Método de Clasificación Automática de Textos Subjetivos

Protocolo de Investigación

M.C Adelina Escobar Acevedo

Antecedentes

De acuerdo con la UNESCO, las lenguas son instrumentos de que disponen los seres humanos para la interacción y la expresión de ideas, sentimientos, conocimientos, memorias y valores.[1] A través de ellas, el hombre tiene la posibilidad de acumular el conocimiento por medio de la comunicación oral y escrita. Es precisamente la escritura, base de la existencia de los textos, consecuencia de la voluntad humana de dar una materialidad perdurable a las experiencias y conocimiento. En los medios escritos, se ha encontrado una forma de preservar y transmitir la culturatanto en el espacio como en el tiempo.

En nuestra era, los textos han evolucionado al medio electrónico, los individuos utilizan las tecnologías de la información y comunicación con el fin de mantener una comunicación e interacción a distancia, aumentando el valor del tiempo, la comodidad y la relativa facilidad de mantener diversas tareas cotidianas a través de los medios tecnológicos. La creación de medios específicos de divulgación del conocimiento ha desencadenado en consecuencia una escalada de textos impresos y electrónicos. Dichos documentos conforman hoy repositorios desmedidos de información en continuo crecimiento.

Debido a la imposibilidad humana de manejar enormes cantidades de textos,nace la necesidad de automatizar procesos que manipulen y organicen tales volúmenes de documentos(Galicia Haro & Gelbukh, 2007). A fin de aprovechar la información contenida en los documentos, surgen distintas líneas de investigación como los sistemas de recuperación de información (*InformationRetrieval*), para buscar datos relevantes en grandes repositorios de documentos (Bolshakou & Gelbukh, 2004); búsqueda de respuestas (*QuestionAnswering*), un tipo de recuperación de información que permite a los usuarios plantear preguntas en lenguaje natural (Aceves Pérez, Villaseñor Pineda, & Montes y Gómez, 2007); generación automática de resúmenes(*Text Summarization*), que pretende extraer las líneas importantes de los documentos a fin de agilizar su lectura. (Villatoro Tello , Villaseñor Pineda, & Montes y Gómez , 2006);clasificación de textos (*Text Categorization*), para asignar etiquetas a los textos de forma automática; entre otros.

[1] http://www.unesco.org/culture/ich/index.php?pg=00136

La clasificación o categorización se define como la tarea de asignar objetos de un universo a dos o más clases predefinidas. Las clases o categorías son las opciones con las que se puede asignar una etiqueta. Para tomar una decisión de la clase a la que pertenecen los objetos, es necesario conocer las características particulares de cada clase. Específicamente la clasificación de textos tiene por objetivo asignar automáticamente la clase apropiada a cada documento. En particular, tiene aplicaciones como filtrar un caudal de noticias para un grupo de interés particular; clasificación de textos de opinión, sentimiento o juicio; atribución de autoría; organizar grandes volúmenes de información de acuerdo a cierta taxonomía, por ejemplo, en bibliotecas, etc.

La clasificación automática de textos en el llamado aprendizaje computacional (o aprendizaje automático) es comúnmente una tarea supervisada, lo que significa que no solo se conocen las categorías sino que debe contarse con un conjunto de entrenamiento. El conjunto de entrenamiento es un grupo de documentos previamente etiquetados con las clases a las que pertenecen. Cada documento es representado por una serie de variables características de la clase. La tarea de clasificación se precede por el aprendizaje de un modelo que usa información del conjunto de entrenamiento, almacenando los valores de las variables junto con la clase a la que pertenecen (Sierra Araujo, y otros, 2006).

Nuevos retos se han presentando ante la tarea de clasificar automáticamente textos subjetivos, de acuerdo a opiniones y sentimientos y no de acuerdo a su temática. Este tipo de clasificación, conocida como*sentimentclassification, sentimentanalysis y opinionmining* ha generado recientes investigaciones debido al gran interés por parte de la comunidadde investigadores del procesamiento del lenguaje natural. Dichas investigaciones se han generado debido a la creciente cantidad de textos generados por los usuarios en Internet y por los posibles beneficios que conllevaría a las empresas yadministraciones públicas en analizar, filtrar o detectar automáticamente las opinionesvertidas por sus clientes o ciudadanos.

Los textos generados en los foros, blogs y demás documentos existentes en Internet que versan sobre opiniones, discusiones, tendencias, características etc. de algún producto o tema común, son una fuente de información que auxilia a un posible consumidor y genera a la vez un nicho de oportunidad a los productores y comercializadores Sin embargo, determinar de forma automática la inclinación del usuario hacia la aceptación o rechazo de algún producto o de características particulares del mismo, es unproblema complejo que requiere de varios pasos para su resolución.

Entre la vasta variedad de aplicaciones que tienen los sistemas de minería de textos subjetivos, se encuentran análisis automatizados de opiniones de obras literarias, películas y demás productos comerciales hasta estudios que consideren el progreso de percepciones ciudadanas sobre organizaciones, figuras políticas, programas culturales y sociales, etc. Debido a las prometedoras aplicaciones existen varias empresas, tanto grandes como PYMES, para las que la minería deopiniones y el análisis de emociones tiene una alta prioridad.

Existen muy diversas direcciones en la investigación sobre extracción de opiniones. Porejemplo, en el estado del arte se encuentran trabajos que intentan identificar palabras queson asociadas a priori con conceptos subjetivos o emotivos (Riloff & Wiebe, 2003); (Kim & Hovy, 2004). Porotro lado, otros trabajos intentan determinar la subjetividad de una oración. Estos trabajos(Wiebe & Riloff, 2005) intentan determinar las frases objetivas, aquellas que afirman o describen un hecho, de aquellas frases subjetivas, las cuáles expresan una opinión. Por último, existen trabajos orientados a la identificación de documentos subjetivos, los cuales clasifican un texto completo como positivo o negativo(Das & Chen , 2001); (Turney, 2002); (Pang et al., 2002)

Objetivos y Metas

Objetivo General

Desarrollar un método semi-supervisado de clasificación de textos subjetivos aplicable a diversos dominios para los idiomas inglés y español.

Objetivos Específicos

- Determinar por medios semi-supervisados un conjunto de atributos (palabras osecuencias de palabras) para la clasificación de segmentos subjetivos a favor y en contra.
- Determinar por medios semi-supervisados un conjunto de patrones de extracción deopiniones por tipo semántico del objeto evaluado.
- Determinar el impacto de uso de n-gramas en los resultados de la precisión, recuerdo y F-mesure de la clasificación subjetiva.
- Evaluar el método con un corpus en idioma español y un corpus en idioma inglés.

Metas

- Ampliar el conocimiento del procesamiento automático del lenguaje humano en especialpara el idioma español.
- Definir nuevos métodos para la extracción de información con mínima supervisiónhumana.
- Cooperación con otros grupos de investigación en el área. Se involucrarán relaciones de colaboración para abordarconjuntamente la problemática del tratamiento del lenguaje humano; y se pondrán adisposición de la comunidad científica nacional e internacional los recursos lingüísticos generados durante el desarrollo del proyecto. Actualmente existe una colaboración con la Universidad de Guanajuato que cuenta con un Cuerpo Académico consolidado y se integra la Universidad Tecnológica del Suroeste del Estado (UTSOE) cuyo cuerpo académico se encuentra en consolidación.

Metodología

En todo proceso de clasificación existen 2 etapas bien delimitadas.

Etapa 1: Consiste en la preparación del corpus e incluye:

La creación del corpus: Implica la recopilación, lectura, segmentación y etiquetado del corpus de entrenamiento y prueba por una o varias personas que determinen a su juicio si el texto debe pertenecer a una clase u otra.-Cabe mencionar que el proyecto se está realizando en colaboración con la Universidad de Guanajuato, quienes han avanzado en la construcción de un corpus subjetivo etiquetado que hasta el momento no existía para el idioma español.

El pre procesamiento: Unifica los textos provenientes de diversas fuentes a una codificación única permitiendo homogeneizar caracteres. Se aplica un programa a cada archivo con el fin de obtener un conjunto de archivos en el cual se han eliminado caracteres indeseados, unificado acentos y mayúsculas etc.

Representación de documentos: Incluye esquemas de pesado y reducción de dimensionalidad, define la forma de representación utilizada basándose en los objetivos de la clasificación. La selección de estos criterios no puede ser al azar, debe probarse y compararse para determinar las ventajas de un esquema particular. La dimensionalidad es

uno de los problemas más fuertes del trabajo con textos debido a que fácilmente pueden obtenerse y trabajarse matrices con miles de columnas y reglones. Los cálculos implicados repercuten fuertemente en los tiempos de solución y por ello debe considerarse alguna forma de reducción de atributos apropiada para el problema.

Construcción del clasificador: Existen clasificadores que han probado reiteradamente su eficacia en el procesamiento de textos, tales como Naive Bayes, vecinos más cercanos y Máquinas de vectores de soporte. Para ser comparables con otros trabajos de investigación debe considerarse la obtención de resultados con los clasificadores conocidos. Una serie de experimentos deben ser realizados antes de iniciar con el desarrollo del nuevo método a fin de establecer una referencia con lo ya existente, dichos experimentos implican evaluar el comportamiento de los métodos tradicionales sobre los corpus recolectados.

Etapa 2: Consiste en el desarrollo del nuevo método.

En esta etapa es importante identificar el conjunto de palabras o secuencia de palabras (n-gramas) que son necesarios y suficientes para la clasificación. En el caso específico de los textos subjetivos debe determinarse el grupo de atributos que marcan la preferencia del texto., para ello se utilizan técnicas de aprendizaje automático y el corpus de textos subjetivos previamente etiquetado, debe considerarse la longitud de los textos, el balance entre clases, la densidad del dominio entre otros aspectos.

Determinar un proceso iterativo que considere la ampliación de los conjuntos y mejore la precisión de la clasificación Para ello se debe establecer conforme a los resultados algún nivel de confianza en los nuevos atributos o ejemplos. El número de iteraciones puede predeterminarse si las características del conjunto favorecen la integración reducida o puede delimitarse a recibir aquellos ejemplos que cumplan criterios rigurosos que incrementen su grado de confianza.

Se probarán diferentes esquemas de integración con variaciones en las restricciones, número de iteraciones y características del conjunto objetivo y se realizarán las evaluaciones de los conjuntos generados de forma automática. Se espera obtener resultados suficientes para reducir la dependencia de dominios en textos subjetivos.

Infraestructura disponible

El presente trabajo se desarrolla dentro de la Universidad Tecnológica del Valle de Toluca, conformando una nueva línea de investigación para la misma y atraer con ello nuevas áreas de oportunidad y de vinculación con otras universidades. Para el procesamiento de textos la Universidad Tecnológica del Valle de Toluca cuenta con únicamente con software libre y una computadora asignada al investigador. Es necesario adquirir una computadora dedicada exclusivamente a este proyecto con capacidades suficientes para el manejo de matrices de alta dimensionalidad, licencias de software como Matlab, Editplus y consumibles,

Es apropiado recalcar el apoyo y entusiasmo que ha demostrado la Universidad Tecnológica del Valle de Toluca permitiendo acceso a los recursos existentes en:

- Biblioteca: Para consultas de propósito general y temas relacionados.
- Servicio de cómputo: Se cuenta con personal capacitado para soporte e instalación de equipo de cómputo y paquetería, además de recursos compartidos (como impresora, servidores, etc.) y acceso a internet.
- Servicios administrativos: Contamos con el servicio secretarial y apoyo a diversas tareas relativas a las cuatro funciones de los Profesores de Tiempo completo que facilitan las labores desempeñadas.

Incidencia del proyecto en el Programa Integral de Fortalecimiento Institucional (PIFI)

Dentro del desarrollo de Tecnologías del Lenguaje se considera dos impactos importantes:

El desarrollo de nuevos métodos de procesamiento de Lenguaje para generación de conocimiento global y la contribución de dicha investigación sobre el idioma específico.Los grupos internaciones de investigadores dedicados a la tarea de procesamiento de lenguaje natural forman redes de colaboración a fin de aprovechar los recursos y avances de otros grupos de investigación alrededor del mundo.

Derivado de lo anterior y dado que ésta es una nueva línea de investigación dentro de la Universidad Tecnológica del Valle de Toluca, se considera una rama precursora cuyos objetivos iniciales son:

- Impulsar el desarrollo de nuevas tecnologías en el área de Procesamiento Natural de Lenguaje Humano.
- Fomentar la colaboración con redes académicas de otras universidades mexicanas y españolas que estén interesadas en el desarrollo de tecnologías del lenguaje español.
- Marcar un antecedente para que los estudiantes aprovechen directamente la participación dentro de la línea de investigación y las colaboraciones con otras instituciones impulsando con ello la investigación de tecnologías en el idioma español e inglés.

Referencias

Aceves Pérez, R., Villaseñor Pineda, L., & Montes y Gómez, M. (2007). Using n-gram Models to Combine Query Translatiosn in Cross-Language Question Answering. *International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2007*, 485-493.

Bolshakou, I., & Gelbukh, A. (2004). *Computational Linguistics.* México: Fondo de Cultura Económica.

Das, S., & Chen , M. (2001). Yahoo form AMazon Opinion Extraction from Small Talk on the Web. *Proceedings of the 8th Asia Pacific Finance Association annual Conference*.

Galicia Haro, S. N., & Gelbukh, A. (2007). *Investigaciones en Análisis Sintáctico para el Español.* México: Instituto Politécnico Nacional.

Kim , S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. *International Conference on Computational Linguistics*, (págs. 1367,1373). Geneva.

Riloff, E., & Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Sierra Araujo, B., Arbelaitz, O., Armañanzas, R., Arruti, A., Bahamonde, A., Borrajo, D., y otros. (2006). *Aprendizaje Automático: Conceptos básicos y avanzados.* (M. Martín Romo, Ed.) España: Pearson Educación S.A.

Turney, P. (2002). Thumb Up or Thumbs Down? Semantic Orientation Applied to Unsaupervised Classification of Reviews. *Proceedings of the 40 th Annual Meeting of the Association for Computational Linguistics*.

Villatoro Tello , E., Villaseñor Pineda, L., & Montes y Gómez , M. (2006). Using Word Sequences to Text Summarization. *LNCS*, 297.

Wiebe, J., & Riloff, E. (2005). Creating Subjetive and Objective Sentence Classifiers form Unnannotated Texts. *CICLing*.