Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 297-298 recibido 21-04-10 revisado 18-05-10 aceptado 20-05-10

Sentitext®: sistema de análisis de sentimiento para el español

Sentitext: A Sentiment Analysis System for Spanish

Antonio Moreno Ortiz
Facultad de Filosofía y Letras
Universidad de Málaga
Campus de Teatinos
29071 Málaga
amo@uma.es

Álvaro Pérez Pozo
Facultad de Filosofía y Letras
Universidad de Málaga
Campus de Teatinos
29071 Málaga
0617539139@uma.es

Sergio Torres Sánchez
ETSIIT
Universidad de Granada
Periodista D. Saucedo Aranda
18071 Granada
sergiot@correo.ugr.es

Resumen: Presentamos *Sentitext*, un sistema de análisis de sentimiento con arquitectura cliente-servidor para el español, cuya característica más distintiva es el estar basado en conocimiento. En primer lugar describimos las fuentes de conocimiento léxico que el analizador utiliza y a continuación describimos el analizador. Mostramos el funcionamiento de la primera aplicación cliente on-line, que hemos desarrollado en Adobe Flex.
Palabras clave: Análisis de sentimiento, minería de opiniones, valoraciones de usuario on-line

Abstract: We present Sentitext, a client-server architecture, sentiment analysis system for the Spanish language whose most distinctive feature is being knowledge-based. First we describe the lexical knowledge sources that the software uses for its analysis, and then we describe the analyzer itself. We demonstrate the first on-line client application, developer in Adobe Flex.
Keywords: Sentiment analysis, opinion mining, on-line user reviews.

1 Sentitext[*]

Sentitext es un conjunto de aplicaciones para el análisis de sentimiento en textos. La aplicación principal ha sido desarrollada con el lenguaje C++ y hace uso de la librería de análisis morfológico de *Freeling* (Atserias et al., 2006). A esta aplicación se accede mediante una aplicación servidor escrita en Python que permite la comunicación con las aplicaciones cliente. En la actualidad hemos desarrollado una aplicación cliente mediante Flex y toda la información referente al análisis está codificada en XML. El cliente muestra el resultado del análisis de forma intuitiva mediante gráficos y texto según la polaridad de los segmentos analizados y otorga una valoración global al texto analizado. Actualmente estamos explorando posibilidades en cuanto a representaciones gráficas de contenidos textuales desde el punto de vista de la afectividad, aspecto en el que no se ha trabajado mucho (Liu et al., 2003). El modo de acceso a la información para el análisis se realiza mediante consultas a una base de datos MySQL.

2 Las fuentes de datos

Sentitext es un software basado en conocimiento, y su funcionamiento se apoya en tres bases de datos. La principal de ellas es la de palabras (*Words*), que contiene más de 10.000 entradas. A cada palabra se le ha asignado una valencia que indica la carga afectiva de la misma; más específicamente, una palabra puede tener valencia -2, -1, 1 o 2, dependiendo de si su carga afectiva es muy negativa, negativa, positiva o muy positiva, respectivamente. Para la inserción de las palabras se ha recurrido a un diccionario de sinónimos con palabras semilla: el diccionario de sinónimos en español de OpenOffice. Las valencias, por su parte, han sido asignadas en base a la semántica de la palabra y con la ayuda de corpus de textos de

[*] Esta herramienta ha sido diseñada por el Grupo de investigación de la UMA *Tecnolengua* (http://tecnolengua.uma.es), con la financiación de El Jardín de Junio S.L.U. mediante convenio OTRI con la UMA Nº 8.06/5.21.3199-1.

ISSN 1135-5948

lengua general; a partir de estos datos, los lexicógrafos que trabajan con Sentitext llegan a un consenso y deciden cual es el valor afectivo más adecuado para la palabra. La segunda fuente de datos son las locuciones o expresiones multi-palabra (*MWords*). Se trata de un lexicón de alrededor de 17.000 entradas.

Tanto esta base de datos como la de palabras están sometidas a continuas actualizaciones, ya sean nuevas inserciones o modificaciones de las entradas ya existentes. Diariamente analizamos un buen número de textos, estudiamos el resultado y empleamos este feedback para mejorar tanto las fuentes de datos como el analizador en sí mismo.

La inclusión de locuciones en el software de SA no es algo que hasta ahora se haya tenido muy en cuenta, pero sin duda mejora enormemente los resultados: "aguar la fiesta" contiene una palabra positiva pero es una expresión negativa, e "ir sobre ruedas" es una expresión negativa formada a partir de palabras neutras. El tercer pilar sobre el que se sustenta el analizador son las reglas de contexto (*CRules*), basadas en las *Context Valence Shifters* de Polanyi y Zaenen (2006). El objetivo de las mismas es tener en cuenta el hecho de que ciertas unidades léxicas modifican la carga afectiva de otras unidades cercanas.

3 El analizador

El proceso de análisis está compuesto de cuatro partes fundamentales:

1. Lematización y etiquetado morfológico del texto.
2. Asignación de valencias: se recorre la lista de unidades léxicas obtenidas y, usando como referencia el lema, se busca la valencia de cada una de las unidades en las bases de datos.
3. Aplicación de las reglas de contexto, consistente en recorrer una vez más la lista de unidades léxicas, y en caso de encontrar un modificador que cumpla las restricciones indicadas para una regla dada (posición, cercanía y naturaleza del elemento a modificar), se transforma apropiadamente la valencia de la unidad modificada. Esta parte es la más delicada, ya que un mismo modificador puede tener asociadas varias reglas de contexto, y un mismo elemento modificado puede ser objetivo de varios modificadores, de forma que es necesario especificar un orden jerárquico o unas prioridades en su aplicación.
4. Finalmente, en la fase de extracción de datos se obtiene información derivada de los análisis anteriores, como el índice afectivo (cantidad de palabras con carga afectiva en relación con el número total de palabras) o el índice global, que intenta dar una idea aproximada de la positividad o negatividad del texto.

La forma de calcular dicho valor global no es algo trivial. En principio, podría pensarse en una media aritmética de las valencias de las unidades léxicas, pero esto presenta varios problemas, entre ellos, que a un texto de gran longitud con una única palabra positiva se le asignaría el máximo valor, o que a un texto completamente neutro se le asignaría el mismo valor global que a uno con la mitad de palabras positivas y la otra mitad negativas.

Actualmente, utilizamos como valor global una media aritmética ponderada, que se somete a una modificación posterior basándose en el índice afectivo: si el texto contiene muchas unidades con valencia afectiva distinta de cero, el valor puede moverse más libremente hacia los extremos, en caso contrario, se tiende a centralizar esta medida.

Bibliografía

Atserias, J. et al. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En Proceed*ings of the fifth international conference on Language Resources and Evaluation*. LREC 2006. Genoa, Italy: ELRA.

Liu, H., Selker, T. & Lieberman, H. 2003. Visualizing the affective structure of a text document. In *CHI '03 extended abstracts on Human factors in computing systems*. Ft. Lauderdale, Florida, USA: ACM, pp. 740-741.

Polanyi, L. & Zaenen, A. 2006. Contextual Valence Shifters. In *Computing Attitude and Affect in Text: Theory and Applications*. Dordrecht, The Netherlands: Springer, pp. 1-10.