ISSN 1870-4069

Representaciones vectoriales de palabras de un corpus de normas de asociación

Jorge Reyes Magaña[1], Helena Gómez Adorno[2], Gemma Bel Enguix[2], Gerardo Sierra[2]

[1] Universidad Autónoma de Yucatán, Facultad de Matemáticas, Mérida, Yucatán, México

[2] Universidad Nacional Autónoma de México, Instituto de Ingeniería, Ciudad de México, México

jorge.reyes@correo.uady.mx, {hgomeza,gbele,gsierram}@iingen.unam.mx

Resumen. Las representaciones vectoriales de palabras son muy eficientes para muchas tareas de procesamiento del lenguaje natural y para su construcción es necesario entrenarlos con una gran cantidad de datos para obtener vectores de buena calidad que permitan realizar las tareas con éxito. En este trabajo, presentamos un método basado en el algoritmo *node2vec* para aprender vectores de palabras usando un grafo que ha sido construido tomando como base un corpus de normas de asociación de palabras en español; el grafo construido contiene en los nodos las palabras y en las aristas diferentes pesos como son frencuencia, tiempo y fuerza de asociación. Los recursos computacionales utilizados por este método son razonables y asequibles. Esto nos permite obtener vectores de palabras de buena calidad incluso desde un corpus pequeño. Evaluamos nuestro método en un corpus de similitud y relacionalidad de palabras, obteniendo resultados comparables a los obtenidos con *word2vec* entrenados en un corpus de mil millones de palabras.

Palabras clave: vectores de palabras, normas de asociación de palabras.

Word Embeddings Learned on Word Association Norms

Abstract. Word embeddings are powerful tools for many natural language processing tasks. In order to obtain embeddings useful for different tasks, it is necessary a large training corpus. In this work, we present a method based on the *node2vec* algorithm to learn word embeddings from a graph built using a corpus of word association norm in Spanish. The nodes of the graph correspond to the words in the corpus, whereas the

edges are weighted with the frequency, time and association strength of a pair of words. The computational resources used by this technique is reasonable and affordable. This allows us to obtain good quality word embeddings even from a small corpus. We evaluated our word vectors in a word similarity and relatedness benchmark, achieving comparable results to those obtained with *word2vec* trained on a billion word corpus.

Keywords: word vectors, word association norms.

1. Introducción

La representación semántica de palabras en un espacio vectorial es un área de investigación muy activa en las últimas décadas. Modelos computacionales como la descomposición de valores singulares (SVD) y el análisis semántico latente (LSA) son capaces de modelar representaciones continuas de palabras (*word embeddings*) a partir de matrices término-documento. Ambos métodos pueden reducir un conjunto de datos de N dimensiones utilizando solo las dimensiones más importantes. Recientemente, *Mikolov et al.* [15] introdujeron *word2vec*, inspirado en la hipótesis distribucional que establece que las palabras en contextos similares tienden a tener significados similares [17]. Dicho método utiliza una red neuronal para aprender representaciones vectoriales de palabras al predecir otras palabras en su contexto. La representación vectorial de la palabra obtenida mediante *word2vec* tiene la asombrosa capacidad de preservar las regularidades lineales entre palabras.

Para construir un modelo de espacio vectorial adecuado y confiable, capaz de capturar la similitud semántica y las regularidades lineales de las palabras, se necesitan grandes volúmenes de texto. Aunque *word2vec* es rápido y eficiente de entrenar, y los vectores pre-entrenados generalmente están disponibles en línea, todavía es computacionalmente costoso procesar mediante este método grandes volúmenes de datos en entornos no comerciales, es decir, en computadoras personales.

La asociación libre es una técnica experimental comúnmente utilizada para descubrir la forma en que la mente humana estructura el conocimiento [6]. En las pruebas de asociación libre, se le pide a una persona que diga la primera palabra que se le viene a la mente en respuesta a una palabra *estímulo* dada. Mediante estos experimentos se obtienen unas compilaciones de relaciones léxicas, llamados Normas de Asociación de Palabras (NAP), que pueden reflejar tanto contenidos semánticos como episódicos [4].

El objetivo de este trabajo es presentar un método para aprender representaciones vectoriales de palabras a partir de nodos de un grafo obtenido a partir de un corpus NAP. Nuestra hipótesis es que los vectores aprendidos de este grafo mapean los contenidos de memoria semántica y episódica en el espacio vectorial, y así aprenden mejores representaciones. *Grover y Leskovec* [9] introdujeron un algoritmo llamado *node2vec* que es capaz de aprender mapeos de nodos a un espacio vectorial continuo teniendo en cuenta las vecindades de la red de los nodos. El algoritmo realiza caminos aleatorios sesgados para explorar diferentes

vecindarios con el fin de capturar no solo los roles estructurales de los nodos en la red, sino también las comunidades a las que pertenecen.

El presente trabajo está organizado de la siguiente manera. En la sección 2, discutimos el trabajo relacionado. En la Sección 3, presentamos el Corpus de Normas de la Asociación de Palabras para el Español Mexicano (NAP). En la sección 4, describimos el marco metodológico para aprender vectores de palabras a partir del NAP. La sección 5, muestra la evaluación de los vectores generados, utilizando como corpus de evaluación conjuntos de datos que contienen similitud de palabras en español. Finalmente, en la sección 6 sacamos algunas conclusiones y señalamos las posibles direcciones del trabajo futuro.

2. Trabajo relacionado

Sinopalnikova y Smrz [19] presentaron un marco metodológico para construir y extender redes semánticas con tesauros de asociaciones de palabras (WAT, por sus siglas en inglés). Además, hacen una comparación de la calidad y la información que ofrece WAT vs. otros recursos lingüísticos. Finalmente, los autores muestran que el WAT es comparable y se puede usar como un corpus balanceado de texto en ausencia de este.

Borge-Holthoefer y Arenas [4] describen un modelo para extraer relaciones de similitud semántica desde información de asociaciones libres (denominado RIM por sus sigla en inglés). Los autores aplican un método basado en redes para descubrir vectores de características en una red de asociaciones libres. Los vectores obtenidos fueron comparados con representaciones vectoriales basadas en LSA y el modelo WAS (Word Association Space). Los resultados de este trabajo indican que RIM puede extraer con éxito vectores de características de palabras desde una red de asociaciones libres.

En los últimos años, *Bel-Enguix et al.* [3] usaron técnicas de análisis de grafos para calcular asociaciones desde grandes colecciones de textos. Por otra parte, *Garimella et al.* [8] presentaron un modelo de asociaciones de palabras sensible al contexto demográfico basado en una arquitectura de redes neuronales con n-gramas no consecutivos. Este método mejoró el funcionamiento de las técnicas genéricas para calcular asociaciones que no tienen en cuenta la demografía del escritor.

En este trabajo se propone el uso de un recurso que recoge normas de asociación de palabras en español de México [1]. Desde este corpus, se aprenden representaciones vectoriales de las palabras.

3. Normas de asociación de palabras (NAP)

Las normas de asociación de palabras (WAN, por sus siglas en inglés), son corpus de asociaciones libres de palabras. Uno de los primeros ejemplos de estas recopilaciones es el que ofrecen *Kent y Rosanoff* [13], quienes usaron el método para estudiar la demencia a partir de 100 palabras estímulo emocionalmente neutras. Los autores llevaron a cabo el primer estudio a larga escala, con 1000

informantes, y concluyeron que existía uniformidad en la organización de las asociaciones, de manera que los adultos sanos comparten redes estables de conexiones de palabras [11].

Muchas lenguas cuentan con recopilaciones de Normas de Asociación de Palabras. En las décadas pasadas se han elaborado algunos trabajos interesantes con gran cantidad de voluntarios. Entre los recursos más conocidos en inglés accesibles desde la web se encuentra el *Edinburgh Associative Thesaurus*[3] (EAT) [14] y la compilación de *Nelson et al.* [16][4].

Para el español, existen algunos corpus de asociaciones libres, entre los que se encuentra el *Corpus de Normas de Asociación de Palabras para el Español de México* (NAP, a partir de ahora) [1], que es el primer recurso de este tipo especialmente recopilado entre hablantes nativos mexicanos del español. El corpus NAP fue elaborado con una muestra de 578 adultos jóvenes, hombres (239) y mujeres (339), con un rango de edad que va desde los 18 a los 28 años, y con un rango de educación de al menos 11 años. El número total de tokens del corpus es 65731, con 4704 palabras diferentes.

Para esta tarea se usaron 234 palabras estímulo, todas ellas sustantivos comunes tomados del *Inventario de Compresión y Producción de palabras MacArthur* [12] de *Jackson-Maldonado et al.* Es importante mencionar que si bien los estímulos son siempre sustantivos, las palabras asociadas son de selección libre, es decir, los informantes pueden relacionar a la palabra estímulo con cualquier palabra sin importar su categoría gramatical.

Los *estímulos* se dividieron en dos listas A y B de 117 palabras cada una. Para cada *estímulo* y sus asociados, los autores investigaron diferentes medidas. Entre ellas, las más relevantes para nuestro trabajo son tiempo, frecuencia y fuerza de asociación.

4. Representaciones distribuidas de palabras sobre el NAP

El grafo que representa el corpus NAP se define formalmente como $G = \{V, E, \phi\}$ donde:

- $V = \{v_i | i = 1, ..., n\}$ es un conjunto finito de nodos de longitud n, $V \neq \emptyset$, que corresponde a los *estímulos* y sus *asociados*.
- $E = \{(v_i, v_j) | v_i, v_j \in V, 1 \leq i, j \leq n\}$, es el conjunto de aristas.
- $\phi : E \rightarrow \mathbb{R}$, es una función de peso sobre los ejes.

Hemos experimentado con grafos dirigidos y no dirigidos. En los grafos dirigidos, cada par de nodos (v_i, v_j) sigue un orden establecido donde el nodo inicial v_i corresponde a la palabra *estímulo* y el nodo final v_j a una palabra asociada. Para el grafo no dirigido, se toman todos los *estímulos* y se conectan con todas las palabras asociadas sin ningún orden de precedencia. Evaluamos tres funciones de peso para los ejes:

[3] http://www.eat.rl.ac.uk/
[4] http://web.usf.edu/FreeAssociation

Tiempo Mide los segundos que el participante tarda en dar una respuesta para cada *estímulo*.

Frecuencia Establece el número de ocurrencias de cada una de las palabras asociadas a un *estímulo*.

Fuerza de asociación Relaciona la frecuencia con el número de respuestas para cada estímulo. Se calcula de la siguiente manera: siendo AW la frecuencia de una palabra determinada asociada a un *estímulo*, y ΣF la suma de las frecuencias de las palabras conectadas el mismo *estímulo* (el número total de respuestas), la fuerza de asociación (AS) de la palabra W a dicho *estímulo* se obtiene con la fórmula:

$$AS_W = \frac{AW * 100}{\Sigma F}.$$

4.1. Node2vec

El algoritmo *node2vec* [9] encuentra un mapeo $f : V \rightarrow \mathbb{R}^d$ que transforma los nodos de un grafo en vectores de d-dimensiones. Define un vecindario en una red $N_s(u) \subset V$ para cada nodo $u \in V$ a través de una estrategia de muestreo S. El objetivo del algoritmo es maximizar la probabilidad de observar nodos subsecuentes en una camino aleatorio de una longitud fija.

La estrategia de muestreo diseñada en *node2vec* permite explorar vecindarios con caminos aleatorios sesgados. Los parámetros p y q controlan el cambio entre las búsquedas en anchura (BFS) y en profundidad (DFS) en el grafo. Así pues, elegir un equilibrio adecuado permite preservar tanto la estructura de la comunidad como la equivalencia entre nodos estructurales en el nuevo espacio vectorial.

En este trabajo, hemos usado la implementación disponible en la web[5] del proyecto *node2vec* con valores por defecto para todos los parámetros. Se ha examinado la calidad de los vectores con diferentes número de dimensiones d.

5. Evaluación de los vectores de palabras

Existen diversos métodos de evaluación para técnicas de vectorización de palabras no supervisadas [18], categorizadas como extrínsecas e intrínsecas. En la evaluación extrínseca, se evalúa la calidad de los vectores de palabras en tareas de procesamiento del lenguaje natural (PLN) [9, 10] y se mide la mejora en el rendimiento en la tarea evaluada. La evaluación intrínseca mide la capacidad de los vectores de palabras de capturar relaciones sintácticas o semánticas [2].

La hipótesis de la evaluación intrínseca es que palabras similares deberían tener representaciones similares. Entonces, para evaluar la similitud, primero se llevó a cabo una visualización de una muestra de palabras usando la proyección T-SNE de los vectores de palabras en un espacio vectorial bi-dimensional. En la Figura 1 se aprecia como se agrupan las palabras que están relacionadas entre

[5] http://snap.stanford.edu/node2vec/

sí. Se muestran los resultados obtenidos con las tres formas de pesado de aristas, y se observa que todas son son capaces de detectar algunas coincidencias en el significado. Las figuras ilustran algunos fenómenos interesantes. Por ejemplo, cuando se toma la frecuencia como peso (la gráfica de abajo), la palabra "pájaro" se dibuja muy cerca de "avión". De aquí se infiere que la característica "volar" es más representativa que "animal' para el modelo. Por su parte, la palabra "Caballo", se representa más cercano a "camioneta" que a otros animales, incidiendo más en su condición de "medio de transporte".

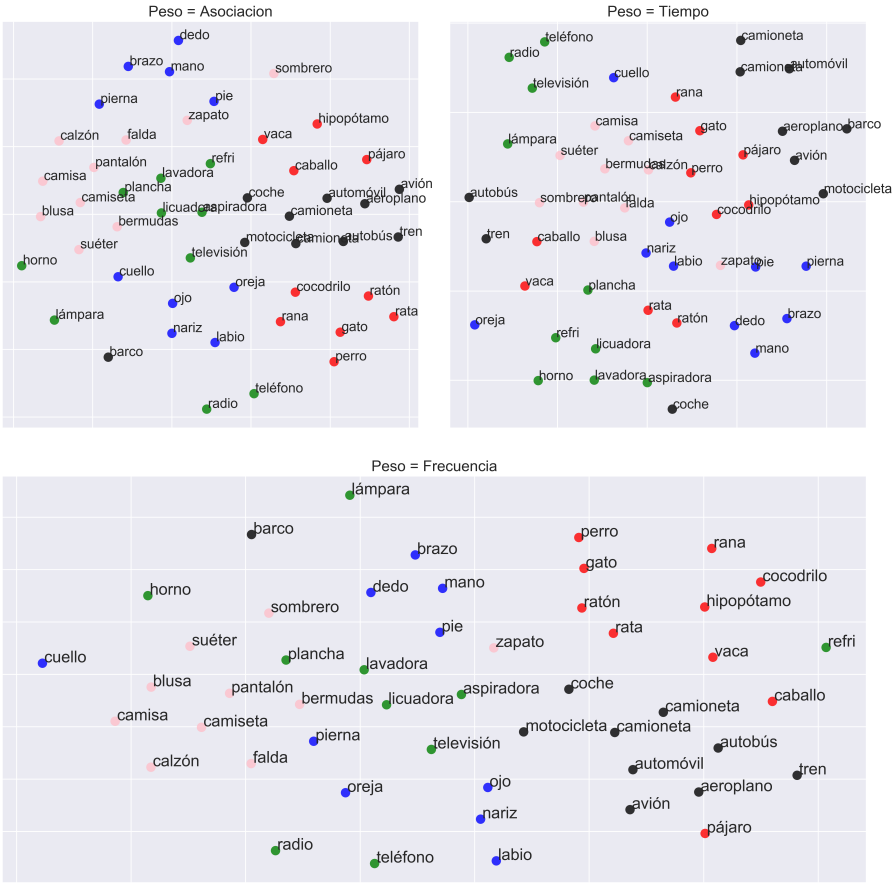

Fig. 1. Proyección de los vectores de palabras para 5 grupos semánticos (de diez palabras cada uno). Los colores están codificados como sigue: animales - rojo, transporte - negro, partes del cuerpo - azul, electrodomésticos - verde y ropa - rosa.

Además, evaluamos la capacidad de los vectores de palabras para capturar las relaciones semánticas mediante una tarea de similitud de palabras. Específicamente, usamos un subconjunto (150 pares de palabras) del corpus *WordSim-*

353 [7] compuesto por pares de términos semánticamente relacionados con puntuaciones de similitud dadas por humanos. *Hassan y Mihalcea* [10] elaboraron una versión de este corpus en español [6].

Nosotros calculamos la similitud coseno entre los vectores del subconjunto de pares de palabras contenidos en el corpus *WordSim-353* y lo comparamos con la similitud dada por humanos. Las Tablas 1 y 2 presentan la correlación de Spearman, en porcentajes, de la similitud dada por etiquetadores humanos, con la similitud obtenida con vectores de palabras (aprendidos del NAP) de diferentes dimensiones aprendidos en los grafos dirigidos y no dirigidos, respectivamente.

Tabla 1. Correlación de Spearman (%) de la similitud coseno calculada con vectores obtenidos del grafo dirigido.

Tamaño del vector	Frecuencia	Asociación	Tiempo
300	-3.07	-3.11	-3.11
200	-1.95	-1.99	-2.03
128	0.88	0.98	0.96
100	**4.61**	**4.61**	**4.63**
50	2.51	2.42	2.39
25	-3.79	-3.89	-3.92

Tabla 2. Correlación de Spearman (%) de la similitud coseno calculada con vectores obtenidos del grafo no dirigido.

Tamaño del vector	Frecuencia	Asociación	Tiempo
300	43.62	43.58	50.77
200	42.89	40.55	44.67
128	39.54	44.01	50.31
100	44.66	44.31	46.50
50	45.60	**47.52**	**53.42**
25	**47.71**	45.75	51.04

Se puede observar que los vectores que se obtienen con los grafos dirigidos no son capaces de trasladar los vecindarios de los nodos al espacio vectorial. En cambio, a causa de la naturaleza no restringida de las aristas, el algoritmo *node2vec* es capaz de caminar diferentes vecindarios en el grafo no dirigido, y por ello consigue mejores representaciones vectoriales.

La tabla 3 muestra la correlación de Spearman entre la similitud coseno obtenida con los vectores pre-entrenados de *word2vec* y la similitud de los humanos (obtenida del corpus *WordSim-353*).

[6] http://web.eecs.umich.edu/ mihalcea/downloads.html

El valor de correlación más alto fue obtenido con los vectores entrenados en el *Spanish Billion Word Corpus* [5] (w2v-1b). Los vectores entrenados con la Wikipedia[7] en español (w2v-wk) obtuvieron resultados similares a los de nuestro método. Los mejores resultados con los vectores entrenados con *node2vec* basados en el NAP se registraron con el grafo no dirigido, considerando el *tiempo* como medida de pesado de las aristas.

Tabla 3. Comparación con vectores pre-entrenados.

Fuente	Tamaño del vector	Correlación de Spearman
w2v-1b	300	**62.20**
w2v-wk	300	53.37
n2v-time	300	50.77
n2v-time	50	53.42

6. Conclusiones

Este artículo propone el uso de corpus de Normas de Asociación de Palabras en lugar de grandes corpus para obtener vectores de palabras. Para ello, se ha aplicado el algoritmo *node2vec* a un grafo construido sobre el NAP para el español de México, una pequeña colección con 4704 nodos.

Los experimentos muestran mejores resultados con grafos no-dirigidos. Se ha otorgado peso a las aristas teniendo en cuenta tres criterios diferentes: *tiempo*, *frecuencia* y *fuerza asociativa*. Los mejores resultados han sido los obtenidos con la categoría tiempo. Visto desde la perspectiva del funcionamiento del sistema *node2vec*, esto no debería ser una sorpresa. Las palabras con una índice más alto de asociación normalmente tienen un tiempo de formulación más breve, y el algoritmo busca los caminos más cortos. Como trabajo futuro, se propone repetir el experimento realizando ajustes a las variables de frecuencia y fuerza asociativa para obtener resultados más concluyentes.

Los resultados que reportamos son comparables a los obtenidos con *word2vec* entrenado con grandes corpus. El rendimiento incluso mejora los resultados alcanzados con *word2vec* entrenados en wikipedia. Sin embargo, algunas estrategias simples ayudarían a mejorar nuestros resultados. Algunas de ellas serían ajustar los parámetros del algoritmo y adaptar el sistema a diferentes tipos de vecindarios para los nodos, que podrían producir diferentes configuraciones de los vectores.

Las evaluaciones realizadas con los vectores generados con el corpus NAP mostraron resultados prometedores respecto a los índices de similitud y relacio-

[7] Vectores de palabras de más de 30 lenguajes: https://github.com/Kyubyong/wordvectors

nalidad, queda como trabajo a futuro la evaluación de estos vectores en alguna tarea de Procesamiento de Lenguaje Natural.

Agradecimientos. Este trabajo ha sido realizado gracias al apoyo de los proyectos: Conacyt FC-2016-01-2225 y PAPIIT IA400117, IN403016.

Referencias

1. Arias-Trejo, N., Barrón-Martínez, J.B., Alderete, R.H.L., Aguirre, F.A.R.: Corpus de normas de asociación de palabras para el español de México [NAP]. Universidad Nacional Autónoma de México (2015)
2. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 238–247 (2014), http://www.aclweb.org/anthology/P14-1023
3. Bel-Enguix, G., Rapp, R., Zock, M.: A graph-based approach for computing free word associations. In: Proceedings of the 9th edition of the Language Resources and Evaluation Conference. pp. 221–230 (2014)
4. Borge-Holthoefer, J., Arenas, A.: Navigating word association norms to extract semantic information. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society (2009)
5. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (March 2016), http://crscardellino.me/SBWCE/
6. De Deyne, S., Navarro, D.J., Storms, G.: Associative strength and semantic activation in the mental lexicon: Evidence from continued word associations. In: Proceedings of the 35th Annual Conference of the Cognitive Science Society. Cognitive Science Society (2013)
7. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proceedings of the 10_{th} International Conference on World Wide Web. pp. 406–414. ACM (2001)
8. Garimella, A., Banea, C., Mihalcea, R.: Demographic-aware word associations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2285–2295 (2017)
9. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining. pp. 855–864. ACM (2016)
10. Hassan, S., Mihalcea, R.: Cross-lingual semantic relatedness using encyclopedic knowledge. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. pp. 1192–1201. Association for Computational Linguistics (2009)
11. Istifci, I.: Playing with words: a study of word association responses. Journal of International Social Research 0 1 (2010)
12. Jackson-Maldonado, D., Thal, D., Fenson, L., Marchman, V., Newton, T., Conboy, B.: Macarthur inventarios del desarrollo de habilidades comunicativas (inventarios): User's guide and technical manual. Baltimore, MD: Brookes (2003)
13. Kent, G.H., Rosanoff, A.J.: A study of association in insanity. Amer J. Insanity 1910(67), 317–390 (1910)

14. Kiss, G., Armstrong, C., Milroy, R., Piper, J.: An associative thesaurus of English and its computer analysis. Edinburgh University Press, Edinburgh (1973)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Computing Research Repository arXiv:1301.3781 (2013), https://arxiv.org/abs/1301.3781
16. Nelson, D.L., McEvoy, C.L., Schreiber, T.A.: Word association rhyme and word fragment norms. The University of South Florida (1998)
17. Sahlgren, M.: The distributional hypothesis. Italian Journal of Disability Studies 20, 33–53 (2008)
18. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 298–307 (2015)
19. Sinopalnikova, A., Smrz, P.: Word association thesaurus as a resource for extending semantic networks. In: Communications in Computing. pp. 267–273 (2004)