APLICACION DE UN MODELO PROBABILISTICO EN LA DESAMBIGUACION MORFOLOGICA

Ramona Miró, Manuel Ollero, Miguel A. Pineda, Pedro Vázquez

Centro de Cálculo
Facultad de Filología
Universidad de Sevilla

resumen:

El análisis automático de un texto plantea, en el nivel morfológico, el problema de la homografía. El estudio de las relaciones sintagmáticas de las unidades homógrafas demuestra que se puede hablar de reglas de distribución que resuelven total o parcialmente esta ambigüedad. La construcción de dichas reglas a partir de un corpus previamente analizado, ofrece la posibilidad de verificarlas y cuantificar sus frecuencias de aparición, así como su rendimiento. De este modo, obtenemos la estructura probabilística en la que se basa el procedimiento de desambiguación morfológica que proponemos.

El desarrollo de este trabajo parte de la necesidad de resolver el problema de la homografía para lograr la automatización de nuestro analizador morfológico, y se enmarca dentro de la llamada Lingüística de Corpus.

APLICACION DE UN MODELO PROBABILISTICO EN LA DESAMBIGUACION MORFOLOGICA

Al abordar la construcciónde un analizador morfológico automático, uno de los problemas más importantes que encontramos es la aparición de palabras homógrafas. En este nivel de análisis nos centramos exclusivamente en la homografía funcional, dentro de la cual distinguimos, como señala S. Marcus, la homografía morfológica (la que presenta la conjugación verbal, por ejemplo) y la léxico-gramatical (cuando a un mismo lexema se le pueden asignar distintas categorías).

El estudio de las relaciones sintagmáticas de las unidades homógrafas demuestra que se puede hablar de reglas de distribución que resuelven total o parcialmente esta ambigüedad. Tras su apariencia desordenada, podemos observar en la lengua una estabilidad frecuencial de determinados hechos lingüísticos que tienen su origen en ciertas propiedades objetivas, derivadas de su propio componente estructural.

A. I. Jinchin en su "método de las funciones arbitrarias" expone que, partiendo de una distribución inicial arbitraria de los datos y tras una reiterada serie

de experimentos, puede demostrarse tanto la estabilidad y el valor numérico de la frecuencia de un determinado fenómeno, como que el valor de la frecuencia no depende de la distribución arbitraria inicial, sino de las particularidades objetivas del fenómeno mismo.

La construcción de estas reglas de distribución a partir de un corpus previamente analizado ofrece la posibilidad de verificarlas y, al mismo tiempo, cuantificar su frecuencia de aparición y su rendimiento. El método estadístico resulta imprescindible para revelar el carácter sistemático de la lengua, puesto que permite determinar ciertas propiedades de regularidad con carácter global.

Los modelos probabilísticos utilizados en la construcción de sistemas expertos han demostrado ser esencialmente adecuados para resolver problemas de incertidumbre. La aplicación de estos modelos en el tratamiento de la desambiguación morfológica nos ofrece un procedimiento válido para la resolución de la homografía.

El desarrollo de este trabajo parte de la necesidad de resolver el problema de la homografía para lograr una total automatización de nuestro analizador morfológico y se enmarca dentro de la llamada Lingüística de Corpus.

El corpus sobre el que trabajamos está constituido por los textos de encuestas de hablantes sevillanos, pertenecientes al nivel popular. Los informantes se distribuyen, en igual proporción, entre las variables sociolingüísticas de edad y sexo.

Es importante destacar que las manifestaciones discursivas que estudiamos pertenecen al lenguaje hablado. Este rasgo confiere al corpus unas características

que lo oponen al lenguaje escrito: la espontaneidad del hablante, la poca sujeción a reglas gramaticales, etc., dando lugar a una gran complejidad sintáctica.

El corpus, que ha sido descrito previamente con un procedimiento semiautomático, tiene una extensión de 117.491 palabras de las cuales 26.231 (el 22.33 por ciento) presentan homografía. Este alto porcentaje justifica la necesidad de abordar este problema, con objeto de llevar a cabo posteriores análisis morfológicos de forma automática, en textos de similares características.

Hemos considerado que el modelo más idóneo para ello es el de reglas probabilísticas; éstas se establecerán partiendo de la situación del elemento homógrafo en el texto y examinando sus relaciones sintagmáticas.