Análisis de la dinámica del contenido semántico de textos

Edgar Altszyler y Pablo Brusco

Laboratorio de Inteligencia Artificial Aplicada, Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, C1428EGA Buenos Aires, Argentina
edgaralts@gmail.com, pbrusco@dc.uba.ar

Resumen El presente trabajo es el primer eslabón de un proyecto en proceso que apunta a analizar la dinámica de distintos conceptos, desde un enfoque semántico, buscando identificar patrones temporales comunes en varios corpus de texto. Como primer paso utilizaremos textos provenientes de libros o subtítulos de películas, sin embargo este análisis puede extenderse a cualquier corpus de texto. En este trabajo en particular estudiaremos la evolución semántica de conceptos a lo largo de las distintas novelas de la saga *Harry Potter* con el objetivo último de analizar variaciones del contenido semántico en textos utilizando herramientas de *Procesamiento de Lenguaje Natural* (PLN). En este contexto, mostraremos una primera aproximación a la comprensión del alcance y de las limitaciones de las herramientas clásicas de PLN para cuantificar la evolución del concepto "oscuridad" con el avance temporal de los libros. La saga de novelas de *Harry Potter* resulta ser un corpus de texto ideal para testear estas herramientas debido a que es de público conocimiento que dicha saga presenta un aumento gradual de la "oscuridad" a medida que se suceden los libros.

Palabras Clave: Latent Semantic Analysis, Text Data Mining, Natural Language Processing.

1. Introducción

En los últimos años se ha incrementado intensamente el volumen de texto escrito digitalizado, especialmente debido al contenido masivo generado en las redes sociales. Tanto en la comunidad científica como en sector privado existe un fuerte interés en la identificación de la información valiosa que se esconde detrás de la gran masa de datos disponible. A raíz de esto, el minado masivo de textos y el Procesamiento de Lenguaje Natural (PLN) están recibiendo una creciente atención [Chen et al., 2014].

Dentro del campo del *Procesamiento del Lenguaje Natural*, la extracción de significados de textos escritos es, al día de hoy, un tema de gran interés. Dentro de esta sub-área, los métodos de cuantificación de distancia semántica entre palabras y/o documentos cumplen un rol fundamental. Existe una gran diversidad de algoritmos capaces de cuantificar la similitud semántica entre palabras,

varios de los cuales parten del supuesto de que las palabras semánticamente relacionadas tenderán a utilizarse en contextos similares, y consecuentemente, a coexistir en documentos. Un algoritmo muy utilizado que parte de estas mismas hipótesis es el *Análisis Semántico Latente* (LSA por sus siglas en inglés) [Deerwester et al., 1990], que describe a cada palabra o conjunto de palabras en un espacio vectorial de dimensión reducida. En el mismo, palabras semánticamente similares se encontrarán espacialmente cerca.

La técnica de LSA toma como entrada la matriz de frecuencias de palabras en los distintos documentos, sobre la cual, suele aplicarse la transformación tf-idf para reducir el peso de palabras muy frecuentes y poco significativas. Luego, a esta matriz se le aplica una descomposición en valores singulares y una reducción de la dimensionalidad, obteniendo así una representación vectorial tanto de las palabras como de los documentos. De esta manera, se podrá cuantificar la similaridad semántica de dos palabras tomando el coseno del ángulo que sustentan las representaciones vectoriales de cada una de ellas, produciendo valores contenidos en el rango de -1 (palabras distantes en su semántica) y 1 (palabras semánticamente similares). Resulta relevante remarcar que debido al proceso de reducción de dimensionalidad, el LSA es capaz de identificar la cercanía de una palabra a un texto incluso en el caso en que el documento no contenga esa palabra.

Si bien el LSA fue desarrollado en un contexto de búsqueda y recuperación de información, también resulta útil para la cuantificación de cercanía entre conceptos y documentos. En el trabajo de Diuk *et al.* [Diuk et al., 2012] se mostró que esta técnica resulta exitosa en la cuantificación de la evolución temporal de conceptos de alto nivel, como lo es el caso de la "introspección". Resulta relevante destacar que la cuantificación de la dinámica temporal de un concepto con LSA resulta un avance significativo respecto a la cuantificación según la frecuencia de aparición de una palabra a lo largo del tiempo [Wolff et al., 1999, Greenfield, 2013]. Esto se debe a que el LSA es capaz de identificar el contenido latente de un concepto en un texto a pesar de que este no contenga la palabra en cuestión.

Siguiendo el mismo lineamiento de trabajo, nos proponemos explorar las capacidades y limitaciones del LSA para estudiar la evolución temporal de conceptos abstractos en corpus de texto. En particular analizaremos la similitud semántica entre el concepto de oscuridad y los conceptos que aparecen en los diferentes libros de la saga de *Harry Potter* en idioma inglés.

2. Metodología

Para la generación de la métrica de distancias del LSA, es de gran relevancia la elección tanto del corpus de entrenamiento como de la dimensión del espacio vectorial. Como es usual en la aplicaciones del LSA, utilizamos el corpus de texto de TASA (Touchstone Applied Science Associates, Inc), que contiene 37651 documentos formados por una colección de material educativo de EEUU y otros textos tales como novelas y noticias. Con respecto a las dimensiones del LSA, Landaur y Dumas [Landauer and Dumais, 1997] estudiaron el efecto de este parámetro sobre el desempeño del mismo en exámenes de idioma, encontrando un máximo alrededor de las 300 dimensiones.

Para el entrenamiento del LSA al corpus de texto crudo se le aplicó un filtrado de palabras frecuentes (se filtraron los "stopwords" del paquete NLTK [Bird et al., 2009] para el idioma inglés) y se llevaron las palabras a su raíz con el algoritmo de Porter [Porter, 1980]. Luego se entrenó un LSA de 300 dimensiones sobre la matriz de frecuencias términos-documentos normalizada según tf-idf. Finalmente, utilizamos la *similitud coseno* como métrica para la medición de distancias entre los distintos conceptos y los libros.

Al igual que al corpus de entrenamiento, a los textos crudos de *Harry Potter* se les aplicó el filtro de palabras frecuentes, se llevaron las palabras a su raíz y se eliminaron las palabras "Potter", "Ron" y "Dark Lord" para evitar aportes espurios en los cálculos de distancia debido a la incapacidad de detectar polisemia del algoritmo.

Luego, con el fin de cuantificar la "oscuridad" presente en los libros de la saga de *Harry Potter* en idioma inglés, seguimos el mismo lineamiento que Diuk et al. [Diuk et al., 2012]. Un punto importante a notar es que la palabra "oscuridad" es polisémica, y hace referencia tanto a "maldad", como a "ignorancia" o a la "falta de luz". En este trabajo nos interesa la primera de las tres acepciones presentadas. Debido a que el LSA describe cada concepto como un punto en el espacio vectorial, este método no es capaz de lidiar con la polisemia, por lo que desglosamos al concepto "oscuridad" en una lista de 12 palabras representativas (*darkness, gloom, desolation, sinful, misery, evil, wickedness, vile, cruelty, harm, demon, malicious*). Las palabras de esta lista las seleccionamos con ayuda del programa *wortsurfer*, que dada una palabra, brinda una red con las palabras semánticamente más cercanas. A partir de las redes semánticas centradas en "darkness" y "evil" seleccionamos las que consideramos que mejor representaban al concepto de "oscuridad".

3. Resultados

Como primer paso, a cada palabra dentro la lista de palabras "oscuras" le calculamos su *similitud coseno* con cada libro. En la Figura 1 mostramos a modo

de ejemplo la "similitud coseno" de cada libro respecto a la palabra "evil" ("mal" en idioma inglés), en particular podemos observar un incremento de la cercanía semántica de los libros sucesivos a esta palabra. Con el fin de cuantificar esta variación, realizamos una regresión lineal de los puntos y tomamos la pendiente como medida de este efecto. De esta manera, pendientes positivas evidencian incrementos de la similitud semántica entre los libros sucesivos y la palabra seleccionada, mientras que pendientes negativas indican una disminución en dicha relación. Cabe remarcar que sólo estamos interesados en capturar el comportamiento a primer orden de las variaciones semánticas, es decir, si globalmente la similitud semántica crece o no. Es por ello que no nos focalizaremos en estudiar la forma funcional especifica que tiene la similitud semántica a lo largo de los libros.

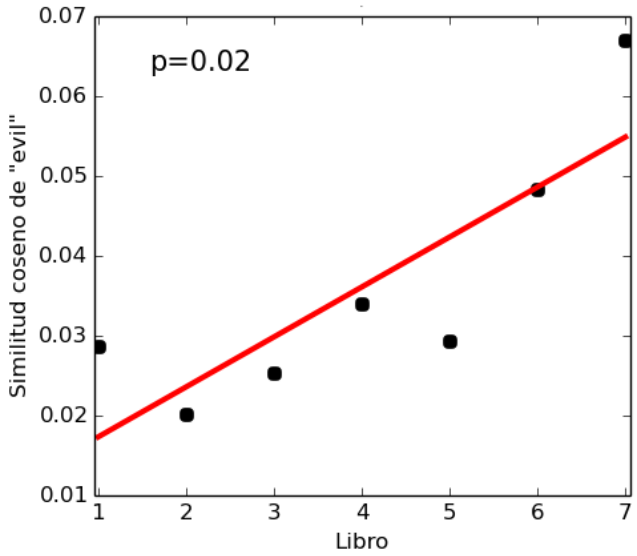

Figura 1. Gráfico de *similitud coseno* entre la palabra "evil" ("mal" en idioma inglés) y cada libro (en puntos negros). Se puede observar un incremento de la cercanía semántica de los libros sucesivos a la palabra "evil". Se realizó una regresión lineal (línea roja), donde la pendiente (m=0.0063) será utilizada para reportar el comportamiento global de la palabra "evil" a lo largo de los libros. En la región superior izquierda se reporta el p-valor (P=0.02) para la hipótesis nula de que la pendiente de la regresión lineal es cero.

Luego, con el fin de estudiar la dinámica del concepto de "oscuridad", calculamos la pendiente de los gráficos de similitud coseno para todas las palabras "oscuras". Representamos en un diagrama de cajas (Figura 2) este conjunto de pendientes, comparadas con las pendientes obtenidas para una lista de pala-

bras de neutras, es decir independientes del concepto de "oscuridad", en este caso utilizamos frutas. De esta manera, podemos cuantificar el incremento de la oscuridad a lo largo de los libros con el valor promedio de las pendientes, obteniéndose así una pendiente promedio de 0.0038. Para mostrar que la distribución de las pendientes de las palabras oscuras tiene una media distinta de cero se utilizó un t-test de muestra única, obteniéndose un p-valor=0.00036. En contraposición, para la distribución de pendientes de las palabras neutras se obtuvo un valor medio de -0.00039 y p-valor=0.48 para el t-test de muestra única, por lo que no resulta significativamente distinto de cero.

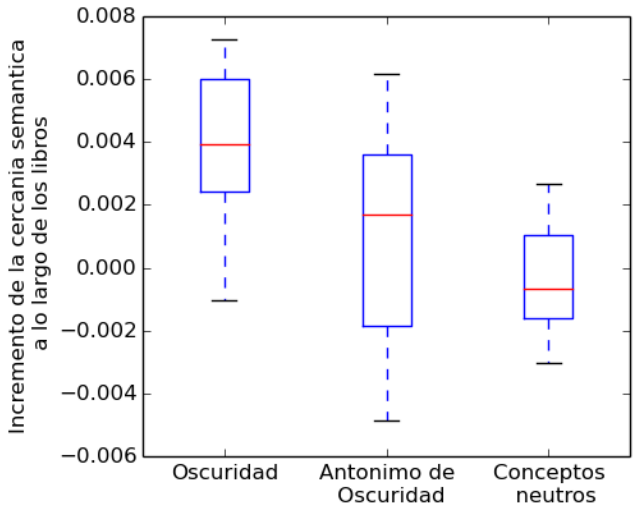

Figura 2. Diagrama de cajas del Incremento de la cercanía semántica a lo largo los libros de las palabras que integran el concepto de "oscuridad", su antónimo y las palabras consideradas neutras. El incremento de la cercanía semántica de una palabra a lo largo de los libros está medido como la pendiente de la regresión lineal del gráfico de *similitud coseno* entre la palabra y cada libro. Los valores medios de las distribuciones de pendientes de las palabras "oscuras", sus antónimos y palabras "neutras"son respectivamente 0.0038, 0.001 y -0.00039, con un p-valor de 0.00036, 0.3 y 0.48 para un t-test de muestra única con hipótesis nula de que la media es cero. Las palabras representativas de los tres conceptos son: Palabras Oscuras: *darkness, gloom, desolation, sinful, misery, evil, wickedness, vile, cruelty, harm, demon, malicious.* Antónimos de Palabras oscuras: *lightness, happiness, joy, admirable, pleasure, virtuous, kindness, friendly, compassion, niceness, benefit, saint, benevolent.* Palabras neutras (frutas): *grape, banana, strawberry, peach, plum, melon, tangerine, mango, pear, watermelon, apple, orange.* En el apéndice se muestran los gráficos de *similitud coseno* entre los libros y las distintas palabras que integran las tres categorías de conceptos.

VI

Finalmente, incluimos en el diagrama de cajas de la Figura 2 las pendientes de los gráficos de similitud coseno para el conjunto de antónimos de las palabras oscuras. En esta figura se puede observar un ligero corrimiento de la distribución de pendientes hacia los valores positivos, con un valor medio de 0.001 y un p-valor=0.3. A pesar de que este corrimiento no resultó significativo cabe remarcar que se ha demostrado que el LSA asigna alto grado de similitud semántica tanto a los conceptos similares como a los opuestos [Landauer, 2002], por lo que podría esperarse un cierto corrimiento de la distribución de pendientes hacia los valores positivos.

4. Conclusiones

En este trabajo nos hemos focalizamos en cuantificar la evolución semántica del concepto de "oscuridad" a lo largo de las distintas novelas de la saga *Harry Potter* (en idioma inglés) utilizando herramientas de Procesamiento de Lenguaje Natural.

La saga de novelas de *Harry Potter* resulta ser un corpus de texto ideal para testear las capacidades y limitaciones de las herramientas clásicas de PLN para cuantificar la presencia y evolución de conceptos abstractos en textos escritos. Esto se debe a que es de público conocimiento que dicha saga presenta un aumento gradual de la "oscuridad" a medida que se suceden los libros.

En este contexto, buscamos cuantificar mediante un algoritmo de LSA la evolución del concepto de "oscuridad" a través de los 7 libros. Para lidiar con el carácter polisémico de la palabra "oscuridad", desglosamos este concepto en una lista de 12 palabras representativas y a cada una de ellas le calculamos su cercanía semántica a cada uno de los 7 libros. Luego, calculamos las pendientes de los gráficos de similitud semántica en función del número de libro y tomamos el promedio de las pendientes como indicador global de la evolución de la cercanía semántica del concepto de "oscuridad". De esta manera, se obtuvo una pendiente promedio de 0.0038, la cual resulta significativamente distinta de cero con p-valor=0.00036 para un t-test de muestra única. Finalmente, la obtención de un incremento significativo en la cercanía semántica entre el concepto de "oscuridad" y los libros sucesivos, resulta acorde al incremento de "oscuridad" esperado, evidenciando la capacidad de este método de cuantificar la evolución semánticas de conceptos abstractos.

Como continuación de la presente línea de investigación nos proponemos establecer un método automático para la selección del conjunto de palabras que representa cada concepto. A su vez, pretendemos extender este enfoque al análisis de distintos conceptos en diversos corpus de texto, pudiendo así identificar patrones semánticos comunes entre distintos textos pertenecientes al mismo género.

Referencias

Bird et al., 2009. Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. Reilly Media, Inc.

Chen et al., 2014. Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile Netw Appl*, 19(171–209).

Deerwester et al., 1990. Deerwester, S. ., Dumais, S. T., Landauer, T., Furnas, G., and Harshman, R. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6).

Diuk et al., 2012. Diuk, C. G., Slezak, D. F., Raskovsky, I., Sigman, M., and Cecchi, G. A. (2012). A quantitative philology of introspection. *Frontiers in integrative neuroscience*, 6.

Greenfield, 2013. Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science*, 24(1722–1731).

Landauer, 2002. Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from lsa. *Psychology of learning and motivation*, 41.

Landauer and Dumais, 1997. Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2).

Porter, 1980. Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3).

Wolff et al., 1999. Wolff, P., Medin, D. L., and Pankratz, C. (1999). Evolution and devolution of folkbiological knowledge. *Cognition*, 73(177 - 204).

Apéndice

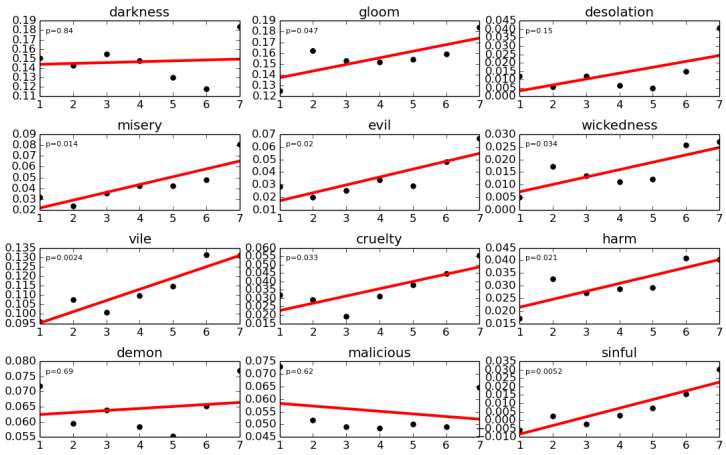

Figura A1. Gráficos de "similitud coseno" entre las palabras que integran el concepto de "oscuridad" y cada libro (en puntos negros). Para cada palabra se realizó una regresión lineal (línea roja) donde la pendiente reporta el incremento de la cercanía semántica a lo largo los libros. En cada caso se muestra el p-valor para la hipótesis nula de que la pendiente de la regresión lineal es cero.

VIII

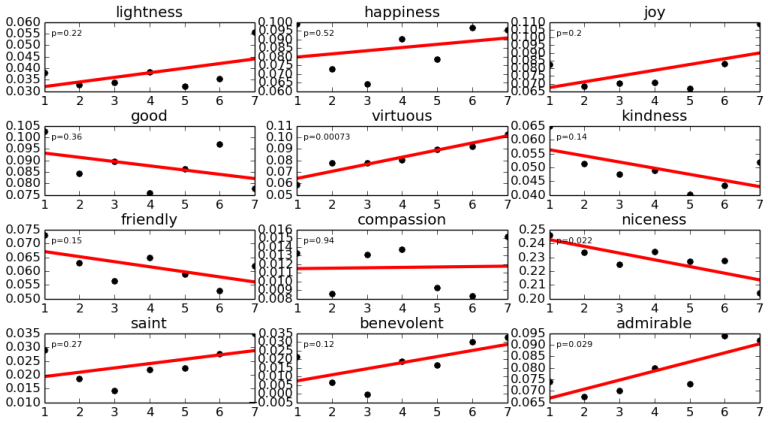

Figura A2. Gráficos de "similitud coseno" entre cada libro y las palabras opuestas a las que integran el concepto de "oscuridad" (en puntos negros). Para cada palabra se realizó una regresión lineal (línea roja) donde la pendiente reporta el incremento de la cercanía semántica a lo largo los libros. En cada caso se muestra el p-valor para la hipótesis nula de que la pendiente de la regresión lineal es cero. La distribución de estas pendientes se encuentran graficadas en el diagrama de cajas de la figura 2.

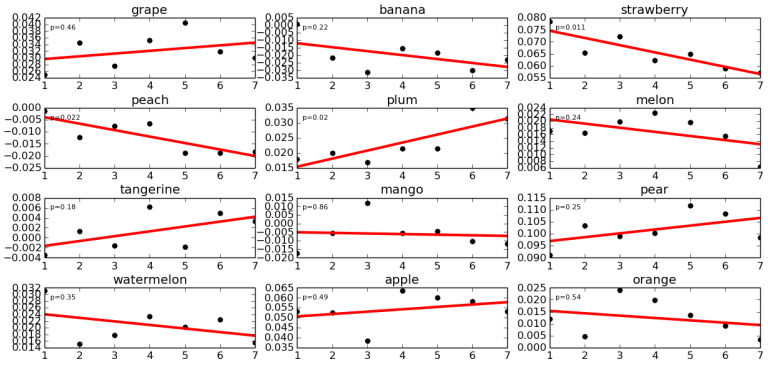

Figura A3. Gráficos de "similitud coseno" entre cada libro y distintas frutas (en puntos negros). Estas palabras fueron seleccionadas como un conjunto neutro, bajo la hipótesis de que no tendrían ninguna dinámica relevante. Para cada palabra se realizó una regresión lineal (línea roja) donde la pendiente reporta el incremento de la cercanía semántica a lo largo los libros. En cada caso se muestra el p-valor para la hipótesis nula de que la pendiente de la regresión lineal es cero. La distribución de estas pendientes se encuentran graficadas en el diagrama de cajas de la figura 2.