Adquisició d'informació lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat*

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Resumen: Tesis doctoral en Lingüísitica realizada por Antoni Oliver González bajo la dirección de la doctora Irene Castellón Masalles (U. de Barcelona) y del doctor Lluís Màrquez Villodre (U. Politècnica de Catalunya). El acto de defensa de la tesis tuvo lugar el 27 de julio de 2004 ante el tribunal formado por los doctores Horacio Rodríguez Hontoria (U. Politécnica de Catalunya), Joan Castellví Vives (U. de Barcelona), Iñaki Alegria Loinaz (U. País Vasco), Toni Badia Cardús (U. Pompeu Fabra) y Ana Maria Fernández Montraveta (U. Autònoma de Barcelona). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad.
Palabras clave: adquisición léxica, morfología computacional, filología eslava

Abstract: PhD Thesis in Linguistics written by Antoni Oliver González under the supervision of Dr. Irene Castellón Masalles (U. de Barcelona) and Dr. Lluís Màrquez Villodre (U. Politècnica de Catalunya). The author was examined in July 27th, 2004 by the commitee formed by Dr. Horacio Rodríguez Hontoria (U. Politécnica de Catalunya), Dr. Joan Castellví Vives (U. de Barcelona), Dr. Iñaki Alegria Loinaz (U. País Vasco), Dr. Toni Badia Cardús (U. Pompeu Fabra) and Dr. Ana Maria Fernández Montraveta (U. Autònoma de Barcelona). The grade obtained was Sobresaliente Cum Laude.
Keywords: lexical acquisition, computational morphology, Slavonic philology

1 *Introducción*

En esta tesis se presentan diversos métodos para la adquisición automática de información léxica y morfosintáctica y de aprendizaje no supervisado de la morfología a partir de corpus sin anotar. Estos métodos se han probado para dos lenguas eslavas: el ruso y el croata; lenguas que se caracterizan por tener una morfología rica y de tipo predominantemente concatenativa. Esta característica se ha aprovechado en el diseño de los algoritmos, que se pueden adaptar fácilmente para funcionar para otras lenguas que presenten una morfología relativamente rica y cuyos principales procesos morfológicos, ya sean sufijales o prefijales, se puedan describir de una manera concatenativa. Se ha realizado una evaluación exhaustiva de los métodos presentados y se ha demostrado que funcionan muy satisfactoriamente para estas lenguas. El hecho que funcionen a partir de corpus sin anotar hace que sean muy interesantes para la creación de nuevos recursos léxicos o para la ampliación de recursos léxicos existentes. Los algoritmos presentados en este trabajo pueden utilizar Internet para buscar información no presente en el corpus, lo que supone que se puedan aplicar los procesos sin la necesidad de recopilar corpus de gran tamaño.

2 *Adquisición léxica a partir de corpus sin anotar*

El objetivo principal del sistema de adquisición léxica es adquirir automáticamente una lista tan completa como sea posible de formas con su lema e información morfosintáctica asociados a partir de un corpus sin anotar. En este apartado de la tesis se presentan dos métodos que se basan principalmente en la coaparición en el corpus de diferentes formas de un mismo paradigma. Hemos denominado el primer método *adquisición por hipótesis-validación*, ya que a partir de las formas del corpus y de las reglas morfológicas que describen la morfología de la lengua se hipotetizan los posibles lemas relacionados y se verifica su existencia en el corpus. El principal problema que presenta este método es la ambigüedad

* Adquisición de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al ruso y al croata

de las reglas morfológicas respecto a la tarea de adquisición, lo que provoca que se confundan formas de un determinado paradigma con lemas de otro paradigma. Esta ambigüedad ha exigido crear unos algoritmos de análisis de las reglas para dividirlas entre ambiguas y no ambiguas previamente al proceso de adquisición. Dos inconvenientes adicionales de este primer método es que es imprescindible que el lema asociado a la forma a adquirir esté presente en el corpus y que no proporciona ningún tipo de información sobre las adquisiciones que no ha podido realizar. El segundo método recibe el nombre de *adquisición por agrupación-comparación* ya que se realizan agrupaciones de las formas del corpus conforme a los paradigmas expresados por las reglas morfológicas. Posteriormente, para cada forma del corpus, se verifica en qué agrupaciones se puede describir y se escoge la que presenta más formas asociadas presentes en el corpus. Este método permite adquirir la información morfosintáctica y el lema asociado a una forma incluso si el lema no está presente en el corpus, y además, proporciona información sobre las adquisiciones que no se han podido llevar a cabo por falta de las formas discriminantes en el corpus. Esta información se puede utilizar para intentar resolver estos casos consultando a corpus más grandes, o bien, como hemos realizado en nuestros experimentos, utilizando buscadores de Internet. Este método permite, además, utilizar los resultados de las búsquedas a Internet para adquirir información sobre formas no presentes en el corpus original. Con este método se ha conseguido una precisión del 91,49% con una cobertura del 77,98% (F_1=84,2) para el ruso.

3 Descubrimiento de la morfología

Para poder aplicar los métodos de adquisición presentados necesitamos conocer la morfología de la lengua tratada. En esta tesis hemos presentado también un algoritmo para el aprendizaje no supervisado de la morfología. Por aprendizaje no supervisado entendemos aquel que precisa como entrada únicamente un texto sin anotar o un conjunto de formas sin ningún tipo de información. El algoritmo presentado pretende servir de asistencia al lingüista encargado de crear las reglas morfológicas y es capaz de descubrir los principales procesos morfológicos tanto flexivos como derivativos. Este algoritmo ha funcionado de manera muy satisfactoria para el ruso y se puede aplicar a otras lenguas que presenten una morfología rica, concatenativa y de tipo sufijal. El algoritmo se puede adaptar fácilmente para la detección de fenómenos prefijales, pero no es directamente aplicable a fenómenos no concatenativos. Una novedad interesante respecto otros algoritmos existentes es el uso de Internet para intentar completar paradigmas detectados de manera incompleta por falta de información en el corpus de trabajo.

4 Conclusiones

En este trabajo se han presentado diversos métodos que nos permiten adquirir tanto información morfológica como léxica de una manera empírica y sin la necesidad de recopilar corpus de gran tamaño. De esta manera se ha podido demostrar que los corpus sin anotar son una buena fuente de conocimiento lingüístico de los que se puede obtener información léxica y morfosintáctica e información sobre los procesos morfológicos. Sobre estas técnicas quedan todavía algunos aspectos por estudiar. Por ejemplo, tanto para el ruso como para el croata ha quedado cierta información sin poder adquirir: categoría animado-inanimado, aspecto verbal, algunas formas variantes coincidentes con formas estándar, etc. Se deben estudiar extensiones de la metodología que permitan adquirir este tipo de información. Relacionado con este tema y también con las posibles mejoras de los resultados obtenidos está la posibilidad de utilizar el contexto de aparición de las formas.

El uso de Internet como corpus de gran tamaño es una posibilidad muy interesante para la Lingüística de Corpus, ya que evita la necesidad de recopilar grandes volúmenes de texto, pero que no está exenta de problemas. Entre estos problemas se puede destacar la presencia de páginas con una calidad lingüística pobre y que pueden contener faltas de ortografía, así como la existencia de páginas devueltas como si fuesen de una determinada lengua pero que en realidad están escritas en otra lengua. Estos hechos pueden tener una influencia negativa sobre los resultados de los algoritmos. Se debe continuar investigando sobre las posibilidades de Internet en la Lingüística de Corpus para aprovechar al máximo las características de los buscadores existentes, o bien crear nuevos buscadores que estén orientados a tareas lingüísticas.