Deep Nominal Semantic Role Labeling with auxiliary linguistic features in view of robustness, generalization in natural language inference

Advait Savant
Prof. Meyers

Abstract*—Assigning general semantic relationships to arguments for a certain predicate in a given sentence is an important yet challenging task in natural language processing [1]. Such an assignment of semantic roles enables us to create a shallow meaning representation for a sentence which can facilitate making further inferences. Motivated by the correlation between syntactic and semantic structures, conventional feature based SRL systems make a heavy use of syntactic features. Over the past seven to nine years, deep learning based approaches have been used to achieve state of the art performances on datasets such as the Propbank. Certain model formulations based on explicitly or implicitly encoding for syntactic structures such as dependency trees or parse trees have been proposed and their utility accessed [2]. In this work, we propose a method to incorporate an auxiliary task based on linguistic considerations in our BERT-based model [3] for semantic role labeling (SRL) on Nombank [4], evaluating the utility of linguistic knowledge in a deep learning pipeline. We analyze the effects of incorporation of linguistic features on the performance and generalization abilities of our model. We analyze our model from the perspective of semantic robustness [5]. We evaluate changes in the learnt representation by virtue of the introduction of auxiliary tasks, expanding on what happens to pre-trained contextualized embeddings in view of fine-tuning [6]. We also develop a multi view learning model which aggregates views based on implicit, symbolic, linguistic feature driven representations, as well as distributed representations based on contextualized word embeddings.

I. Introduction

The following is the link to the repository. https://github.com/adisav17/Deep-Semantic-Role-Labeling-with-Auxilary-tasks

Extracting predicate-argument structure (PAS) of a sentence can help us understand how participants relate to events, states and other propositions. A PAS can be viewed as a relation between a predicate and and an argument. Together a set of such PAS relations is one way to represent the meaning of a sentence. PAS relations for verbal predicates are typically labeled with distinct roles such as agent, instrument or patient.[1] Answering the question of who did what to whom in the context of a sentence is pertinent in view of natural language understanding. Assignment of thematic roles like agent and patient provides a way to capture semantic commonality between sentences and extrapolate general patterns for information extraction across sentences. PropBank [7] provides manual annotation of semantic roles with respect to verbal predicates. Using essentially the same framework, NomBank [8][4] provides manual annotation of semantic roles with respect to noun predicates.

In the context of statistical learning and computational linguistics, We consider the modeling problem as follows. For some sentence, nominal predicate pair $(S[w_1,,,,w_n],v)$, we predict a sequence of tags y such that y_i will give the semantic role label of w_i, if w_i is an argument with respect to v. Development of a useful feature space based on linguistic considerations, using statistical learning to set up a model and extrapolate patterns in the mapping from the feature space to the input space, using the model to make predictions on semantic role labels, are a useful line of inquiry.

AM Does this mean that for each sentence of length N and L possible labels, there are N X N X L possible relations? Each of these triples would be either TRUE of FALSE.

AD For each sentence of length N and a given predicate P, we have (S,P) of length (N,1). Given P, for each word in the sentence, we assign potential labels (semantic roles). This will be a N X L mapping. Each of the pairs (w_i, l_j) given (S,v) would either true or false.

Advancements in deep learning for natural language inference have pivoted a transition in the field. Deep learning approaches learn hierarchical, structured parameterized transformations in order to obtain a useful feature space. In the last five to six years, stacked attention mechanism based transformers have been used for state of the art performance, these are known to model distributed representations which capture notions of syntax and semantics.

Given the current architectural and modeling paradigms, there are considerations about the limitations of these methods and directions to improve on the same. There is work done on the use of novel architectures, integration of syntactic, semantic structures and parse trees in task specific transformer models [9][10]. Work on the analysis of the nature of representations learnt by large language models can help with interpretability and understanding the information captured by large language models can motivate further nuanced approaches.

Traditional NLP consists of analytical frameworks rooted in principles of linguistics such as morphology, syntactic parsing, combinatorial grammars, used in order to interpret, process and generate language. As the field of deep learning in NLP expands and matures, fueled by it's success in detecting complex patterns and long range dependencies in language, we

[1]We assume a framework in which these roles are numbered, rather than named. The numbered roles can be defined for each predicate.

critically evaluate the role and efficacy of linguistic features in a deep learning framework.

Our line of inquiry includes robustness in NLP, out of distribution generalization and interpretability of large language models and consequences for our modeling choices.

II. Semantic roles for Noun Predicates

Semantic Role Labeling (SRL), a concept with roots tracing back to the works of Panini around 2400 years ago[8], is instrumental in identifying and labeling the semantic relationships within sentences. The NomBank project annotated approximately 115,000 noun predicates and their respective arguments. It built upon the annotation schema earlier applied to verbs in the PropBank project. While earlier SRL systems predominantly focused on nouns directly morphologically related to verbal predicates, NomBank embarks on a broader scope incorporating 16 to 19 distinct classes of nouns that may not directly link to verbs, but share a similar frame or argument set. These classes in some respects, resemble the FrameNet approach, in which a shared set of arguments are marked for a set of predicates in a similar semantic domain or "frame". In this work, we explore partitive nouns, a category indicating a group or part of ARG1, highlighting several subclasses within this group. These nouns are characterized by their transparency, wherein the ARG1 frequently acts as the semantic head, drawing a parallel with the function of conjuncts in coordinate conjunctions. For example, in *"They passed a variety of tests."*, the partitive noun *variety* takes the ARG1 *tests*. *tests* is the content word of the phrase *variety of tests*, as evidenced by selection restrictions of the verb *passed*.

For every labeled predicate in a specific portion of the Penn Treebank 2 Wall Street Journal (WSJ) corpus, a distinctive representation of the encompassing sentence was devised by [4]. This representation encapsulated details derived from both the PTB and NomBank datasets, fostering the generation of varied datasets aligned with different categories of NomBank predicates. In the envisioned task setup, the system embarks with a foundational layer of "golden" data: syntactic parses and Part-Of-Speech (POS) tags sourced from the (manually annotated) Penn Treebank, chunk identifiers ,beginning-input-output (BIO) tags for with phrase information, along with the identification of the predicate noun and its respective categories from NomBank, supplemented with **support** verbs as specified in the NomBank guidelines.[2]

III. Deep learning for nominal SRL

We model SRL as a sequence labeling task, given a nominal predicate.[3] This is analogous to the token classification problem, where for every word token, we predict an argument role or lack thereof, with respect to the predicate. For a given input sequence, and for a given predicate, we predict the argument referencing tags over the sentence as the output sequence. We can model a probability distribution of the output sequence given the input sequence and the predicate.

A language model will assign a probability distribution over a sequence of words. Models consisting of recurrent neural networks with gates added for selectively reading, writing and forgetting information drove the use of deep learning for language modeling.

AM The previous sentence is a long run-on sentence and should be replaced by a fe simpler sentences. Currently, I cannot understand it.

Deep learning based language models can assume a factorization and a conditional dependence in the input sequence and the subsequent intermediate representations in order to induce an efficient parametrization. The utility of a language model is that we can sample from a distribution in order to predict the next word given a sequence of words. Bi directional language models, in order to make a prediction at a particular time step, consider the context from both sides of the input sequence, in our case, the input sentence.

BERT[11], namely, bidirectional encoder representations through transformers, is a language model which has successive multi headed attention blocks[12], with inputs being the outputs from the previous attention block, forming a stack of attention blocks. This enables us to model complex non linear hierarchical distributed representations for an encoder model which is intended to create contextual representations and a decoder which can be tuned for any downstream task. The outputs of the final attention block are aggregated, again with a linear transformation, to produce a single output. BERT is designed to produce representations which enable further use in downstream tasks.

The baseline of the modeling process for our task using a BERT-base model is as follows [13]. In the fine tuning phase, for each word in the input sentence, we can collect the word representations produced by a forward pass of BERT. We will add a long short term memory network layer for sequential representations, on the top this we will add a linear layer to make the predictions. The model be trained to predict the argument tag for every word, learning a representation tuned for our task.

IV. On Large language models, fine tuning, robustness and generalization paradigms

In this section, we define certain concepts in transfer learning, semantic robustness, generalization, nuances in current deep learning based natural language inference models, and the interpretability of Bert in order to motivate our modeling choice.

AM This should be reworded. What does "notions" mean – metrics? subprocesses? ...

A. Issues in NLI

Some limitations of language models are discussed in [14][15]. Deep learning based language models require a high volume of data to train on, they have high sample complexity and data insufficiency. Hyper-parameter tuning and architecture search can be computationally expensive. Fine-tuned models can also be seen to exhibit poor generalization

[2]In NomBank, a support verb links together a NomBank predicate and argument. For example, in "Mary took a walk", **took** is a support verb of the noun *walk*. The subject of *took* is the ARG0 of the nominalization *walk*.

[3]SRL can be divided into at least 2 subtasks: 1) finding a predicate; 2) finding an argument of that predicate. For our purposes, we are assuming that the first task has already been solved.

capabilities in natural language inference tasks. A fine-tuned model on a particular data set may have a degradation in it's performance on a different data set for a similar downstream task. Linguistic elements(morphemes, words, phrases) can be combined in various ways in order to produce novel sense making sentences (systematicity). Thus the set of semantically well-formed sentences is exponentially larger than the set of linguistic elements. Deep learning NLP models are known not to capture this property very well. [16]

Fine tuned models for natural language inference can rely on spurious correlations while making predictions, even though the associations may not hold meaningful causal relationships. To elaborate, patterns in the training data, e.g. based on structural, lexical biases, which would not be linguistically meaningful but would be statistically significant if they correlate with the output variable, can be picked up by the model for making predictions. Annotation artifacts [15], unintentionally present in the data could be exploited by the model and taken as cues. [17] Models might adopt certain heuristic short cuts based on lexical patterns, e.g, based on lexical overlap or constituent structure, focusing on training data statistics in order to achieve high performance. This is not beneficial from a generalization perspective and does not enable the model to faithfully represent the underlying data generating process in a predictive manner. We analyze our model in view of these issues.

B. Domain adaptation and model generalization

We a have a task T Which consists of finding a mapping f between an input feature space X and an output feature space Y. A domain consists of the input feature space X and an underlying distribution P(X) over the feature space. Given a source domain D_S and a task T_S, transfer learning concerns itself with how to leverage the knowledge gained in the performance of the task T_S in order to perform the task T_T over a domain D_T. In domain adaptation, the concerned input feature space is the same, i.e. the set of possible inputs is the same, but the distribution over the input space changes. There is a domain shift in the data distribution on which the model is trained on and the data distribution on which performance is desired. Every machine learning model can be construed as consisting of (i) an inductive bias, a set of assumptions which the model makes regarding the input to output relationship; (ii) a hypothesis, a function which describes an input to output mapping and an hypothesis space; and (iii) the set of all possible hypotheses which can be represented by a model [18]. Given a certain hypothesis $h(x)$ and an output y. We would want our model to minimize the expected cost $E_{D\sim x,y}[Cost(h(x), y)]$ for the data coming from a data generating distribution D.

The complexity of the hypothesis space governs the ability of the hypothesis space to contain the desired optimal hypothesis. If our hypothesis space is large enough, the odds of the best possible hypothesis in our hypothesis space h_o being the desired optimal hypothesis h_* are high. But as our hypothesis space grows larger and more flexible, the capacity of a given model to estimate the best possible hypothesis in a given hypothesis space can reduce. There is an intrinsic trade off between hypothesis space complexity and model estimation capability. This trade off relates closely with deep learning and overfitting [19].

AM I can understand the previous sentence with effort, but it would be more readable as 3 sentences instead of one enormous one.

AD Noted Prof. Broken the sentence into separate sentences

For a certain hypothesis space and a chosen hypothesis, the total error/cost coming from $h_* - h_{ours}$ is a combination of the approximation error coming from $h_* - h_o$ error and the estimation error coming from $h_o - h_{ours}$. Deep learning based language models tend to be proven to overfitting, as is evident by the literature on model pruning [20], techiques on regularization and optimization, we work on the need to enhance domain adaptation and generalization capabilities.

C. Semantic Robustness

Robustness refers to the ability of the model to maintain stable performance when faced with various kinds of input perturbations, noise, or changes in the data distribution. Let $f(x;\theta)$ be a machine learning model parameterized by θ, where x is an input data point. Let D be the data distribution from which training samples are drawn. A model is said to exhibit local continuous robustness if, for small perturbations δ in the input x, the output of the model does not change significantly, i.e.,

$$\forall \delta : \|\delta\| \leq \epsilon,\ \|f(x;\theta) - f(x+\delta;\theta)\| \leq \gamma$$

where ϵ is a small positive value defining the magnitude of the perturbation and γ is a tolerance parameter which defines how much output change is allowed. A notion of local discrete robustness is more suitable for analysis in natural language inference since language is discrete. [5]

Dense continuous word embeddings based on the distributional hypothesis can be used to define notions of discrete continuous robustness. As an example, for a sentence s containing a word w, let w' be word corresponding to the closest vector representation in the neighborhood of w in the embedding space. Sentence s' is formed by replacing w with w' in s. Contingent on the context, under robustness considerations, one plausibly could expect f(s) and f(s') to be similar [5]. Based on linguistic rules in view of task considerations, [5] develop a notion of semantic robustness in view of having a definition of robustness which is linguistically meaningful. Robustness in machine learning models is of utility in view of generalization, handling real world noise [21][22]. We analyze our model in view of the same.

AM In the following section, you claim that on the one hand, ML models are black boxes, but on the other hand, you would to explain/understand pieces of these black boxes. This would seem like a contradiction. Perhaps, you are trying to define equivalence based on output, i.e., X = Y if X and Y generate the same something (same strings, same correct cases, same incorrect cases, same ???

AD The ML model is indeed a black box in the following sense. After the model is trained, when the model makes an inference, the process itself does not give us any information on why the inference is made. The computations of the forward pass do not entail information on explanatory factors. While the model is a black box, there is a critical need for understanding what information is captured in the learned parameters and develop frameworks to explain why a model made a certain inference. These frameworks are not explicit but developed implicitly. They call it 'opening the black box.' In our case, we aim to analyze the learned representations in view of syntax. Our analysis and assertions are based upon the processes used to probe the parameters of the model, this is the lens we use. Hence, we say that we interpret the model.

AM Please try to use non-xarv references, e.g., I changed the Kevin Clark reference. In general, only peer reviewed papers are valid referencees.

AD Noted Prof. Will update the references

D. Interpretability of BERT

Models make predictions as black boxes. Understanding and interpreting the information represented in the learned parametrizations, the explainability of model outputs and a quantification of uncertainty in model outputs can be of practical utility. [14] [23]. Consider the following analysis on the interpretability of BERT [24]. For every attention head, [24] treat it as a no-training-required classifier. For some input word and an attention head, we look at the most attended to word based on the attention weights distribution. We evaluate attention heads across layers for their ability to represent various syntactic relations. It is observed that across initial layers, the spread of attention is broad, while up the hierarchy of layers, there is a focused distribution of attention. Certain attention heads across higher layers do well in capturing certain syntactic relations. For example, some heads have direct objects attending to their words, some heads have prepositions attending to their objects, some heads have following coreferent words attending to the preceding coreferents. [25] uses edge probing, and tests the BERT model across various language processing tasks like POS tagging and dependency pars- ing. It observes a certain progression based on their language processing task evaluation metrics, i.e, POS tags processed earliest, followed by constituents, dependencies, semantic roles, and coreference. If the optimization process of the model proceeds in a way that it picks up implicit representations of syntax without any inductive bias in a free functional optimization, one could make conjecture that the information on syntax is correlated and useful for an arbitrary language processing task in the context of predictive modeling. In [26] [27], we see that the use of feature engineering driven by linguistic considerations alongside a deep learning based natural language sequence labeling pipeline is seen to deliver improvements in performance. External syntactic information having a discrete feature space integrated with the distributed word vector representations has been of utility for benchmark performance in the recent years [2][9].

V. Experimentation and analysis

We consider the line of inquiry established in the previous section to describe our modeling choice, in view of performance, robustness and domain adaptation.

[6] compare a fine tuned model and a pre-trained model, based on structured probing and edge probing, analyze changes in the representations of linguistic features for a fine-tuned model. They argue that while linguistic features may not be incorporated into predictions, models can tend to rely on heuristics and annotation artifacts for predictions, information regarding linguistic features is still available in the model's representations with a syntactic sub-space [28] being present. We introduce changes in the training procedure in order to explicitly make the model incorporate linguistic features for predictive modeling. We study how such a scheme effects the representations learned by the model, further analyzing the nature of encoded representations in large language models.

We introduce linguistic features based on BIO tags, POS tags, the directed parse tree distances to the predicate in our model pipeline. We construct an auxiliary task in which we make our bert model predict these features through a parameterization which has a shared base representation with the downstream model predicting semantic roles and certain separate auxiliary parameters used for the prediction of the linguistic features. Since the base parameterization is shared, we push the model to learn a representation which can perform these tasks simultaneously. We give a hyperparameter weight to the auxiliary task in the cost function and the corresponding errors are backpropagated for parameter updates via gradient descent in conjunction with the errors from the downstream task cost function. For input to the auxiliary task, we give it a weighted combination of selected bert layers such that the weights are themselves learned. The model can decide how much importance to give to the inputs from each layer for the auxiliary task. [29] use such a paradigm for their downstream task. We select the initial layers of bert in view of observations in [25].

We consider the BERT-based model with and without predicate embeddings, with and without positional embeddings. We experiment with predicate indicator embeddings as well. We use the binary cross entopy loss for the prediction of ARG1 labels. A softmax and cross entropy loss for prediction of BIO tags, POS tags and a mean squared error loss for the prediction of the directed parse tree distance. We experiment with long short term memory and feed forward layers, various hyperparameters for the predicate embedding dimension, the hidden size of the lstm, the dropout rate, a custom weight value to scale the loss function for imbalanced classes etc.

We test for performance, generalization abilities of our model. Given the nature of partitive nouns and percent nouns

in the Nombank [4], shifting from assigning semantic roles for partitive nouns, to assigning semantic roles to percent nouns, can be viewed as domain adaptation, wherein, the nature of task and the input space is the same, but the distribution over the input space changes, so we have a domain shift. Shifting from assigning semantic roles for partitive nouns to assigning semantic roles for attribute nouns and relational nouns, can be viewed as out-of-distribution generalization, since, our input space is changing with regards to the category of the noun. In view of prior linguistic knowledge of the mechanisms which assign semantic roles to nouns[4], we find it fit to formulate task A, assigning semantic roles for partitive nouns, and task B, assigning semantic roles for relational nouns, as different tasks.

Linguistic features seen to help generalization in natural language inference and improve performance on the F-score. For the entire partitive group Nombank task, For the BERT-based model with predicate indicator embeddings and positional embeddings attached to it, we get an F-score of 83.8. We call this model 1. For the BERT-based model trained with an auxiliary model tuning on linguistic features, we get an F-score of 84.6. We call this model 2.

Given a model 1 and model 2 trained on the partitive task, we make the predict on the percent task. We get an F-score of 90.6 with model 1. We get an F-score of 96.1 with model 2. If we re-train the models for a few epochs on the percent task, we get an F-score of 92.3 with model 1. We get an F-score of 97.2 with model 2.

Domain adaptation can be viewed in terms of a model based interpretation and feature based interpretation. We take into perspective a Bert base model and our auxiliary model. Each model will assume an implicit distribution over the input data. The inductive bias that is introduced by our model being in terms of linguistic features, holds in general across data sets. In other words, the mechanisms by which we assign BIO tags or POS tags to the words in sentences, generate parse trees for sentences, does not strongly depend on the corpus being a penn treebank or a brown corpus and can be better seen as an implicit property of language which our model needs to capture.

The auxiliary task we introduce alters our hypothesis space and constrains it to prefer some hypotheses over others [30]. These hypotheses represent models which have a better grasp over the implicit properties of language as it assigns a distribution over a sentence. We notice a convergence in the loses for the both the downstream and the auxiliary task. The inductive bias added by the auxiliary task enables the model to learn representations which explain more than one task. Such a multi task learning approach[31] can enable the model to pick up robust features which are pertinent in view of natural language inference and since different tasks have different noise patterns, can mitigate the effects of data-dependent noise. Preferring representations suited across tasks, from the predicate task to the partitive task, as the input data distribution shifts, the auxiliary model is able to better represent the distributional shift, as evident in the performance on domain adaptation.

Based on the principles of multi-view learning [18], we also develop an ensemble model. We have the underlying data generating process which produces a sentence and the corresponding semantic roles as input output pairs. For every model, when we make a choice of representation to define the input and output data, we are modeling a view of the data generating process. There can be multiple views, each having there own caveats, which attempt to capture the underlying data generating process. The performance of a model is proportional to the potency of a view to faithfully represent the data generating process. If we have multiple views, each view can better capture certain aspects of the process as compared to another one, and it can be of benefit to aggregate the information obtained from multiple views in order to construct a better model of the data generating process. For the multi view ensemble model, we get an F-score of 87.2.

In our context, the BERT-based model has input data consisting of contextualized distributed word vector representations and our auxiliary task. We also have a feature based model which uses the linguistic features as input data representations rather than an auxiliary task to drive parameterizations and makes predictions on the semantic roles based on the aggregation of contextualized word vectors and linguistic features as input. This enables us to combine information from two different views of the data generating process, each being related to the underlying mechanism of generation in their own merits and having the capacity to represent and traverse a corresponding hypothesis space.

We continue to document our results as we conduct a representational similarity analysis[32] in view our auxiliary task based fine-tuned model, a standard fine tuned model and a pre-trained model; for in-distribution and out-of-distribution data, expanding on [6]. We aim to document our results soon. We aim to derive insights and analyze results viz-a-viz the nature of encoded information representations in large language models. We are in process of testing our model for robustness and aim to critically document the results. [5]

VI. Conclusion and further work

We aim to understand the impact of linguistic features on nominal SRL in a deep learning framework and evaluate the utility of the same. We show concrete improvements in performance, tests for domain adaptation. We demonstrate the utility of an ensemble model. We present arguments for our modeling choices.

The section is in progress. We could potentially use our task as part of the set of tasks for model agnostic meta learning in order to simultaneously model all Nombank categories and evaluate the utility of the same in a generalized SRL system which works across predicate categories.

References

[1] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[2] Qingrong Xia et al. “Syntax-aware neural semantic role labeling”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 7305–7313.

[3] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *NAACL-HLT (1)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. ISBN: 978-1-950737-13-0. URL: http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DevlinCLT19.

[4] Adam Meyers et al. “The NomBank Project: An Interim Report”. In: *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004, pp. 24–31. URL: https://aclanthology.org/W04-2705.

[5] Emanuele La Malfa and Marta Kwiatkowska. *The King is Naked: on the Notion of Robustness for Natural Language Processing*. 2022. arXiv: 2112.07605 [cs.CL].

[6] Amil Merchant et al. “What Happens To BERT Embeddings During Fine-tuning?” In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 33–44. DOI: 10.18653/v1/2020.blackboxnlp-1.4. URL: https://aclanthology.org/2020.blackboxnlp-1.4.

[7] M. Palmer, D. Gildea, and P. Kingsbury. “The Proposition Bank: An Annotated Corpus of Semantic Roles”. In: *Computational Linguistics* 31.1 (2005), pp. 71–106. URL: http://www.cs.rochester.edu/~gildea/palmer-propbank-cl.pdf.

[8] Adam Meyers et al. “The NomBank project: An interim report”. In: *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*. 2004, pp. 24–31.

[9] Devendra Singh Sachan et al. “Do syntax trees help pre-trained transformers extract information?” In: *arXiv preprint arXiv:2008.09084* (2020).

[10] Yau-Shian Wang, Hung-Yi Lee, and Yun-Nung Chen. “Tree transformer: Integrating tree structures into self-attention”. In: *arXiv preprint arXiv:1909.06639* (2019).

[11] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[12] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[13] Peng Shi and Jimmy J. Lin. “Simple BERT Models for Relation Extraction and Semantic Role Labeling”. In: *ArXiv* abs/1904.05255 (2019).

[14] Marta Garnelo and Murray Shanahan. “Reconciling deep learning with symbolic artificial intelligence: representing objects and relations”. In: *Current Opinion in Behavioral Sciences* 29 (2019), pp. 17–23.

[15] Suchin Gururangan et al. “Annotation Artifacts in Natural Language Inference Data”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 107–112. DOI: 10.18653/v1/N18-2017. URL: https://aclanthology.org/N18-2017.

[16] Jake Russin et al. “Compositional generalization in a deep seq2seq model by separating syntax and semantics”. In: *arXiv preprint arXiv:1904.09708* (2019).

[17] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference*. 2019. arXiv: 1902.01007 [cs.CL].

[18] Fuzhen Zhuang et al. “A comprehensive survey on transfer learning”. In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[20] Davis Blalock et al. “What is the state of neural network pruning?” In: *Proceedings of machine learning and systems* 2 (2020), pp. 129–146.

[21] Philipp Koehn et al. “Statistical Significance Tests for Machine Translation Evaluation”. In: *Proceedings of EMNLP*. Association for Computational Linguistics. 2004.

[22] Dan Hendrycks and Thomas Dietterich. “Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations”. In: *arXiv preprint arXiv:1807.01697* (2018).

[23] Lorenzo Ferrone and Fabio Massimo Zanzotto. “Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey”. In: *Frontiers in Robotics and AI* 6 (2020), p. 153.

[24] Kevin Clark et al. “What Does BERT Look at? An Analysis of BERT’s Attention”. In: (Aug. 2019), pp. 276–286. DOI: 10.18653/v1/W19-4828. URL: https://aclanthology.org/W19-4828.

[25] Ian Tenney, Dipanjan Das, and Ellie Pavlick. “BERT rediscovers the classical NLP pipeline”. In: *arXiv preprint arXiv:1905.05950* (2019).

[26] Yufei Wang et al. “How to best use syntax in semantic role labelling”. In: *arXiv preprint arXiv:1906.00266* (2019).

[27] Minghao Wu, Fei Liu, and Trevor Cohn. “Evaluating the utility of hand-crafted features in sequence labelling”. In: *arXiv preprint arXiv:1808.09075* (2018).

[28] John Hewitt and Christopher D. Manning. “A Structural Probe for Finding Syntax in Word Representations”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4129–4138. DOI: 10.18653/v1/N19-1419. URL: https://aclanthology.org/N19-1419.

[29] Ian Tenney et al. “What do you learn from context? probing for sentence structure in contextualized word representations”. In: *arXiv preprint arXiv:1905.06316* (2019).

[30] Sebastian Ruder. “An overview of multi-task learning in deep neural networks”. In: *arXiv preprint arXiv:1706.05098* (2017).

[31] Rich Caruana. “Multitask learning”. In: *Machine learning* 28 (1997), pp. 41–75.

[32] Hamed Nili et al. “A toolbox for representational similarity analysis”. In: *PLoS computational biology* 10.4 (2014), e1003553.