Un modelo para detectar la similitud semántica entre textos de diferentes longitudes

Darnes Vilariño, Mireya Tovar, Beatriz Beltrán, Saúl León

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Avenida San Claudio, 14 Sur, Ciudad Universitaria.
Puebla México.
\{darnes,mtovar,bbeltran,saul\}@cs.buap.mx
http://nlp.cs.buap.mx

Resumen En el presente trabajo se desarrolla un modelo para resolver el problema de similitud semántica entre textos de diferente longitud. Se propone extraer características léxicas, características basadas en conocimiento y características basadas en corpus, con el objetivo de desarrollar un modelo de aprendizaje supervisado. El modelo fue desarrollado utilizando regresión logística de la herramienta Weka. Los resultados obtenidos sobre los datos ofrecidos en el marco del Semeval 2014, han sido buenos para dos tipos de corpora.

Palabras clave: Similitud Semántica, Información mutua, Análisis semántico latente

1. Introducción

La similitud semántica textual tiene como objetivo capturar cuando el sentido de dos textos es similar. Este concepto difiere de encontrar el grado de similitud textual, que está solamente interesado en medir el número de componentes léxicas que comparten ambos textos, es decir, el valor de similitud textual puede variar cuando uno de los textos no está completo y no se logra con la poca información medir cuanto realmente comparten, pero no logra medir la similitud de ambos textos en cuanto al sentido que se desea expresar.

Encontrar la similitud semántica entre pares de textos se ha convertido en un gran reto para los especialistas en Procesamiento de Lenguaje Natural (PLN), ya que se puede aplicar en diferentes tareas de PLN, tales como máquinas de traducción, construcción automática de resúmenes, atribución de autoría, pruebas de lectura comprensivas, recuperación de información y muchas otras, que necesitan medir el grado de similitud entre dos textos dados.

Esta problemática se torna aún más compleja cuando se desea encontrar la similitud semántica textual entre textos de diferentes tamaños ya que ambos textos no ofrecen la misma cantidad de información y no se logra descubrir facilmente si el sentido del texto más pequeño es el mismo sentido del texto de mayor longitud. En particular esta problemática se ve reflejada cuando se desea

encontrar el grado de similitud entre un párrafo y una sentencia, una sentencia y una frase, una frase y una palabra y una palabra y un sentido. Esta tarea se presentó en el marco de la conferencia internacional Semeval 2014, como la tarea 3, que se denominó similitud semántica entre niveles [8]. El modelo que se desarrolla debe brindar un grado de similitud considerando los siguientes niveles:

- *4* Los dos textos tienen significados muy similares y las ideas más importantes, conceptos o acciones en el texto más grande están representadas en el texto más pequeño.
- *3* Ambos textos comparten muchas ideas importantes, conceptos o acciones, pero sin embargo lo que se expresa en el texto de menor longitud es similar, pero no idéntico a lo que se expresa en el texto de mayor longitud.
- *2* Ambos textos tienen significados diferentes, pero los conceptos, ideas o acciones en el texto de menor longitud tienen relación con el sentido que ofrece el texto de mayor longitud.
- *1* Los dos textos expresan ideas, conceptos y acciones completamente diferentes, pero es posible que se puedan encontrar juntos en un texto que hable del mismo tema.
- *0* Ambos textos no significan lo mismo, ni hablan del mismo tema.

El artículo está estructurado de la siguiente manera: en la sección 2 se discuten las diferentes metodologías desarrolladas para ofrecer el grado de similitud semántica entre textos, en la sección 3 se presentan las características seleccionadas, en la sección 4 se discute como se representa la información extraída de ambos textos y se presenta el modelo propuesto. La sección 5 muestra las características de las colecciones de datos, y el análisis de los resultados obtenidos. Por último en la sección 6 se presentan las conclusiones de la investigación y el trabajo a futuro.

2. Trabajo Relacionado

La búsqueda del grado de similitud semántica entre textos se ha propuesto como tarea dentro del marco de la conferencia internacional Ejercicios de Evaluación Semántica [1], [2] por lo que se le está dedicando bastante atención en los últimos años. Muchos de los modelos desarrollados han hecho énfasis fundamentalmente en la búsqueda de las características que comparten ambos textos, para lograr con ello detectar si ambos textos poseen un sentido similar. Se han propuesto modelos markovianos, se han propuesto modelos de traducción automática, buscando la alineación de ambos textos y se ha introducido en investigaciones la búsqueda de ciertos patrones sintácticos que comparten. A continuación se presentan alguno de los trabajos desarrollados.

En la investigación reportada en [15], se busca establecer un modelo semántico que permita identificar los argumentos implícitos en los textos. Se plantea que a pesar de ser una tarea fácil para un lector humano, resulta complicado para las computadoras, debido a que no hay una manera de indicarle a éstas, que un argumento puede ser inferido varias veces en el texto. En este trabajo se propone

una aproximación por inducción que explota la información complementaria obtenida de un par de textos comparables. Esto significa que se desarrollan modelos que permiten alinear ambos textos, esta es una técnica que se ha utilizado para el desarrollo de diccionarios estadísticos en las tareas de traducción automática. La metodología propuesta en este trabajo ha permitido encontrar la similitud entre textos de longitudes similares con un grado de precisión de un 83 %.

En el trabajo desarrollado por [4], se propone un modelo Markoviano para determinar la cantidad de información que comparten dos sentencias de una longitud aproximada. En el marco de esta investigación se desarrollan un conjunto de reglas que permiten inferir cuando dos textos poseen el mismo significado. La precisión reportada para los datos del test a partir de los modelos construidos fue de 73 %, con los datos de prueba que se ofrecieron en la tarea 6 de la conferencia [1].

De igual modo, se puede hacer mención al trabajo [18], en el cual se muestra un método computacional que identifica las metáforas en textos sin restricción respecto a su interpretación, es decir se convierte al problema de encontrar el parecido entre textos, como encontrar si un texto es el parafraseo del otro. Se define a la interpretación en metáforas como una tarea de encontrar una parárasis literal para una palabra usada metafóricamente, además de introducir el concepto de parafraseado inverso simétrico como un criterio para la identificación de metáforas. Esto lo logra haciendo experimentos en los que se manejan relaciones de verbo-sujeto y sujeto-objeto indirecto, sin embargo la precisión que logran usando los datos del test y training en la tarea 6 del SemEval 2013 es solamente del 66 %.

La investigación presentada por [16], plantea que para lograr medir el grado de similitud semántica textual entre textos estos deben ser representados no con términos unipalabras, sino con términos multipalabras. Para encontrar entonces la similitud en este tipo de representación acuden a traducciones automáticas, para lo que usan a la herramienta PanLex, la cual le permite la creación de un diccionario estadístico. Si la traducción es posible, esto quiere decir que es equivalente un término en un texto, con una expresión multipalabra en el otro. El sistema presentado ofreció una precisión del 66 % en el marco del Semeval 2013.

Otro trabajo al que se puede hacer mención es el desarrollado por [17], cuyo objetivo es caracterizar los errores que se producen al realizar conteo de palabras. Muestra que los conteos léxicos pueden ser significativamente mejorados mediante el empleo de medidas de ambigüedad. Es decir encontrar sobre el texto aquellas palabras ambigüas, eliminar aquellas que no están totalmente relacionadas con el contexto. Esta propuesta no brindó resultados significativos, ya que solamente alcanzó el 50 % de precisión.

Otra manera de abordar esta tarea es considerandola como un problema de Question Answering, donde uno de los textos es la pregunta y el otro es la respuesta dada, es este el sentido del trabajo desarrollado por [3] donde se propone un modelo que mide el grado de similitud en funcin de que si la

respuesta logra responder a la pregunta. El modelo propuesto con los datos de entrenamiento y prueba dados en el Semeval 2013, ofreció una precisión del 73 % .

La propuesta desarrollada en la presente investigación tiene como objetivo fundamental medir el comportamiento de diferentes características y un modelo para expandir los textos de menor longitud. A continuación se discuten las características empleadas para representar a ambos textos.

3. Extracción de Características

Básicamente se han usado tres tipos diferentes de características: léxicas, basadas en conocimiento y basadas en Corpus. La primera de ellas cuenta la frecuencia de ocurrencia de los n-gramas de caracteres, *skip*-gramas[1], palabras y algunas relaciones léxicas como sinónimos e hiperónimos. Adicionalmente se han incluído otras dos características: El coeficiente de similitud de Jaccard entre dos textos, expandiendo cada término con el conjunto de sinónimos tomados de WordReference [5], y la similitud coseno entre los dos textos representados cada uno por la bolsa de n-gramas y *skip*-gramas de caracteres. En esta propuesta no se incluye ningún proceso de desambiguación después de expandir con los sinónimos de cada una de las palabras que conforman los textos.

Tabla 1. Características usadas para encontrar la similitud semántica textual

Característica	**Tipo**
n-gramas de caracteres ($n = 2, \cdots, 5$)	Léxica
skip-gramas de caracteres($skip = 2, \cdots, 5$)	Léxica
Número de palabras que comparten	Léxica
Número de sinónimos que comparten	Léxica
Número de hiperónimos que comparten	Léxica
Coeficiente de Jaccard con expansión de sinónimos	Léxica
Similitud Coseno con n-gramas y *skip*-gramas de caracteres	Léxica
Similitud de palabra de Leacock & Chodorow	Basada en conocimiento
Similitud de palabra de Lesk	Basada en conocimiento
Similitud de palabra de Wu & Palmer	Basada en Conocimiento
Similitud de palabra de Resnik's	Basada en conocimiento
Similitud de palabra de Lin	Basada en conocimiento
Similitud de palabra de Jiang & Conrath	Basada en conocimiento
Métrica de Rada Mihalcea usando Información mutua	Basada en Corpus
Métrica de Rada Mihalcea usando LSA	Basada en Corpus

El segundo conjunto de características considera las 6 medidas de similitud de palabras ofrecidas por la herramienta NLTK: Leacock & Chodorow [10],

[1] Son conocidos como n-gramas dispersos porque consideran un salto en cierto número de caracteres.

Lesk [11], Wu & Palmer [20], Resnik [14], Lin [12], y Jiang & Conrath[2] [7]. En este caso, se determina la similitud semántica entre dos textos como el máximo valor de similaridad obtenido entre los pares de palabras. El tercer conjunto de características considera dos medidas basadas en corpus, ambas medidas utilizan la métrica de similitud semántica textual ofrecida por Rada Mihalcea [13]. La primera usa Información mutua (PMI) [19] para el cálculo de la similitud entre pares de palabras, mientras que la segunda utiliza análisis semántico latente(LSA) [9] (implementado en el entorno estadístico R). Para esta investigación los valores de PMI y LSA fueron obtenidos en base a un corpus construído,con Europarl, el proyecto Gutenberg y el thesauro de OpenOffice. Todas las características utilizadas se pueden ver en la Tabla 1.

4. Metodología propuesta

Se extraen las características descritas anteriormente de los datos de entrenamiento, con el objetivo de desarrollar un modelo de clasificación. Se construye un vector para cada par de textos. Este vector es introducido en Weka para construir un modelo de clasificación basado en regresión logística.[6]

Dos de los corpus ofrecidos poseen textos de diferentes longitudes. Encontrar la similitud semántica en este caso se vuelve una tarea complicada, es por ello que se decide utilizar algún mecanismo de expansión de los textos de menor longitud. Para encontrar la similitud semántica entre frases y palabras, se expanden las palabras con los términos relacionados obtenidos de Flickr. Cuando se desea encontrar la similitud entre palabra y sentido, se expande el sentido utilizando la taxonomía de Wordnet, en ninguno de los casos se introduce un proceso de desambiguación, para eliminar aquellos términos no relativos al contexto con el que se está trabajando.

5. Resultados experimentales

A continuación se describen los datos que permitieron validar el modelo desarrollado.

5.1. Conjunto de Datos

Se dispone de un corpus conformado por 2,000 pares de textos para entrenamiento y 2,000 pares de texto para prueba. El conjunto de datos considera 500 pares para cada nivel, es decir 500 pares para párrafo-sentencia, 500 pares para frase-sentencia y así sucesivamente. Una descripción completa del conjunto de datos empleado se encuentra en el artículo que describe esta tarea en el marco del Semeval 2014. [8]

[2] Herramienta de Python para el procesamiento de Lenguaje Natural; http://www.nltk.org/

5.2. Resultados Obtenidos

Los resultados obtenidos se muestran en la Tabla 2. Nuestra aproximación, denominada *BUAP*, obtuvo un rendimiento por arriba del promedio reportado en la conferencia (*Promedio*). Cabe aclarar que dicho promedio se obtiene sobre el conjunto completo de ejecuciones reportadas en SemEval 2014, a pesar de que en esta tabla solamente se muestran los tres primeros resultados. El porcentaje de mejora de nuestro sistema con respecto al promedio se muestra en el último renglón de la tabla. Como puede apreciarse se ha obtenido un buen comportamiento del modelo desarrollado, cuando se calcula la similitud semántica entre párrafos y sentencias y entre sentencias y frases. Sin embargo la expansión de las palabras utilizando Flickr sobre el corpus de frase-palabra no ofreció resultados buenos, se considera que se debe a que la expansión realizada no está relacionada con el dominio del corpus, se introducen términos de caracter muy general. Cuando se expanden los sentidos al trabajar con los pares palabra- sentido, nuevamente se considera que se introduce un grado fuerte de ambieguedad, lo que provoca que el comportamiento también sea inapropiado.

Tabla 2. Resultados obtenidos en la Tarea 3 del Semeval 2014

Equipo	**Sistema**	**Párr-Sentencia**	**Sentencia-frase**	**Frase-Palabra**	**Palabra-Sentido**	**Rango**
SimCompass	run1	0.811	0.742	0.415	0.356	1
ECNU	run1	0.834	0.771	0.315	0.269	2
UNAL-NLP	run2	0.837	0.738	0.274	0.256	3
BUAP	**run**	**0.805**	**0.714**	**0.142**	**0.194**	**10**
Promedio	-	0.728	0.651	0.198	0.192	11-12
run - Promedio		8 %	6 %	-6 %	0 %	-

6. Conclusiones y Recomendaciones

En este artículo se presenta el modelo para resolver el problema planteado en la tarea 3 del Semeval 2014. Las características utilizadas ofrecen un buen comportamiento para detectar el grado de similitud semántica entre párrafo- sentencia y sentencia-frase, sin embargo la metodología de expansión propuesta para detectar el grado de similitud semántica entre los pares frase a palabra y palabra a sentido no fue correcta. Los resultados obtenidos fueron extremadamente bajos, lo que nos indica que deben utilizarse mecanismos de expansión diferentes. Se está trabajando en extraer de la web documentos donde aparezcan las palabras, para construir vectores representativos de cada una de ellas considerando 5 términos a la derecha y 5 términos a la izquierda, con esto se pretende detectar aquellos términos relacionados con la palabra que se desea expandir.

Referencias

1. Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *sem 2013 shared task: Semantic textual similarity. In *2nd Joint Conference on*

*Lexical and Computational Semantics (*SEM)*, pages 32–43, Atlanta, Georgia, USA, 2013.

2. Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012)*, pages 385–393, 2012.
3. Alexis Palmer Andrea Horbach and Manfred Pinkal. Using the text to evaluate short answers for reading comprehension exercises. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 520–524, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.
4. Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. Montague meets markov: Deep semantics with probabilistic logical form. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*Sem-2013)*, Atlanta, GA, June 2013.
5. Maya Carrillo, Darnes Vilariño, David Pinto, Mireya Tovar, Saul León, and Esteban Castillo. Fcc: Three approaches for semantic textual similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SemEval 2012)*, pages 631–634, Montréal, Canada, 2012.
6. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.
7. Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING'97*, pages 19–33, 1997.
8. David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland, 2014.
9. Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284, 1998.
10. C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *MIT Press*, pages 265–283, 1998.
11. Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26. ACM, 1986.
12. Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
13. Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 775–780, 2006.
14. Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995.
15. Michael Roth and Anette Frank. Automatically identifying implicit arguments to improve argument linking and coherence modeling. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*Sem-2013)*, Atlanta, GA, June 2013.

16. Bahar Salehi and Paul Cook. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*Sem-2013)*, Atlanta, GA, June 2013.
17. H. Andrew Schwartz, Johannes Eichstaedt, Lukasz Dziurzynski, Eduardo Blanco, Margaret L. Kern, Stephanie Ramones, Martin Seligman, and Lyle Ungar. Choosing the right words: Characterizing and reducing error of the word count approach. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*Sem-2013)*, Atlanta, GA, June 2013.
18. Ekaterina Shutova. Metaphor identication as interpretation. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*Sem-2013)*, Atlanta, GA, June 2013.
19. Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, pages 491–502. Springer-Verlag, 2001.
20. Zhibiao Wu and Martha Stone Palmer. Verb semantics and lexical selection. In James Pustejovsky, editor, *ACL*, pages 133–138. Morgan Kaufmann Publishers / ACL, 1994.