

Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias

Emanuel García Pérez

October 31, 2014

1 INTRODUCCIÓN

2 ESTADO DEL ARTE

3 CLASIFICACIÓN DE OPINIONES BASADA EN DEPENDENCIAS

- Propuesta Base
- Tratamiento de la Intensificación
- Tratamiento de las oraciones adversativas
- Tratamiento de la Negación
 - Identificación del alcance de la negación
 - Modificación de la polaridad

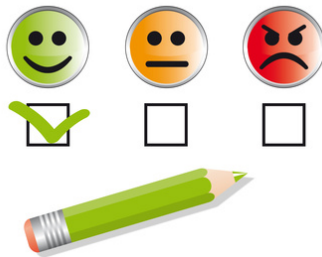
4 RESULTADOS EXPERIMENTALES

- Implementación
- Evaluación

5 CONCLUSIONES Y TRABAJO FUTURO

6 BIBLIOGRAFÍA

- En los últimos años el alcancen de los blogs, foros y redes sociales han hecho que millones de usuarios los usen para expresar sus opiniones sobre diversos temas.
- Esta gran variedad de opiniones y críticas que abundan en la web son de gran utilidad para que vendedores y fabricantes conozcan el impacto de sus productos en los consumidores.



Debido a las ventajas que presenta toda esa información y la complejidad que conlleva su análisis es necesario la búsqueda de soluciones eficientes capaces de monitorizar este flujo de información. La minería de opinión(MO) se enfoca en el tratamiento automático de información con opinión, lo que permite extraer la polaridad(positiva, negativa, neutra, mixta) de un texto.

Polaridad en textos

La clasificación de la polaridad en textos es un problema de gran relevancia en la MO, este ha sido abordado desde dos enfoques:

- Concebir esta tarea como un proceso genérico de clasificación, y a partir de un conjunto de datos de entrenamiento, donde los textos son anotados con su respectiva polaridad, se construye un clasificador mediante aprendizaje automático(AA).
- Apoyarse en la orientación semántica(OS) de las palabras, cada término que expresa opinión es anotado con un valor que representa su polaridad.

Sistemas de MO

La mayoría de los sistemas de MO existentes se enfocan en tratar únicamente textos en inglés. Para textos en español el, quizás, sistema más relevante es "The Spanish SO Calculator", desarrollado en la universidad Simon Fraser de Canadá.

Spanish SO CA

- Resuelve la OS almacenada a nivel individual en: adjetivos, sustantivos, verbos y adverbios.
- Trata modificadores de la polaridad, como la negación o intensificadores(" muy", " poco", " bastante").
- Detecta y excluye el sentimiento reflejado en el contenido no fáctico del texto(expresiones condicionales o subjuntivas.)
La forma más común de tratar estas construcciones lingüísticas es a nivel léxico, y "The Spanish SO CA" no es la excepción.

Negación

Respecto a la negación tenemos lo siguiente:

- Taboada, 2011: utiliza información morfológica para identificar el alcance de la negación.
- Yang, 2008: considera el alcance de la negación como los términos a la derecha de la negación.
- Fernández Anta, 2012: se emplea una heurística que asume que los tres elementos a continuación de una negación son los que deben cambiar su polaridad.

Intensificación

Para los intensificadores tenemos:

- Fernández Anta, 2012: considera de nuevo que los tres términos a la derecha son los que deben variar su polaridad.
- Taboada, 2011: además de los intensificadores propiamente dichos, trata como tales aspectos del discurso como la conjunción "pero" o las mayúsculas.

La propuesta utilizada se basa en obtener la estructura sintáctica del texto para tratar las construcciones lingüísticas e identificar los elementos de la frase que están implicados en ellas.

Trabajos anteriores (Jiu, yuMeng, 2009) han mostrado los beneficios de utilizar la estructura sintáctica de la frase en textos con ocurrencias de términos negativos.

Otro problema que enfrentan los sistemas de MO es la calidad ortográfica de los textos. Si estos provienen de la web, se debe tener en cuenta que a menudo sus autores omiten acentos, letras, vocablos, etc., o utilizan abreviaturas no reconocidas y oraciones agramaticales.

La solución más utilizada es usar patrones heurísticos para adaptar el texto.

En contraste a las propuestas léxicas dominantes, se propone utilizar la estructura sintáctica de la frase para obtener la OS de un texto. Inicialmente se debe hacer un preprocesamiento de los textos que contemple los siguientes aspectos:

- Unificación de expresiones compuestas que actúan como una sola unidad de significado("a menos que").
- Normalización de los signos de puntuación.

Posteriormente se debe segmentar el texto en oraciones y tokenizar cada una, para después hacer el etiquetado morfosintáctico de cada palabra del texto.

Diagram illustrating the structure of the sentence "Ese ordenador es rápido" (That computer is fast) in Spanish, showing the relationship between the root and the modifiers:

```

graph LR
    ROOT ---|sentence| Ese
    ROOT ---|sentence| ordenador
    ROOT ---|sentence| es
    Ese ---|spec| ordenador
    ordenador ---|suj| es
    es ---|atr| rapido
  
```

The diagram shows the following components and relationships:

- ROOT** (Root)
- Ese** (specifier)
- ordenador** (subject)
- es** (copula)
- rápido** (attribute)

The relationships are indicated by arrows and labels:

- spec** (specifier): connects **Ese** to **ordenador**.
- suj** (subject): connects **ordenador** to **es**.
- atr** (attribute): connects **es** to **rápido**.

Finalmente para realizar el análisis semántico, se utiliza SODcictionariesV1.11Spa(Brooke, Tofiloski, Taboada, 2009), un conjunto de diccionarios de polaridad para adjetivos, sustantivos, verbos, adverbios e intensificadores. Cada término tiene asociado un valor entre -5 y 5, siendo -5 lo más negativo y 5 lo más positivo. El valor asignado a cada palabra corresponde con una OS genérica, independientemente del dominio o contexto donde se use.

Ej: rápido(adj):= 2, recomendar(verb):= 2

Diccionario	Nº términos
adjetivos	2,049
sustantivos	1,324
verbos	739
adverbios	548
intensificadores	157

Los valores numericos asociados a los intensificadores tiene un significado distinto, representan el porcentaje(positivo o negativo) por el que modifican el sentimiento de la expresión a la que afectan.

Se determina la polaridad de un texto a partir de la combinación de la OS de sustantivos, adjetivos, verbos, y adverbios, sin considerar la estructura sintáctica del texto (construcciones lingüísticas complejas).

Los intensificadores son términos o expresiones que modifican la polaridad de ciertas palabras. Podemos clasificarlos en dos tipos:

- Amplificadores: permiten aumentar la polaridad; ("muy", "bastante").
- Decrementadores: permiten disminuir la polaridad; ("poco", "en absoluto").

Para modelar esta construcción, se asocia a cada intensificador un factor de ponderación. Se utiliza el árbol de dependencias para determinar la parte de la frase que se ve afectada por tal modificación, considerando las dependencias anotadas en Ancora.

Si hay varios intensificadores presentes, se combinan sus porcentajes de intensificación antes de que actúen sobre el término afectado.

Ej: "muy" := 0.25, "en absoluto" := -1




"muy rápido"—OS: $2 \cdot (1 + 0.25) = 2.5$

"en absoluto muy rápido"—OS: $2 \cdot (1 + (-1 + 0.25)) = 0.5$

Los nexos adversativos contraponen hechos expresados en dos oraciones. En MO este tipo de frases se emplean para restringir o excluir opiniones, que puede ser considerado como un caso especial de intensificación.

Un árbol de dependencias permite identificar con precisión la oración subordinada y la subordinante.

Bibliografía

-  Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación. Kuna, H., Rey, M., Martini, E., Solonezen, L., Podkowa. Revista Latinoamericana de Ingeniería de Software 2014.
-  THE AR-INDEX: COMPLEMENTING THE H-INDEX. Bihui Jin. National Library of Science CAS, Beijing, 100080, China. 2005-2007.
-  <http://www.core.edu.au/>