Creado el 17/09/2013
Última versión: 17/09/2013
Maite Taboada, Simon Fraser University, Canadá
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Aplicaciones de las TLH/ANÁLISIS DE LA OPINIÓN

GENERACIÓN DE RECURSOS PARA EL ANÁLISIS DE OPINIÓN

Después de conocer los rudimentos del análisis de opinión y hasta alguna aplicación real, toca generar los recursos necesarios para poder abordar la tarea con éxito.

Transcripción del vídeo

¡Hola! Mi nombre es Maite Taboada, y hoy voy a presentar una lección sobre la generación de recursos para el análisis de la opinión.

En el análisis de la opinión generalmente hay dos enfoques básicos: el primero utilizando diccionarios; el segundo mediante el aprendizaje automático, en esencia esta es una tarea de aprendizaje automático supervisado.

En ambos casos, es necesario tener importantes recursos. En el primero, es necesario tener diccionarios que se utilizan para extraer la opinión. Y en el segundo, es necesario tener un corpus con un mínimo nivel de anotación, que es lo que nos permite aprender, para textos nuevos.

Vamos a empezar entonces con el primer caso, el sistema que necesita de diccionarios.

En un sistema basado en diccionarios, lo primero que hay que hacer es crear un diccionario. El diccionario incluye una palabra y su polaridad, y también se puede incluir el grado. Por ejemplo, “fantástico” podemos decir que su polaridad es positiva, es un adjetivo positivo. Y como grado, si queremos añadir eso además, podemos decir que es un grado de 5 positivo, en el sistema por ejemplo en el que 5 es el máximo positivo.

En un sistema de este tipo, una vez que tenemos textos nuevos, lo que hacemos es buscar palabras en el texto que estén en el diccionario, y se agregan los valores del texto. Y al final damos un valor final para el texto: decimos que el texto contiene opinión positiva o negativa.

Por ejemplo, un extracto de una reseña de una película. El texto dice: “Cuando algo me gusta tanto, utilizaría palabras como sublime, inigualable, imprescindible…” Dejo que terminen de leer…

Bien, en este texto, hemos señalado palabras que en un contexto aislado pueden expresar opinión positiva o negativa, por ejemplo “gusta” es una palabra positiva. La hemos puesto en verde. “Usadas”

puede ser negativa. Y en este caso vemos que hay, en general, más palabras positivas que negativas. La conclusión de un sistema basado en diccionarios sería que este texto es positivo, pero con algunos negativos.

Si vemos un sistema un poquito más avanzado, que incorpora frases, negación y expresiones hechas, nos indica que “gusta” no es solamente “gusta”, sino “gusta tanto”, es decir, es positivo a un nivel más avanzado. “Tan usadas” también es negativo a un nivel más avanzado. Si vamos al final del texto, incorporando negación, “defraudará”, que era negativo anteriormente, ahora nos damos cuenta de que es positivo, porque tiene una negación, “no os defraudará”. El texto también incluye expresiones hechas, como “hay que verla”, que en un contexto de reseñas es positivo; “no perdáis la oportunidad”, “perdáis” tiene una negación delante, pero además “perder una oportunidad”, o “no perder una oportunidad” en este contexto, es también una recomendación.

Entonces, en este tipo de sistemas lo que es necesario es un diccionario. Un posible diccionario para el texto anterior sería de esta manera. “Gusta” es una palabra positiva, “sublime” es positiva, “inigualable” también es positiva. Sería necesario, por supuesto, incorporar lematización: decir que “gusta” es una forma del verbo “gustar”, igual que “perdáis” o “defraudará”. Lo que el diccionario generalmente incluiría sería el lema, la forma base de la palabra. Con adjetivos, igualmente, “usadas” tendríamos la misma forma para “usadas”, “usados”, “usado” y “usada”, y el diccionario incorporaría solamente el lema y tendríamos que usar un lematizador para saber a qué lema pertenece la palabra “usadas”.

Entonces, el sistema lo que hace es ir al diccionario, extraer la polaridad y hacer una agregación, que puede ser un promedio simplemente del número de palabras positivas y negativas.

Recursos adicionales que se necesitan, por supuesto, son una lista de operadores de negación: “no, nunca, nadie”, incluidas algunas palabras negativas, de polaridad negativa.

Expresiones que tienen polaridad convencional, como “hay que verla”, e intensificadores. En el texto anterior teníamos “tanto” y “tan”.

Otras posibilidades de diccionarios más avanzados sería un diccionario que, además de la polaridad, incluye el grado de intensidad. “Gusta” podríamos decir, de nuevo en un sistema donde 5 es el máximo, que tiene un grado de 3, “sublime” de 5, y por contrario, “usadas”, por ejemplo, un poco más abajo, negativo de -2.

La cuestión es, ¿cómo se crean estos diccionarios? Una opción es crearlos automáticamente, y la siguiente es manualmente. Si son automáticos, se extrae información de colocaciones típicas, o se puede minar un diccionario existente de sinónimos.

Manualmente, se añaden valores positivos o negativos a palabras que se consideran relevantes.

Bien, observemos el primer caso, diccionarios creados automáticamente, utilizando diccionarios existentes de sinónimos y antónimos. Por ejemplo, el diccionario online Wordreference.com. Para la palabra “inteligente”, buscamos la palabra “inteligente” porque sabemos, por ejemplo, que es una palabra positiva, y una vez que encontramos “inteligente”, podemos construir el diccionario con todas estas palabras positivas que son sinónimos de “inteligente”: “listo, capaz, perspicaz”, etc. Y además podemos incluir palabras como “tonto” e “ignorante”, con polaridad negativa, porque están en la lista de antónimos para “inteligente”. Y así, de esta manera, se puede extender un diccionario y crear un

diccionario bastante grande.

La otra forma de crear un diccionario automáticamente es usando información de corpus, utilizando palabras semilla, en inglés se denominan “seed words”. Si por ejemplo sabemos que “inteligente” es una palabra positiva, buscamos en qué contextos se encuentra. Una forma muy fácil de encontrar contextos es con conjunciones coordinativas, como “y” y “pero”. Y sabemos que “y” coordina dos palabras que generalmente suelen tener la misma polaridad, y por el contrario, “pero” coordina palabras que tienen diferente polaridad.

Si sabemos, de nuevo, que “inteligente” es positivo, algo coordinado con “inteligente” con la conjunción “y” es positivo, y, por el contrario, algo coordinado con la conjunción “pero” es negativo. Es un sistema bastante sencillo que se puede extender a palabras no solamente con conjunciones, sino palabras en el vecindario, en el contexto inmediato de palabras semilla que sabemos son positivas o negativas.

La otra manera de crear diccionarios es crear diccionarios manualmente. Es una labor muy parecida a otras tareas lexicográficas, basada de nuevo en diccionarios existentes, utilizando textos en dominios de interés y asegurando consistencia con diferentes anotadores. Se ha demostrado que crear diccionarios manualmente no da más error que diccionarios creados automáticamente, y que no es una labor que lleve demasiado tiempo. Los diccionarios manuales son perfectamente válidos para este tipo de tareas

Otra cuestión en el sistema basado en diccionarios es cómo elaborar sistemas que funcionen para más de un idioma, es decir, ¿es posible traducir un diccionario de un idioma a otro? Y la respuesta es sí. Generalmente es necesario corregirlo manualmente, pero utilizando diccionarios bilingües, utilizando incluso herramientas de traducción automática, como Google Translate, es posible crear diccionarios bastante fiables, que solamente necesitan de un poco de corrección manual.

Hay diccionarios existentes que se pueden utilizar, y la mayor parte de ellos están disponibles de manera gratuita.

El diccionario del General Enquirer es un diccionario muy antiguo que se utilizó en principio para hacer análisis del contenido, pero que últimamente se ha utilizado para hacer extracción de la opinión.

El siguiente es el Multi-Perspective Question Answering Dictionary. Es un diccionario creado en la Universidad de Pittsburgh para hacer extracción de la subjetividad. En inglés, pero que se puede traducir fácilmente al castellano.

SentiWordNet es un diccionario también en inglés, basado en minar la base de datos WordNet.

El Macquarie Semantic Orientation Lexicon es un diccionario extraído de un diccionario de sinónimos.

Y finalmente el diccionario del Semantic Orientation Calculator es un diccionario especialmente enfocado a las reseñas, y que está disponible en inglés y en castellano.

Bien, hemos terminado la parte que habla de diccionarios y de sistemas basados en diccionarios. Vamos a hablar un poco de corpus y para qué se utilizan

Si se emplea un sistema basado en diccionarios, el corpus es muy útil para elaborar el diccionario, y para comprobar su cobertura y eficacia.

Si se emplea un sistema de aprendizaje automático, el corpus es necesario para entrenar el sistema.

El aprendizaje automático es un aprendizaje automático supervisado que necesita de un set de ejemplos para el aprendizaje, equilibrado, que contenga textos u oraciones, y anotado, o etiquetado, con su polaridad, de ahí viene la parte supervisada, es decir, se incluyen textos positivos y negativos, generalmente de números similares. Entonces se puede construir un clasificador automático que aprende los rasgos característicos de cada una de las clases.

Los rasgos en el aprendizaje automático son varios, generalmente se emplean partes del discurso, unigramas y puntuación. Y los tipos de clasificadores son diferentes. Es un tema para otra lección, qué tipo de clasificadores se pueden utilizar.

Los recursos necesarios son: un corpus, un set de aprendizaje, y un set de test, generalmente se separan el uno del otro. Se reserva una parte para hacer pruebas y para hacer evaluación, pero también se puede hacer una validación cruzada, utilizando solamente el set de aprendizaje y dividiéndolo en partes.

El set de aprendizaje, por supuesto, puede ser textos enteros u oraciones. En cualquier caso, anotados con su polaridad, como mínimo, y con el grado, si eso es lo que se quiere aprender.

Hay varios corpus para el aprendizaje. En inglés el Movie Review Database contiene 2,000 reseñas de películas, divididas igualmente entre positivas y negativas.

El Simon Fraser University Review Corpus contiene reseñas de películas, libros, y algunos objetos de consumo, en inglés y en español.

Y el Corpus de cine es un corpus de reseñas de películas en castellano.

En resumen, los sistemas basados en diccionarios necesitan diccionarios y corpus para hacer tests, para hacer evaluación. Y los sistemas de aprendizaje automático necesitan corpus para el aprendizaje y un clasificador.

Muy bien, este es el final de la lección.

Referencias

Estas son las fuentes en las que me he basado para confeccionar esta lección:

- Opinion Mining and Sentiment Analysis, Bo Pang, Lillian Lee; Journal Foundations and Trends in Information Retrieval archive Volume 2 Issue 1-2, January 2008, Pages 1-135, Now Publishers Inc. Hanover, MA, USA doi>10.1561/1500000011
- Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Peter D. Turney; Proceeding ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Pages 417-424 Association for Computational Linguistics Stroudsburg, PA, USA ©2002 doi>10.3115/1073083.1073153
- SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining Andrea Esuli Stefano Baccianella, and Fabrizio Sebastiani. Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta, European Language Resources Association ELRA, (May 2010)