Propuesta de Proyecto de Investigación
Maestría en Ciencias y Tecnologías de la Información

01 de noviembre del 2022

1. **Nombre del proyecto**
   - Exploración de redes léxicas automáticas para determinar el deterioro de la memoria semántica y de trabajo presentes en adultos mayores de México

2. **Responsable(s)**

   - Dr. Benjamin Moreno Montiel, Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana – Unidad Iztapalapa, Cubículo T-106, bmoreno@izt.uam.mx, opelo1209@gmail.com
   - Dr. Ricardo Marcelín Jiménez, Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana – Unidad Iztapalapa, Cubículo T-326 Bis, calu@xanum.uam.mx

3. **Área(s) de conocimiento relacionada(s) con el proyecto**
   Aprendizaje Maquinal, Lingüística Computacional, Procesamiento de Lenguaje Natural, Minería de Datos, Teoría de Grafos.

4. **Descripción del proyecto**
   - Contexto:
     Siempre que revisamos la estructura de un Lenguaje (el español o de programación) nos viene a la mente los tres principales componentes, el léxico, el sintáctico y el semántico. Es importante cada uno de estos componentes en lenguajes de programación, ya que de estos depende la correcta estructura de un programa. Cuando aprendemos la fase del analizador lexicográfico conocemos el diccionario de nuestro lenguaje, ya que las palabras reservadas, identificadores y símbolos especiales son expresadas por expresiones regulares. Después conocemos al analizador sintáctico que nos enseña a definir variables, la estructura de una sentencia selectiva e iterativa, entre otras cosas que establecen el orden de un lenguaje mediante diagramas de sintaxis. Finalmente, el proceso que tiene relación directa con la lógica de lo que queremos expresar con el componente léxico y el sintáctico, recibe el nombre de analizador semántico.
     Este pequeño ejemplo del proceso de compilación de un lenguaje de programación nos permite introducir de forma superficial los componentes de cualquier lenguaje hablado. En Lingüística Computacional existen las redes léxicas [1, 2], las cuales representan las relaciones que tienen un conjunto de palabras para denotar la organización de la memoria semántica, el procesamiento del lenguaje y el acceso al léxico. Podemos ver una similitud con respecto al analizador lexicográfico que manejamos en cualquier lenguaje de programación, sin embargo, la complejidad de las redes léxicas es que solo representan una diminuta porción de la complejidad que tiene el cerebro de los seres humanos.
     En la actualidad existen diversas formas de como explorar estas redes léxicas, tomando métodos tradicionales cómo tareas de decisión léxicas [3, 4] o de nombramiento [5, 6], normas de asociación de palabras (NAP) [7, 8] y priming [9, 10]. Sin embargo, en años recientes la teoría de grafos ha tenido aporte en la Lingüística Computacional desarrollando una técnica llamada *small worlds* [11], basados en el uso de grafos y coeficientes de correlación.
     Los modelos que mencionamos anteriormente permiten construir graficas de las redes léxicas, las cuales pueden ser analizadas con estadísticas y métricas para observar las relaciones que existen en la gráfica. Estas métricas tienen el objetivo de entender el comportamiento léxico de los hablantes, permitiendo desarrollar modelos predictivos de la accesibilidad y la similitud dentro del lexicón mental (sucesión de las palabras ordenadas). Algunas medidas comunes sobre gráficas de relaciones léxicas son, estadísticas sobre la gráfica [12], el diámetro [13] y el coeficiente de agrupamiento [14].

Entre tantos métodos para la exploración de redes léxicas y métricas para evaluar las gráficas que se obtienen, la pregunta sería: ¿Podremos utilizar este análisis para conocer algún deterioro en la parte semántica de un ser humano, y dicho en un término formal, este análisis puede determinar algún factor de envejecimiento típico en la memoria semántica de los seres humanos?

- Motivación:
  En Lingüística Computacional existen dos tipos de memorias, la semántica y la de trabajo. La memoria semántica se construye a partir de las redes léxicas, las cuales pueden ser estudiadas mediante técnicas de asociación libre de palabras y la relación que existe entre una palabra estimulo y su respuesta. Durante el envejecimiento típico estas redes se conservan ante la aparición de déficits cognitivos, además se tiene registrado que más de un 60% de las redes léxicas típicas se encuentran tanto en adultos mayores como en adultos jóvenes.
  Esto no pasa con la memoria de trabajo, ya que al envejecer esta sufre una lentificación de las habilidades lingüísticas. Esta memoria es considerada una memoria de corto plazo y en un estado normal permite desarrollar tareas cognitivas complejas como la comprensión del lenguaje, la lectura, las habilidades matemáticas, el aprendizaje o el razonamiento. El envejecimiento típico se relaciona con un desfase entre la producción y la comprensión lingüística, algo que ha sido detectado en la memoria de trabajo, ya que las asociaciones léxicas dependen de un criterio sintáctico y semántico.
  Con base a lo anterior, podemos ver que existe un área de investigación muy interesante en donde las técnicas del Aprendizaje Maquinal y Minería de Datos pueden ser de mucha ayuda. Esto lo mencionamos porque hemos tenido un trabajo pasado en el cual se fusionaros estas dos áreas que parecieran muy distantes, generando el primer lugar en una de las tareas establecidas en el International Workshop on Semantic Evaluation [15],
- Aporte esperado al área de conocimiento:
  En este proyecto pretendemos explorar cada una de las técnicas de exploración que mencionamos a grandes rasgos en el contexto de este proyecto, sin embargo, no mencionamos ninguna forma en como generarlas y basándose en el título del proyecto que proponemos, generarlas automáticamente. Esto nos permite presentar el principal aporte que esperamos obtener con este proyecto, ya que proponemos utilizar la representación vectorial de cada una de las palabras para formar redes léxicas utilizando el factor de similitud SimLex, que por sus siglas en ingles denota la similitud léxica de pares de palabras.
  Como segundo aporte queremos explorar un escenario de clasificación sobre graficas de las redes léxicas que presenten o no deterioros en la memoria semántica y de trabajo, en base a estudios previos y las métricas que mencionamos en la sección del contexto de esta propuesta de proyecto de investigación.
  Sabemos que esto representará todo un reto, ya que se romperá el molde del análisis tradicional sobre las redes léxicas, pero los resultados pueden ser satisfactorios y novedosos para el área de Lingüística Computacional, Procesamiento de Lenguaje Natural y Aprendizaje Maquinal.

5. Objetivos

- Objetivo general: Diseñar un modelo de evaluación de redes léxicas automáticas para determinar el deterioro de la memoria semántica y de trabajo presentes en adultos mayores de México
- Objetivos particulares:
  - Revisión del estado del arte sobre métodos para la evaluación de redes léxicas y sus métricas aplicadas sobre las gráficas obtenidas a partir de la formación de estas.
  - Incorporar la noción de representación vectorial de palabras y el factor de similitud SimLex, para automatizar la creación de las redes léxicas en base a las NAP´s.
  - Establecer un esquema de clasificación sobre las redes léxicas automáticas para determinar cuando una persona puede o no presentar deterioro en su memoria semántica y de trabajo.
  - Realizar pruebas para validación de la propuesta para obtener las tasas de aprendizaje del sistema propuesto.

6. Metodología

Los pasos que necesitamos para desarrollar este proyecto son:

1. Revisión del estado del arte sobre las técnicas de exploración sobre redes léxicas, para determinar cuál es la mejor opción para implementar en este proyecto de investigación.
2. Revisión del estado del arte sobre las principales métricas para evaluar las gráficas que se obtienen con las redes léxicas.
3. Revisión del estado del arte sobre las principales representaciones vectoriales de palabras que se tiene reportadas en la literatura.
4. Modelar y caracterizar las redes léxicas para formar la base de datos del aprendizaje supervisado.
5. Selección de los clasificadores del aprendizaje supervisado que sean los más adecuados manejar las bases de datos de redes léxicas que generemos.
6. Establecer el esquema de clasificación que aplicaremos a los datos que podamos obtener de las redes léxicas automáticas que generemos.
7. Finalmente se plantea utilizar análisis ROC, curvas de aprendizaje Lift y F1 score para obtener la tasa de aprendizaje del sistema propuesto,

7. Calendarización de actividades

Actividad	Trimestre 23-I	Trimestre 23-P	Trimestre 23-O	Trimestre 24-I
Desarrollo del primer borrador de la tesis de los idóneos resultados				
Revisión del estado del arte sobre técnicas de exploración, métricas de evaluación y representación vectorial de palabras.				
Creación de la base de datos de las redes léxicas automáticas.				
Presentación de avances del proyecto.				
Selección de los clasificadores del aprendizaje supervisado.				
Creación del modelo de clasificación del deterioro de la memoria semántica y de trabajo				
Presentación de avances del proyecto.				
Desarrollo de la versión final del sistema de clasificación de redes léxicas automáticas para determinar si se tiene o no un deterioro en la memoria semántica y de trabajo.				
Pruebas y resultados finales para obtener los principales aportes del proyecto.				
Redacción de la Tesis Final				
Presentación de avances del proyecto.				
Examen de Grado				

8. Infraestructura necesaria y disponible

Equipos: Laptop HP OMEN 15-ce0xx o el uso de servidores AWS para el desarrollo del proyecto.

Lenguajes de Programación: C, C++, C#, Python, Matlab y Octave.

9. Lugar de realización

Cubículos 326-Bis y T-106 de la Universidad Autónoma Metropolitana – Unidad Iztapalapa. Sin embargo, en la medida de que la pandemia siga por todo el 2023, se puede realizar un trabajo virtual en casa entre los asesores y el alumno que participe en el Proyecto.

10. Entregables

- Se promoverá la presentación del trabajo en un congreso internacional.
- ICR en los diferentes formatos establecidos por el posgrado para cada Proyecto de Investigación.

11. Referencias bibliográficas básicas

1. Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operation. Journal of Experimental Psychology, 90, 227–234. https://doi.org/10.1037/h0031564.
2. Solé, R., Redes complejas: Del genoma a Internet, isbn = 9788490663332, Metatemas, https://books.google.com.mx/books?id=W4q3DAAAQBAJ, 2016, Tusquets Editores S.A.
3. Alario, F. X., & Ferrand, L. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. L'année Psychologique, 98(4), 659–709. https://doiorg/10.3406/psy.1998.28564
4. Ferrand, L., & New, B. (2003). Semantic and associative priming in the mental lexicon. Mental lexicon: some words to talk about words. Hauppauge, NY: Nova Science Publisher.
5. Hutchison, K., Balota, D., Cortese, M., & Watson, J. A. (2008). Predicting semantic priming at the item level. Quarterly Journal of Experimental Psychology (2006), 61(7), 1036–1066. https://doi.org/10.1080/17470210701438111
6. Lucas, M. (2000). Semantic priming without association: a meta-analytic review. Psychonomic Bulletin & Review, 7(4), 618–630. https://doi.org/10.3758/BF03212999
7. Alario, F. X., & Ferrand, L. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. L'année Psychologique, 98(4), 659–709. https://doiorg/10.3406/psy.1998.28564
8. Macizo, P., Gómez-Ariza, C., & Bajo, M. T. (2000). Associative norms of 58 Spanish for children from 8 to 13 years old. Psicológica, 21, 287–300.
9. Arias-Trejo, N., & Plunkett, K. (2013). What's in a link: Associative and taxonomic priming effects in the infant lexicón. Cognition, 128, 214–227. https://doi.org/10.1016/j.cognition.2013.03.008
10. Mani, N., & Plunkett, K. (2008). Phonological priming in infancy. En Paper presented at the CogSci 2008: 30th Annual Meeting of the Cognitive Science Society.Washington, DC, USA.
11. Atchiston, J. (2012). Words in the Mind: An introduction to the mental lexicon. Oxford: Wiley-Blackwell.
12. Sims, A. D. (2020) Infectional Networks, Graph-theoretic Tools for Inflectional Typology. *En Proceedings of Society for Computation in Linguistics*.
13. Li, W., Lin, Y., & Liu, Y. (2007). The structure of weighted small-world networks. Physica A: Statistical Mechanics and Its Application, 376, 708–718.
14. Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslovak Mathematical Journal, 25(4), 619–633.
15. Helena Gómez-Adorno, Gemma Bel-Enguix, Jorge Reyes-Magaña, Benjamin Moreno, Ramón Casillas, Daniel Vargas, MineriaUNAM at SemEval-2020 Task 3: Predicting Contextual Word Similarity Using a Centroid based Approach and Word Embeddings. *Proceedings of the 14th International Workshop on Semantic Evaluation*, pages 150–157 Barcelona, Spain (Online), December 12, 2020.