Word Sense Disambiguation (WSD), Neural networks, Kohonen model, Learning Vector Quantization algorithm, Information retreival, Multiword recognition, Fusion collection problem in cross-lingual information retrieval

1 Introducción

Tanto el Procesamiento del Lenguaje Natural (PLN) con las Redes Neuronales Artificiales (RNA) son dos áreas fundamentales dentro de la Inteligencia Artificial. Sin embargo, y a pesar de la gran cantidad de trabajos realizados en ambas disciplinas, los intentos por combinarlas han sido muy escasos.

Por una parte, los trabajos que incorporan aprendizaje automático en los sistemas de PLN son numerosos, y por otra, las RNA se han aplicado a un gran número de problemas con características muy similares a los del PLN. Sin embargo, curiosamente el número de estudios que hacen uso de RNA en sistemas de PLN es muy reducido. Más sorprendente aún, cuando los resultados obtenidos en los pocos trabajos existentes ponen de manifiesto que el uso de un enfoque neuronal constituye una buena alternativa para la construcción de sistemas PLN basados en aprendizaje.

El objetivo principal de esta tesis consiste en demostrar que es posible aprovechar las ventajas y características que presentan las RNA para abordar con éxito el desarrollo e implementación de sistemas que traten el lenguaje de manera automática.

Para ello, se propone un formalismo común basado en un modelo neuronal para resolver diversas tareas de PLN. Concretamente se tratan tres tareas:

- La categorización de texto
- La resolución de la ambigüedad léxica
- La recuperación de información.

Mientras que para las dos primeras tareas se desarrollan sistemas completos para la

recuperación de información se abordan dos problemas concretos relacionados con este tipo de sistemas:

- El reconocimiento de términos multipalabra
- La fusión de colecciones

El primero de los problemas se trata desde una perspectiva monolingüe mientras que el segundo se aborda para un ambiente multilingüe.

El esquema neuronal utilizado se basa en el modelo de Kohonen y más concretamente en su versión supervisada: el algoritmo de aprendizaje por cuantificación vectorial o algoritmo LVQ (Learning Vector Quantization algorithm). Se demuestra que es posible adaptar dicho algoritmo para resolver aplicaciones reales del procesamiento del lenguaje natural presentándolo como un método robusto, flexible y efectivo. Los experimentos realizados ponen de manifiesto que el algoritmo LVQ se adapta fácilmente a los distintos escenarios utilizados y que los resultados obtenidos son comparables, y en muchos casos superan a los métodos tradicionales utilizados para resolver cada uno de los problemas estudiados.

2 Estructura de la tesis

La tesis comienza exponiendo los objetivos perseguidos así como la motivación para el desarrollo de la misma.

El capítulo 2 presenta una introducción al PLN y se definen las tres tareas que se estudiarán en capítulos posteriores, así como otras aplicaciones estrechamente relacionadas. A continuación, se estudian los modelos de representación de información, prestando mayor atención al modelo de espacio vectorial puesto que será la base para otros capítulos. Asimismo, se presentan las medidas de evaluación y algunas técnicas habituales para aumentar la efectividad de las aplicaciones. Por último, se comentan algunos recursos lingüísticos que se integrarán en nuestros sistemas.

El capítulo 3 trata de manera general las redes neuronales artificiales. Se dan algunas definiciones y se exponen las principales características de las redes neuronales artificiales. Después se presentan los distintos tipos de arquitecturas de red y se enumeran algunas de las muchas aplicaciones posibles de estos modelos. El capítulo finaliza comentando algunos trabajos recientes que ponen de manifiesto la relación entre las RNA y el PLN.

El modelo neuronal de Kohonen que sirve de base para el desarrollo de nuestros sistemas, se describe en el capítulo 4. En primer lugar, se estudian las características y fundamentos del aprendizaje competitivo. El resto de capítulo presenta el modelo de mapas autoorganizativos (versión no supervisada) y el algoritmo LVQ (versión supervisada). Para ambos esquemas, se presenta su arquitectura, los algoritmos de entrenamiento y evaluación, así como algunos trabajos que estudian la aplicación de cada modelo a algunas tareas concretas del PLN.

El capítulo 5 estudia la aplicación del algoritmo LVQ a la categorización de texto. Una vez descrita la tarea concreta, se desarrollan dos experimentos. El primero consiste en categorizar un recurso multilingüe (la Biblia políglota) y el segundo utiliza la colección Reuters para construir una red capaz de aprender las categorías de la colección.

El capítulo 6 está dedicado al uso del algoritmo LVQ en tareas de desambiguación. En primer lugar, se describe la tarea de la resolución de la ambigüedad léxica de las palabras y se comenta el modelo de representación utilizado. A continuación, se explica como se integran los recursos lingüísticos utilizados para establecer el entorno experimental.

El capítulo 7 trata la aplicación del algoritmo LVQ para resolver dos problemas relacionados con la recuperación de información cuya resolución mejora el rendimiento de los sistemas. Se describen los sistemas de información tanto monolingüe como multilingüe. El resto del capítulo se dedica a resolver los dos problemas propuestos, el primero de ellos (problema del reconocimiento de multipalabras) en un ambiente monolingüe mientras que el segundo (problema de la fusión de colecciones) se aborda desde una perspectiva multilingüe.

En el capítulo 8 se resumen las principales aportaciones y se exponen las principales líneas de trabajo futuro a desarrollar. Asimismo, se incluye una recopilación de los trabajos publicados en revistas y congresos nacionales e internacionales durante el desarrollo de esta memoria y relacionadas directamente con ella.