Advanced Course: Analyzing the Cognitive Plausibility of Deep Language Models

Lisa Beinborn and Willem Zuidema
l.beinborn@uva.nl, zuidema@uva.nl
Institute for Logic, Language and Computation
Universiteit van Amsterdam

Abstract

Computational models of language serve as an increasingly popular tool to examine research hypotheses related to language processing. In the last decade, distributional representations which interpret words, phrases, sentences, and even full stories as a high-dimensional vector in semantic space have become very popular. Whereas the most commonly known model *word2vec* only provides standardized representations for isolated words (Mikolov et al., 2013), more recent models interpret words in context. In this tutorial, we introduce the participants to the functionality of state-of-the-art contextualized language models and discuss methods for evaluating their cognitive plausibility. We explore a range of evaluation scenarios using cognitive data and provide practical examples. In a second step, we aim at opening up the blackbox and introduce methods for analyzing the hidden representations of a computational model.

1 Introduction

Computational models of language serve as an increasingly popular tool to examine research hypotheses related to language processing. Representing language in a computational way forces us to operationalize our underlying assumptions in a concrete, implementable and falsifiable manner.

In the last decade, distributional representations which interpret words, phrases, sentences, and even full stories as a high-dimensional vector in semantic space have become very popular. They are obtained by training language models on large corpora to optimally encode contextual information. Whereas the most commonly known model *word2vec* only provides standardized representations for isolated words (Mikolov et al., 2013), more recent models interpret words in context which can better account for the polysemy of words.

Computational language models are optimized for engineering objectives and are trained on terabytes of data. The architectures violate basic human processing constraints by using cognitively implausible tricks such as look-ahead and backpropagation. The meaningful combination of word representations into representations for phrases, sentences and stories still remains an unsolved problem (Repplinger et al., 2018). Sentence-based models have tremendously improved in the last years and can be tuned to perform surprisingly well on tasks such as question answering and textual entailment (Conneau and Kiela, 2018). However, diagnostic approaches have shown that these models "spectacularly fail on a wide range of linguistic phenomena" (Wang et al., 2018).

In this tutorial, we introduce the participants to the functionality of state-of-the-art contextualized language models and discuss methods for evaluating their cognitive plausibility. We explore a range of evaluation scenarios using cognitive data and provide practical examples. In a second step, we aim at opening up the blackbox and introduce methods for analyzing the hidden representations of a computational model. A shorter version of this tutorial has been accepted for the *41st Annual Meeting of the Cognitive Science Society (2019)*.

2 Scientific Background

The tutorial is based on recent developments in two interdisciplinary research areas: evaluating computational models with cognitive data, and analyzing the interpretability of neural language models.

2.1 Evaluating Computational Models With Cognitive and Neural Data

As computational language models are trained on human-generated text, their performance is inherently optimized to simulate human behavior. Although novel architectural solutions attract notable interest in the research community, the ultimate benchmark for a model is the ability to approximate human language processing abilities. Models are supposed to reach a gold standard of human annotation decisions (Resnik and Lin, 2010) and the difficulty of a task is often estimated by the inter-annotator agreement (Artstein and Poesio, 2008) or by error rates of human participants (Beinborn et al., 2014). While these product-oriented evaluations focus on a final outcome, procedural measures of response times (Monsalve et al., 2012) or eye movements

(Hollenstein et al., 2019; Barrett et al., 2018) are analyzed to provide deeper insights on sequential phenomena like attention or processing complexity. As neural network models are inspired by neuronal activities in the human brain, it is particularly interesting to analyze similarities and differences between distributed computational representations and low-level brain responses.

Electroencephalography (EEG) measures can be used to study specific semantic or syntactic phenomena (Hale et al., 2018; Fyshe et al., 2016; Sudre et al., 2012) and compare the processing complexity of computational models to brain responses, for example, with respect to the N400 and P600 effects (Frank et al., 2013). Signals with higher spatial resolution like magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) are often used for experiments that aim at distinguishing the representations for different linguistic stimuli. A computational model learns to identify differences in the brain signal and to discriminate between the responses for abstract and concrete words (Anderson et al., 2017), for different syntactic classes (Bingel et al., 2016; Li et al., 2018), for levels of syntactic complexity (Brennan et al., 2016), and many other linguistic categories. (Mitchell et al., 2008) have shown that it is not only possible to distinguish between semantic categories, but that a model can even learn to distinguish which word a participant is reading based on the representational relations between words. Many computational word representations (lexical, distributional, multimodal) have been tested on this task (Abnar et al., 2018; Anderson et al., 2017; Bulat et al., 2017; Xu et al., 2016). Recently, new datasets using longer linguistic stimuli such as sentences (Pereira et al., 2018) and even full stories (Jain and Huth, 2018; Dehghani et al., 2017; Brennan et al., 2016; Wehbe et al., 2014a) are emerging. In some experiments, it has been shown that contextualized representations obtained from recurrent neural networks (Wehbe et al., 2014b; Jain and Huth, 2018) seem to represent the continuous stimuli slightly better than models that represent sentences as a conglomerate of context-independent word representations (Dehghani et al., 2017; Pereira et al., 2018). However, these results are hard to generalize because they have been tested only on a single dataset and do not account for differences between models and evaluation metrics (Gauthier and Ivanova, 2018; Beinborn et al., 2019).

2.2 Interpretability of Neural Language Models

Deep language models are extremely successful on a wide range of tasks (e.g., Devlin et al., 2018; Peters et al., 2018). Unfortunately, the inner workings of a model are hard to trace for human users, because of the high dimensionality of the activation vectors and weight matrices, the large number of non-linear transformations, and the complex effects of 'gates', 'self-attention', 'query-key-value', 'transformer heads' and 'positional embeddings' in modern neural language models. For use in cognitive science, this lack of interpretability is a major concern: what good is it to have model predictions that correlate well with human data, if we don't understand the model?

Recently, new methods for interpreting and analyzing neural networks are being developed, such as diagnostic classifiers (e.g., Hupkes et al., 2018), representational stability analysis (e.g., Abnar et al., 2019), layer-wise relevance propagation (e.g., Arras et al., 2017) and contextual decomposition (Murdoch et al., 2018), and a variety of other analysis and visualisation techniques (Belinkov and Glass, 2019). These methods enable us to gain a better understanding of the internal ("hidden") representations of these models. In addition, comparative techniques like representational similarity analysis (Kriegeskorte et al., 2008) allow us to directly compare computational processing models with brain activation signals observed during human language processing by abstracting from the technical architecture.

3 Course Plan

In the following, we provide details on the structure of the course, the target audience, and the requirements.

3.1 Structure

The tentative structure of the tutorial is planned as follows:

Monday We start with an introduction to contextualized language models and provide an overview of different architectures and training objectives. Participants will be introduced to bidirectional long-short-term memory network (LSTM) architectures like *Elmo* (Peters et al., 2018), inference-based language models like *InferSent* (Conneau et al., 2017) and models inspired by a translational approach like the transformer model *Bert* (Devlin et al., 2018).
Main responsibility: Lisa Beinborn.

Tuesday We first discuss aspects of cognitive plausibility (e.g., multimodal grounding, compositionality, incremental processing). We then introduce experimental psycholinguistic data (eye-tracking, fmri, eeg) and focus on evaluating the processing similarity to humans using different evaluation paradigms.
Main responsibility: Lisa Beinborn.

Wednesday We organize a hands-on practice session that combines knowledge from the first two days. Participants get practical experience in evaluating the ability of a contextualized language model to predict psycholinguistic data. The session will be based on jupyter notebooks derived from our code projects *Cognitive NLP* and *Language–Brain Encoding with fMRI*.
Main responsibility: Nora Hollenstein supervised by Lisa Beinborn.

Thursday We discuss the challenges for interpreting neural language models and introduce methods for analyzing the hidden representations of a neural language

model such as diagnostic classifiers, representational similarity and stability analysis, contextual decomposition, layer-wise relevance propagation and related analysis and visualization methods.
Main responsibility: Willem Zuidema.

Friday We conclude with a hands-on practice session that combines the introduced methods. Participants learn how to analyze and visualize the similarity between computational representations of language and human representations of language using representational similarity analysis. The session will be based on jupyter notebooks derived from our code project on *Representational Stability Analysis*.
Main responsibility: Nora Hollenstein supervised by Willem Zuidema.

3.2 Target Audience and Requirements

Our interdisciplinary topic is situated in the Language and Computation (LaCo) field and addresses a range of backgrounds: i) computer science students who want to learn more about human language processing ii) cognitive science and linguistics students who want to learn more about computational modeling of language iii) computational linguists who want to learn more about analysis methods with cognitive data. Familiarity with the basic concepts of deep learning is beneficial for the course, but we will introduce the most important concepts and not go too deep into mathematical foundations.

4 Expertise

We have background knowledge in computational linguistics, machine learning, and cognitive science and can build on years of research on approaching natural language processing from a psycholinguistic point of view. Our recent joint work (Abnar et al., 2019) combines expertise in evaluating language models with fMRI data (Beinborn et al., 2019) with comparative methods for interpreting and analyzing neural networks (Hupkes et al., 2018). Nora Hollenstein complements our expertise with her knowledge on using eye tracking and EEG data for analyzing the cognitive plausibility of NLP models.

Lisa Beinborn Lisa Beinborn's interdisciplinary research profile has been shaped by studies in renowned international institutes in Amsterdam, Darmstadt, Saarbrücken, Barcelona, Bolzano, Hamburg, Frankfurt, and Trento. Her research focuses on human-centered natural language processing and she has worked on modeling compositionality, multilingual and multimodal modeling, phonetic modeling, and adjusting language representations for a target group (Beinborn and Choenni, 2019; Beinborn et al., 2018; Repplinger et al., 2018; Beinborn et al., 2016, 2014, 2013). She is now working in the *Language in Interaction* project dedicated to bridging neuroscientific, psycholinguistic and computational models of language and analyzes the cognitive plausibility of language models by using human processing data (Beinborn et al., 2019; Abnar et al., 2019). She has been teaching lectures on *Deep Learning for NLP*, *Knowledge-based Systems*, and *Artificial Intelligence* and has supervised multiple interdisciplinary student projects. For more details, check *https://beinborn.eu/*.

Willem Zuidema Willem Zuidema is an associate professor in cognitive science and computational linguistics at the Institute for Logic, Language and Computation of the University of Amsterdam. He conducts research on analyzing and interpreting cognitive and computational models of language, coordinates the *Cognition, Language and Computation lab*, and teaches in the interdisciplinary masters programs *Brain & Cognitive Sciences*, *Artificial Intelligence*, and *Logic*. He has a long term interest in the neural basis of language, and thanks to that cognitive interest, was an early contributor to deep learning in NLP, with work on neural parsing published as early as 2011 (Borenszstajn and Zuidema, 2011), and pioneering contributions on tree-shaped neural networks, including the TreeLSTM (Le and Zuidema, 2015). In 2016, his group introduced Diagnostic Classification (Veldhoen et al., 2016; Hupkes et al., 2018), a key interpretability approach in deep learning for NLP. For more details, check *http://illc.uva.nl/LaCo/clclab/*.

Nora Hollenstein Nora Hollenstein is a PhD candidate in Natural Language Processing at ETH Zurich. The focus of her work lies in enhancing and evaluating NLP applications with cognitive processing signals such as eye-tracking and brain activity recordings (Hollenstein et al., 2019; Barrett et al., 2018). After an MSc in Artificial Intelligence from the University of Edinburgh, she worked at IBM for a few years on various IBM Watson projects before starting her PhD. She can build on teaching experience in *Ontology Learning* at the Institute for Computational Linguistics, University of Zurich, and will teach *Language Technology and Human Cognition* in 2020. For more information check *https://norahollenstein.github.io/*.

References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 57–66. Association for Computational Linguistics.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. In *Proceedings of the ACL-Workshop on Analyzing and Interpreting Neural Networks for NLP*, page to appear.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. *EMNLP 2017*, page 159.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312.

Lisa Beinborn, Samira Abnar, and Rochelle Choenni. 2019. Robust evaluation of language-brain encoding experiments. *International Journal for Computational Linguistics and Applications*, page to appear.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339. Association for Computational Linguistics.

Lisa Beinborn and Rochelle Choenni. 2019. Semantic drift in multilingual representations.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association of Computational Linguistics*, 2(1):517–529.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2016. Predicting the spelling difficulty of words for language learners. In *Proceedings of the NAACL-Workshop on Innovative Use of NLP for Building Educational Applications*, pages 73–83.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Joachim Bingel, Maria Barrett, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fMRI – with an application to pos induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 747–755.

Gideon Borensztajn and Willem Zuidema. 2011. Episodic grammar : a computational model of the interaction between episodic and semantic memory in language processing. In *Proceedings of the Conference of the Cognitive Science Society*, pages 507–512.

Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1091. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the 2018 Conference on Language Resources and Evaluation*. European Language Resource Association.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Morteza Dehghani, Reihane Boghrati, Kingson Man, Joe Hoover, Sarah I Gimbel, Ashish Vaswani, Jason D Zevin, Mary Helen Immordino-Yang, Andrew S Gordon, Antonio Damasio, et al. 2017. Decoding the neural representation of story meanings across languages. *Human brain mapping*, 38(12):6096–6106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2013. Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 878–883.

Alona Fyshe, Gustavo Sudre, Leila Wehbe, Nicole Rafidi, and Tom M Mitchell. 2016. The semantics of adjective noun phrases in the human brain. *bioRxiv*.

Jon Gauthier and Anna Ivanova. 2018. Does the brain represent words? An evaluation of brain decoding studies of language understanding. *arXiv:1806.00591*.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1 (Long Papers)*, pages 2727–2736. Association for Computational Linguistics.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing NLP with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fMRI. *bioRxiv*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Phong Le and Willem Zuidema. 2015. Compositional Distributional Semantics with Long Short Term Memory. In *Proceedings *SEM*.

Jixing Li, Murielle Fabre, Wen-Ming Luh, and John Hale. 2018. The role of syntax during pronoun resolution: Evidence from fMRI. In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 56–64. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408. Association for Computational Linguistics.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *ICLR*.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North-American Association for Computational Linguistics*.

Michael Repplinger, Lisa Beinborn, and Willem Zuidema. 2018. Vector-space models of words and sentences. *Nieuw Archief voor Wiskunde: Machine Learning*, 5.

Philip Resnik and Jimmy Lin. 2010. Evaluation of NLP systems. *The handbook of computational linguistics and natural language processing*, 57:271–295.

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463.

Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. 2016. Diagnostic classifiers: revealing how neural networks process hierarchical structure. In *Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (at NIPS)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP"*, pages 353–355. Association for Computational Linguistics.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, 9(11):e112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014b. 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)*. Association for Computational Linguistics.

Haoyan Xu, Brian Murphy, and Alona Fyshe. 2016. Brainbench: A brain-image test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2017–2021.