Estimadores computacionales de la predicción humana durante la lectura

Kamienkowski JE[123], Bengolea Monzón G[1], Bianchi B[1], and Ferrer L[13]

[1] Laboratorio de Inteligencia Artificial Aplicada, Depto. de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina
[2] Depto. de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina
[3] CONICET, Argentina

Abstract. El estudio de los movimientos oculares es de gran interés para la neurociencia, ya que reflejan procesos cognitivos que subyacen a la ejecución de cualquier tarea visual, en particular a la lectura. Una variable muy importante que determina estos movimientos es la Predictibilidad. Esta variable representa la predicción que va haciendo el lector sobre la palabra que viene a continuación. Actualmente, en campos como la neurolingüística, esta variable no se estima a partir del texto sino de las respuestas de otros lectores completando la palabra que sigue a un contexto dado. En paralelo, el desarrollo del campo de procesamiento de lenguaje natural ha tenido estimar este tipo de predicciones de manera automática como una de sus metas. Uno de los ejemplos más simples y a la vez más exitosos son los n-grams. Donde la probabilidad de predecir una palabra se construye a partir de la aparición del contexto en un corpus grande de texto, que representa el conocimiento del lenguaje que tiene el lector. Este modelo puede extenderse utilizando el texto recientemente leído para actualizar dichas probabilidades (*cache n-gram*). En este trabajo se propone estimar la predictibilidad de una palabra de forma automática a partir de distintas variantes de modelos de lenguaje. Para nuestro conjunto de datos mostramos que la nueva predictibilidad automática es igual de efectiva que la predictibilidad humana para explicar los movimientos oculares, siendo mucho más sencilla, barata y rápida de obtener por no requerir experimentos que involucren una gran cantidad de personas.

1 Introducción

El lenguaje es uno de los elementos distintivos de los seres humanos, y ha sido investigado desde diversos puntos de vista y disciplinas. En particular, el lenguaje escrito hoy es abordado desde la lingüística, la neurociencia, la ciencias de la visión, la psicología, y más recientemente la computación. En el presente trabajo nos proponemos realizar un esfuerzo más en acercar los campos de la neurociencia y la computación, en particular el procesamiento de lenguaje natural y el estudio de los movimientos oculares durante la lectura, con el objetivo de implementar modelos de lenguaje que emulen el comportamiento humano.

La retina del ojo humano posee una región en el centro del campo visual con muy alta resolución llamada fóvea, de apenas unos grados de ángulo visual, y una periferia de menor resolución [Pal99]. Estas restricciones provenientes de la distribución misma de fotoreceptores en la retina, hace que, por ejemplo, no sea posible leer y ni siquiera distinguir palabras en la periferia [Toe92]. Entonces, para poder adquirir información de calidad de una imágen extensa, o de todo un texto, es necesario desplazar la mirada a través de toda escena, deteniéndose en ciertos puntos para poder adquirir información. De hecho, el patrón de movimientos oculares observado en casi cualquier tarea que realice un humano, esta compuesto por *sacadas* —movimientos rápidos, de entre 200 y 1000 grados/segundo— y *fijaciones* —detenciones de aproximadamente 250ms—. En trabajos fundacionales en el campo de los movimientos oculares, se ha mostrado que los puntos

en los se detiene la mirada no son aleatoreos sino que responden a propiedades de la escena y de la tarea a realizar [Yar67,Rol15]. En particular en la lectura, la mirada se detiene especificamente en las palabras, y el tiempo que cada palabra es fijada depende fuertemente de una gran cantidad de variables de la palabra y del contexto [Jus80,Ray98,Kli06]. Una de las variables más importantes que afectan los tiempos de lectura es la *Predictibilidad*, que representa la capacidad del lector de inferir la palabra que viene a continuación, y por ende acelerar su procesamiento —es decir, disminuir el tiempo de fijación—. La predictibilidad en el texto es además una variable que pone en evidencia mecanismos de control cognitivo sobre la percepción, modificándola según conceptos ya procesados o almacenados. Estos procesos son responsables por ejemplo, de que el texto sea percibido completo, y no se tenga la sensación de que se está viendo sólo una pequeña porción instante a instante [Ray98]. Estas predicciones se construyen en base a claves semánticas, gramaticales o mnemónicas.

En los estudios de lectura, la forma estándar de estimar la predictibilidad de una palabra en un texto es reclutando voluntarios que, en el laboratorio, realicen un experimento llamado *Cloze-Task* [Tay53], en el que se les muestra el texto hasta una determinada palabra y se les pide que lo completen con la palabra que esperan que vaya a continuación. Al agregar las respuestas de muchos participantes se obtiene una distribución empírica de la probabilidad de aparición de una palabra dado un contexto previo. Entre las desventajas de este método se encuentran que resulta muy costoso, en tiempo y esfuerzo, y que debe repetirse para cada palabra en cada contexto.

Por otro lado, desde el campo de Procesamiento de Lenguaje Natural, se vienen desarrollando con mucho éxito modelos de lenguaje para diversas aplicaciones, como transcripción automática de habla, teclados predictivos, etc. Estos modelos tienen muchas veces una tarea similar al humano completando un cloze task, predecir la palabra que viene a continuación en base al contexto previo. El avance de la tecnología (mejores algoritmos, mayor rapidez de cómputo y capacidad de almacenamiento) y la explosión en la cantidad de información disponible hace que el alcance de estas tecnologías sea cada vez mayor

Uno de los modelos más simples, pero a la vez más exitosos, se basa en asumir que la probabilidad de aparición de una palabra depende sólo de las $n-1$ anteriores, estimándola a partir de un corpus grande de texto donde aparezca muchas veces ese contexto. Estos modelos se conocen como n-grams, y son el método más usual para hacer modelado de lenguajes. En un modelo n-gram, la probabilidad $p(w_1^l)$ de observar una oración w_1^l, donde w_j^k denota la secuencia $w_j, w_{j+1}, \ldots, w_k$, se define como:

$$p(w_1^l) = \prod_{i=1}^{l} p(w_i \mid w_1^{i-1}) \approx \prod_{i=1}^{l} p(w_i \mid w_{i-(n-1)}^{i-1}) \tag{1}$$

donde $p(w_i \mid w_1^0) = p(w_i)$. Es decir, asumimos que podemos aproximar bien la probabilidad de la aparición de la i-ésima palabra utilizando solamente las $n-1$ anteriores (o tantas como haya disponibles hasta el comienzo de la oración).

A su vez, podemos estimar las probabilidades condicionales $p(w_i \mid w_{i-n+1}^{i-1})$ utilizando las frecuencias relativas:

$$p_{ML}(w_i \mid w_{i-n+1}^{i-1} t) = \frac{c(w_{i-n+1}^{i})}{c(w_{i-n+1}^{i-1})} \tag{2}$$

Donde $c(\alpha)$ denota la cantidad de veces que aparece la cadena α en el corpus de entrenamiento. Este estimador p_{ML} es el denominado de Máxima Verosimilitud (*Maximum Likelihood*). Cuanto mayor es la profundidad de un n-gram, mayor es el contexto usado y mayor la expresividad del modelo, pero también mayor la dificultad para estimar los parámetros del modelo, cuyo número crece en forma exponencial. Dada la cantidad de combinaciones es muy probable que en

la práctica aparezcan n-grams no vistos, a los cuales el estimador de Máxima Verosimilitud les asigna incorrectamente una probabilidad nula, para lo cual existen distintas técnicas de suavizado [Goo01].

Partiendo de estos modelos es posible hacer diversas modificaciones, adaptándolos a cada uso. Por ejemplo, es posible introducir una memoria a corto término sobre el texto recientemente leído. Esta modificación, propuesta originalmente por Kuhn y De Mori (1990) para problemas de Reconocimiento del Habla, se denomina *cache*, y desde entonces es una de las variantes más comunes para n-grams [Kuh90,Goo01]. Brevemente, la implementación consiste en entrenar un modelo n-gram de baja profundidad a medida que se va procesando el texto de prueba (*modelo dinámico*), para luego ir utilizándolo palabra a palabra en combinación con un modelo n-gram ya entrenado previamente (*modelo estático*) [Wan08,Goo01]. Se pueden distinguir dos aspectos lingüísticos importantes que explicarían por qué funcionan: el realzamiento de palabras recientes y el aprendizaje de vocabulario nuevo. Esto implica que estos modelos resultan más provechosos cuanto más difieren el corpus de entrenamiento y el texto de prueba.

En la literatura pueden encontrarse algunos pocos estudios que exploran la relación entre modelos de lenguaje, la predictibilidad cloze y los movimientos oculares [Bia14b,Ong08,Smi11]. En un estudio previo hemos mostrado correlaciones estadísticamente significativas entre la predictibilidad de una palabra y la distancia semántica al contexto —medida con *Latent Semantic Analysis*—. Los coeficientes obtenidos resultaron menores a 0.1, mostrando que si bien la distancia semántica captura cierta información de la predictibilidad de la palabra, esta conforma sólo una parte de la estimación realizada por los humanos [Bia14b]. Este estudio fue realizado en el mismo corpus de textos largos que el presente trabajo (cuentos). En un estudio previo en oraciones aisladas, Ong y Kliegl (2008) ya habían mostraron correlaciones entre predictibilidad y la distancia semántica al contexto —medida con co-ocurrencia—. Sin embargo, al incluir esta variable en el análisis de los movimientos oculares, los autores concluyen que los lectores no usan esta información para guiar el movimiento de los ojos [Ong08]. Finalmente, un único estudio explora un modelo de lenguaje basado en n-grams [Smi11], en el mismo toman pares contexto-palabra existentes en un corpus grande, sobre el cual tienen estimado un modelo n-gram simple. Luego, a partir de estos pares generan un corpus de oraciones que evaluan en un experimento tipo cloze task, de esta manera estudían la correlación entre el modelo computacional y la predictibilidad humana. Si bien muestran correlaciones significativas, exploran fundamentalmente las diferencias a la hora de predecir los tiempos de lectura [Smi11].

Este trabajo se propone avanzar en la confluencia de estas dos áreas de investigación —el estudio de los movimientos oculares durante la lectura y Procesamiento de Lenguaje Natural— con el objetivo específico de generar modelos de lenguaje que permitan estimar la predictibilidad de una palabra en forma automática. En particular, esperamos reemplazar a las medidas obtenidas con experimentos cloze-task, muy costosas, por estimaciones computacionales de la Predictibilidad.

2 Métodos

2.1 Obtención y Tratamiento de los Datos

A lo largo del trabajo se utilizarán dos corpus distintos: el de entrenamiento y el de evaluación. El corpus de entrenamiento es el utilizado para entrenar los modelos de lenguaje. Mientras que el de evaluación es un corpus más chico que contiene las mediciones de los movimientos oculares y las respuestas a los cloze-task. Los textos de entrenamiento son una recopilación de 2082 libros en castellano —los cuales no incluyen a los utilizados para evaluación—, que combinados suman

aproximadamente 107 millones de palabras [4]. Para la evaluación utilizamos un recopilación de 7 cuentos, de los cuales la mitad de las palabras estaban anotadas con entre 10 y 20 respuestas por palabra en un *cloze-task* implementado *online* [Bia14a,Bia14b]. Estos cuentos habían sido seleccionados por su idoneidad para los experimentos de movimientos oculares: su extensión (aproximadamente 3000 palabras) es suficiente para evaluar efectos dentro un mismo texto (e.g. cómo afecta a la predictibilidad la repetición y la posición en el texto) al mismo tiempo que se logra un gran porcentaje de completitud de los experimentos por parte de los sujetos. También se los seleccionó por casi no presentar diálogos y utilizar un español argentino o neutro que coincide con el idioma de los sujetos de experimentación [Bia14a,Car13]. El detalle de la curación de los textos de entrenamiento y evaluación puede encontrarse en [Ben16].

2.2 Métricas de Performance

Existen varias formas de medir la aptitud de un modelo. En este trabajo utilizamos tres: la *perplejidad*, que es una forma estándar de medir rendimiento de modelos de lenguaje, la correlación de las probabilidades asignadas a los targets con las probabilidades asignadas por el modelo, que nos permite evaluar modelos de lenguaje, y el *Akaike Information Criteria* (AIC) obtenido a partir de la implementación de modelos lineales mixtos (LMMs) para explicar los movimientos oculares.

En la literatura, la forma más común para evaluar un modelo de lenguaje es la perplejidad, que es una medida que intenta capturar cuan bien el modelo predice nuevas muestras del lenguaje que se busca reconocer. El mejor modelo se considera aquel que tiene menor perplejidad en un texto de prueba.

Dado que sólo se quiere calcular la perplejidad sobre las mismas palabras (w) para las que existe un valor del cloze-task (targets) y que se están utilizando n-grams, el cálculo de la perplejidad se reduce a la siguiente ecuación (ec. 3):

$$PP(w) = 10^{-\frac{1}{|T|}\sum_{i:w_i \in T} log_{10} p(w_i | w_{i-n+1}^{i-1})} \tag{3}$$

donde T es el conjunto de targets.

Otra métrica posible es calcular directamente la correlación de las probabilidades asignadas a los targets del texto de evaluación por los experimentos cloze-task con las probabilidades dadas por un modelo. Finalmente, una medida muy utilizada de la calidad relativa de un modelo estadístico *para un conjunto dado de datos* es el Criterio de Información de Akaike (AIC).

$$AIC = 2K - 2LogLikelihood \tag{4}$$

donde K es el número de covariables del modelo. En nuestro caso se utilizarán modelos lineales mixtos con la duración de las miradas como variable dependiente y un conjunto de variables explicativas [Bat15], que incluyen factores fijos de las palabras y de los movimientos oculares y factores aleatoreos como el texto leído o el lector. Los mismos fueron elegidos en base a estudios previos contruyendo un modelo base [Bat15,Bia14a,Ben16,KamXX] y se comparó el resultado obtenido al incluir la predictibilidad estimada a partir de los modelos de lenguaje o de los experimentos cloze-task. Es importante destacar que una de las virtudes de la medida AIC es que pondera la bondad de ajuste del modelo y su complejidad, al agregar el término $2K$ (ec. 4).

[4] Estos textos fueron cedidos generosamente por el grupo de Procesamiento de Habla de la Facultad de Ingeniería de la Universidad de Buenos Aires (Argentina) para la realización de este trabajo.

2.3 Modelos de Lenguaje

Para la implementación de los n-grams se utilizó la librería SRILM [Sto02]. Para la implementación del modelo *cache* se utilizaron unigramas estimados sobre todo el texto ya leído (dentro de cada cuento).

Tanto la perplejidad, como los modelos lineales de los movimientos oculares —que usan la variable predictibilidad transformada con una función *logit()* [KamXX,Kli06]—, no están bien definidos si la predictibilidad toma valores iguales a cero. Esto ocurre cuando ningún participante acertó la palabra en el experimento, así como cuando la palabra no apareció previamente en el modelo *cache* o cuando el contexto nunca fue visto en el corpus para los n-grams. Para ello se utilizan métodos de suavizado [Che96]. En el caso de los n-grams, se exploraron diversos métodos de suavizado ya implementados en SRILM, sin observar grandes diferencias en el desempeño de los modelos. En los resultados presentados se utilizó suavizado de Katz [Sto02].

Para la predictibilidad estimada a partir de las respuestas en el experimento cloze y para el modelo *cache* se utilizó un suavizado aditivo 5. Este consiste en sumarle un δ a la cantidad de ocurrencias de una palabra w ($c(w)$), tanto si fue vista como si no lo fue.

$$p_{aditivo}(w) = \frac{c(w) + \delta}{\delta V + N} \tag{5}$$

donde N es el tamaño del corpus y V es el tamaño del vocabulario (i.e. la cantidad de palabras distintas). En el caso del cloze, se considera la cantidad de respuestas distintas en la posición del target como el vocabulario y la cantidad de respuestas totales dadas por los sujetos en la posición del target como el tamaño del corpus en la ecuación. En la implementación del modelo *cache* se define el vocabulario y el tamaño del corpus a partir del texto leído hasta el momento. El valor óptimo de δ se determina experimentalmente por separado para las probabilidades cloze y las *cache*.

Para el modelo de lenguaje final, se combinan las probabilidades obtenidas del modelo *cache* con las del modelo n-gram utilizando una interpolación lineal (ec. 6), introduciendo un segundo parámetro (λ_{cache})[5].

$$p(w_i \mid w_1^i) = \lambda_{cache}\, p_{cache}(w_i) + (1 - \lambda_{cache})\, p_{ngram}(w_i \mid w_{i-n+1}^{i-1}) \tag{6}$$

3 Resultados y Discusión

3.1 Análisis de Profundidad

Se comenzó explorando la profundidad de los n-grams utilizando perplexity como medida (Figura 1A). Se puede observar que las mejoras son importantes hasta $n = 3$, luego apenas mejora para 4 y finalmente son constantes a partir de allí. Inspeccionando las probabilidades resulta que a partir de 3 son en su mayoría *idénticas*. El hecho de que no se degrade la probabilidad para valores altos de profundidad, por no encontrar o encontrar muy pocos ejemplos de casi cualquier secuencia, se debe a que el método de suavizado aplicado utiliza la técnica de retroceso (*backoff*). Brevemente, cuando una secuencia de profundidad n no es encontrada el corpus se busca la de $n - 1$ y así sucesivamente, corrigiendo las probabilidades al combinar modelos de distinto orden [Che96,Sto02].

Estos resultados coinciden además si se compara cómo se comportan las predicciones del modelo n-gram contra las de los humanos, la correlación entre ambas aumenta a medida que sube la

[5] Se exploró también una interpolación geométrica pero con peores resultados.

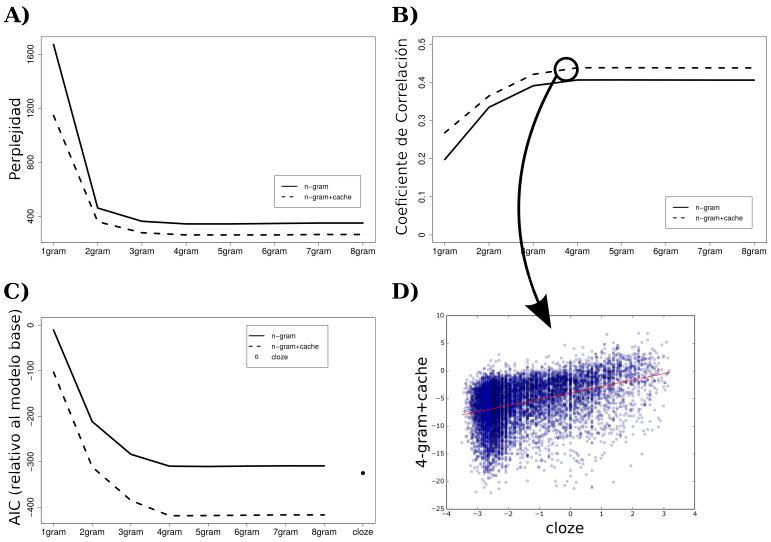

Fig. 1. Evaluación de n-grams de distinta profundidad, con y sin cache, para A) Perplejidad, B) Correlación con las respuestas en el *cloze-task*, y C) Bondad del ajuste de los movimientos oculares (AIC). D) Datos crudos de la estimación de la predictibilidad humana y computacional (4-gram con *cache*), ambos transformados con la función *logit()*.

profundidad hasta 4, a partir del cual deja de mejorar (Figura 1B,D). Así como también, al momento de generar modelos que expliquen los movimientos oculares durante la lectura (Figura 1C). Por lo que $n = 4$ resultó ser la profundidad óptima para los n-grams.

3.2 Análisis Parámetros Cache

Se implementó un modelo *cache* usando unigrama con suavizado aditivo. En una primera exploración con parámetros de λ_{cache} (factor de interpolación con el modelo estático) y δ_{cache} (parámetro de suavizado) similares a la biliografía ($\lambda_{cache} = 0.1$ y $\delta_{cache} = 0$), se encontró que $n = 4$ seguía siendo una profundidad óptima para desarrollar el modelo estático (Figura 1A, línea punteada). Luego se realizó un barrido más fino de los parámetros λ_{cache} y δ_{cache}, buscando aquellos valores que minimicen el AIC del Modelo Lineal Mixto obtenido. Se obtuvo un δ_{cache} de 0.0007, sugiriendo que casi no es necesario aplicar suavizado para las probabilidades del *cache*, lo cual es razonable ya que al estar interpolando con el modelo estático las probabilidades finales nunca son cero. Mientras que λ_{cache} óptimo resultó igual a 0.21, lo cual muestra un peso importante del modelo dinámico en la probabilidad final. Se obtuvieron resultados muy similares para la correlación con la probabilidad cloze (con $\delta_{cloze} = 1$), por ejemplo al variar el λ_{cache}, el λ_{cache} óptimo es 0,17. El aporte del *cache* es evidente para todas las profundidades (Figura 1).

El peso óptimo para el λ_{cache} obtenido con ambas medidas (correlación y AIC) está relacionado con el peso relativo que los lectores le asignan a la historia reciente, y tratándose de unigramas, al léxico utilizado en el texto. Por ello, es esperable que las palabras de contenido (adjetivos, verbos y sustantivos) tengan un λ_{cache} mayor que las palabras de función (determinantes, conjunciones, numerales, artículos y preposiciones). Está hipótesis fue verificada con correlaciones en el presente cuerpo de datos (donde los valores óptimos fueron $\lambda_{cache}(contenido) = 0.25$ versus $\lambda_{cache}(funcion) = 0.16$), así como también evidencia en la diferencia entre el aporte significativo del *cache* al ajuste de los movimientos oculares para palabras de contenido (Figura 2A) y no significativo para las de función (Figura 2B). Cabe destacar que tanto el modelo general de lenguaje,

como la predictibilidad no son buenos predictores de los movimientos oculares (Figura 2B). Este resultado es también consistente con observaciones previas en estudios de análisis de textos, en los cuales se muestra que la repetición de una palabra de función (unigrama) no es específica de un texto, mientras que la repetición de una palabra de contenido sí denota una relevancia dentro del mismo [Mon10]. Resultados similares también se obtienen al estudiar la influencia de la repetición de una palabra en los movimientos oculares, en los cuales se observa que la duración de los movimientos oculares decrece significativamente con la repetición de la palabra, y que este efecto es más fuerte para palabras de contenido de baja frecuencia que de alta frecuencia en el léxico [KamXX].

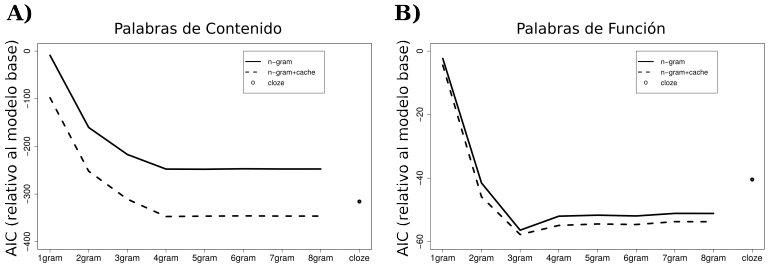

Fig. 2. Evaluación del AIC de n-grams de distinta profundidad, con y sin cache, para palabras de A) contenido (adjetivos, verbos y sustantivos) y B) función (determinantes, conjunciones, numerales, artículos y preposiciones)

Por último, evaluamos limitar el número de palabras utilizadas para conformar el cache, es decir la memoria del modelo. Como podía esperarse, la performance del modelo mejora monotónicamente cuanto más contexto usa, hasta utilizar todo el texto disponible hasta el momento. Sin embargo, es interesante que a partir de las 1000 palabras hacia atrás ya no presenta mejoras importantes. De 0 a 100 palabras en el cache la PP disminuye más de un 10% , mientras que de 1000 a 3000 lo hace en menos de un 1%. Este resultado podría tener implicancias muy interesantes en la comprensión del proceso de lectura.

4 Conclusiones

En el presente trabajo mostramos que es posible construir sustitutos computacionales de la predictibilidad humana a partir de modelos de lenguaje simples. En particular, para este corpus, los mejores resultados fueron obtenidos con un modelo n-gram de profundidad 4, interpolado linealmente con un modelo *cache* de profundidad 1. Para el modelo *cache* no parece necesario realizar un suavizado, mientras que sí lo es para el modelo n-gram, aunque los distintos métodos de suavizado no presentaron diferencias significativas [Ben16]. El factor de interpolación óptimo hallado es de aproximadamente 20% para palabras de contenido. Este factor resultó menor para palabras de función, como era esperado ya que estas no son representativas del texto particular que se está leyendo. Al usar la predictibilidad obtenida con este modelo para predecir los movimientos oculares se logra un ajuste levemente mejor del modelo que al usar la predictibilidad basada en los experimentos cloze. Esto se puede deber, entre otras cosas, a que las probabilidades cloze son obtenidas con pocos ejemplos, mientras que las probabilidades automáticas están basadas en un gran corpus de datos.

El programa a futuro incluye explorar la adición de nuevas fuentes de información al modelo, como las categorías gramaticales y la modelización de una distancia semántica entre palabras, aunque estas no hayan sido co-ocurrido explicitamente en una frase anteriormente.

References

[Bat15] Bates D, Mächler M, Bolker B, Walker S *Fitting Linear Mixed-Effects Models Using lme4*. Journal of Statistical Software 67 (2015): 1–48. 2015.

[Ben16] Bengolea Monzón G *Estudio de Predictibilidad en Textos Naturales: Experimentos comportamentales masivos y de movimientos oculares.*, Tesis de Licenciatura en Cs. de la Computación (FCEyN, UBA), 2016.

[Bia14a] Bianchi B *Estudio de Predictibilidad en Textos Naturales: Experimentos comportamentales masivos y de movimientos oculares.*, Tesis de Licenciatura en Cs. Biológicas (FCEyN, UBA), 2014.

[Bia14b] Bianchi B, Carrillo F, Fernández Slezak D, Kamienkowski JE, Shalom DE *Human and computer estimations of Predictability of words on written language.* 15º Simposio Argentino de Inteligencia Artificial, pags. 99-106, Sociedad Argentina de Informática, 2014.

[Car13] Carbajal MJ *Organización de procesos cognitivos en la lectura natural.*, Tesis de Licenciatura en Cs. Físicas (FCEyN, UBA), 2013.

[Che96] Chen SF, Goodman J *An empirical study of smoothing techniques for language modeling.* Proceedings of the 34th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1996.

[Goo01] Goodman J *A Bit of Progress in Language Modeling.* 2001.

[Jus80] Just MA, Carpenter PA *A theory of reading: From eye fixations to comprehension.* Psychological Review, Vol 87(4): 329-354, 1980.

[KamXX] Kamienkowski JE, Carbajal MJ, Bianchi B, Sigman M, Shalom DE *Cumulative Repetition Effect while Reading Stories: An Eye Movements and Linear-Mixed Models study on a Spanish Corpus.*, Under Review

[Kli06] Kliegl R, Nuthmann A, Engbert R *Tracking the mind during reading: The influence of past, present, and future words on fixation durations.* Journal of Experimental Psychology: General, 135,13-35. 2006.

[Kuh90] Kuhn R, De Mori R *A Cache-Based Natural Language Model for Speech Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990.

[Mon10] Montemurro MA, Zanette DH *Towards the quantification of the semantic information encoded in written language.* Advances in Complex Systems 13.02: 135-153, 2010.

[Ong08] Ong JKY, Kliegl R *Conditional co-occurrence probability acts like frequency in predicting fixation durations*, Journal of Eye Movement Research, 2(1), 2008.

[Pal99] Palmer SE *Vision science: Photons to phenomenology*, MIT press Cambridge, MA, 1999.

[Ray98] Rayner K *Eye movements in reading and information processing: 20 years of research.* Psychological Bulletin, Vol 124(3): 372-422, 1998.

[Rol15] Rolfs M *Attention in active vision: A perspective on perceptual continuity across saccades*, Perception, Vol 44 (8-9):900-919, 2015.

[Smi11] Smith NJ, Levy R *Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing*, Proceedings of the 33rd annual conference of the Cognitive Science Society: 1637–1642, 2011.

[Sto02] Stolcke A *SRILM – An Extensible Language Modeling Toolkit.* Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver, 2002.

[Tay53] Taylor WL *Cloze procedure: a new tool for measuring readability.*, Journalism and Mass Communication Quarterly, 30(4), 1953.

[Toe92] Toet A, Levi DM *The two-dimensional shape of spatial interaction zones in the parafovea*, Vision research, 32(7):1349–1357 1992.

[Yar67] Yarbus AL *Eye movement and vision*, Ed:Plenum Press, New York, 1967.

[Wan08] Wandmacher T, Antoine JY *Methods to integrate a language model with semantic information for a word prediction component.* EMNLP'2007 Conference (Prague), 2008.