LIAA Laboratorio de Inteligencia Artificial Aplicada

ICC

Deep neural network-based Language Models for estimating word-Predictability

Alfredo Umfurer[1], Juan Kamienkowski[1,2], Bruno Bianchi[1],

1. Laboratorio de Inteligencia Artificial Aplicada, Instituto de Ciencias de la Computación, FCEyN, UBA - CONICET 1

2. Departamento de Física, FCEyN, UBA.

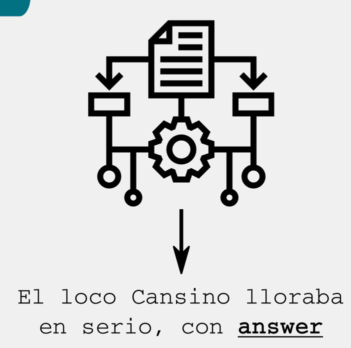
alfredoumfurer@gmail.com

SAN2020 XXXV ANNUAL MEETING OCTOBER 7-9

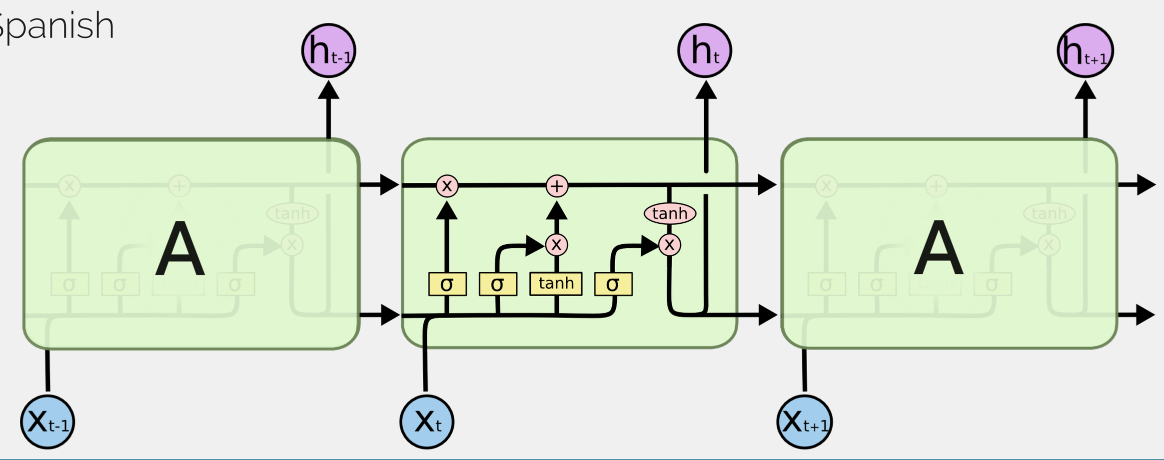
Introduction

- Word-Predictability (i.e. how predictable is a word in a context) is one of the most important variables to explain human behavior and information processing during reading.
- Predicting future words correctly can speed up and ease its processing.
- Understanding how predictions are generated during reading would allow us to better understand more general cognitive mechanisms.
- To estimate it an experiment called Cloze-Task [**see posters #255 Bianchi et al.** and **#115 Peroni et al.**] is used.
- Cloze-Predictability is known to correlate with Eye Movements metrics [Kliegl et al. (2006)].
- Since Cloze-task is an expensive experiment it is difficult to deeply explore predictions. Thus, computer-based Predictability can become a useful tool for the neurolinguistic field.
- A good model has to have similar properties to the cloze-task values. In [Bianchi et al. (2020)] used some models to estimate achieving a partial modeling and use it to model the human eye movements while reading.
- In recent years, advances in computing capacity and data availability allowed us to use more complex neural models, such as Deep Neural Networks.
- Within these models, Recurrent Neural Networks (RNN) are useful to resolve problems involving sequential data.
- The introduction of Long Short-Term Memory (LSTM) models solved problems like the vanishing gradient (i.e. losing long-term information in long sequences), and has allowed great advances in the Natural Language Processing field (NLP)

Computational models of Predictability



- N-Grams and embeddings (LSA and FastText) were previously used to estimate the next word predictability.
- Here we will use Recurrent Neural Network (RNN) to try to improve our previous results.
- LSTM is a RNN architecture with feedback connections, which means that information about previous words can be used to predict the predictability of the next word.
- Like any neural network architecture, LSTM models have multiple parameters that are optimized while training it.
- Here we used AWD-LSTM [Merity, S., 2017; Merity, S., 2018]. It consists of three layers of LTSM cells with dropout layers.
- Given the size of the model, small datasets are not enough to get good results since they are likely to get overfitted, so they are trained first on a large corpus and **fine-tuned using the target corpus.**
- For the training we used 444.571 documents from Spanish wikipedia, and later this model was fine tuned using a corpus of 2081 books in Spanish



Eye movements analisys

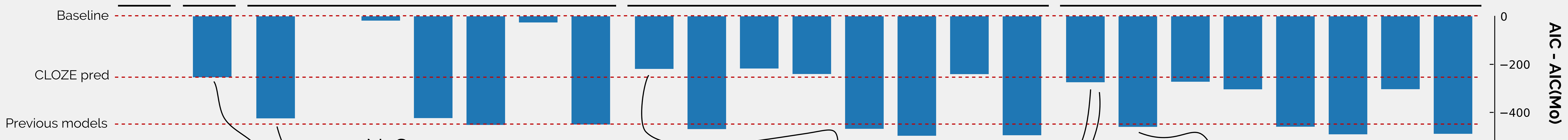
- While reading our eyes move across words, stopping on them for a variable amount of time.
- Time spent on a word (i.e, Gaze Duration -GD-) reflects processing cost.
- Processing cost depends on word and text properties.
- We analyzed the dependency between GD and some of these properties using Linear Mixed Models (LMM).
- Analyzing LMM t-values for each fixed effect and their changes when fitting nested models allows us to better understand how these properties are involved in brain processes that underlie reading.

Eye Movement and Computational Models

Baseline Linear Mixed Model: `log(GD) ~ LinePos + SentencePos + TextPos + LaunchSite + log(Freq) × inv(Length) + (1|Subj) + (1|word)`

	BL + ClozePred		Baseline Model + Computational Pred																						
	Baseline Model	Cloze Model	Previous Computational Models							AWD-LSTM Models: without finetunning								AWD-LSTM Models: with finetunning							
	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	FM9	FM10	FM11	FM12	FM13	FM14	FM15	FM16
length (inv)	-18.15	-18.56	-19.10	-17.89	-17.42	-19.10	-18.24	-16.93	-18.06	-19.41	-19.62	-19.27	-18.61	-19.66	-18.75	-18.26	-18.60	-19.28	-19.42	-19.22	-18.41	-19.47	-18.54	-18.14	-18.41
freq (log)	-10.83	-10.60	-1.93	-8.40	-11.49	-2.04	-2.74	-8.71	-2.32	-1.73	0.92	-1.34	-2.42	0.44	0.15	-1.61	0.16	-2.56	-0.22	-2.39	-3.31	-0.67	-0.99	-2.64	-0.92
inv length:Freq	16.98	17.17	15.68	17.09	16.30	15.54	14.86	16.44	14.85	14.71	14.71	14.69	13.96	14.53	13.89	14.04	13.85	14.38	14.73	14.29	13.50	14.54	13.83	13.53	13.78
CLOZE pred		-16.23																							
4-gram + cache			-21.02			-20.93	-21.26		-21.03		-16.08			-16.12	-16.29		-16.21		-13.84			-13.90	-13.89		-13.85
LSA (w = 9)				-2.17		0.68		-3.15	-0.41			-0.62		1.04		-1.65	-0.06			0.14		1.15		-1.00	0.01
FastText(w=50)					4.43		5.39	4.98	5.36				4.84		5.44	5.07	5.34				5.65		5.75	5.73	5.63
AWD-LSTM										-14.97	-6.83	-14.83	-15.09	-6.88	-6.87	-14.85	-6.86	-16.76	-6.11	-16.62	-17.13	-6.18	-6.42	-16.87	-6.41
Resudials + CLOZE model	-16.14	0.00	-9.47	-15.99	-16.55	-9.48	-9.87	-16.39	-9.87	-11.63	-8.26	-11.61	-12.02	-8.28	-8.66	-11.99	-8.66	-9.87	-8.06	-9.87	-10.22	-8.07	-8.41	-10.22	-8.41

t-values: 20, 15, 10, 5, 0, −5, −10, −15, −20



N-Gram
The ngram model (i.e. 4-word transition probability) with the addition of the local frequency generates a better model (with greater t-value and AIC). But after fitting the Cloze Predictability effect to the model residuals it is clear that it is not modeling the whole effect. It also makes the Frequency effect drop on its t-value.

Cloze-Pred
The effect of Cloze Predictability shows a negative relation with Gaze Duration. We would like to find a computer-based Predictability that mimics this effect

FineTuning effect
Comparing both models of AWD-LSTM (before and after performing the fine tuning with an small corpus of texts similar to the test corpus) it is possible to observe an enhanced estimation of Predictability. Not only on an increase in the (absolute value) of the t-value and the AIC, but also in a lower decrease for the Frequency effect.

AWD-LSTM + N-Gram
The AWD-LSTM model seems to capture similar aspects of word predictions than the Ngram model, but the RNN-based model has a behaviour more similar to the Cloze Predictability.

Semantics-based models
Both LSA and FastText model Natural Language by word embeddings, that represents words in a multidimensional semantic space. Modelling the Predictability as the semantic closeness to its context (the previous w words) makes little difference to the LMMs.

* Table shows Fixed effect's t-values. Absolute values greater than 2 corresponds to significant effects.

* Saccade LaunchSite, Relative positions in text, line and sentence were also used as fixed effects. no significant changes were observed for any moldes. Visit **LINK** for the full table.

Conclusions

- AWD-LSTM captures some of the Cloze Predictability effect.
- How much Cloze Predictability effect is captured by the AWD-LSTM depends on its training:
  - Training only with a general corpus (like Wikipedia) generates a model that captures much of the Frequency Effect.
  - When fine tuning the model with a more specific corpus (similar to those in the test corpus), the drop in the Frequency Effect is lower.
- The AWD-LSTM effect is partially overlapped with the previously observed N-Gram effect.
- The N-Gram model has the advantage of being completely transparent on its functioning. But it also tends to capture the whole Frequency Effect, despite being trained only on the specific corpus.

References:

Merity, S., Keskar, N. S., & Socher, R. **(2017)**. Regularizing and optimizing LSTM language models. arXiv preprint arXiv:1708.02182.

Merity, S., Keskar, N. S., & Socher, R. **(2018)**. An analysis of neural language modeling at multiple scales. arXiv preprint arXiv:1803.08240.

Kliegl, R., Nuthmann, A., & Engbert, R. **(2006)**. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. Journal of Experimental Psychology: General, 135(1), 12–35.

Bianchi B., Monzón G. B., Ferrer L., Slezak D. F., Shalom D. E. & Kamienkowsk J. E. **(2020)** Human and computer estimations of Predictability of words in written language. Sci Rep 10, 4396 (2020). https://doi.org/10.1038/s41598-020-61353-z