Método de clasificación automática de textos en base a keywords utilizando información semántica. Aplicación a Historias Clínicas

Roque E. López
Ingeniería de Sistemas
Universidad Nacional San Agustín
Arequipa, Perú
Email: rlopezc27@gmail.com

***Abstract*—En este artículo se presenta un método de clasificación de textos médicos en base a keywords, el cual utiliza información semántica. El clasificador propuesto consta de dos etapas: En la primera, se extraen keywords para cada clase de enfermedad del conjunto de entrenamiento, posteriormente se realiza un ranking de los keywords considerando la relación semántica que comparten. En la segunda etapa se calculan las similitudes entre la historia clínica a clasificar y los keywords de cada clase, eligiendo la clase más similar. Los resultados experimentales de este trabajo de tesis son alentadores, pues superan los resultados de métodos tradicionales, tales como Naive Bayes y Rocchio.**

***Keywords*-Clasificación de Documentos; Relación Semántica; Procesamiento del Lenguaje Natural**

I. Introducción

En los últimos años, la producción de textos en formato digital en Internet ha crecido en grandes proporciones. En esta inmensa cantidad de información, se pueden encontrar diversos tipos de documentos: noticias, libros, tutoriales, reportes médicos, entre otros. Estos grandes volúmenes de información han despertado el interés de varias áreas de la computación, una de ellas, el Procesamiento del Lenguaje Natural. El objetivo del Procesamiento del Lenguaje Natural (NLP por su siglas en inglés Natural Language Processing) es desarrollar modelos computacionales del lenguaje natural (lenguaje humano) para su análisis y generación [1].

En Internet existe una gran variedad de información. Acceder y analizar estas grandes cantidades de información de forma manual, es una tarea casi imposible, costosa en tiempo y recursos. Para utilizar de forma eficiente estos datos, en el área de Procesamiento del Lenguaje Natural se han desarrollado diversos métodos automáticos, tales como Clasificación de Documentos, Búsqueda de Documentos, Traducción Automática, Generación de Resúmenes Automáticos, etc. La clasificación automática de documentos ha sido una de las áreas con más estudios realizados en los últimos años. Esto se debe a su importancia, tanto en la industria, como en el ámbito académico. En la industria, la clasificación de textos, es importante debido a la gran cantidad de documentos que deben ser procesados y clasificados para un mejor análisis. Es importante en el ámbito académico, pues otras áreas de investigación dependen de ella, tales como: Búsqueda de Respuestas (Question Answering), Detección de Subjetividad (Subjective Detection), etc.

La clasificación automática de textos, también conocida como categorización de textos, es la tarea de asignar un documento dentro de un grupo de clases o categorías predefinidas [2]. La mayoría de los algoritmos y métodos propuestos para clasificar documentos se basan en información estadística tales como: frecuencia de aparición de palabras en el documento, frecuencia de aparición en las categorías, etc. Si bien es cierto, esta información es útil, podría complementarse con otro tipo de información para mejorar los resultados de la clasificación automática de textos.

En lo referente a historias clínicas, cada documento presenta varias secciones en las cuales se plasma, de forma resumida, información referente a antecedentes fisiológicos, patológicos, exámenes clínicos, diagnóstico, tratamiento, indicaciones, seguimiento médico, etc. Cada documento se caracteriza por tener poco texto y también porque existen muchas palabras que aparecen frecuentemente en todas las categorías, por ejemplo los términos: paciente, dolor, enfermedad, medicamento, etc. Estas palabras no aportan información importante a la enfermedad a la cual pertenecen. En este contexto, utilizar solo información estadística no ayudaría a clasificar correctamente las historias clínicas, con lo cual se vislumbra un problema en el rendimiento del clasificador.

En este trabajo de tesis se propone una solución alternativa, la cual buscará mejorar la clasificación de documentos médicos aprovechando la relación semántica existente entre los keywords de una historia clínica. Dicha relación semántica es extraída de la ontología de conceptos biomédicos UMLS (Unified Medical Language System). Para la evaluación del rendimiento del clasificador utilizó el corpus OHSUMED[1], una colección de documentos médicos en los cuales cada

[1]Los datos se pueden descargar desde la página: http://trec.nist.gov/data/t9_filtering.html

documento tiene asignado un tipo de enfermedad.

El resto del documento se organiza de la siguiente manera. En la sección 2 se presentan algunos trabajos relacionados con este tema de investigación. La sección 3 describe el método propuesto para la clasificación automática de historias médicas. Los experimentos y resultados se encuentran en la sección 4. Finalmente, en la sección 5 se exponen las conclusiones de este trabajo.

II. Trabajos Relacionados

Un ejemplo de clasificación de textos en el ámbito médico se da en [3], donde se aplica el método Naive Bayes para la tarea de clasificación de triajes médicos.

En [4], los textos utilizados describen las razones por la cual un paciente es internado. En este trabajo se utiliza una red bayesiana construida manualmente y las probabilidades se actualizan en la fase de entrenamiento.

En [5], tres clasificadores (K-Nearest Neighbor, relevance feedback y el clasificador bayesiano independiente) se aplican para asignar automáticamente códigos ICD-9 (*International Classification of Diseases, ninth revision*). Ellos muestran que la combinación de estos clasificadores obtienen el mejor rendimiento en la clasificación. [6] presenta un método basado en aprendizaje supervisado, donde el clasificador se entrena en un corpus formado por historias clínicas etiquetadas.

III. Método Propuesto

El método presentado en este artículo, a diferencia de los enfoques mencionados en la sección 2, además de considerar información estadística, toma en cuenta la relación semántica existente entre los keywords de una historia clínica. En esencia, el método propuesto consta de 2 etapas: Etapa de Entrenamiento y Etapa de Clasificación. La Figura 1 muestra la arquitectura del enfoque propuesto.

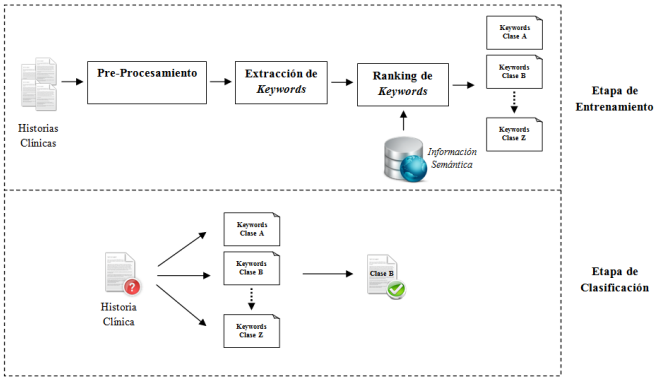

Figure 1. Arquitectura del Método Propuesto

A. Etapa de Entrenamiento

Un clasificador automático de textos requiere un conjunto de documentos clasificados manualmente, llamado *conjunto de entrenamiento* [7]. El objetivo de esta etapa consiste en extraer automáticamente los keywords más relevantes de cada tipo de enfermedad que existe en el conjunto de entrenamiento. Para tal efecto, se realizan tres pasos en esta etapa: pre-procesamiento de las historias clínicas, extracción de keywords y ranking de keywords.

- *Pre-procesamiento de las Historias Clínicas:*
Este paso tiene como finalidad eliminar las partes de las historias clínicas que no sean importantes, es decir, que no aporten significado [8] [9]. En este paso se eliminan los *stopwords* (pronombres, preposiciones, conjunciones, etc.). También se realiza una eliminación de los símbolos de puntuación.

- *Extracción de Keywords:*
Los keywords de un documento son las palabras que en forma precisa y compacta representan el contenido de un documento [10]. Algunas palabras, como por ejemplo, *medicamento, paciente, dolor*, etc., aparecen frecuentemente en todas las historias clínicas, y estas palabras no aportan información importante sobre la clase (enfermedad) a la cual pertenecen. Por este motivo, en este paso se propone un mecanismo de extracción de keywords en el cual, el keyword indique la importancia que tiene éste para una clase, y al mismo tiempo sea discriminante para las demás clases. Con este mecanismo, una palabra tendrá mayor valor para una clase, cuando más veces aparezca en ella y menos en las demás.

El peso de la palabra i-ésima $w_{i_{clase}}$, en relación a la clase se calcula como:

$$w_{i_{clase}} = tf_i \cdot log(\frac{N_{clases}}{n_{i_{clases}}}) \quad (1)$$

donde tf_i es el número de historias clínicas en la clase en los que la palabra i-ésima aparece, este valor es normalizado entre el total de documentos en la clase; N_{clases} es el total de clases; y $n_{i_{clases}}$ es el número de clases que tienen historias clínicas con la i-ésima palabra. En base a esta información estadística, para cada historia clínica del conjunto de entrenamiento se extraen las 5 palabras con mayor peso, las cuales serán consideradas los keywords de la historia clínica.

- *Ranking de Keywords:*
Una vez obtenidos los 5 keywords para cada historia clínica del conjunto de entrenamiento, el siguiente paso consiste en realizar un ranking considerando la relación semántica que existe entre los keywords. La relación semántica indica cuánto dos conceptos o términos son relacionados en una ontología con todas las relaciones entre ellos: relaciones lexicales (Hiperonimia y Sinonimia) y relaciones funcionales (tales como: es-un-tipo-de, es-parte-de, es-un-ejemplo-de, etc.) [11].

Si dos conceptos o términos, tienden a ocurrir juntos más a menudo de lo habitual, esto es un indicativo de que la relación semántica entre los términos es más fuerte. Por ejemplo, las palabras *próstata* y *micción*, tienen más relación que las palabras *próstata* y *digestivo*.

Para realizar el ranking de keywords, se utilizó la modificación del algoritmo PageRank usada en [12]. El algoritmo de PageRank [13] construye un grafo en base a las páginas web (nodos) y los enlaces (aristas) de entrada y salida de las mismas. El PageRank es un valor numérico que representa la relevancia que una página web tiene en Internet. En nuestro

caso, el PageRank representa la importancia de un keyword en el conjunto de entrenamiento.

A diferencia del algoritmo PageRank original, la modificación utilizada incluye un peso entre los nodos. En este escenario, la importancia de un keyword depende de los keyords que lo recomiendan y de la relación semántica entre ellos. El algoritmo PageRank modificado se muestra en la ecuación 2:

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{W_{i,j}}{|Out(V_j)|} S(V_j) \quad (2)$$

Donde $S(V_i)$ es el PageRank del keyword V_i. d es un factor de amortiguación que tiene un valor entre 0 y 1. $S(V_j)$ son los valores de PageRank que tienen cada una de los keywords que se encuentran en una misma historia clínica con V_i. $In(V_i)$ son los keywords que referencian a V_i. $Out(V_j)$ es el número total de enlaces salientes del keyword V_j.

$W_{i,j}$, el peso de la arista que enlaza los keywords V_i y V_j se calcula como:

$$W_{i,j} = tf_{i,j} * UMLS_{V_i,V_j} \quad (3)$$

Donde $tf_{i,j}$ es el número de veces de ocurrencias de los keywords V_i y V_j en una misma historia clínica. $UMLS_{V_i,V_j}$ es el peso asignado por la ontología UMLS, el cual corresponde a la relación semántica entre dichos keywords.

B. Etapa de Clasificación

Para clasificar nuevas historias clínicas, se estima la similitud entre el nuevo documento médico y los keywords de cada categoría (enfermedad). La categoría que obtenga un índice mayor de similitud es la categoría a la cual se asigna la historia clínica.

La idea de calcular la similitud consiste en conocer qué tantas características comparten la nueva historia clínica con los keywords, y no sólo eso, sino también saber si las características que comparten son importantes o no. En [14] se denota *intersección pesada* a esta forma de comparar documentos con las clases y se define como:

$$similitud(d,k) = \sum_{i \in d} w_{i_{doc}} \cdot w_{i_{clase}} \quad (4)$$

donde d es el documento que se quiere clasificar, k es el conjunto de keywords de la categoría k, $w_{i_{clase}}$ es el peso del keyword i-ésimo de la clase k, $w_{i_{doc}}$ representa el peso de la palabra i-ésima en el documento, que en este caso es la frecuencia de aparición de la palabra.

IV. Experimentos y Resultados

Para los experimentos se utilizó el corpus OHSUMED [15], el cual está conformado de 50.216 documentos médicos escritos en inglés. Generalmente los 10.000 primeros se utilizan para la etapa de entrenamiento y los 10.000 restantes se usan para la etapa de evaluación [16]. Este corpus contiene historias médicas de 23 diferentes enfermedades cardiovasculares. Se realizaron experimentos para evaluar la utilidad de la relación semántica en la clasificación de textos. También se realizó una comparación de la tasa de aciertos con los métodos de Naive Bayes y Rocchio.

En la Figura 2 se muestra una comparación de la tasa de aciertos del método propuesto utilizando dos tipos de rankings de keywords. El Ranking Simple utiliza el algoritmo PageRank, mientras que el Ranking Semántico, utiliza la modificación del PageRank (ver ecuación 2). El Ranking Semántico considera la relación semántica extraída de la ontología UMLS. Los resultados de la Figura 2 muestran que la relación semántica ayuda a mejorar la tasa de aciertos en la clasificación de historias clínicas.

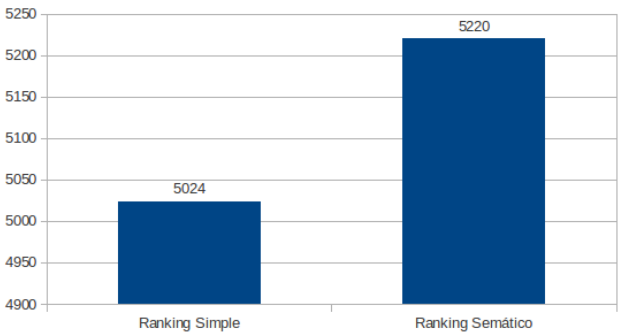

Figure 2. Comparación de Rankings

Los resultados mostrados en la Figura 3, indican que el método propuesto obtiene la mayor tasa de aciertos en la clasificación automática de historias clínicas. En los textos de las historias clínicas aparecen términos médicos que son importantes para cada categoría, por ejemplo: *indigestión, esofágico, acidez*, son significativos para la clase *sistema digestivo*, estas palabras aparecen con poca frecuencia en el texto. Sin embargo el clasificador propuesto toma en cuenta la importancia semántica que tiene un término en una clase, a diferencia de los enfoques Naive Bayes y Rocchio. Es por eso que en el método propuesto los keywords *indigestión, esofágico, acidez*, tienen más importancia para la clase *sistema digestivo* que para las demás, y es en base a estos términos importantes que se mejora la tasa de aciertos.

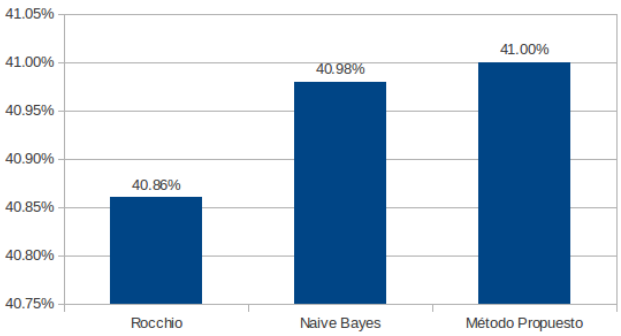

Figure 3. Comparación de Métodos

V. Conclusiones

En este trabajo se presentó un método para clasificar historias clínicas el cual mejora los resultados del método Naive Bayes y Rocchio. Este método, además de considerar información estadística, toma en cuenta la relación semántica que existe entre los keywords de los documentos médicos.

Los puntos más importantes a resaltar del presente artículo son: en primer lugar, que el método propuesto obtuvo resultados aceptables en la clasificación automática de historias clínicas. En segundo lugar, que la utilización de la relación semántica entre los keywords puede ayudar a mejorar el rendimiento de un clasificador de documentos médicos.

Cabe resaltar que este artículo es un resumen de un trabajo de tesis que se encuentra en la etapa final de escrita. Entre los principales pasos futuros se destacan: (1) Realizar comparaciones específicas de las clases en base a precisión, recall y f-measure. (2) Utilizar un analizador morfosintáctico como primer filtro en la extracción de keywords.

References

[1] B. Akshar, V. Chaitanya, and R. Sanga, *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi, 1996.

[2] R. M. Coyotl, "Clasificación automática de textos considerando el estilo de redacción," *Tesis de Maestría, Instituto Nacional de Astrofísica Óptica y Electrónica, México*, 2007.

[3] R. Olszewski, "Bayesian classification of triage diagnoses for the early detection of epidemics," *Proceedings of the FLAIRS Conference, pp. 412-416*, 2003.

[4] W. Chapman, L. Christensen, M. Wagner, P. Haug, O. Ivanov, J. Dowling, and R. Olszewski, "Classifying free-text triage chief complaints into syndromic categories with natural language processing," *Artificial Intelligence in Medicine 33, pp. 31-40*, 2005.

[5] L. Larkey and W. Croft, "Combining classifiers in text categorization," *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 289-297*, 2006.

[6] A. Argraw, A. Hulth, and B. Megyesi, "General-purpose text categorization applied to the medical domain," *DSV Research report - 2007-016*, 2007.

[7] C. Figuerola, J. Berrocal, A. Zazo, and E. Rodríguez, "Algunas técnicas de clasificación automática de documentos," *pp. 1-3*, 2004.

[8] R. López, D. Barreda, J. Tejada, and L. Alfaro, "Clasificación automática de historias clínicas basada en prototipos utilizando técnicas de procesamiento de lenguaje natural," in *Proceedings of the 10th Jornadas Peruanas de Computación*, 2011.

[9] R. López, D. Barreda, J. Tejada, and L. Alfaro, "Método supervisado orientado a la clasificación automática de historias clínicas," in *Proceedings of the 23rd Encuentro Chileno de Computación*, 2012.

[10] X. Jiang, Y. Hu, and H. Li, "A ranking approach to keyphrase extraction," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. ACM, 2009, pp. 756–757.

[11] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, ser. AAAI'06. AAAI Press, 2006, pp. 1419–1424.

[12] R. López, D. Barreda, J. Tejada, and E. Cuadros, "Mfsrank: An unsupervised method to extract keyphrases using semantic information," in *Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, I. Batyrshin and G. Sidorov, Eds. Springer Berlin Heidelberg, 2011, vol. 7094, pp. 338–344.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University, Technical Report;, 1998.

[14] J. D. Alvarez, "Clasificación automática de textos usando reducción de clases basada en prototipos," *Tesis de Maestría, Instituto Nacional de Astrofísica Óptica y Electrónica, México*, 2009.

[15] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "Ohsumed: an interactive retrieval evaluation and new large test collection for research," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '94. New York, NY, USA: Springer-Verlag New York, Inc., 1994, pp. 192–201.

[16] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of ECML-98, 10th European Conference on Machine Learning*, C. Nédellec and C. Rouveirol, Eds., no. 1398. Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998, pp. 137–142.