Natural Language Processing: Does it work like our brain?

Luis Miguel Garcia Marin
Fulda University of Applied Science
luis-miguel.garcia-marin@informatik.hs-fulda.de

Contents

ABSTRACT

In this paper we introduce ourselves into the Natural Language Processing (NLP) world, and we discuss if it is possible to understand the functioning of the new discovered technics in NLP with our understanding on the functioning of our brain with its habitual processing of natural language. And in order to do it, it is used a variety of inputs like brain image records with some recent models in Deep Learning.

1. INTRODUCTION

1.1 How everything started

At the ACL Conference in 1987, Don Walker, Jane Robinson and Karen Spark Jones were talking about when NLP research began [1]. Fred Thompson said that he began in 1954 and others, like Martin Key, started out too in the 50's. Work in the field has concentrated first on one problem, then on another, sometimes because solving problem X depends on solving problem Y, but sometimes just because problem Y seems more tractable than problem X. This means that the focus was not to solve NLP the same way that our brain does, because this was not tractable, so the problems were tried to solve in a lexico-logical way at first.

1.2 Nowadays

This triggered that NLP has been traditionally a complex problem to solve [2]. However, Recurrent Neural Networks and two significant advances - one in 2017 and another in 2019 - brought substantial improvements to NLP. In 2017, a new form of deep learning model called Transformer [3] made it possible to parallelize ML training more efficiently, resulting in vastly improved accuracies. In 2019, Google introduced Bidirectional Encoder Representations from Transformers (BERT) [4], which improves the above Transformer architecture. Straightaway BERT helped achieve state-of-the-art performance [5] on several NLP tasks such as reading comprehension, text extraction, sentiment analysis, etc. These two advancements meant that NLP could easily outdo average humans in many tasks and in some cases, even exceed the performance of subject matter experts.

1.3 Our question

But does recurrent neural network propagation of information works similar as our brain does? We propose to look at brain activity of subjects reading naturalistic text as a source of additional information for interpreting neural networks. And to do this we use brain imaging recordings of subjects reading complex natural text to interpret word and sequence embeddings from 4 recent NLP models - ELMo, USE, BERT and Transformer-XL, and study how their representations differ across layer depth, context length, and attention type.

2. STATE OF THE ART

Most work investigating language in the brain has been done in a controlled experiment setup where two conditions are contrasted [6]. These conditions typically vary in complexity (simple vs. complex sentences), in the presence or absence of a linguistic property (sentences vs. lists of words) or in the presence or absence of incongruities (e.g. semantic surprisal) [6]. A few researchers instead use naturalistic stimulus such as stories [7]. Some use predictive models of brain activity as a function of multi-dimensional features spaces describing the different properties of the stimulus [8].

A few previous works have used neural network representations as a source of feature spaces to model brain activity. Wehbe et al. [8] aligned the MEG brain activity we use here with a Recurrent Neural Network (RNN), trained on an online archive of Harry Potter Fan Fiction. The authors aligned brain activity with the context vector and the word embedding, allowing them to trace sentence comprehension at a word-by-word level. Jain and Huth [9] aligned layers from a Long Short-Term Memory (LSTM) model to fMRI recordings of subjects listening to stories to differentiate between the amount of context maintained by each brain region. Other approaches rely on computing surprisal or cognitive load metrics using neural networks to identify processing effort in the brain, instead of aligning entire representations [10].

There is also a little work that evaluates or improves NLP models through brain recordings. Søgaard [11] proposes to evaluate whether a word embedding contains cognition-relevant semantics by measuring how well they predict eye tracking data and fMRI recordings. Fyshe et al. [12] build a non-negative sparse embedding for individual words by constraining the embedding to also predict brain activity well and show that the new embeddings better align with behavioral measures of semantics.

3. METHODOLOGY

We investigate how the representations of all four networks (ELMO, BERT, USE and T-XL) change as we provide varying lengths of context. We compute the representations x in each available intermediate layer ([1, 2] for ELMo; [1, 12] for BERT; the layer is the output embedding for USE; and [1, 19] for T-XL). We compute x for word w by passing the most recent k words through the network.

3.1 fMRI and MEG data

We use fMRI and MEG data which complement each other very well. fMRI is sensitive to the change in oxygen level in the blood which is a consequence to neural activity, it has high spatial resolution (2-3 mm) and low temporal resolution (multiple seconds). MEG measures the change in the magnetic field outside the skull due to neural activity, it has low spatial resolution (multiple cm) and high temporal resolution (up to 1KHz). We use fMRI data published by Wehbe et al. [8]: 8 subjects read chapter 9 of Harry Potter and the Sorcerer's stone [13] which was presented one word at a time for a fixed duration of 0.5 seconds each, and

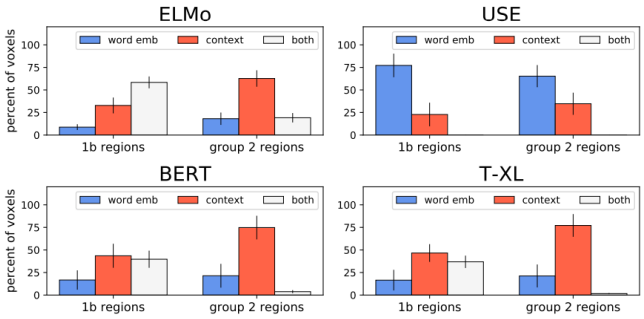

Figure 1. Amount of group 1b regions and group 2 regions predicted well by each network-derived representation: a 10-word representation corresponding to the 10 most recent words shown to the participant (Red) and a word-embedding corresponding to the last word (Blue). White indicates that both representations predict the specified amount of the regions well (about 0.7 threshold).

45 minutes of data were recorded. The fMRI sampling rate (TR) was 2 seconds. The same chapter was shown by Wehbe et al. [14] to 3 subjects in MEG with the same rate of 0.5 seconds per word. Details about the data and preprocessing can be found in the supplementary materials.

3.2 Encoding models

For each type of network-derived representation x, we estimate an encoding model that takes x as input and predicts the brain recording associated with reading the same k words that were used to derive x. We estimate a function f, such that f(x) = y, where y is the brain activity recorded with either MEG or fMRI. We follow previous work [8] and model f as a linear function, regularized by the ridge penalty. The model is trained via four-fold cross-validation and the regularization parameter is chosen via nested cross-validation.

3.3 Evaluation of predictions

We evaluate the predictions from each encoding model by using them in a classification task on held-out data, in the four-fold cross-validation setting. The classification task is to predict which of two sets of words was being read based on the respective feature representations of these words [8, 14]. This task is performed between sets of 20 consecutive TRs in fMRI (accounting for the slowness of the hemodynamic response), and sets of 20 randomly sampled words in MEG. The classification is repeated a large number of times and an average classification accuracy is obtained for each voxel in fMRI and for each sensor/timepoint in MEG. We refer to this accuracy of matching the predictions of an encoding model to the correct brain recordings as "prediction accuracy". The final fMRI results are reported on the MNI template, and we use pycortex to visualize them [15].

3.4 Proof of concept

Since MEG signals are faster than the rate of word presentation, they are more appropriate to study the compo-

nents of word embeddings than the slow fMRI signals that cannot be attributed to individual words. We know that a word embedding learned from a text corpus is likely to contain information related to the number of letters and part of speech of a word. We show in section 4 of the supplementary materials that the number of letters of a word and its ELMo embedding predict a shared portion of brain activity early on (starting 100ms after word onset) in the back of the MEG helmet, over the visual cortex. Indeed, this region and latency are when we expect the visual information related to a word to be processed (Sudre et al., 2012). Further, a word's part of speech and its ELMo embedding predict a shared portion of brain activity around 200ms after word onset in the left front of the MEG sensor. Indeed, we know from electrophysiology studies that part of speech violations incur a response around 200ms after word onset in the frontal lobe (Frank et al., 2015). We conclude from these experiments that the ELMo embedding contains information about the number of letters and the part of speech of a word. Since we knew this from the onset, this experiment serves as a proof of concept for using our approach to interpret information in network representations.

4. HEADINGS

First level headings are in Times 12 pt bold, centered with 1 line of space above the section head, and 1/2 space below it. For a section header immediately followed by a subsection header, the space should be merged.

4.1 Second Level Headings

Second level headings are in Times 10 pt bold, flush left, with 1 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

4.1.1 Third Level Headings

Third level headings are in Times 10 pt italic, flush left, with 1/2 line of space above the section head, and 1/2 space below it. The first letter of significant words is capitalized.

Using more than three levels of headings is strongly discouraged.

5. FLOATS AND EQUATIONS

5.1 Equations

Equations should be placed on separated lines and numbered. The number should be on the right side, in parentheses.

$$r = \sqrt[13]{3} \tag{1}$$

Always refer to equations like this: "Equation (1) is of particular interest because..."

5.2 Figures, Tables and Captions

All artwork must be centered, neat, clean and legible. Figures should be centered, neat, clean and completely legible. All lines should be thick and dark enough for purposes of reproduction. Artwork should not be hand-drawn. The proceedings will be distributed in electronic form only, therefore color figures are allowed. However, you may want to check that your figures are understandable even if they are printed in black-and-white.

String value	Numeric value
Hej SMC!	2023

Table 1. Table captions should be placed below the table, like this.

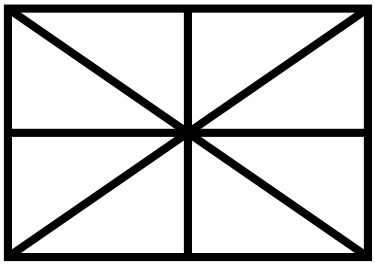

Figure 2. Figure captions should be placed below the figure, exactly like this.

Numbers and captions of figures and tables always appear below the figure/table. Leave 1 line space between the figure or table and the caption. Figure and tables are numbered consecutively. Captions should be Times 10pt. Place tables/figures in the text as close to the reference as possible, and preferably at the top of the page.

Always refer to tables and figures in the main text, for example: "see Fig. 2 and Table 1". Figures and tables may extend across both columns to a maximum width of 17.2cm.

Vectorial figures are preferred, e.g., eps. When using `Matlab`, export using either (encapsulated) Postscript or PDF format. In order to optimize readability, the font size of text within a figure should be no smaller than that of footnotes (8 pt font-size). If you use bitmap figures, make sure that the resolution is high enough for print quality.

5.3 Footnotes

You can indicate footnotes with a number in the text [1], but try to work the content into the main text.Use 8 pt font-size for footnotes. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a 0.5 pt horizontal rule.

6. CITATIONS

All bibliographical references should be listed at the end, inside a section named "REFERENCES". References must be numbered in order of appearance. You should avoid listing references that do not appear in the text.

Reference numbers in the text should appear within square brackets, such as in [16] or [16–18]. The reference format is the standard IEEE one. We highly recommend you use BibTeX to generate the reference list.

[1] This is a footnote example.

7. CONCLUSIONS

Please, submit full-length papers. Submission is fully electronic and automated through the Conference Web Submission System. Do not send papers directly by e-mail.

Acknowledgments

At the end of the Conclusions, acknowledgements to people, projects, funding agencies, etc. can be included after the second-level heading "Acknowledgments" (with no numbering).

8. REFERENCES

[1] K. S. Jones, "Natural language processing: a historical review," *Current issues in computational linguistics: in honour of Don Walker*, pp. 3–16, 1994.

[2] ——, "Natural language processing: an overview," *International encyclopedia of linguistics (ed W. Bright)*, pp. 3–59, 1992.

[3] Z. Wang, Y. Ma, Z. Liu, and J. Tang, "R-transformer: Recurrent neural network enhanced transformer," *arXiv preprint arXiv:1907.05572*, 2019.

[4] U. Kamath, K. L. Graham, and W. Emara, "Bidirectional encoder representations from transformers (bert)," *Transformers for Machine Learning; Chapman and Hall/CRC: New York, NY, USA*, pp. 43–70, 2022.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[6] A. D. Friederici, "The brain basis of language processing: from structure to function," *Physiological reviews*, vol. 91, no. 4, pp. 1357–1392, 2011.

[7] J. Brennan, Y. Nir, U. Hasson, R. Malach, D. J. Heeger, and L. Pylkkänen, "Syntactic structure building in the anterior temporal lobe during natural story listening," *Brain and language*, vol. 120, no. 2, pp. 163–173, 2012.

[8] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell, "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses," *PloS one*, vol. 9, no. 11, p. e112575, 2014.

[9] S. Jain and A. Huth, "Incorporating context into language encoding models for fmri," *Advances in neural information processing systems*, vol. 31, 2018.

[10] S. L. Frank, L. J. Otten, G. Galli, and G. Vigliocco, "The erp response to the amount of information conveyed by words in sentences," *Brain and language*, vol. 140, pp. 1–11, 2015.

[11] A. Søgaard, "Evaluating word embeddings with fmri and eye-tracking," in *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, 2016, pp. 116–121.

[12] A. Fyshe, P. P. Talukdar, B. Murphy, and T. M. Mitchell, "Interpretable semantic vectors from a joint model of brain-and text-based meaning," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2014. NIH Public Access, 2014, p. 489.

[13] J. Rowling, *HARRY POTTER and The Sorceres Stone*. Bloomsbury Publishing Plc, 2012.

[14] L. Wehbe, A. Vaswani, K. Knight, and T. Mitchell, "Aligning context-based statistical models of language with brain activity during reading," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 233–243.

[15] J. S. Gao, A. G. Huth, M. D. Lescroart, and J. L. Gallant, "Pycortex: an interactive surface visualizer for fmri," *Frontiers in neuroinformatics*, p. 23, 2015.

[16] A. Someone, B. Someone, and C. Someone, "The title of the conf. paper," in *Proc. Int. Conf. Sound and Music Computing*, Porto, 2009, pp. 213–218.

[17] X. Someone and Y. Someone, *The Title of the Book*. Springer-Verlag, 2010.

[18] A. Someone, B. Someone, and C. Someone, "The title of the journal paper," in *J. New Music Research*, 2008, pp. 111–222.