



LICENCIATURA EN CIENCIAS DE DATOS
Propuesta de tema de tesis

Modelos de lenguaje como estimadores de los movimientos oculares durante la lectura

Existe una vasta literatura que respalda la hipótesis de que los movimientos oculares durante la lectura reflejan procesos cognitivos. Por ejemplo, las palabras fuera de contexto, poco frecuentes o nuevas, requieren más tiempo de lectura. En paralelo, los modelos de lenguaje actuales pueden predecir la próxima palabra en un texto, y se ha investigado su capacidad para predecir atributos de los movimientos oculares, como la duración de la fijación. El objetivo de esta investigación es analizar cómo distintas variables textuales predicen atributos de los movimientos oculares, utilizando modelos computacionales como N-gram, Word2Vec, FastText, y GPT-2. También se estudiarán efectos simples como la repetición de palabras y su posición en el texto. Los análisis se realizarán con modelos lineales mixtos y modelos aditivos generales. Los datos provienen de un experimento de lectura de cuentos cortos con 76 estudiantes, resultando en 1018 ensayos con aproximadamente 800 palabras cada uno. Los datos están curados y el código para calcular medidas clásicas ya está disponible.

Palabras clave: Modelos de Lenguaje, Lectura, Movimientos Oculares, Modelos Lineales Mixtos

Conocimientos deseables

Si lx estudiante cursó materias optativas como Procesamiento de Lenguaje Natural, Aprendizaje Automático, entre otras mejor, si no lo veremos durante la tesis. No se requieren conocimientos específicos.

¿Qué podría aprender quien realice esta tesis?

Lx estudiante aprenderá a trabajar con datos de comportamiento humano (movimientos oculares) reales, modelos de lenguaje y embeddings de palabras, y modelos lineales mixtos. Además, se acercará a contenidos relacionados a aspectos cognitivos de la lectura.

Dirección de la tesis

*Kamienkowski, Juan Esteban
ICC, UBA-CONICET; DC, FCEyN, UBA*

Contacto: juank@dc.uba.ar

Más información en el pdf a continuación.

Modelos de lenguaje como estimadores de los movimientos oculares durante la lectura

Fermín Travi, Bruno Bianchi, Diego Fernández Slezak, Juan E Kamienkowski (Laboratorio de Inteligencia Artificial Aplicada, Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires - CONICET).

Preludio

El año pasado, estuvimos tomando datos de eye-tracking (seguimiento ocular) durante la lectura de 20 cuentos cortos (~800 palabras) en español Rioplatense. El experimento se dividió en dos sesiones de una hora cada una (en días distintos), donde en cada sesión se leían 10 cuentos. Un total de 76 estudiantes participaron del experimento, resultando en 1018 ensayos distintos.

Los datos ya se encuentran curados y el código para calcular distintas medidas clásicas sobre cada palabra (cuánto tiempo se la vio, si se regresó a ella, etc.) está disponible. Dado que es el dataset de lectura en español más grande hasta el momento, nos gustaría replicar algún resultado previo para validar los datos.

Propuesta

Trasfondo: Hay una amplia literatura relacionada a los movimientos oculares durante la lectura que soporta la hipótesis de los ojos como ventana a la mente. Más precisamente, las medidas derivadas a partir de los movimientos oculares reflejan procesos cognitivos diversos. Por ejemplo, se ha encontrado que si una palabra a leer está fuera de contexto y es impredecible, se le dedica más tiempo de fijación (es decir, hay una relación directa entre la predictibilidad de la próxima palabra y el tiempo que se le dedica al leerla). De esto se deriva que palabras poco frecuentes también sean miradas por más tiempo, lo cual se atribuye al mayor esfuerzo cognitivo requerido para su comprensión. Más aún, estudios previos han demostrado que el tiempo de lectura sobre una palabra decrece tras ocurrencias sucesivas en un mismo texto.

Paralelamente, en las últimas décadas se desarrollaron modelos de lenguaje muy precisos, capaces de generar abstracciones semánticas precisas. Dada una parte de una oración o texto, estos modelos pueden determinar cuál es la siguiente palabra más probable (dentro de su vocabulario). En otras palabras, a cada una de las palabras de su vocabulario le es asignada una probabilidad de que ella sea la próxima.

En los últimos años, se ha comenzado a evaluar la capacidad de estos modelos para predecir distintos atributos de los movimientos oculares, como el tiempo que el ojo se detiene a analizar cada palabra (duración de la fijación o de la mirada) [Bianchi et al. (2020), Hofmann et al. (2022)].

Objetivo: Se busca analizar cómo distintas variables estimadas a partir del texto pueden predecir distintos atributos de los movimientos oculares comúnmente utilizados como estimadores de procesos cognitivos. En particular, se analizará la predictibilidad de la siguiente palabra a partir de distintos modelos computacionales entrenados en texto (como N-gram, Word2Vec, FastText, RNN y modelos más avanzados como GPT-2) [Bianchi et al. (2020), Hofmann et al. (2022)]. También se estudiarán los efectos de variables más simples como repetición sobre las medidas de lectura o la posición de la palabra en el texto [Kamienkowski et al. (2018)]. Los métodos estadísticos utilizados serán modelos lineales mixtos y modelos aditivos generales.

Bibliografía:

[Bianchi et al. (2020)] Bianchi, B., Bengolea Monzón, G., Ferrer, L., Fernández Slezak, D., Shalom, D. E., & Kamienkowski, J. E. (2020). *Human and computer estimations of predictability of words in written language*. Scientific reports, 10(1), 4396.

[Hofmann et al. (2022)] Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). *Language models explain word reading times better than empirical predictability*. Frontiers in Artificial Intelligence, 4, 730570.

[Kamienkowski et al. (2018)] Kamienkowski, J. E., Carbajal, M. J., Bianchi, B., Sigman, M., & Shalom, D. E. (2018). *Cumulative repetition effects across multiple readings of a word: Evidence from eye movements*. Discourse Processes, 55(3), 256-271.