Using LSTM-based Language Models and human Eye Movements metrics to understand next-word predictions.

Utilización de modelos de lenguaje basados en redes LSTM y movimientos oculares para la comprensión del proceso de predicción de palabras futuras

Alfredo Umfurer[1,2], Juan Kamienkowski[1,2,3], and Bruno Bianchi[1,2,@]

[1] Departamento de Computación, FCEyN, UBA
[2] Laboratorio de Inteligencia Artificial Aplicada, Instituto de Ciencias de la Computacion, CONICET-UBA
[3] Maestría en Exploración de Datos y Descubrimiento del Conocimiento, FCEyN, UBA
bbianchi@dc.uba.ar

Abstract. Modern Natural Language Processing models can achieve great results on linguistic tasks. For example, LSTM-based models can generate abstractions to make predictions about upcoming words. This ability opens a window in the cognitive neuroscience field.
It is known that the probability that a reader knows a word before reading it (i.e., cloze-Predictability) impacts on the time spent on it. Nevertheless, little is known about when or how these predictions are made.
Here, we trained LSTM-based models to predict future words and used their predictions to replace cloze-Predictability in statistical models from the neuroscience field. We found that the LSTM-Predictability can model eye movements with high overlap with both cloze-Predictability and the lexical frequency. Also, this performance varies depending on the training corpus. This study is a step forward in understanding how our brain performs predictions during reading.

Resumen. Los modelos actuales de Procesamiento del Lenguaje Natural son capaces de alcanzar excelentes resultados en tareas lingüísticas. Por ejemplo, los modelos basados en redes LSTM pueden generar abstracciones para hacer predicciones sobre las palabras futuras. Dicha habilidad abre una ventana en el campo de la neurociencia cognitiva.
Se sabe que la probabilidad de que un lector sepa una palabra antes de leerla (variable denominada cloze-Predictability) impacta en el tiempo que el lector se posa sobre ella. Sin embargo, poco se sabe acerca de cuándo o cómo estas predicciones son realizadas.
Aquí, entrenamos modelos basados en LSTM para predecir palabras futuras y usar sus predicciones para reemplazar la cloze-Predictability en

Received April 2022; Accepted Jun 2022; Published July 2022

modelos estadísticos del campo de la neurociencia. Observamos que la LSTM-Predictability puede modelar movimientos oculares con una alto solapamiento tanto con cloze-Predictability como con la frecuencia léxica. Además, este rendimiento varía en función del corpus de entrenamiento. Este estudio es un paso más hacia la comprensión de cómo nuestro cerebro realiza predicciones durante la lectura.

Keywords: LSTM · Eye Movements · Linear Mixed Models · Reading.

1 Introduction

The Natural Language Processing (NLP) field has witnessed a rapid evolution during the last couple of decades. This evolution has allowed to achieve the resolution of a great number of computational-linguistics tasks. Part of the advances performed in the last decade were made by the use of Recurrent Neural Networks (RNN), in particular, Long Short-Term Memory networks, first introduced in 1997 [11] and popularized some years ago after beating several competitions [23, 9, 21]. This architecture helped to solve some of the issues presented in the RNN. In the classical (vanilla) RNN, the cell state is calculated from the product of all the previous states. This makes the gradients grow or decrease exponentially, generating the vanishing and exploding gradients, respectively. To solve these issues, in the LSTM architecture, the update of the context information is done by the use of gates that allow the old information to be forgotten. This scheme allows the model to retain context information through long sequences of words, minimising gradient issues.

Additionally, in the last couple of years, a new type of architecture was introduced, generating the latest revolution in the field: the Transformers [34]. The main difference between Transformers and RNN is that the former processes the whole sequence of words at once, and not one-by-one like the latter ones. This allows not only to avoid vanishing and exploding gradient issues, but also to speed up the learning process, and thus, to generate bigger models, producing more complex language abstractions. Nowadays many models are based in the Transformer architecture, like BERT [7], ELMO [25], or the GPT family [26, 27, 6], alternating the number of sequential layers of Transformer modules, the presence of encoding and decoding layers, etc.

On the one hand, these advances allow software companies to develop applications with a great impact on our daily use of technologies. Automatic subtitles and closed captions, text translation between a large set of languages, chatbots for customer service that can answer more than pre-defined questions, etc., are now commonly found for final users on several online applications. On the other hand, they can also help us to better understand how our brain processes language during cognitive tasks like reading or listening. From this crossover between NLP and cognitive neuroscience, it could also be possible to generate a better understanding of how the NLP models work. Previous knowledge from neuroscience and brain representations of language can allow us to understand how these models generate the abstractions.

In the Psycholinguistic field, brain mechanisms involved in natural reading are studied by relating word properties with behavioural and physiological data acquired from readers. For example, the eye-tracking technique is based on recording the position of the reader's gaze on a screen during text presentation [17, 28, 19]. With this information, the time expended by the reader on each word (i.e., Gaze Duration –GD–) is analysed as the reflection of their processing cost. It has been shown that GD correlates with word properties like word length (in characters), lexical frequency, position in the sentence and text, and Predictability, among others [19, 18, 2, 28]. Nowadays, these analyses are performed using Linear Mixed Models (LMM). The LMM allow us to understand how each covariable relates to GD, taking into account the variance introduced by subjects or the selected material for the experiment (random effects). Thus, by doing this type of analysis, it is possible to understand which features are used by our brains to process the information.

Most of the previously introduced variables can be easily calculated from the text itself. For example, the word length and word position in the sentence and in the text can be calculated by simple algorithms. Other of these variables are estimated from an independent corpus, like the lexical frequency [29, 8]. But the Predictability is a subjective variable, that depends on readers and not only on the words and the text. This variable is defined as the probability of knowing the word based only on its previous context. It is usually assessed by performing an independent experiment named cloze-task. In this experiment, incomplete contexts are presented to participants who must answer the most probable word that continues it [30]. Then, the cloze-Predictability for a given word in a given context is defined as the proportion of participants that correctly answered it. The quality of the Predictability estimation is closely related to the number of participants on the cloze-task. More participants not only implies more data itself but also a better definition of the scale. For example, if the cloze-task is resolved by 5 participants, cloze-Predictability will only take values of 0, 0.2, 0.4, 0.6, 0.8, or 1. As the number of participants increases, the number of possible cloze-Predictability values also increases. Additionally, this estimation is only valid for a given word in the context in which it was tested. When a new experiment with a different text corpus is planned, a new cloze-task is needed. Thus, the cloze-Predictability is an expensive variable, which results in the fact that several experiments are performed using the same text corpus several times. Finding a computational replacement that behaves like the human-estimated Predictability would allow us to expand the possibilities of renewing the stimuli in the psycholinguistic experiments.

In the last decades, researchers have made several attempts to model cloze-Predictability using simple computational models, but until now, they have not reached conclusive results [24, 12, 2, 13, 1]. In 2008, Ong and Kliegl [24] analysed how the conditional co-occurrence probability of a word given its context, measured by its frequency on internet search engines (Google, Yahoo!, MSN), replaced the cloze-Predictability in eye movements models. They found that their

measure predicts fixation duration more similarly to lexical frequency than to predictability.

More recently, Hofmann and colleagues [12, 13] used NLP algorithms for next-word predictions. In these studies, they trained N-grams, Topic Models (LDA), and Recurrent Neural Networks with a corpus from Wikipedia and movie subtitles, adding the resulting probabilities to statistical models with eye movement variables (single fixation duration and gaze duration).After analysing how much variance these probabilities accounted for in each model they conclude that computational algorithms can explain these eye movements variables better than the original cloze-Predictability. Finally, Algan [1] showed how a LSTM-based Predictability correlates with cloze-Predictability in Turkish.

In 2020, Bianchi and colleagues [2] showed that N-gram probabilities and semantic similarities from different distributional semantics algorithms (LSA, word2vec, FastText) can partially replace the cloze-Predictability in LMM using the GD as the dependent variable. In this study, they not only analysed how computational-Predictabilities predicted GD, but also how much variance these estimations left for the cloze-Predictability. They found that the count-based algorithm N-gram is the one that better models the next-word probability, capturing a relevant part of the cloze-Predictability effect. They also found that this variable also capture much of the variance from the lexical frequency effect. Similarity metrics from all the other three algorithms were not able to explain a relevant amount of the GD variance in the LMM. For these embedding-base models, most of the observed effects were based on taking variance from the Frequency effect. From these results, they conclude that, on one hand, the N-gram language model was their best approach to mimic the cloze-Predictability. On the other hand, that it is important to take into account how much each computational estimation relies on the Frequency effect. As far as we know, this is the only precedent of this type of analysis performed in Spanish.

All together these results shows that language models, even a really simple one like N-gram, are able to capture part of the nature of the human-Predictability. In contrast, metrics from other models results in less accurate estimations. Language Models are designed to model the text a probability distribution, where the occurrence of a word depends on the previous elements of the sentence. These are trained to retain information about the context and predict the upcoming word. For example, the N-gram model estimates probabilities based only on the co-occurrence probability of the exact same chain of the last n words. Modern language models, like LSTM-based language models, add layers of abstractions and the possibility of using more previous context to solve this task.

In the present work, we aim to use the ASGD Weight-Dropped LSTM (AWD-LSTM) model [21] to comprehend how predictions are performed by analysing how they relate to GD and other word properties. To compare with previous results, we use an available corpus of short stories with Gaze Duration and close-Predictability measured for each word. As we previously stated, modelling Predictability with computational algorithms will not only ease the psycholinguistic

experiments but will also allow a better understanding of the human brain predictions. Even more, since the state-of-the-art NLP algorithms are highly opaque to the understanding of how they generate language abstractions, this crossover with the neuroscience field could also be an opportunity for the Artificial Intelligence field. The knowledge from cognitive neuroscience and psycholinguistics could help us to better understand how these complex neural networks achieve their amazing results.

2 Methods

In the present study we will analyse how the cloze-Predictability (Figure 1A) impacts on the Gaze Duration during natural reading of short stories (Figure 1B) and how predictions from two different LSTM-based language models replace this variable (Figure 1C).

2.1 Eye movements

Eye movements were recorded from thirty-six native Spanish readers. All participants had normal or corrected-to-normal vision. Each participant completed sessions of 2 hours reading between three and four stories out of eight possible texts from the *Buenos Aires Corpus* [18]. Stories were presented on a PC monitor (black Courier New bold font; 0.448 letter width; 0.318 minimum letter height; grey background). Each one was presented across several screens, with 10 lines of text at a time (double-spaced, 1.68 interline spacing; 55 character maximum per line). Subjects were instructed to read at their own rate, moving forward or backward in the screen sequence by pressing the right and left arrow keys, respectively. Texts were assigned pseudo-randomly to participants to achieve a similar number of readings of each text. Subjects answered five questions regarding the contents of each text, which were used to determine their comprehension level. We obtained an average of 4.7 correct answers and a minimum of 3.

Gaze position was recorded at a sampling rate of 1kHz with a video-based eye tracker (EyeLink 1000 from SR research). A chin rest that was aligned with the centre of the screen prevented head movements. The participant's gaze was calibrated with a standard 13-point grid for both eyes. Two nine-point validations were run before and after each text. Based on these validations, the best calibrated eye was selected for each participant. Presentation of stimuli was developed using Matlab [4] and Psychotoolbox [5].

Then, gaze position was used to calculate the Gaze Duration (GD) for each word. This variable, also called First Pass Reading Time, is defined as the total time spent on a word before leaving it for the first time, i.e., the addition of all fixations on a word during the first pass, without counting future refixations. This eye movement variable will be used as the dependent variable in the statistical models used in this study. All this data is publicly available from Bianchi et al. [2].

[4] (http://www.mathworks.com/)

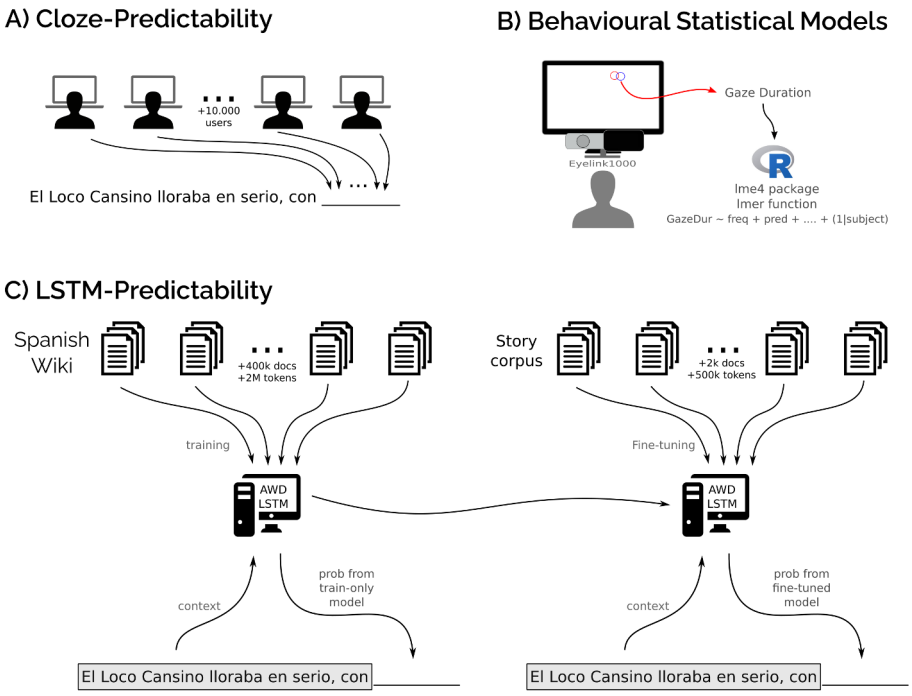

Fig. 1. Experimental designs: **A)** The human Predictability was estimated from the online responses of several participants to a web cloze-task experiment. Each participant had to complete one of every 30 words, and the text was uncovered as they responded. **B)** Eye movements were recorded in separate participants that read three of the eight texts in the lab. The eye movement measures (Gaze duration) were analysed using Linear Mixed Models. **C)** AWD-LSTM architecture was trained on a large Wikipedia corpus and fine-tuned with a smaller corpus of texts from a similar domain as the tested short stories (A,B).

2.2 Cloze Task

The cloze-task is performed by presenting uncompleted texts to participants who have to answer the next most probable word for that context. Then, the Predictability of each word is estimated as the probability of correctly guessing it in the cloze-task. The corpus from Bianchi et al. comprises cloze-Predictability from more than 1000 participants (16 ± 8 per word) collected online [2]. It was performed using a custom-made web page where participants logged-in to find one of the eight selected stories randomly assigned. After finishing a story, participants were allowed to either close the experiment or continue with a new randomly assigned text. Participants that closed the experiment could return to the following stories at any moment.

2.3 AWD-LSTM predictability

In the present study all the LSTM-based language models were fitted using the AWD-LSTM architecture [22]. This architecture is a variant of the LSTM model that introduces a regularisation method called *DropConnect* that avoids the over-fitting of the gates without penalising the training speed. Additionally, it also uses other regularisation methods such as embedding dropout, Variational dropout between layers, and L2 regularisation of the weights.

For this the implementation provided by the fastAI library[5] was trained to get the next-word probability (from now on LSTM-Predictability). The model consisted of three stacked LSTM layers with 400, 1152, and 400 dimensions respectively, and multiple dropout layers, as described by Merity et al. [22]. Each word was represented as 400 dimensional embeddings in the input and the output layers.

Following the method proposed by Howard and collaborators [16], we trained the model in two phases. Firstly, the model was trained using a corpus taken from the Spanish Wikipedia. This corpus has a total of 444,571 documents and 2,751,415 tokens. Under the hypothesis that this training would result in predictions biased towards an encyclopaedic style, and given that the testing corpus is composed of narrative stories, a fine-tuning with a corpus of a closer domain was performed in a second stage. This small corpus consisted of 2,081 Spanish narrative stories with 535,068 tokens [2].

In the first phase, we trained the model for 10 epochs, using a One-Cycle policy with a maximum learning rate of 0.002. In the second phase, we replaced the encoder layer and trained it for two steps (*max learning rate = 0.026*) while keeping the remaining layers unchanged to avoid the effect known as *catastrophic forgetting*. After that, all the parameters were tuned on eight epochs (*max learning rate = 0.0026*).

Both versions, the one trained with Wikipedia-only and its fined-tuned version, were used independently to perform next-word prediction for each word from the story corpus used in the eye tracking experiment previously presented. This data is publicly available at [6].

2.4 Statistical analysis

The logit of Predictability measures (both cloze and LSTM) were used as co-variables in successive LMM with Log-transformed Gaze Duration (GD) as the dependent variable. This model also included as additional co-variables a set of previously described text properties (launch position, the inverse of the word length, the logarithm of lexical frequency, their interaction, and the position in the line, the text, and the sentence). Subject and text identifiers, and the fixated word as a string, were used as random effects. Text variables and the results for the LMM with the ngram-Predictability as co-variable were publicly available by Bianchi et al. [2].

[5] www.fast.ai
[6] reading.liaa.dc.uba.ar

The main outputs of the LMM are the estimates of the slopes and their errors (SD) for each of the fitted fixed factors. Using them, t-values are calculated as the ratio between each slope and its SD. These values represent how far away from zero the slopes are. As our models are fitted with a high number of instances, the distributions of the used co-variables can be considered as normal, and thus, absolute t-values larger than 2.0 are considered significant with $\alpha < 0.05$ [3]. Each significant effect implies a linear relation between that co-variable and the dependent variable. Since the estimate of the slope of a LMM co-variable depends on its scale, and each estimation of Predictability has a different range of values, we based our analyses mostly on the effect t-value, which are standardised.

To analyse how each LSTM-Predictability mimics the cloze-Predictability, residuals of the corresponding LMM are analysed. That is, after fitting a LMM with each LSTM-Predictability as co-variable, residuals of the fixed effects will be used in a new LMM with cloze-Predictability as the only fixed effect, conserving the random structure. For this procedure, we used the *remef* function [14] implementation for R.

To compare the performance of the hierarchically built models on the data the Akaike Information Criterion (AIC) was used. This estimator is calculated as the log likelihood of the model, compensated by the number of fixed effects [33]. The smaller it gets, the better the model can explain the data, compensating the number of variables to avoid over-fitting.

3 Results

A series of Linear Mixed Models (LMM) with different combinations of co-variables were fitted to analyse how the next-word probability from two AWD-LSTM models (LSTM-Predictability) mimics the cloze-Predictability. The baseline model (Fig. 2A, M0) comprised a set of previously described co-variables: launch position of the eye, the inverse of the word length in characters, the logarithms of the lexical frequency, the positions in the line, the text and the sentence, and the interaction between length and frequency. All these variables showed significant effects, as expected from previous studies [18, 2]. Subsequently, the cloze-Predictability was added in a a new model (Fig. 2A, M1), showing a clear and significant negative effect on GD. The addition of this co-variable generated negligible changes in the co-variables effects of the baseline model.

Results from both AWD-LSTM models were added as co-variables in independent LMM. Firstly, we used the output of a LSTM model trained only with a Spanish Wikipedia corpus, a big but not specific corpus for the task (Fig. 2A, M2). The t-value of the Wikipedia-Only LSTM-Predictability on the LMM ($t = -14.97$) was similar to the cloze-Predictability t-value on M1 ($t = -16.23$). Some co-variables from the baseline model showed changes in their effects. This change is particularly substantial on the lexical frequency.

Going one step further, we compared not only how this LSTM-Predictability explain the GD on the LMM, but also how much of the explained variance by this co-variable overlap with the variance explained by other possible co-variables,

like the cloze-Predictability. Then, to observe if the cloze-Predictability can be explained by the results from this LSTM, the residuals of the LMM (M2 Wiki-Only) were fitted into a new LMM with cloze-Predictability as the only fixed effect (Residuals + cloze-Predictability). This analysis showed that the effect of human Predictability remains significant ($t = -11.63$, Fig. 2B, M2). This implies there is still variance associated with cloze-Predictability left after fitting the model with the LSTM results. That is, the AWD-LSTM model trained only

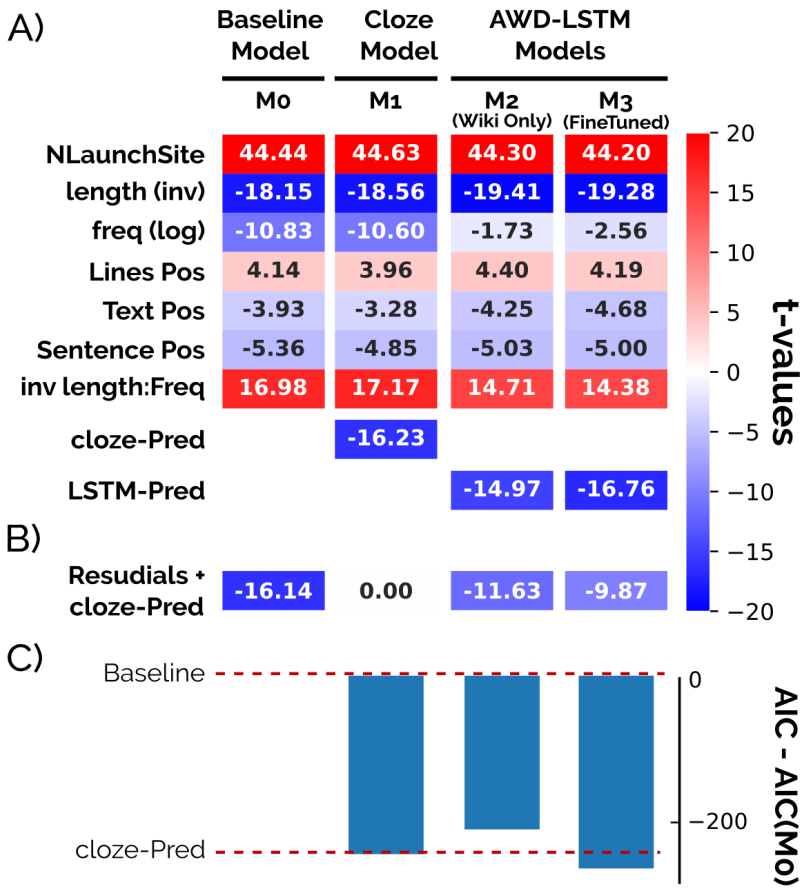

Fig. 2. A) t-values from four LMMs with different sets of co-variables. **M0**: baseline models. **M1**: baseline model and Cloze-Predictability variable. **M2**: baseline model and LSTM-Predictability trained only with Wikipedia. **M3**: baseline model and LSTM-Predictability trained with Wikipedia and fine-tuned with a story corpus. **B)** t-values for the cloze-Predictability effect on a Linear Mixed Model fitted on the residuals of each of models on A. **C)** AIC values for each of the fitted models on A relative to the M0 AIC.

with a Wikipedia corpus can only partially model the cloze-Predictability effect on Gaze Duration. Moreover, the drop in the significance of the frequency effect shows that a part of its effect comes from the lexical frequency and not from the cloze-Predictability.

Secondly, the output of an AWD-LSTM model trained with Spanish Wikipedia and fine-tuned with a corpus of stories was included as a co-variable (Fig. 2A, M3). The t-value for this metric on the LMM was closer to the cloze-Predictability than the wiki-Only estimation ($t = -16.76$). Contrary to the observed result for the M2, the frequency effect remained significant, although largely decreased. Furthermore, the cloze-Predictability effect on the residuals of the LMM in M3 was smaller than in M2 ($t = -9.87$, Fig. 2B, M3), suggesting that the fine-tuning improved the LSTM performance.

These effects also have an impact on the goodness-of-fit of the fitted LMMs, estimated with the Akaike Information Criterion (AIC) for each model relative to M0 and M1 (Fig. 2C). M2 in particular showed an increase in the absolute AIC relative to M0 while decreasing relative to M1. This indicates a better fit than the baseline model but worse than the Cloze Model (M1). Meanwhile, M3 showed a slight improvement in the overall fit relative to M1.

In a previous study Bianchi and collaborators [2] explored, among others, the output of a 4-gram model as co-variable in the same corpus. For this estimation they also found significant effects on the LMM, and a decrease in the frequency effect (Fig. 3A, M4). To compare this ngram-Predictability (4-gram+cache) with the fine-tuned LSTM-Predictability an additional model using this two co-variables together was fitted (Fig. 3A, M5). By doing this, the t-values of both computational-Predictability effects decrease but remains significant. The LSTM-Predictability goes from -16.76 to -6.11 and ngram-Predictability from -21.02 to -13.84. This indicates that there is partial overlap between the variance they explain in the model, but they still explain different aspects of GD. Additionally, the drop in the Frequency effect is complete, showing that both computational-Predictabilities relies on the lexical frequency. Finally, there was a significant effect of cloze-Predictability when fitting the residuals of the LMMs (Fig. 3B, M5). This suggests that the effect of cloze-Predictability on Gaze Duration cannot be fully explained by any these computational models nor both together.

4 Conclusions

On the last decades LSTM networks have allowed great advances in Natural Language Processing tasks. Their large number of internal parameters and their internal architecture that avoids the problem of the vanishing and exploding gradients allow them to learn complex interactions while maintaining context information. These advances in the NLP field open a window to the cognitive neuroscience field to better understand how our brain processes language.

In the present work, we explored how LSTM models natural language using its output to mimic a human-based linguistic variable. The cloze-Predictability

is a commonly used variable in psycholinguistic research to study how our brain process language. This variable is known to correlate with behavioural (e.g. Fixation Duration) [19] and electrophysiological metrics (e.g., scalp potentials) [20]. For this study, we replaced it with the LSTM-Predictability on the statistical models that are used to understand eye movements during reading.

Using a text corpus from the Spanish Wikipedia we trained a LSTM model (trained-only model). This model was then fine-tuned with a small corpus of narrative texts (fine-tuned model). Both models were used to estimate how probable

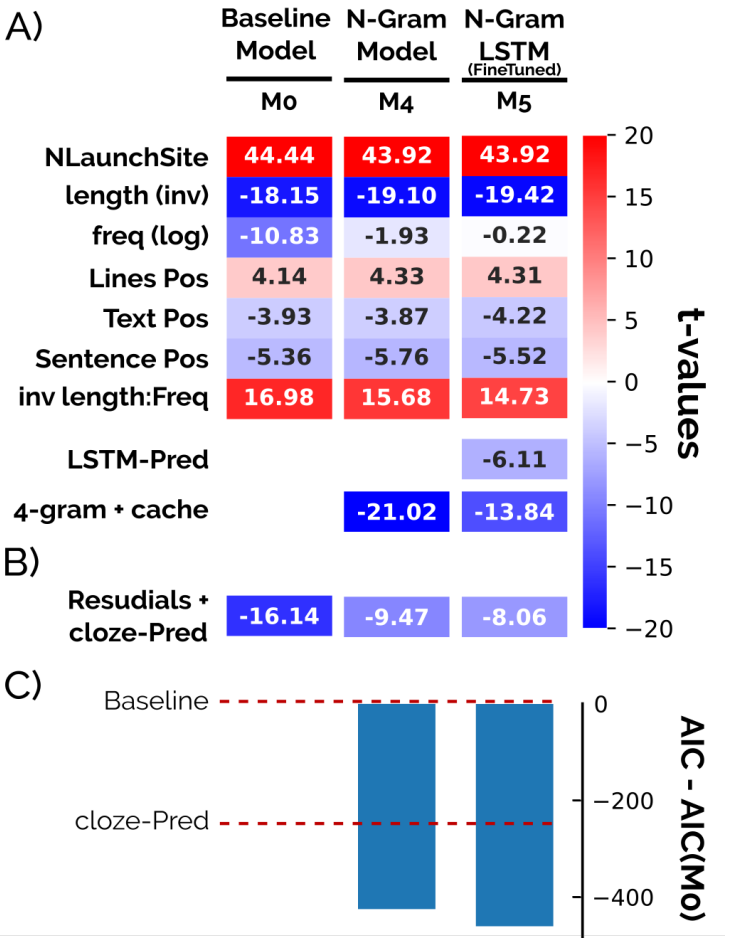

Fig. 3. A) t-values from two more LMMs with different sets of co-variables. **M0**: baseline models. **M4**: baseline model and Ngram model from [2], **M5**: baseline model, Ngram model, and LSTM-Predictability from the fine-tuned model. **B)** t-values for the cloze-Predictability effect on a Linear Mixed Model fitted on the residuals of each of models on A. **C)** AIC values for each of the fitted models on A relative to the M0 AIC.

is each word on a set of 8 stories (LSTM-Predictabilities), previously used by Bianchi et al. [2] for a similar analysis. LSTM-Predictabilities were used as co-variables in a independent LMM with other linguistic properties as co-variables and the Gaze Duration as a dependent variable.

LSTM-Predictabilities from both, the trained-only and the fine-tuned models, showed significant effects, partial overlap with the cloze-Predictability, and an improvement in the LMM goodness of fit (measured with the AIC) relative to baseline model. These results shows that these complex neural networks are able to predict future words in a similar manner than humans. The partial overlap with cloze-Predictability indicates that these predictions are not exactly like human predictions, but that their nature shares some relation on how they affect eye movement variables.

Additionally, in both LMMs a decrease in the frequency effect was observed. On the trained-only model, there was a mayor decrease of the lexical frequency effect, which became non-significant. The fine-tuned model generated a less prominent decrease and the frequency effect remained significant. Thus, to predict future words, LSTM seems to rely on lexical frequency more than humans. This overlap with the frequency effect, which is not present on the cloze-Predictability variable, was previously observed for the conditional co-occurrence metric analysed by Ong and Kliegl [24] and for the N-gram model by Bianchi and collaborators [2]. Thus, this states a clear difference between how humans and these computational models predict future words. These computational-Predictabilities are generated, at least partially, based on the lexical Frequency of the words. Interestingly, the difference between how trained-only and fine-tuned models interacted with the Frequency effect implies that training on a corpus with certain linguistic similarities to the testing corpus could minimise this undesirable issue.

Thus, we conclude that in order to achieve a good replacement of the cloze-Predictability it is important to consider training or fine-tuning the computational model on a corpus similar to the tested one. A general corpus, such as Wikipedia, will lead to rely the predictions mostly on word Frequency. Additionally, our results on the AIC metric show that a computational-Predictability that generates a better goodness of fit does not imply that the former one is better for explaining brain processes underlying predictions. That is, a computational-Predictability variable can explain more variance than the cloze-Predictability, but part of the explained variance could come from other co-variables and brain processes.

The comparison between the AWD-LSTM model presented here and the N-gram model implemented by Bianchi showed that they explained different aspects of the cloze-Predictability, with some degree of overlap. The comparison with simpler and more transparent models may also serve as a way to understand complex models, like LSTM. Despite the fact that the N-gram model can be improved, for example, by adding information about grammatical properties of words [4], the text processing needed for this (like Part-of-Speech tagging) is highly expensive and not robust, while modern NLP algorithms, like AWD-LSTM, can infer this information implicitly. Additionally, algorithms based on

neural networks have more hyperparameters (embedding size, number of layers, etc.) that were not explored in the present study and may allow future improvements.

In this line, future work must be aimed at improving the LSTM-Predictability based on AWD-LSTM and other LSTM architectures, experimenting with different parameters on the training and testing phases. First, using a larger corpus for the specific fine-tuning may result in a better replacement of the cloze-Predictability, allowing us to further explore how LSTM predictions are performed. Secondly, experimenting with the amount of information used by the LSTM to predict future words would give more insight on how long dependencies are used by the model and, also, by the brain. Moreover, these analyses could be extended to more modern models, like transformers based models.

As previously stated, transformers are the results of removing the recurrent aspect of the RNN and keeping only the attention mechanism [34]. This mechanism allows neural networks to learn which elements of sequences are important to attend to and the magnitude of that attention. Thus, transformers can be thought as a simplification of RNN. Nevertheless, after a couple of years of its development, the complexity of transformer architecture has increased exponentially, mainly based on the parallel and serial stacking of attentional heads and layers. Thus, state-of-the-art transformer-based architectures are highly expensive to train from scratch, and it is necessary to use pre-trained models. This could result a limitation when studying Spanish readers. Nevertheless, there is some resources, like small- and medium-size Spanish or multilingual GPT-2 pre-trained models available in on-line repositories[7]. These general propose versions can be fine-tuned for specific domains and tasks, achieving similar results to the ones available for English. In future works we aim to fine-tune one of these versions with a corpus from narrative stories and corpus of other domains to further investigate the objectives of this work.

This work is another step in the dialogue between NLP and Neuroscience, using cognitive and physiological measures to understand NLP and vice versa, that will which both fields [32, 10, 15, 31].

Acknowledgements

The authors were supported by the Consejo Nacional de Investigaciones Cientificas y Técnicas (CONICET), the Universidad de Buenos Aires (UBA), and the Programa de Pasantías from the Departamento de Ciencias de la Computación (Facultad de Ciencias Exactas y Naturales, UBA). The research was supported by the UBA (20020190100054BA), the National Agency of Promotion of Science and Technology (PICT 2018-2699) and the CONICET (PIP 11220150100787CO).

[7] huggingface.co

References

1. Algan, A.C.: Prediction of words in Turkish sentences by LSTM-based language modeling. Master's thesis, Middle East Technical University (2021)
2. Bianchi, B., Monzón, G.B., Ferrer, L., Slezak, D.F., Shalom, D.E., Kamienkowski, J.E.: Human and computer estimations of predictability of words in written language. Scientific reports **10**(1), 1–11 (2020). https://doi.org/https://doi.org/10.1038/s41598-020-61353-z
3. Bianchi, B., Shalom, D.E., Kamienkowski, J.E.: Predicting known sentences: Neural basis of proverb reading using non-parametric statistical testing and mixed-effects models. Frontiers in human neuroscience **13**, 82 (2019). https://doi.org/https://doi.org/10.3389/fnhum.2019.00082
4. Bilmes, J., Kirchhoff, K.: Factored language models and generalized parallel backoff. In: Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers. pp. 4–6 (2003)
5. Brainard, D.H., Vision, S.: The psychophysics toolbox. Spatial vision **10**(4), 433–436 (1997)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., Carreiras, M.: Espal: One-stop shopping for spanish word properties. Behavior research methods **45**(4), 1246–1258 (2013)
9. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE (2013). https://doi.org/https://doi.org/10.1109/ICASSP.2013.6638947
10. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron **95**(2), 245–258 (2017)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
12. Hofmann, M.J., Biemann, C., Remus, S.: Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, eegs and eye movements. In: Cognitive approach to natural language processing, pp. 197–215. Elsevier (2017). https://doi.org/https://doi.org/10.1016/B978-1-78548-253-3.50010-X
13. Hofmann, M.J., Remus, S., Biemann, C., Radach, R., Kuchinke, L.: Language models explain word reading times better than empirical predictability. Frontiers in Artificial Intelligence **4** (2021)
14. Hohenstein, S., Kliegl, R.: remef (remove effects)(version v0. 6.10) (2013)
15. Hollenstein, N., Torre, A., Zhang, C.: Cognival: Framework for cognitive word embedding evaluation. arXiv:1909.09001 (2019)
16. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
17. Just, M.A., Carpenter, P.A.: A theory of reading: From eye fixations to comprehension. Psychological review **87**(4), 329 (1980). https://doi.org/https://psycnet.apa.org/doi/10.1037/0033-295X.87.4.329

18. Kamienkowski, J.E., Carbajal, M.J., Bianchi, B., Sigman, M., Shalom, D.E.: Cumulative repetition effects across multiple readings of a word: Evidence from eye movements. Discourse Processes **55**(3), 256–271 (2018). https://doi.org/https://doi.org/10.1080/0163853X.2016.1234872
19. 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**(1), 12 (2006). https://doi.org/https://psycnet.apa.org/doi/10.1037/0096-3445.135.1.12
20. Kutas, M., Hillyard, S.A.: Reading senseless sentences: Brain potentials reflect semantic incongruity. Science **207**(4427), 203–205 (1980). https://doi.org/10.1126/science.7350657
21. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing lstm language models. arXiv preprint arXiv:1708.02182 (2017). https://doi.org/https://arxiv.org/abs/1708.02182v1
22. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing lstm language models. arXiv preprint arXiv:1708.02182 (2017)
23. Märgner, V., Abed, H.E.: Icdar 2009 arabic handwriting recognition competition. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 1383–1387 (2009). https://doi.org/https://doi.org/10.1109/ICDAR.2009.256
24. Ong, J.K., Kliegl, R.: Conditional co-occurrence probability acts like frequency in predicting fixation durations. Journal of Eye Movement Research **2**(1) (2008). https://doi.org/https://doi.org/10.16910/jemr.2.1.3
25. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arxiv 2018. arXiv preprint arXiv:1802.05365 **12** (1802)
26. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. preprint (2018)
27. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
28. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological bulletin **124**(3), 372 (1998). https://doi.org/https://psycnet.apa.org/doi/10.1037/0033-2909.124.3.372
29. Sebastián-Gallés, N.: LEXESP: Léxico informatizado del español. Edicions Universitat Barcelona (2000)
30. Taylor, W.L.: "cloze procedure": A new tool for measuring readability. Journalism quarterly **30**(4), 415–433 (1953). https://doi.org/https://doi.org/10.11772F107769905303000401
31. Toneva, M., Wehbe, L.: Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Advances in Neural Information Processing Systems. pp. 14954–14964 (2019)
32. Ullman, S.: Using neuroscience to develop artificial intelligence. Science **363**(6428), 692–693 (2019)
33. Vaida, F., Blanchard, S.: Conditional akaike information for mixed-effects models. Biometrika **92**(2), 351–370 (2005)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)