Analyzing and Interpreting Neural Networks for NLP: A Report on the First BlackboxNLP Workshop

Afra Alishahi
Tilburg University

Grzegorz Chrupała
Tilburg University

Tal Linzen
Johns Hopkins University

1 Introduction

Neural networks have rapidly become a central component in language and speech understanding systems in the last few years. The improvements in accuracy and performance brought by the introduction of neural networks has typically come at the cost of our understanding of the system: what are the representations and computations that the networks learn?

In October 2018, we organized a workshop called BlackboxNLP[1] as part of the Empirical Methods in Natural Language Processing (EMNLP 2018) conference to bring together researchers who are attempting to peek inside the neural network black box, taking inspiration from machine learning, psychology, linguistics and neuroscience. The workshop generated a lot of interest withing the NLP community: we received a total of 75 submissions, and more than 600 EMNLP attendees signed up for the workshop. The topics presented and discussed in the workshop were diverse, representing the variety of methodologies, resources and techniques currently being used for analyzing the inner-workings and knowledge acquired by neural networks. The following themes, however, were dominant:

- Systematic manipulation of input to neural networks and investigating the impact on their performance, often through developing annotated and specialized datasets;
- Testing whether interpretable knowledge can be decoded from intermediate representations acquired by neural networks, often through passing them to *diagnostic classifiers* or other downstream tasks;
- Proposing modifications to neural network architectures to make their knowledge state or generated output more explainable;
- Examining the performance of the network on simplified or formal languages.

In this report, will briefly review some representative studies in each category, and provide some thoughts for future research.

2 Input manipulation

A number of studies focus on the impact of input on the performance of the model, and use this as a diagnostic method for identifying the important characteristics that affect the model's decisions. A straightforward approach applies simple preprocessing techniques to the training data of a model, and examines how the performance of the model changes.

For example, Søgaard, de Lhoneux, and Augenstein (2018) focus on the impact of punctuation on syntactic parsing, and show that neural dependency parsers are more sensitive to punctuation-free input than the previous generation of parsers. In a similar vein, Camacho-Collados and Pilehvar (2018) investigate the impact of various text preprocessing techniques, such as tokenizing, lemmatizing and chunking, on the performance of neural models of text classification and sentiment analysis, and suggest preprocessing guidelines for training word embeddings.

[1] `https://blackboxnlp.github.io/2018/`

2.1 Datasets

Alternatively, some develop diagnostic resources containing carefully constructed input examples which vary in their degree of (linguistic) complexity or in providing different types of challenges to a model. Diverse NLI Collection (DNC, (Poliak et al., 2018)) is such a case, which presents a large collection of datasets of natural language inference examples that represent a range of semantic phenomena. Another example is the General Language Understanding Evaluation (GLUE, (Wang et al., 2018)), a benchmark for a number of different natural language understanding tasks with training and test data from different domains and an evaluation platform. GLUE also provides a diagnostic dataset including manually-annotated examples that cover a range of linguistic phenomena.

Burns, Nematzadeh, Grant, Gopnik, and Griffiths (2018) present a study that uses such a resource, a dataset based on three reasoning tasks that are inspired by theory-of-mind experiments on children (Nematzadeh, Burns, Grant, Gopnik, & Griffiths, 2018). They use this dataset to test a memory network model on a question answering task and evaluate its reasoning capacity and generalizability. Similarly, Sommerauer and Fokkens (2018) use an existing dataset of semantic properties and augment it with negative examples for each property, which they use to examine which semantic properties are incorporated in word embeddings. They train supervised neural classifiers to identify each semantic property in a word embedding, and compare their performance to another model that uses vector cosine similarity. Their preliminary results suggest that interaction-based properties are captured better by neural models, whereas perceptual properties are not.

Overall, these studies vary in their objective and the type and scale of the resource they create and/or use, but there is no doubt that identifying various linguistic phenomena and challenges and providing carefully annotated test cases for each is the first step towards standardizing the evaluation of the new generation of analysis techniques, a point we will come back to in the last section.

3 Analysis techniques

Many of the papers submitted to the workshop proposed new methods for analysis of neural representations or applied existing methods to particular models and data. Below we review the main families of analytical methods represented at the workshop.

3.1 Diagnostic auxiliary models

A commonly applied technique consists in using neural representations as features, and training a predictive model on top to predict information of interest such as particular linguistic features. If the model is able to predict this information with high accuracy, the inference is that the neural representation encodes it. It is variously known as

- auxiliary task (Adi, Kermany, Belinkov, Lavi, & Goldberg, 2017),
- diagnostic classifier (Hupkes, Veldhoen, & Zuidema, 2018),
- probing (Conneau, Kruszewski, Lample, Barrault, & Baroni, 2018) or,
- decoding (Alishahi, Barking, & Chrupała, 2017)

The submission of Giulianelli, Harding, Mohnert, Hupkes, and Zuidema (2018) was especially illustrative of the focus of the BlackboxNLP workshop, by looking *under the hood* of neural language models, and it received the best paper award. The authors apply the *diagnostic classifier* approach to answer the question of how these models keep track of subject-verb agreement in English. They decode number information from internal states of a language model and analyze how, when, and where this information is represented. Additionally they use the mistakes of the diagnostic classifier to nudge the hidden states of the network in the direction which would make the classifier more likely to be correct and show that this intervention also improves the results on the language modeling task.

Zhang and Bowman (2018b, 2018a) use the diagnostic classifier framework to investigate which primary training tasks are best suited for transfer learning of syntactic information. They decode

POS and CCG tags from hidden layers of recurrent models trained on language modeling, translation, auto-encoding, and on the skip-thought objective (Kiros et al., 2015). They find that the language modeling task learns representations which overall encode syntax best; another interesting finding is that randomly initialized LSTM states enable quite a high accuracy of diagnostic POS and CCG tag classification: they ascribe it to the fact that the hidden states of these random networks encode the identity of words around the current position. This result suggests that successful diagnostic classification should be interpreted with caution and not automatically taken as proof that a particular neural model has *learned* to encode syntax: it could be that the diagnostic classifier itself is doing the learning.

Spinks and Moens (2018) propose a method for evaluating text encoders which is an interesting variation on the basic idea of diagnostic auxiliary model. It is different from the canonical diagnostic classifier setup in that the diagnostic model in this case is not a simple classifier but rather a GAN-based image generation model which attempts to produce an image given a representation of its caption. The intrinsic quality of those images as well as their alignment with the text is then assessed to determine how well the given text encoding represents visually important information. The authors apply the technique to a dataset of captioned medical X-rays and evaluate several simple text encoding approaches.

3.1.1 Correlation Analysis

Saphra and Lopez (2018a, 2018b) introduce an original twist on the idea of an auxiliary model. Instead of using a diagnostic classifier to probe learned neural language model representations, they train parallel recurrent models to do POS, semantic and topic tagging. They then measure the correlation between the activations of the language model and those of the tagger, using the technique of Singular Value Canonical Correlation Analysis (SVCCA) (Raghu, Gilmer, Yosinski, & Sohl-Dickstein, 2017). They use this method to examine how representations evolve over time and find that POS tag information is acquired before semantic or topics information.

3.2 Nearest Neighbors and Kernels

A natural approach to analyzing neural representations is based on retrieving neighboring exemplars in the induced vector space. The application of kernel methods is a related approach, in that kernels are functions which encode some notion of pairwise similarity between exemplars.

Wallace, Feng, and Boyd-Graber (2018) combine two classes of techniques commonly used in neural model interpretation: nearest neighbors and sensitivity analysis. The authors apply the Deep k-Nearest Neighbor model (Papernot & McDaniel, 2018) to perform neural text classification in an interpretable fashion. The Deep k-Nearest Neighbor model classifies a test input based on the labels of the training examples which are nearest to it in representation space, as an alternative to the standard softmax method. Wallace et al. derive a metric of model uncertainty, *conformity* based on this exemplar-based classification, and compute input feature importance as a change in conformity when a feature is removed (features in this case are words). They show that conformity as a metric of uncertainty leads to feature importance assignments which better agree with human perception compared to using the confidence of a softmax-based classifier. They also demonstrate that the method confirms some known artifact in the SNLI dataset.

Madhyastha, Wang, and Specia (2018a, 2018b) investigate the hypothesis that image captioning systems work by exploiting similarity in multimodal feature space, i.e. essentially they retrieve captions of training images close to a given test image in representation space, and assemble a caption for the given image by re-using elements of the captions of the retrieved images. The authors support their conjecture by holding the caption generation portion of the captioning system constant while varying image representations. It turns out that performance is in many cases comparable across substantially different image representations. This may be interpreted as evidence against the fact that the specifics of these representations play a major role in generating the captions, as long as the representations encode a reasonable notion of image similarity.

Croce, Rossini, and Basili (2018) focus on explainability of neural classifiers by providing the user with examples which motivate the decision. They adapt the technique of *layerwise relevance propagation* (Bach et al., 2015) to Kernel-based Deep Architectures (Croce, Filice, Castellucci, & Basili, 2017) in order to retrieve such examples. In essence, in this architecture a vectorial input for a given structured symbolic input (such as a parse tree) is build based on kernel evaluations between the input and a subset of training examples known as landmarks. With layerwise relevance propagation, the network decision can be traced back to the landmarks which had most influence on it. Based on this technique the authors build variants of explanatory models for question classification and semantic role argument classification. The models provide human-readable justification for their decisions which were submitted for evaluation to human annotators. Additionally, qualitative evaluation shows that the explanations were able to capture semantic and syntactic relations among inputs and landmarks.

3.3 Saliency and attention

Another well-represented approach to neural network analysis was the visualization of feature saliency and attention maps. Feature saliency refers to the importance of input features in influencing the output of a neural model. For models dealing with written language, these would typically be words, n-grams or similar patterns. Attention mechanisms provide a model-internal way to extract similar information. Attention visualization may be especially relevant for translation, where links between input and output patterns can be highlighted.

These type of approaches are often grouped into two loose families:

- The patterns can be used to illuminate a particular decision made by a model; this has been called *prediction interpretability* (Jacovi, Sar Shalom, & Goldberg, 2018) or *local introspection* (Krug & Stober, 2018).
- Typical or canonical patterns which tend to lead to a particular decision can be extracted and used to provide an explanation of what the model has learned in general; this has been called *model interpretability* or *global introspection*.

3.3.1 Saliency in Convolutional Networks

The submission of Jacovi et al. (2018) is a illustrative example of the use of feature saliency mapping. The work deals with convolutional neural models for text classification and refines common assumptions about how CNNs work on discrete sequences. Specifically, they show that max-pooling induces a thresholding behavior such that pattern activations below a certain value do not affect prediction, that a single filter often acts as a detector for several semantic classes of n-grams, and that filters can also detect negative patterns. They show application of their findings to model interpretability and prediction interpretability.

3.3.2 Saliency in Recurrent Networks

Krug and Stober (2018) tackle the audio domain which has some inherent difficulties for interpretability: while humans find it easy to understand visual input patterns such as portions of images for vision or n-grams for written language, the input used in speech recognition such as waveforms or even spectrograms may be relatively hard to interpret even for domain experts. In this study they analyze the activation patterns of a fully convolutional model which transcribe speech represented as spectrograms into character sequences. They apply techniques previously found useful in interpreting image classification. Firstly, for global introspection they apply regularized activation maximization which finds a synthetic stimulus which maximally activates a particular class (in this case a letter). They also introduce a technique which averages and normalizes all spectrogram frames predicted as the same letter. Secondly, for local introspection they experiment with layerwise relevance propagation (Montavon, Lapuschkin, Binder, Samek, & Müller, 2017) and sensitivity analysis (Gevrey, Dimopoulos, & Lek, 2003). From qualitative analysis the authors conclude that local introspection rarely results in interpretable patterns. For global introspection, the aligned frame averaging technique yields somewhat interpretable letter-specific patterns.

Gupta and Schütze (2018) discover salient patterns and use them for interpreting the representations and decisions of a custom neural architecture named Connectionist Bidirectional RNN (Vu, Adel, Gupta, et al., 2016) applied to relation classification. They introduce two visualization techniques: firstly, they plot the network's cumulative prediction score as a function of the position in the sentence such that steep changes in the score correspond to salient words. Secondly, they aggregate the patterns corresponding to individual examples of a single given class to visualize the set of common patterns which lead the network to predict this class. They carry out a qualitative analysis of the discovered salient n-grams.

Verwimp, Van Hamme, Renkens, and Wambacq (2018a, 2018b) track how long information is retained in the states of an LSTM language model by computing and decomposing the gradient matrix of the state with respect to input word embeddings, with a certain delay. The authors observe retention of input for up to 30 steps, with increased selectivity for longer delays; they also see some word classes, such as pronouns, retained for longer than others.

Hiebert, Peterson, Fyshe, and Mehta (2018) also tackle the LSTM language model and discover interpretable input patterns by applying HDBSCAN clustering (Campello, Moulavi, & Sander, 2013) to the model's hidden states. The technique reveals interpretable clusters which display both character-level and grammatical structure patterns. Inspired by the patterns discovered, the authors propose a method for extracting word-embeddings from a character-level model, showcasing a practical application of model analysis.

Poerner, Roth, and Schütze (2018) aim to assign interpretable synthetic rather than corpus-attested input patterns to individual dimensions (neurons) in neural representations. The challenge with textual data is the discrete nature of the input: they use the Gumbel softmax trick (Jang, Gu, & Poole, 2017) to generate word sequences which maximize activations for particular neurons. They apply this method to the Imaginet architecture of Chrupała, Kádár, and Alishahi (2015) and confirm one of the findings in Kádár, Chrupała, and Alishahi (2017): that the language model part of the Imaginet architecture is more sensitive to function words than the visual part, which tends to ignore them. They also carry out a separate quantitative evaluation of the synthetic patterns vs corpus-attested in terms of achieved maximum activation.

3.3.3 Attention mechanisms

Two submissions (Raganato & Tiedemann, 2018; Mareček & Rosa, 2018) analyze the internal workings of the encoder component of the Transformer model (Vaswani et al., 2017). The Transformer is not autoregressive; instead it uses an encoding of absolute word position as well as a number of (self-)attention heads to encode structural information. This model has recently become a state-of-the-art neural architecture for translation. Given its novelty, not much is known about what exactly enables its superior performance. Raganato and Tiedemann (2018) apply two main techniques: deriving dependency relations from the attention weights, and diagnostic classification of syntactic and semantic tags. The highlights of their findings are: firstly, a significant amount of syntactic information is encoded at each layer, and secondly, semantic information is easier to decode from higher layers. Additionally they demonstrate the transferability of the encoder weights from low to high resource language pairs.

4 Explainable architectures

Another trend observed in the workshop was to use neural network architectures that are easier to analyze and understand. Such architectures are often more explicit in the types of linguistic representations they use; for example, they use latent variables for incorporating linguistic information such as lexical categories or (sub-)tree structures, or are specifically trained to learn structured output. Alternatively, specialized attention mechanisms are employed which provide more insight into the linguistic features that models rely on for performing their final task.

4.1 Latent variables for capturing linguistic representations

In his invited talk, Graham Neubig presented a series of works on developing more explainable architectures that can be trained in an unsupervised or semi-supervised fashion, and allow for the emergence of interpretable linguistic structure. Specifically, Zhou and Neubig (2017) propose an architecture called *multi-space variational encoder-decoder* which uses discrete (e.g. POS, tense, person) and continuous (e.g. lemma) latent variables for transforming labeled sequences in a supervised or semi-supervised fashion. In a similar approach, Yin, Zhou, He, and Neubig (2018) propose a variational auto-encoding model for semantic parsing which learns meaning representations in unlabeled data as tree-structured latent variables. In both these cases the inclusion of the latent variables improves the performance in the down-stream tasks, but equally importantly, such structured variables make it easier to analyze the linguistic representations that models extract from labeled and unlabeled data.

Peters, Niculae, and Martins (2018) also suggest sparse and structured latent computation as a mechanism for improving interpretability in neural networks. Their goal is to use various techniques for transforming the dense inner-representations of neural networks into sparse representations, which show what part of input the model bases its decision on. Various techniques are suggested for this purpose, for example using regularization techniques in order to incorporate prior assumptions (Niculae & Blondel, 2017), or learning latent structure predictors such as parses or aligners (Niculae, Martins, Blondel, & Cardie, 2018). Trifonov, Ganea, Potapenko, and Hofmann (2018) apply similar techniques to sentence embeddings in order to make them more sparse and therefore easier to interpret.

For tasks such as sentiment analysis where negations and contrastive phrases change the polarity of the sentence, it is more informative to interpret the task as a series of incremental inferences. Tutek and Šnajder (2018) propose a special attention mechanism which makes such interpretation possible. Their iterative recursive attention model relies on a non-linear transformation of the representations from previous steps to build a recursive representation of the input sequence in an incremental manner.

4.2 Generating interpretable output

As an attempt to improve the *global interpretability* of a trained model and to provide explicit explanations for its output, Sushil, Suster, and Daelemans (2018) propose a technique for generating a set of if-then-else rules that explain the prediction of a class label based on the most important features extracted from the input. They do this by first analyzing the feature weights and highlighting the most informative ones in a trained model, then mapping them to a set of discrete features that represent either a positive, a negative, or no correlation with a class label. They induce a set of rules from this reduced feature space that best explains the model's predictions.

Stahlberg, Saunders, and Byrne (2018) generate interpretable output by training a neural machine translation model to incorporate explicit word alignment information in the representation of the target sentence: the model is trained to generate a target sentence in parallel with its alignment with the source sentence (as a linear sequence of operations).

Harbecke, Schwarzenberg, and Alt (2018) borrow explanation techniques from Computer Vision called PatternNet and PatternAttribution (Kindermans et al., 2017), which estimate the signal and the noise in the input to a model. They apply these techniques to a CNN trained for text classification, and retrieve neuron-wise signal contributions in the input vector space. They align the aggregated contribution scores with the input text and show that the model especially pays attention to bigrams, and ignores stop words.

5 Linguistics and formal language theory

5.1 What do neural networks learn about language?

Neural network representations cannot be read off from network's hidden state. A number of papers in the workshop followed an alternative approach, in which the representations used by a network are inferred from the network's behavior on examples that illustrate an interpretable linguistic phenomenon. This approach can leverage a number of experimental paradigms developed to study human linguistic behavior; given that human internal representations are arguably even harder to read off than those of neural networks, such paradigms are quite developed.

5.1.1 Agreement as a probe into structural representations

Ravfogel, Goldberg, and Tyers (2018) trained RNNs to predict the agreement features of a verb in Basque; perfect accuracy on this task requires identifying the subject of the verb, which in turn requires sophisticated syntactic representations (Linzen, Dupoux, & Goldberg, 2016). In Basque, which differs from English in a large number of properties, accuracy was substantially lower than in earlier studies on English. While the difference may be due to a number of factors, an intriguing possibility is that the inductive biases of popular neural architectures are better suited to English than to other languages.

Dhar and Bisazza (2018) test whether syntactic representations in an LM can transfer across languages. They trained an RNN LM on a corpus formed by concatenating a large French corpus and a small Italian corpus, and evaluated the model on the Italian agreement dependencies collected by Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018). They found that adding the large French corpus provides a modest improvement over training on the smaller Italian corpus alone, but the bilingual model is less effective at resolving agreement dependencies than a model trained on a monolingual Italian corpus of a matched size, suggesting that cross-linguistic transfer is limited even across related languages.

5.1.2 Structural dependencies beyond agreement

A number of studies investigated whether RNN LMs show awareness of linguistic phenomena beyond subject-verb agreement; studying these more complex phenomena often requires more experimental ingenuity than subject-verb agreement does. Jumelet and Hupkes (2018) investigate whether an RNN LM shows sensitivity to the restrictions on the distribution of negative polarity items (NPIs). NPIs (such as *any*) can only occur in a limited set of licensing environments (Giannakidou, 2011); for example, *I did not eat any cookies* is grammatical, but **I ate any cookies* is not. They found that the model was able to detect the environments in which NPIs are allowed to appear; however, this ability degraded with the distance between the NPI and the word that signals the licensing environment (Linzen et al., 2016 reported an analogous distance-dependent approximation of relative clauses).

Wilcox, Levy, Morita, and Futrell (2018) study the representation of filler-gap dependencies in RNN LMs. To illustrate what those dependencies are, consider the verb *devoured*, which typically takes a direct object, but is missing in the sentence *I know what the lion devoured* ____ *at sunrise*; intuitively, it can be thought of as having been displaced to the left of *the lion* and replaced with the question word *what*. Wilcox et al. (2018) found that two RNN LMs showed significant awareness of such gaps, as well as of some (but not all) of the restrictions on their distribution ("island constraints", Ross, 1967). Rønning, Hardt, and Søgaard (2018) analyzed bidirectional RNNs trained to resolve the antecedent of a "sluiced" question word, as in *If this is not practical, explain why* ____; here the phrase *this is not practical* need to be interpreted again following the question word, even though it does not appear again. They report both behavioral analyses on challenging cases and analysis of the network's activation; for example, they show that the distance between the question word (*why*) and the network's activation is minimal at the edge of the antecedent, and conclude that the network reactivates the question word when it encounters the antecedent (recall that the network processes the sentence both from right to left and from left to right).

5.1.3 Other phenomena

Bacon and Regier (2018) propose a new diagnostic task dataset for measuring the syntactic sophistication of vector sentence representations generated by neural networks, across four languages. The task measures whereby the tense of the main clause verb be decoded using a linear classifier (following Conneau et al., 2018; see also section 3.1). Extending previous work, they sample sentences that have multiple verbs that conflict in their tense; for example, the main verb in *we know who won* is in the present tense (*know*), but the past-tense embedded clause verb *won* serves as a distractor. This hypothesis-driven approach, in which test sentences are selected based on their linguistic complexity rather than sampled based on their corpus frequency, characterizes much of the linguistically-informed neural network analysis work.

Wei, Pham, O'Connor, and Dillon (2018) propose to evaluate the grammaticality of the output of a neural machine translation model using a precision HPSG grammar; if a sentence can be parsed by this grammar, it is necessarily grammatical. The majority of the sentences in a French-to-English translation dataset were parseable. A preliminary error analysis of the unparseable sentences shows that about 20% of them contained agreement errors; future analysis of other types of errors can provide insight into the limitations of the language model learned by the decoder.

By contrast with the preceding studies, which were concerned with the syntactic awareness of sentence-level models in which words were the basic unit, Kementchedjhieva and Lopez (2018) investigate the morphosyntactic generalizations acquired by character-based language models, which often do not have any explicit representations of words. Qualitatively, they show that samples from a character-based model trained on English text include real words, and pseudowords that made up of real English morphemes (such as *indicatements* or *breaked*), but also nonwords that do not conform to English phonotactics (*ouctromor*). Perhaps surprisingly, they discovered a highly interpretable recurrent unit, which recognizes the ends of words and sub-word units. They use diagnostic tasks to determine whether the model's internal representations contain sufficient information for morphological segmentation and part-of-speech tagging. Finally, they argue that the model shows some sensitivity to morpheme's selectional restrictions (e.g., *-ment* is more likely to attach to a verb than to other parts of speech). At the same time, there appears to be a limit to the amount of information that can be gleaned by a character RNN; Vania and Lopez (2018) show that explicit morphological case annotation improve the performance of a dependency parser compared to a model trained only on characters.

5.2 Synthetic languages

Synthetic training data, which can be carefully controlled, can illuminate the learning capabilities of an architecture (or, more precisely, the combination of an architecture and a learning algorithm). Networks trained on synthetic data can be easier to analyze, and more extensive experiments can be run without large computational power—indeed, in the early days of neural networks, those were the only experiments that were feasible (Elman, 1990; Hochreiter & Schmidhuber, 1997). If an architecture is unable to learn a particular phenomenon given a large amount of synthetic data, we may be justified in being pessimistic about its ability to learn it with natural language. Of course, the converse is not always true: the inductive bias of the architecture may be sufficient for learning from the distribution of constructions in a particular synthetic data set, but not from the distribution of natural language, where critical cases that disambiguate two hypotheses about the language may be rare.

In his invited talk, Yoav Goldberg discussed a study in this vein. It is a well-known theoretical result that RNNs are Turing complete (Siegelmann & Sontag, 1995); yet this result relies on the assumption that the system can process infinite precision numbers, whereas in practice, of course, all neural networks are implemented in finite-precision computers. Weiss, Goldberg, and Yahav (2018) show that this assumption has important consequences. In particular, fine-grained differences in the details of the recurrence equations between GRUs and LSTMs lead to large differences in their ability to count;[2] such an ability to count is essential for recognizing formal languages such as $a^n b^n$, and may be relevant for the processing of recursive embedding in natural language. This result, obtained using a synthetic language, is particularly striking because GRUs and LSTMs are not typically considered to obtain substantially different performance in applications.

Two groups trained RNNs with different recurrent units to recognize Dyck languages, which are simple context-free language that requires matching opening and closing brackets (Sennhauser & Berwick, 2018; Skachkova, Trost, & Klakow, 2018). The studies find that the RNNs are able to deal with shorter strings, but generalize poorly to strings that are longer than those on which they were trained. Sennhauser and Berwick (2018) conclude from their results that LSTMs' success in practical applications relies on approximations of natural language grammar, which may work well in most practical cases, but do not conform to the theoretical analysis of the grammar of the language (which is at least context-free for most natural languages).

Going beyond the classic test of string recognition in a formal language, a number of papers instead trained networks to assign an interpretation to a string in the language. Paperno (2018) reports that LSTMs are able to learn compositional interpretation rules, but only if their order of application follows the sequential order of the sentence, and they require extensive training to learn to do so. Likewise, sequence-to-sequence LSTMs trained to translate commands to sequences of actions on the SCAN dataset (Lake & Baroni, 2018) do not learn systematic compositional rules (Loula, Baroni, & Lake, 2018); Bastings, Baroni, Weston, Cho, and Kiela (2018) propose a modification to the SCAN dataset that makes it even more challenging and exposes poor generalization performance in sequence-to-sequence networks. Taken together, these synthetic language studies suggest that RNNs, while often very effective for cases that are similar to those they were trained on, do not generalize in the way that a human would.

5.3 Injecting linguistics into neural networks

Classic language technologies are based on symbolic representations that are readily interpretable to a human with basic training in linguistics; such symbolic representations include constituency

[2]Specifically, a GRU cell's memory is constrained to be between −1 and 1, whereas an LSTM is not constrained in that way

parses, logical formulas or meaning graphs. A number of studies explored ways to reincorporate computational elements from linguistics or formal language theory into neural network architectures; in addition to their potential to improve interpretability, such hybrid neural networks may have inductive biases that are more appropriate to language learning than are the biases of sequential RNNs (see section 5.2).

Hao et al. (2018) present a systematic analysis of the behavior of one such architecture: stack-augmented neural networks (Grefenstette, Hermann, Suleyman, & Blunsom, 2015), an architecture motivated by the observation that adding a stack to a finite automaton enables it to recognize context-free languages (Chomsky, 1962). An attractive feature of this architecture is that its utilization of the stack can be visualized in a straightforward and interpretable way. Along with interpretable visualizations for a number of synthetic languages, Hao et al. also find that the external stack is underutilized when the controller is an LSTM rather than a simpler neural network. This case study shows that it is not sufficient to add a linguistically motivated component to the architecture; it is crucial to ensure that the model cannot circumvent that component.

A number of recently proposed architectures incorporate constituency parses into neural networks: words are composed based on the parse rather than from left to right as in a standard RNN. While early architectures required the parse to be provided as part of the input, more recent architectures learn to parse sentences without supervision in the form of parse trees. Prior to the workshop, Williams, Drozdov, and Bowman (2018) showed that the trees induced by some of these models did not conform to linguistic intuitions and were inconsistent across run. By contrast, in an abstract presented at the workshop, Htut, Cho, and Bowman (2018) analyzed the trees produced by one of these grammar-induction networks—the Parsing-Reading-Predict Network (Shen, Lin, Huang, & Courville, 2018)—and found that these trees were much more consistent with gold parses than the trees produced by previous models. A notable feature of this work is that it is a replication study that addresses issues with the experimental design of the paper that originally proposed the model.

6 Future trends and outlook

This first edition of the BlackboxNLP workshop brought together a large amount of recent work on issues related to the analysis and interpretability of neural models, and thus allows for a preliminary assessment of the trends and potential points of focus for the future.

Evaluation One point which stood out to us is that there is no consensus on the best way of evaluating the different analytical techniques which are being introduced. A number of submissions resorted to qualitative evaluation to see whether the conclusions reached via a particular approach have face validity and match pre-existing intuitions. While this is often a necessary and helpful first step we believe that going forward it needs to be supplemented by more rigorous quantitative evaluation in order for the field as a whole to make measurable progress and become accepted as part of mainstream NLP. Some papers carried out quantitative evaluation of explanations of model decisions by human annotators. As currently practiced it can have its own issues. Specifically, when an explanation matches what a human would see as a reasonable basis of a particular decision, it does not necessarily follow that this was the basis that caused the model to make this decision. We see developing agreed-upon approaches to the evaluation of analytical and explanatory techniques as a major challenge for the field in the immediate future.

Benchmarks A related issue concerns datasets and benchmarks. Some popular datasets for evaluation have already emerged such as the agreement datasets (Linzen et al., 2016; Gulordava et al., 2018). As we have seen in sections 2 and 5.1 the workshop has seen some more contributions in that space. In future we would also like to see formal shared or unshared tasks to facilitate further progress. Again, we see further developments and gradual standardization here as crucial for the field.

Neuroscience We saw several disciplinary traditions represented at the workshop, including NLP, computer vision, speech processing, formal linguistics and psycholinguistics. The workshop also featured an inspiring invited talk on modeling language representations in the human brain with neural network models by Leila Wehbe. In future we would be very interested in welcoming neuroscientists to join the BlackboxNLP community in larger numbers and collaborate on answering fundamental questions about language in human and artificial brains.

BlackboxNLP 2019 The second edition of the workshop will be co-located with ACL 2019 in Florence and will take place on August 1. In order to ensure continuity and combine it with future sustainability the organizers of this upcoming edition include two of the current authors (Tal Linzen and Grzegorz Chrupała) as well as two new members (Dieuwke Hupkes and Yonatan Belinkov). The invited speakers will be Arianna Bisazza and Ari Morcos. We look forward to the new developments at BlackboxNLP 2019 and the field as a whole.

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. International Conference on Learning Representations (ICLR).

Alishahi, A., Barking, M., & Chrupała, G. (2017). Encoding of phonology in a recurrent neural model of grounded speech. In Proceedings of the 21st conference on computational natural language learning (conll 2017) (pp. 368–378).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), e0130140.

Bacon, G., & Regier, T. (2018). Probing sentence embeddings for structure-dependent tense. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 334–336). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5440`

Bastings, J., Baroni, M., Weston, J., Cho, K., & Kiela, D. (2018). Jump to better conclusions: Scan both left and right. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 47–55). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5407`

Burns, K., Nematzadeh, A., Grant, E., Gopnik, A., & Griffiths, T. (2018). Exploiting attention to reveal shortcomings in memory models. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 378–380). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5454`

Camacho-Collados, J., & Pilehvar, M. T. (2018). On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 40–46). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5406`

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pacific-asia conference on knowledge discovery and data mining (pp. 160–172).

Chomsky, N. (1962). Context-free grammars and pushdown storage.

Chrupała, G., Kádár, À., & Alishahi, A. (2015). Learning language through pictures. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers) (Vol. 2, pp. 112–118).

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 2126–2136). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/P18-1198`

Croce, D., Filice, S., Castellucci, G., & Basili, R. (2017). Deep learning in semantic kernel spaces. In Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers) (Vol. 1, pp. 345–354).

Croce, D., Rossini, D., & Basili, R. (2018). Explaining non-linear classifier decisions within kernel-based deep architectures. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 16–24). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5403

Dhar, P., & Bisazza, A. (2018). Does syntactic knowledge in multilingual language models transfer across languages? In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 374–377). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5453

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179–211.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling, 160(3), 249–264.

Giannakidou, A. (2011). Negative and positive polarity items: Variation, licensing, and compositionality. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), Semantics: An international handbook of natural language meaning (pp. 1660–1712). Berlin: Mouton de Gruyter.

Giulianelli, M., Harding, J., Mohnert, F., Hupkes, D., & Zuidema, W. (2018). Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 240–248). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5426

Grefenstette, E., Hermann, K. M., Suleyman, M., & Blunsom, P. (2015). Learning to transduce with unbounded memory. In Advances in Neural Information Processing Systems 28 (pp. 1828–1836).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers) (Vol. 1, pp. 1195–1205).

Gupta, P., & Schütze, H. (2018). Lisa: Explaining recurrent neural network judgments via layer-wise semantic accumulation and example to pattern transformation. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 154–164). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5418

Hao, Y., Merrill, W., Angluin, D., Frank, R., Amsel, N., Benz, A., & Mendelsohn, S. (2018). Context-free transductions with neural stacks. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 306–315). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5433

Harbecke, D., Schwarzenberg, R., & Alt, C. (2018). Learning explanations from language data. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 316–318). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5434

Hiebert, A., Peterson, C., Fyshe, A., & Mehta, N. (2018). Interpreting word-level hidden state behaviour of character-level lstm language models. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 258–266). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5428

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Htut, P. M., Cho, K., & Bowman, S. (2018). Grammar induction with neural language models: An unusual replication. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 371–373). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5452

Hupkes, D., Veldhoen, S., & Zuidema, W. (2018). Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. Journal of Artificial Intelligence Research, 61, 907–926.

Jacovi, A., Sar Shalom, O., & Goldberg, Y. (2018). Understanding convolutional neural networks for text classification. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 56–65). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5408

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In

International conference on learning representations.

Jumelet, J., & Hupkes, D. (2018). Do language models understand anything? on the ability of lstms to understand negative polarity items. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 222–231). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5424

Kádár, A., Chrupała, G., & Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. Computational Linguistics, 43(4), 761–780.

Kementchedjhieva, Y., & Lopez, A. (2018). 'indicatements' that character language models learn english morpho-syntactic units and regularities. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 145–153). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5417

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., & Dähne, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. arXiv preprint arXiv:1705.05598.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294–3302).

Krug, A., & Stober, S. (2018). Introspection for convolutional automatic speech recognition. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 187–199). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5421

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Proceedings of the 35th international conference on machine learning, ICML 2018, stockholmsmässan, stockholm, sweden, july 10-15, 2018 (pp. 2879–2888).

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. Transactions of the Association of Computational Linguistics, 4(1), 521–535.

Loula, J., Baroni, M., & Lake, B. (2018). Rearranging the familiar: Testing compositional generalization in recurrent networks. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 108–114). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5413

Madhyastha, P. S., Wang, J., & Specia, L. (2018a). End-to-end image captioning exploits distributional similarity in multimodal space. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 381–383). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5455

Madhyastha, P. S., Wang, J., & Specia, L. (2018b). End-to-end image captioning exploits distributional similarity in multimodal space. In Proceedings of the british machine vision conference (bmvc).

Mareček, D., & Rosa, R. (2018). Extracting syntactic trees from transformer encoder self-attentions. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 347–349). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5444

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211–222.

Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., & Griffiths, T. L. (2018). Evaluating theory of mind in question answering. arXiv preprint arXiv:1808.09352.

Niculae, V., & Blondel, M. (2017). A regularized framework for sparse and structured neural attention. In Advances in neural information processing systems (pp. 3338–3348).

Niculae, V., Martins, A. F., Blondel, M., & Cardie, C. (2018). Sparsemap: Differentiable sparse structured inference. arXiv preprint arXiv:1802.04223.

Paperno, D. (2018). Limitations in learning an interpreted language with recurrent models. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 384–386). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5456

Papernot, N., & McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv preprint arXiv:1803.04765.

Peters, B., Niculae, V., & Martins, A. F. T. (2018). Interpretable structure induction via sparse attention. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 365–367). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5450`

Poerner, N., Roth, B., & Schütze, H. (2018). Interpretable textual neuron representations for nlp. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 325–327). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5437`

Poliak, A., Haldar, A., Rudinger, R., Hu, J. E., Pavlick, E., White, A. S., & Van Durme, B. (2018). Collecting diverse natural language inference problems for sentence representation evaluation. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 337–340). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5441`

Raganato, A., & Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 287–297). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5431`

Raghu, M., Gilmer, J., Yosinski, J., & Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Advances in neural information processing systems (pp. 6076–6085).

Ravfogel, S., Goldberg, Y., & Tyers, F. (2018). Can lstm learn to capture agreement? the case of basque. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 98–107). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5412`

Rønning, O., Hardt, D., & Søgaard, A. (2018). Linguistic representations in multi-task neural networks for ellipsis resolution. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 66–73). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5409`

Ross, J. R. (1967). Constraints on variables in syntax (Unpublished doctoral dissertation). MIT.

Saphra, N., & Lopez, A. (2018a). Language models learn pos first. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 328–330). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5438`

Saphra, N., & Lopez, A. (2018b). Understanding learning dynamics of language models with svcca. arXiv preprint arXiv:1811.00225.

Sennhauser, L., & Berwick, R. (2018). Evaluating the ability of lstms to learn context-free grammars. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 115–124). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5414`

Shen, Y., Lin, Z., Huang, C.-W., & Courville, A. (2018). Neural language modeling by jointly learning syntax and lexicon. In 2018 International Conference on Learning Representations.

Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. Journal of Computer and System Sciences, 50(1), 132–150.

Skachkova, N., Trost, T., & Klakow, D. (2018). Closing brackets with recurrent neural networks. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 232–239). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5425`

Søgaard, A., de Lhoneux, M., & Augenstein, I. (2018). Nightmare at test time: How punctuation prevents parsers from generalizing. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 25–29). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5404`

Sommerauer, P., & Fokkens, A. (2018). Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 276–286). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/W18-5430`

Spinks, G., & Moens, M.-F. (2018). Evaluating textual representations through image generation. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 30–39). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5405

Stahlberg, F., Saunders, D., & Byrne, B. (2018). An operation sequence model for explainable neural machine translation. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 175–186). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5420

Sushil, M., Suster, S., & Daelemans, W. (2018). Rule induction for global explanation of trained models. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 82–97). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5411

Trifonov, V., Ganea, O.-E., Potapenko, A., & Hofmann, T. (2018). Learning and evaluating sparse interpretable sentence embeddings. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 200–210). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5422

Tutek, M., & Šnajder, J. (2018). Iterative recursive attention model for interpretable sequence classification. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 249–257). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5427

Vania, C., & Lopez, A. (2018). Explicitly modeling case improves neural dependency parsing. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 356–358). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5447

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998–6008).

Verwimp, L., Van Hamme, H., Renkens, V., & Wambacq, P. (2018a). State gradients for rnn memory analysis. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 344–346). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5443

Verwimp, L., Van Hamme, H., Renkens, V., & Wambacq, P. (2018b, Sep). State gradients for rnn memory analysis. Interspeech 2018. Retrieved from http://dx.doi.org/10.21437/Interspeech.2018-1153 doi: 10.21437/interspeech.2018-1153

Vu, N. T., Adel, H., Gupta, P., et al. (2016). Combining recurrent and convolutional neural networks for relation classification. In Proceedings of naacl-hlt (pp. 534–539).

Wallace, E., Feng, S., & Boyd-Graber, J. (2018). Interpreting neural networks with nearest neighbors. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 136–144). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5416

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (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 (pp. 353–355). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5446

Wei, J., Pham, K., O'Connor, B., & Dillon, B. (2018). Evaluating grammaticality in seq2seq models with a broad coverage hpsg grammar: A case study on machine translation. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 298–305). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5432

Weiss, G., Goldberg, Y., & Yahav, E. (2018). On the practical computational power of finite precision rnns for language recognition. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 740–745). Association for Computational Linguistics.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler–gap dependencies? In Proceedings of the 2018 EMNLP workshop BlackboxNLP:

Analyzing and interpreting neural networks for NLP (pp. 211–221). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5423

Williams, A., Drozdov, A., & Bowman, S. R. (2018). Do latent tree learning models identify meaningful structure in sentences? Transactions of the Association for Computational Linguistics, 6, 253–267. Retrieved from http://aclweb.org/anthology/Q18-1019

Yin, P., Zhou, C., He, J., & Neubig, G. (2018, July). StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In The 56th annual meeting of the association for computational linguistics (acl). Melbourne, Australia. Retrieved from https://arxiv.org/abs/1806.07832v1

Zhang, K., & Bowman, S. (2018a). Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. arXiv preprint arXiv:1809.10040.

Zhang, K., & Bowman, S. (2018b). Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp (pp. 359–361). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W18-5448

Zhou, C., & Neubig, G. (2017, July). Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction. In The 55th annual meeting of the association for computational linguistics (acl). Vancouver, Canada. Retrieved from http://www.phontron.com/paper/zhou17acl.pdf