Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 127-128 recibido 13-01-11 revisado 15-02-11 aceptado 24-02-11

Categorización semi-supervisada de documentos usando la Web como corpus

Semi-supervised categorization of documents using the Web as corpus

Rafael Guzmán Cabrera
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n, Valencia
guzmanc@ugto.mx

Resumen: Tesis doctoral en reconocimiento de formas e inteligencia artificial realizada en la Universidad Politécnica de Valencia por Rafael Guzmán Cabrera bajo la dirección de los doctores Paolo Rosso y Manuel Montes y Gómez (INAOE, México). La defensa de la tesis tuvo lugar el 24 de noviembre ante el tribunal formado por los doctores Manuel Palomar Sanz (Universidad de Alicante), Paloma Martínez Fernández (Universidad Carlos III de Madrid), Luis Villaseñor Pineda (INAOE, México), Grigori Sidorov (Instituto Politécnico Nacional, México) y Antonio Molina Marco (Universidad Politécnica de Valencia). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad.
Palabras clave: Categorización, Semi-supervisado, Web, Corpus.

Abstract: PhD thesis in pattern recognition and artificial intelligence written by Rafael Guzmán Cabrera at the Universidad Politécnica de Valencia under the joint supervision of Dr. Paolo Rosso and Dr. Manuel Montes y Gómez (INAOE, México). The author was examined on november 24th, 2009 by the committee formed by Manuel Palomar Sanz (Universidad de Alicante), Paloma Martínez Fernández (Universidad Carlos III de Madrid), Luis Villaseñor Pineda (INAOE, México), Grigori Sidorov (Instituto Politécnico Nacional, México) and Antonio Molina Marco (Universidad Politécnica de Valencia). The grade obtained was Sobresaliente Cum Laude (highest mark)
Keywords: Categorization, Semi-supervised, Web, Corpus.

1 Introduccion

La mayoría de los métodos para la categorización automática de documentos están basados en técnicas de aprendizaje supervisado y, por consecuencia, tienen el problema de requerir un gran número de instancias de entrenamiento. Con la finalidad de afrontar este problema, en esta tesis se propone un nuevo método semi-supervisado para la categorización de documentos, el cual considera la extracción automática de ejemplos no etiquetados de la Web y su incorporación al conjunto de entrenamiento. Los ejemplos no etiquetados que se incorporan al conjunto de entrenamiento son seleccionados por medio de un método basado en aprendizaje automático. Este modelo incremental permite la selección sólo de los mejores ejemplos no etiquetados en cada iteración.

Para la selección de las instancias no etiquetadas se utilizan medidas tanto a nivel local (clase) como a nivel global (corpus).

2 Método propuesto

La finalidad de desarrollar un método semi-supervisado para llevar a cabo la tarea de categorización automática de documentos es el poder incorporar información no etiquetada proveniente de la Web al conjunto de entrenamiento.

En la figura 1, se muestra el esquema general del método propuesto. El objetivo de la etapa de adquisición de corpus es descargar ejemplos no etiquetados de la Web. Estos ejemplos son descargados para cada clase de manera independiente; para llevar a cabo esta tarea se construyen una serie de peticiones formadas por las palabras relevantes de cada clase.

ISSN 1135-5948

Llamamos palabras relevantes al conjunto de palabras que supera un umbral en las medidas local y global, es decir aquellas que permiten distinguir a una clase de otra.

La finalidad de la etapa Aprendizaje semi-supervisado es incrementar el tamaño del conjunto de entrenamiento para mejorar la precisión en la categorización.

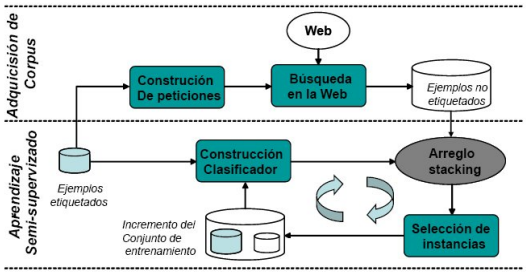

Figura 1 Método propuesto

El método propuesto es independiente del dominio y del lenguaje, su funcionamiento resulta más adecuado en aquellos escenarios en los cuales no se cuenta con suficientes instancias de entrenamiento manualmente etiquetadas.

La evaluación experimental del método se llevó a cabo por medio de cuatro experimentos en diferentes tareas y en dos diferentes idiomas. Tres de estos experimentos corresponden a la tarea de categorización, tanto temática como no-temática, de documentos y un experimento sobre la desambiguación del sentido de las palabras. En todos los casos se formaron conjuntos de entrenamiento y prueba, donde el conjunto de prueba nunca fue visto por el conjunto de entrenamiento. La incorporación de información no etiquetada, descargada de la Web, al conjunto de entrenamiento permitió, en todos los escenarios evaluados, mejorar la exactitud de referencia. Específicamente el método desarrollado fue evaluado en los siguientes escenarios:

(i) Categorización de noticias sobre desastres naturales (en español), caracterizado por tener muy pocos ejemplos de entrenamiento manualmente etiquetados (menos de diez).

(ii) Categorización de la distribución ModApte de Reuters (en inglés), la característica de este experimento es que la colección de documentos es grande, con diez clases y un alto grado de traslape.

(iii) Atribución de autoría de poemas (en español), es este caso se trata de la tarea de atribución de autoría, en la cual no queda claro cuáles son los atributos que deben ser utilizados para entrenar al sistema de categorización automática, por estar más bien relacionados con el estilo de escritura del autor.

(iv) Desambiguación del sentido de las palabras, usando sustantivos de la colección SemEval (en inglés). En este caso se lleva a cabo la tarea de WSD como una tarea de categorización, en la cual las clases corresponden a los diferentes sentidos que una palabra polisémica puede tomar.

Los resultados obtenidos en cada uno de estos experimentos nos permiten ver la efectividad de incorporar datos no etiquetados descargados de la Web al conjunto de entrenamiento.

Los corpora utilizados en el presenta trabajo de tesis se encuentran disponibles[1]. Una descripción detallada del método aplicado en las diferentes tareas de categorización evaluadas, así como los resultados obtenidos en cada una de ellas se encuentra en (Guzmán-Cabrera et al., 2009)

Bibliografía

Di Nunzio Giorgio M., Using scatterplots to understand and improve probabilistic models for text categorization and retrieval, International Journal of Approximate Reasoning, Elsevier, volume 50, Issue 4, pp: 581-594, 2009.

Glenn J. and Myatt N., Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, John Wiley, 2006.

Grieve, J., Quantitative Authorship Attribution: An Evaluation of Techniques, Literary and Linguistic Computing, volume 22, issue 3, pp: 251-270, 2007.

Guzmán-Cabrera R., Montes y Gómez M., Rosso P. y Villaseñor-Pineda L., Using the Web as corpus for self training text categorization. Journal of information Retrieval. ISSN: 1386-4564, No. 10791, 2009.

Lee C. and Lee G., Information gain and divergen-ce-based feature selection for machine learning-based text categorization, Information Processing and Management: an International Journal, volume 42, issue 1, pp: 155-165, 2006.

Sebastiani F., Classiffication of text, automatic, The Encyclopedia of Language and Linguistics, volume 2, pp: 457-463, Elsevier, Science Publishers, 2006.

[1] http://inaoep.mx/~mmontesg/resources