Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 149-150 recibido 1-10-09 revisado 08-02-10 aceptado 05-03-10

Análisis léxico robusto

Robust Lexical Analysis

Juan Otero Pombo
Departamento de Informática, Universidade de Vigo
Escola Superior de Enxeñaría en Informática, Edificio Politécnico
Campus de As Lagoas s/n, 32002 - Ourense
jop@uvigo.es

Resumen: Tesis doctoral de Informática realizada por Juan Otero Pombo bajo la dirección de los doctores Manuel Vilares Ferro (Universidade de Vigo) y Jorge Graña Gil (Universidade de A Coruña). La defensa tuvo lugar el 4 de junio de 2009 ante el tribunal formado por los doctores Guillermo Rojo Sánchez (Universidade de Santiago de Compostela), José Gabriel Pereira Lopes (Universidade Nova de Lisboa, Portugal), Jean-Éric Pin (Laboratoire D-Informatique Algoritmique, Fondement CNRS, Francia), Leo Waner (Universidad Pompeu Fabra) y Víctor Manuel Darriba Bilbao (Universidade de Vigo). La calificación obtenida fué Sobresaliente *Cum Laude* con mención de Doctor Europeo. Se puede obtener más información acerca de esta tesis en `http://www.grupocole.org`.
Palabras clave: Corrección ortográfica, tokenización, etiquetación morfosintáctica, recuperación de información

Abstract: PhD Thesis in Computer Science written by Juan Otero Pombo under the supervision of Dr. Manuel Vilares Ferro (Universidade de Vigo, Spain) and Dr. Jorge Graña Gil (Universidade de A Coruña, Spain). The author was examined on 4th June, 2009 by the commitee formed by Dr. Guillermo Rojo Sánchez (Universidade de Santiago de Compostela, Spain), Dr. José Gabriel Pereira Lopes (Universidade Nova de Lisboa, Portugal), Dr. Jean-Éric Pin (Laboratoire D-Informatique Algoritmique, Fondement CNRS, France), Dr. Leo Waner (Universidad Pompeu Fabra, Spain) and Dr. Víctor Manuel Darriba Bilbao (Universidade de Vigo, Spain). The grade obtained was *Sobresaliente Cum Laude*, with an European Doctor mention. Further information is available at `http://www.grupocole.org`.
Keywords: Spelling correction, tokenization, part of speech tagging, information retrieval

1. Introducción

Aunque en los últimos años se han realizado importantes avances, los fundamentos teóricos del Procesamiento del Lenguaje Natural (PLN) se encuentran todavía en constante evolución. Resulta por tanto de especial interés el desarrollo de tecnología de base, imprescindible para abordar tareas de mayor nivel como la traducción automática, elaboración automática de resúmenes, *búsqueda de respuestas* y, en particular la *recuperación de información* (RI). En este sentido, y como primer paso en esta escala de problemas, adquiere especial relevancia el desarrollo y mejora de técnicas que permitan manejar el léxico. Más concretamente, nos hemos centrado aquí en el desarrollo de un marco común que nos permita representar y resolver las ambigüedades presentes en este nivel de análisis.

2. Objetivos

El objetivo principal de esta tesis ha sido el desarrollo y evaluación de la tecnología de base necesaria para el PLN, más concretamente en el ámbito del análisis léxico, la corrección ortográfica y la etiquetación.

Nuestros mayores esfuerzos se han centrado en el desarrollo de un nuevo método de corrección ortográfica regional sobre *autómatas finitos* (AF) como alternativa a los métodos de corrección global clásicos, integrando las técnicas desarrolladas en la herramienta de etiquetación MrTagoo con el fin de aprovechar la información morfosintáctica contextual embebida en el modelo estocástico que subyace en dicha herramienta, y determinar el grado de idoneidad de las alternativas

ISSN 1135-5948

de corrección obtenidas.

Por otra parte, la minimización del coste computacional, tanto desde el punto de vista espacial como temporal, ha sido prioritaria a lo largo de todo el proyecto mediante el uso de tecnología de estado finito y la integración de los métodos implementados sobre una estructura de datos común, aplicando técnicas de programación dinámica que redundan en un importante ahorro al evitar la repetición innecesaria de cálculos.

De este modo, hemos desarrollado una herramienta de análisis léxico robusto capaz de manejar los tres tipos de ambigüedades que pueden surgir en esta fase: La ambigüedad morfosintáctica, que surge cuando a una unidad léxica le pueden ser asignadas diferentes etiquetas morfosintácticas; la ambigüedad segmental, que aparece cuando es posible dividir el texto en unidades léxicas de más de un modo; y la ambigüedad léxica, que es la que introducen los métodos de corrección ortográfica cuando ofrecen varias alternativas de corrección.

3. Resultados

En primer lugar, hemos desarrollado un nuevo método de corrección ortográfica regional sobre AFs cuya característica diferencial radica en el concepto de región que nos permite delimitar el área de reparación de una palabra errónea, en contraposición con los métodos de corrección globales que aplican las operaciones básicas de reparación en todas las posiciones de la palabra sin tener en cuenta el punto en que el error es detectado. Para estimar la viabilidad del método desarrollado, hemos implementado también un método de corrección global que nos ha servido como referencia a la hora de evaluar el rendimiento, la cobertura y la precisión de nuestra propuesta. Los resultados preliminares corroboraban no sólo que el rendimiento ofrecido por nuestra técnica de corrección regional era superior al que arrojaba el método global, sino también que la diferencia entre ambos crecía al mejorar la localización del primer punto de error. Además, el método regional ofrecía un menor número de alternativas de corrección debido a que acotaba la zona del AF a explorar a la región en la que se detecta el error. Otro aspecto a tener en cuenta era el *ratio* de acierto del método. En el caso de nuestro método regional es del 77 % frente al 81 % del global, aunque cabía esperar que la integración de información lingüística, tanto desde el punto de vista semántico como sintáctico, debería reducir de forma significativa esta diferencia de precisión, que es menor del 4 %, o podría incluso eliminarla.

Nuestro siguiente paso, consistió en comprobar si la pérdida de precisión del método regional podía ser compensada en un entorno de corrección contextual, ya que el hecho de que éste devolviese un menor número de alternativas podría repercutir de forma positiva en la precisión del sistema global. Para ello hemos integrado los algoritmos implementados en una herramienta de etiquetación morfosintáctica capaz de manejar ambigüedades de segmentación o de tokenización. Nuestros experimentos no han corroborado la hipótesis inicial, pero han servido para evidenciar que el incremento del rendimiento del método regional en términos de espacio y tiempo respecto al global era aún mayor cuando aplicábamos estas técnicas en un entorno de corrección contextual. Esto era debido a que, además de resultar más eficiente desde el punto de vista computacional, el algoritmo regional ofrece un menor número de alternativas de corrección. Esto nos anima a continuar en la búsqueda de técnicas y heurísticas que nos permitan determinar cuando es posible optar por una corrección regional.

Finalmente, hemos realizado pruebas con el fin de verificar la utilidad práctica de nuestra propuesta en un entorno de *Recuperación de Información* en el que las consultas presentan errores. Para ello, hemos comparado tres métodos. El primero, consiste en expandir las consultas con todas las alternativas de corrección. El segundo, aplica nuestro corrector contextual para determinar cuál de las alternativas obtenidas encaja mejor en el contexto de la palabra errónea. El tercero, evita la aplicación de métodos de corrección ortográfica al utilizar n-gramas para la indexación y la recuperación. Los resultados de estos experimentos revelan que el uso de técnicas de corrección ortográfica mejora el rendimiento de los sistemas de RI basados en extracción de raíces. Sin embargo, cuando no estén disponibles los recursos necesarios para aplicar estas técnicas, o resulte de interés un sistema independiente del idioma, el uso de n-gramas permite obtener mejoras significativas. Un factor a tener presente es la longitud de las consultas y los documentos, así como el *ratio* de error presente en las primeras.