Resolución de la ambigüedad léxica mediante aprendizaje por cuantificación vectorial*

Manuel García Vega
Departamento de Informática
Universidad de Jaén
mgarcia@ujaen.es

Resumen: Tesis doctoral en Informática Realizada por Manuel García Vega y dirigida por el Doctor L. Alfonso Ureña López (Univ. de Jaén). El acto de defensa de tesis tuvo lugar en Jaén en diciembre de 2006 ante el tribunal formado por los doctores Miguel Toro Bonilla (Univ. de Sevilla), Manuel Palomar Sanz (Univ. de Alicante), Lidia Moreno Boronat (Univ. Politécnica de Valencia), Andrés Montoyo Guijarro (Univ. de Alicante) y María Teresa Martín Valdivia (Univ. de Jaén). La calificación obtenida fue Sobresaliente *Cum Laudem* por unanimidad.
Palabras clave: Desambiguación, Redes neuronales, LVQ

Abstract: PhD thesis in Computer Science written by Manuel García Vega under the supervision of Dr. L. Alfonso Ureña López (Univ. of Jaén). The author was examined in December 2006 in Jaén by the committee formed by Miguel Toro Bonilla (Univ. of Sevilla), Manuel Palomar Sanz (Univ. of Alicante), Lidia Moreno Boronat (Univ. Politécnica of Valencia), Andrés Montoyo Guijarro (Univ. of Alicante) y María Teresa Martín Valdivia (Univ. of Jaén). The grade obtained was *Sobresaliente Cum Laudem*.
Keywords: WSD, Neural Nets, LVQ

1. Introducción

La desambiguación del sentido de las palabras (Word Sense Disambiguation) es el problema de asignar un sentido determinado a una palabra polisémica, usando su contexto. Este problema ha sido de interés, prácticamente desde el comienzo de la informática, en los años 50. La desambiguación es una tarea intermedia y no un fin en sí misma. En particular, es muy útil, a veces imprescindible, para muchos problemas del PLN, como por ejemplo la recuperación de información, la categorización de textos, la traducción automática...

Los objetivos de este trabajo son:

1. Implementar un desambiguador del sentido de las palabras basado en el Modelo de Espacio Vectorial optimizando los pesos de los vectores del entrenamiento usando la red neuronal LVQ (Learning Vector Quantization) del modelo neuronal supervisado de Kohonen.
2. Proponer un método uniforme de integración de recursos que sirvan para el entrenamiento de la red. Los parámetros de la red LVQ han sido optimizados para el problema de la desambiguación.

En este trabajo se ha demostrado que las redes neuronales, concretamente los modelos de Kohonen, resuelven brillantemente el problema de la resolución de la ambigüedad léxica, aportando robustez, porque la red LVQ es insensible a pequeños cambios observándose unos resultados homogéneos independientemente del entrenamiento; flexibilidad, porque es fácilmente aplicable a cualquier tarea de PLN; escalabilidad, porque pueden introducirse multitud de textos de entrenamiento para ajustarlo a cualquier dominio y efectividad, porque los resultados obtenidos son comparables y en muchos casos superan a los métodos tradicionales utilizados para resolver los mismos problemas.

Se ha calculado los parámetros óptimos de configuración de la red LVQ para la tarea de desambiguación, maximizando la precisión, el *recall* y la cobertura.

Se han integrado el corpus SemCor y la base de datos léxica WordNet. Además, se ha aportado un método de integración automática de cualquier corpus.

* Este trabajo ha sido parcialmente financiado por los proyectos FIT-150500-2002-416, FIT-150500-2003-412 y TIC2003-07158-C04-04

Los experimentos realizados muestran el buen comportamiento de esta red para el problema concreto de la desambiguación.

2. Estructura de la tesis

La estructura sigue un esquema clásico, introduciendo el problema, la motivación y las contribuciones obtenidos.

En el capítulo 2 se describe detalladamente el problema de la desambiguación y la terminología que es comúnmente usada. Así mismo, se describen con detalle los recursos lingüísticos que se usan, concretamente corpus de textos y bases de datos léxicas. A continuación, se explican las principales medidas para la evaluación de los sistemas desambiguadores. Se describe la organización Senseval que actualmente es el principal medio de evaluación para cualquier sistema de resolución de la ambigüedad léxica y se describen los principales métodos de desambiguación, así como los mejores desambiguadores presentados en las tres ediciones de Senseval.

El capítulo 3 trata de manera general las redes neuronales artificiales, clasificándolas según diferentes criterios, definiendo sus partes principales y describiendo sus características más importantes. A continuación, se detallan los principales métodos de entrenamiento, haciendo hincapié en su carácter supervisado o no supervisado. Se sigue con la cuantificación vectorial, como base matemática del aprendizaje LVQ. El modelo de Kohonen es presentado a continuación, enlazando la cuantificación vectorial y el aprendizaje competitivo para producir la red neuronal LVQ.

El capítulo 4 está dedicado a describir el desambiguador. Comienza explicando el modelo del espacio vectorial, que da el soporte matemático a la red neuronal. Se detalla la integración del modelo matemático con la red LVQ y cómo se ha de realizar el entrenamiento. A continuación, se incluyen en el entrenamiento las fuentes lingüísticas disponibles. En primer lugar, el corpus SemCor, con el que se hace un experimento para comprobar su validez. Después, se detalla cómo integrar WordNet en el entrenamiento y se experimenta con los datos que aporta. Continúa con la integración de ambos recursos a la vez. Para terminar, se estudian a fondo los distintos parámetros de la red LVQ para optimizar su comportamiento.

En el capítulo 5 se evalúa el desambiguador que se ha construido. Primero, se simula una participación en la competición Senseval-2 en la tarea de *English Lexical Sample* y posteriormente se detalla la participación en Senseval-3, concretamente a las tareas de *English Lexical Sample* y de *English All Words*.

En el capítulo 6 se detallan las conclusiones, se explican las principales aportaciones presentadas en esta memoria, así como las líneas futuras de investigación como continuación de este trabajo. Finalmente, se incluye una recopilación de trabajos publicados en revistas y congresos nacionales e internacionales durante el desarrollo de esta memoria y relacionadas con ella.

3. Aportaciones de la investigación

Las principales contribuciones de este trabajo de investigación son:

- Se ha propuesto un desambiguador basado en el modelo neuronal de Kohonen, usando la red LVQ.
- Proponemos un desambiguador que puede afinarse tanto para precisión como para *recall*, ajustando adecuadamente un cierto valor *umbral* para la probabilidad de acierto del sentido desambiguado.
- Se ha afinado el algoritmo LVQ para una mayor efectividad en el problema de la resolución de la ambigüedad léxica, fundamentando el cálculo en el comportamiento del desambiguador con experimentos ya contrastados.
- El desambiguador propuesto es muy robusto, mostrando un comportamiento homogéneo en los distintos experimentos realizados donde los dominios semánticos de los textos objeto de estudio eran muy diversos.
- Se ha construido un desambiguador independiente de la lengua, siempre y cuando se disponga de los recursos necesarios: lexicón y textos etiquetados en la lengua objeto.
- Se ha definido un método de integración de recursos lingüísticos heterogéneos para su uso como entrenamiento de la red LVQ, que permite la incorporación de información específica en cualquier dominio semántico.