Interpreting and Controlling Linguistic Features in Neural Networks' Representations

Tomasz Limisiewicz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic
limisiewicz@ufal.mff.cuni.cz

Abstract

In recent years, Neural Networks proved their usefulness in computing and processing vector representations to solve various NLP tasks. The key objective of the thesis is to provide a better explanation of the neural network's representation of language, also known as embeddings. We hypothesize that the understanding of representations is necessary to improve the models and alleviate issues hindering models' performances. We investigate the spatial distribution of vector representation and correlation with language features (syntax, lexicon, morphology) to answer which phenomena are encoded by a model. We aim to indicate which parts of the representation are relevant to a specific type of information. We propose a method that allows disentangling embedding spaces into parts that encode specific linguistic features. Our findings generalize to multiple diverse languages. In future work, we intend to analyze the encoding of higher-level linguistic features. We want to apply our findings to cope with the unwanted behaviors of language models, such as the unprovoked generation of toxic or biased texts.

1 Introduction

The recent wave of deep learning research in Natural Language Processing produced the models that achieve state-of-the-art results for diverse NLP tasks. The contributing factor to the success was transfer learning. In this approach, models trained on unannotated data (language models, neural machine translators) were applied to solve, Question Answering, Syntactic Parsing, Natural Language Inference, Named Entity Recognition, etc. (Kondratyuk and Straka, 2019; Sun et al., 2020; Yan et al., 2021, *inter alia*). The success of such models is not yet fully explained. Their internal mechanism is unknown to the users and the developers; such models are commonly referred to as *black boxes*.

The interpretability analysis closely follows the works on developing new models. The concept of interpretability is crucial for applications where the model's decision needs to be justified, for instance, health care, finance. The understanding of the algorithms also raises the social trust toward new technologies (Molnar, 2020). Therefore, the first aim of our research is the interest in learning the inner workings and the reason behind the success of neural networks in NLP. The second motivation, not less important, is recognizing deep models' shortcomings and finding paths for improvement. In our work we mostly concentrate on post-hoc (or extrinsic) interpretation defined by Lipton (2018) explaining the functioning of a model after training, "without sacrificing its predictive performance".

We perceive the hidden representations of pretrained neural networks as the way a model understands the language. This understanding is learned by seeing a large amount of data. It covers linguistic features of the texts (e.g., syntax, lexicon, semantics) and non-linguistic ones such as factual information about the world and various biases present in the training corpora.

The former features are more interesting from the linguistic point of view. Nevertheless, we cannot disregard the representation of non-linguistic aspects. Deep language models convey significant factual information (Brown et al., 2020; Khashabi et al., 2020) and tend to correctly answer the questions like: *"George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?"* answer: *"dry palms"*[1] (Clark et al., 2018; Mihaylov et al., 2018). Notwithstanding, the ability to non-linguistic features makes the model prone to biases present in the data. There is no mystery that contemporary

[1] In the evaluation model needs the most probable option out of four.

systems require vast corpora that cannot be manually pruned. This leads to the problem that the models are typically trained on lower quality data that could include nonfactual information, discriminatory texts, and toxic language. The biases can easily leak to the output of the models (Bolukbasi et al., 2016; Manzini et al., 2019).

Obtaining good language representations requires extensive corpora to optimize the deep model. Such corpora are available for English and a handful of other languages. For other less resourceful languages, we need to somehow transfer the representations from high-resource languages. An efficient solution to this problem is training a model on monolingual corpora in multiple languages together. Such models benefit from seeing more data and acquire linguistic knowledge in multiple languages (Pires et al., 2019). From an interpretability perspective, we may question how the model's representations vary across languages. The analysis could improve the representations for low-resource languages.

1.1 The Motivations of the Thesis

In the thesis, we want to continue our work on interpretations of neural networks' representations. The main directions of my current and future work are:

A Which linguistic and non-linguistic features are encoded by the neural networks?

B We aim to analyze multilingual representations. How do such representations vary across languages? Will we observe similar patterns for diverse languages?

C Can we disentangle the representation and filter out specific parts related to specific features?

D Can our findings be applied to filter out biases present in the representation? Will it improve the fairness of NLP models?

1.2 Structure of the Proposal

Sections 2 to 4 presents the main areas of previous research: architectures and types of the most prominent neural models; the methods of analyzing models' representations; biases and other unwanted non-linguistic features manifested by the models. In Section 5, I present my previous work on the topic and set a plan for future research in Section 6. Section 7 reiterates the main points and concludes the proposal.

2 Deep Models in NLP

The employment of neural networks to obtain the representation of language has a long history (Bengio et al., 2003; Mikolov et al., 2013a, *inter alia*). The networks needs to numerically express correlation found in corpora, such representations are especially suitable for processing and statistical analysis. The major breakthrough was the application of neural language models to obtain contextual representation. The first such model was ELMo proposed by Peters et al. (2018). Contextual embeddings encode not only one word but also its context. The ELMo was based on *bidirectional recurrent neural network*, this architecture was recently supplanted by more potent Transformer architecture (Vaswani et al., 2017). The Transformer based successor of ELMo - BERT improved results across various NLP tasks. Currently, the backbone of transfer learning in NLP are language models. They are trained on extensive corpora that do not need human curation. To a lesser extent, representations of neural machine translation systems are used for transfer learning (McCann et al., 2017), their main disadvantage is the requirement of parallel data that are harder to obtain than monolingual corpus.

Transformer language models can be divided into two families: *masked* (or auto-encoding) and *auto-regressive* models. The *masked* models (e.g., BERT (Devlin et al., 2019), XLNet (Yang et al., 2019))are trained to predict specific words inside the sentence based on all the other words in the sentence, and the *auto-regressive* models predict the word that follows after a sequence of given or previously generated words.

2.1 Auto-encoding Models

The first Transformer-based masked language model for pre-trained contextual embeddings was BERT, introduced by Devlin et al. (2019). The pre-trained model generates useful sentence representations and may be fine-tuned for many natural language processing tasks. This approach led to outstanding results on multiple NLP tasks. Among other: natural language inference (Bowman et al., 2015) and question answering (SQUaD by (Rajpurkar et al., 2016)).

A range of models has come after BERT, all

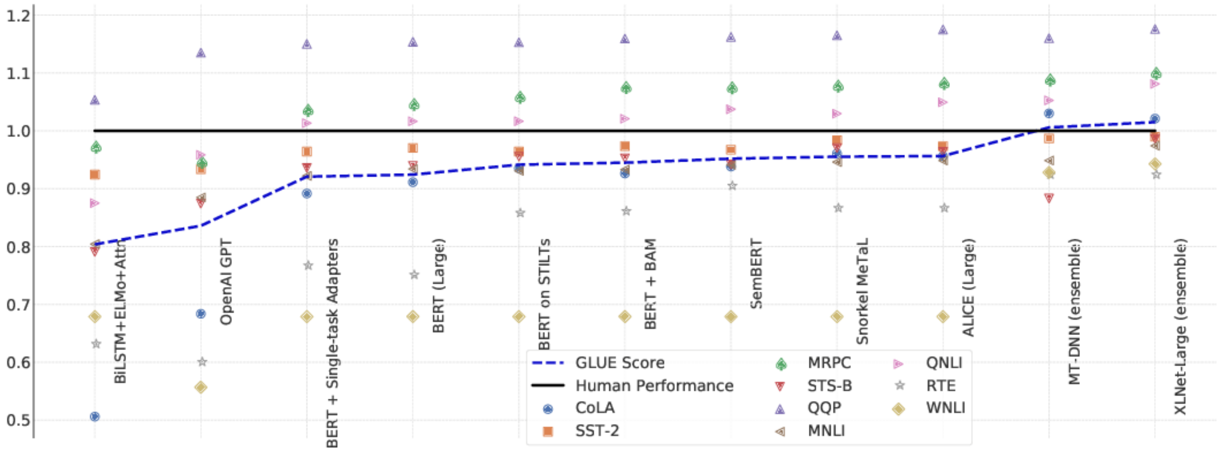

Figure 1: GLUE benchmark (Wang et al., 2018) performance of the pre-trained systems compared to the human performance (1.0 on the y-axis). Reprint from Wang et al. (2019a).

based on the same architecture, just with some additional twists that improve the performance in various ways. The most important are the following:

- **XLNet** Yang et al. (2019) enabled learning bidirectional contexts by maximizing the expected likelihood over all permutations of words and therefore overcame the pretrain–finetune masking discrepancy, which is present in BERT.
- **RoBERTa** Liu et al. (2019b) is a model similar to BERT trained by Facebook using much more training data than the original BERT
- **DistillBERT** Sanh et al. (2020) promises to deliver comparable results to BERT at a significantly lower computational cost.
- **Albert** Lan et al. (2019) is claimed to be as good as BERT with fewer parameters.

The subsequent models improve the results on benchmarks. They also base on Transformer architecture. The key factor contributing to the improvements were additional data, a higher number of parameters, more efficient initialization.

The transfer-learning performance of the models is measured by GLUE benchmark (Wang et al., 2018). It consists of nine language understanding tasks. Fig. 1 shows the gradual improvement of subsequent pre-trained models. The average score surpassed the non-expert human performance. Therefore, the authors introduced a new evaluation benchmark – SuperGLUE with more challenging tasks Wang et al. (2019a).

2.2 Auto-regressive Models

The notable examples of auto-regressive transformer models are:

- **GPT, GPT-2** Radford et al. (2019) large-scale unsupervised Transformer-based language models called GPT and GPT-2, which are able to generate coherent paragraphs of text.
- **GPT-3** Brown et al. (2020) model that builds upon GPT and GPT-2. It outperforms preceding models due to a significantly larger pre-training dataset and a larger number of parameters.
- **CTRL** Keskar et al. (2019) is a large language model. Predictions are conditioned on control codes allowing to set desired properties of the generated text explicitly.

Auto-regressive language models are closely related to the contemporary neural machine translation systems (NMT). They share the same architecture and auto-regressive prediction method. The main difference is that in NMT input and output of the network are in different languages. Some of the Transformer-based translation models are publicly available (Ng et al., 2020; Sennrich et al., 2016).

2.3 Multilingual models

The previously mentioned models were trained mainly on English data. We can obtain the representations in other languages in two ways: training a new model specifically for the language or training a multilingual model. The latter approach requires less computation and can benefit languages with a small amount of data available. From an interpretability perspective, the multilingual models allow us to easily investigate whether the observations about the model's representations generalize across languages. Because of these reasons, we will focus on shared multilingual representations.

The notable examples of Transformer-based multilingual models are:

- **MBERT** Devlin et al. (2019) trained on Wikipedias in 100 languages. The architecture of the model and training method is the same as in original BERT.

- **XLM** Lample and Conneau (2019) improves initialization and training method of MBERT.

- **XLM-RoBERTa** Conneau et al. (2020) uses enhancements introduced in RoBERTa and XLM. Model is trained on a much larger dataset than previous ones – cleaned Common Crawl (Wenzek et al., 2020).

Similarly to the monolingual Transformers, the subsequent models achieve improvement on a variety of cross-lingual benchmarks.

3 Interpretation of Neural Networks' Representations

With the recent success of pre-trained models in NLP, a significant focus was put on interpreting their representations. Latent representations of neural networks encode specific linguistic features. Recently, a lot of focus was devoted to finding correspondence from the internal representations of the networks to existing linguistic abstraction. Among others, they include coreference, syntactic structure, morphology.

Two main approaches to explaining the inner workings of neural networks are behavioral and structural analysis. The former one investigates the sensitivity of the model's output to the targeted changes of input. This way, it is possible to investigate the model's behavior in specific situations (Ribeiro et al., 2020). The structural analysis asks how linguistic features are encoded in the inner representations of the network. This approach directly works with the parameters of the network (Belinkov and Glass, 2019). We will devote more focus to structural analysis.

3.1 Probing

One of the most popular methods of analysis is probing. The parameters of the pre-trained network are fixed, the output word representations are fed to a simple neural layer. This simple layer is optimized for a linguistic task (e.g., POS tagging) to evaluate whether the necessary feature is encoded in the representation.

The number of probing experiments rose with the advent of multilayer RNNs and Transformers trained for language modeling and machine translation.

Belinkov et al. (2017a) probe a recurrent neural machine translation system with four layers to predict part of speech tags (along with morphological features). They use Arabic, Hebrew, French, German, and Czech to English pairs.

The introduction of ELMo (Peters et al., 2018) brought a remarkable advancement in transfer learning from the RNN language model to a variety of other NLP tasks. The authors examined POS capabilities of the representations and compared the results with the neural machine translation system CoVe (McCann et al., 2017), which also uses RNN architecture.

Another comprehensive evaluation of morphological and syntactic capabilities of language models was conducted by Liu et al. (2019a). Probing was applied to a language model based on the Transformer architecture (BERT) and compared with ELMo and static word embeddings (Word2Vec Mikolov et al. (2013a)).

Apart from morphosyntax, semantic information was also probed for in other experiments. (Teichert et al., 2017; Rudinger et al., 2018) examined encoding of semantic proto-roles, (Bjerva et al., 2016) optimized probe for semantic tagging. Those experiments showed that the probes had problems with uncovering deep syntactic relations.

3.2 Encoding Structure

Extraction of dependency structure is demanding because instead of predicting single tokens, every pair of words needs to be evaluated.

Blevins et al. (2018) propose a feed-forward layer on top of a frozen RNN representation to predict whether a dependency tree edge connects a pair of tokens. They concatenate the vector representation of each of the words and their element-wise product. Such a representation is fed as an input to the binary classifier. It only looks at one pair of tokens at a time, therefore predicted edges might not form a valid tree.

Another approach, induction of the whole syntactic structures from latent representations, was proposed by Hewitt and Manning (2019). Their syntactic probing is based on training a matrix that is used to transform the output of the network's layers (they use BERT and ELMO). The objective of the probing is to approximate dependency tree distances between tokens [2] by the L2 norm of the difference of the transformed vectors. Where authors consider linear transformation, i.e., embeddings are multiplied by gradient-optimized matrix. Probing produces the approximate syntactic pairwise distances for each pair of tokens. The minimum spanning tree algorithm is used on the distance matrix to find the undirected dependency tree. The best configuration employs the 15th layer of BERT large. It induces treebank with 82.5% undirected UAS on Penn Treebank with Stanford Dependency annotation (relation directions and punctuation were disregarded in the experiments). The result for BERTis significantly higher than for ELMO, which gave 77.0% when the first layer was probed.

The paper also describes an alternative method of approximating the syntactic depth by the L2 norm of latent vector multiplied by a trainable matrix. The estimated depths allow prediction of the root of a sentence with 90.1% accuracy when representation from the 16th layer of BERTlarge is probed.

3.3 Encoding Multilingualism

The subsequent paper by Chi et al. (2020) applies the setting of Hewitt and Manning (2019) to the multilingual language model MBERT. They train syntactic distance probes on 11 languages and compare UAS of induced trees in four scenarios: 1. training and evaluating on the same languages; 2. training on a single language, evaluating on a different one; 3. training on all languages except the evaluation one; 4. training on all languages, including the evaluation one. They demonstrate that the transfer is effective as the results in all the configurations outperform the baselines[3]. Even in the hardest case – zero-shot transfer from just one language, the result is at least 6.9 percent points above the baselines (for Chinese). Nevertheless, for all the languages, no transfer-learning setting can beat the training and evaluating a probe on the same language.

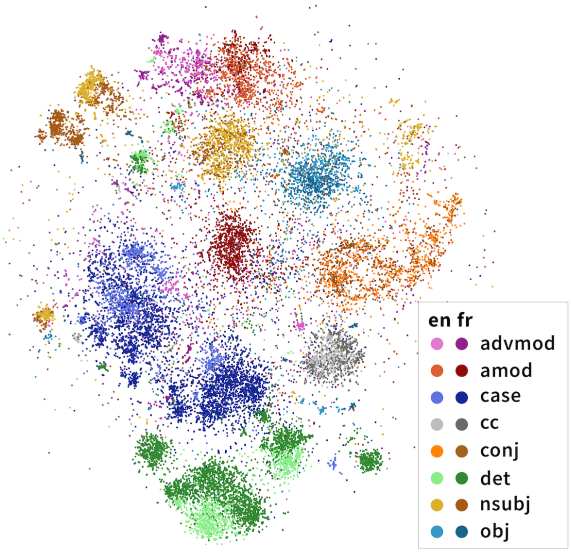

Figure 2: Two-dimensional t-SNE visualization of probed MBERT embeddings. Analysis of the clusters shows that embeddings encode information about the type of dependency relations and, to a lesser extent, language. Reprint from Chi et al. (2020).

The paper includes an analysis of intrinsic features of the BERT's vectors transformed by a probe. Noticeably, the vector differences between the representations of words connected by dependency relation are clustered by relation labels, see figure 2.

Multilingual BERTembeddings are also analyzed by Wang et al. (2019b). They show that even for the multilingual vectors, the results can be improved by projecting vector spaces across languages. They use a Biaffine Graph-based Parser by Dozat and Manning (2017), which consists of multiple RNN layers. Therefore, the experiment is not strictly comparable with probing as most of the syntactic information is captured by the parser and not by the embeddings. The article compares different types of vector representations fed as an input to the parser. It is demonstrated that cross-lingual transformation on MBERT embedding improves the results significantly in LAS of parser trained on English and evaluated on 14 languages; on average, from 60.53% to 63.54%. In comparison to other cross-lingual representations, the proposed method outperforms transformed static embeddings (FastText with SVD Bojanowski et al. (2017)) and also slightly outperforms contextual embeddings (XLM

[2] Tree distance is the length of the tree path between two tokens

[3] There are two baselines: right-branching tree and probing on randomly initialized MBERT without pretraining

Most weight over layers	Gravity center of layer attention	Linguistic features	Layer attention distribution shape
11–13	11.7	Part of speech	peaked
11–17	13.1	Constituency syntax	very peaked
12–17	13.8	Dependency syntax	very peaked
13–18	13.6	Semantic roles	peaked
14–20	13.2	Named entities	flat
13–22	12.7	Semantic proto-roles	very flat
15–22	12.8	Semantic relations	very flat
16–20	15.8	Coreference	peaked

Table 1: Distribution of linguistic features in the 24-layer BERT, interpreted from Tenney et al. (2019). For each feature type, we list an estimate of the range of layers on which it is captured significantly more than on other layers, together with the "center of gravity" of the layer attention, and a note how peaked or flat the distribution of the layer attention weights is. Reprint from Mareček et al. (2020)

Lample and Conneau (2019)).

3.4 Beyond Probing

Probing is by far the most popular method of analyzing the information encoded in pre-trained neural networks. However, it has been criticized for introducing too much supervision and increasing the risk that the knowledge for the task is learned by the probe and not retrieved from the underlying model (Hewitt and Liang, 2019).

There are methods of analyzing the network parameters that do not require supervision for a downstream task. One of them is the direct analysis of the patterns in Tranformer's attention matrices. Mareček and Rosa (2019) found that in some layer, all the words in a phrase of constituency tree are attended by its governor. Other works showed that in a few heads, high attention values correlated with the presence of dependency edges between the words (Voita et al., 2019; Clark et al., 2019; Vig and Belinkov, 2019).

3.5 Where is Linguistic Information Encoded?

It was observed that the encoding of linguistic features varies across layers. The study of Tenney et al. (2019) showed that starting from the input, subsequent layers of the network tend to capture aspects of language in a similar order as they appear in the traditional language processing pipeline (Manning et al., 2014). Table 1 shows the range of layers in which specific linguistic phenomenon was the most salient. Additionally, in other works, the authors also observed in Transformer models that typically, the initial layer captures morphological information (Belinkov et al., 2017b), the intermediate one encodes syntax (Blevins et al., 2018; Hewitt and Manning, 2019), and the later ones learn semantic information (Jawahar et al., 2019).

The recent state of knowledge in intractability is summarized in surveys on probing (Belinkov and Glass, 2019) and interpretation of BERT's representations (Rogers et al., 2020).

4 Fairness of NLP Models

The reliance on black-box models raises concerns about various biases acquired from raw scrapped Internet data.

4.1 Toxicity in Language Models

In text generation, biases are especially severe and raise questions about the reliability of deep learning systems. Gehman et al. (2020) have created REALTOXICITYPROMPTS, a dataset of 100 thousand prompts in English. The data can use to assess the Language Model's vulnerability to generate a rude, disrespectful response. Identification of toxic texts can be performed with commercially developed PERSPECTIVE API.[4] The authors of the API defines *toxicity* as "a rude, disrespectful, or unreasonable comment; likely to make people leave a discussion." The current challenge is to prevent the generation of toxic languages and improve the accuracy of toxicity identification. The recent approaches to solving the former issue include:

- Word filtering (Raffel et al., 2020). The approach requires the creation and maintenance of large vocabularies of disallowed words.

[4] https://www.perspectiveapi.com

Finnish source	English translation
Hän on siivooja.	She is a cleaner.
Hän on johtaja.	He is a leader.
Hän on varhaiskasvatuksen opettaja.	She is an early childhood education teacher.
Hän on presidentti.	He is the president.

Table 2: Sentences automatically translated from Finnish, which does not have grammatical gender, to English with widely used Google's Neural Machine Translation System (Wu et al., 2016). The predictions of gender in English correlate with stereotypical gender roles.

The choice of words is debatable because, in many cases, the context of a word is detrimental to toxicity.

- Additional phase of LM pre-training on non-toxic data (Gururangan et al., 2020; Keskar et al., 2019). The approach requires a robust method for non-toxic text selection and additional, computationally extensive epochs of pre-training.
- Altering inference algorithm to penalize toxic predictions of a model (Dathathri et al., 2020; Schick et al., 2021; Gehman et al., 2020).

All of the approaches are partially effective and display various disadvantages. Therefore, there is space for further improvement in the field.

4.2 Biases in Machine Translation Systems

Other problems are relevant for machine translation. The notion of grammatical gender is significantly different in languages. In some, it is prevalent (German, French, Czech, Polish, etc.), and other languages (Hungarian, English, etc.) do not denote gender. Gender is often mistranslated due to relying upon spurious correlations present in language corpora instead of a clear indication from the context.

Stanovsky et al. (2019) analyze gender bias of state-of-the-art academic MT systems in translation to eight languages with grammatical genders (Spanish, French, Italian, Russian, Ukrainian, Hebrew, Arabic, German). The challenge set follows Winograd Scheme principles (Winograd, 1972). Both works conclude that machine translation systems exhibit spurious correlations (gender bias), e.g., the word "doctor" is typically translated to male form and the word "nurse" to female form even if different gender is indicated in their contexts. Moreover, many systems perform better when translating male names of professions. Noticeably, the symptoms of gender bias can be observed in widely-used industrial MT systems (Table 2).

Comprehensive surveys on bias in NLP systems were recently conducted by Blodgett et al. (2020) and B et al. (2021).

5 Results

In my current and past research, I have focused on gaining and extending the knowledge on interpreting and analyzing linguistic information encoded in Transformer models, mainly BERT. The experiments were performed for English and other languages. The work up to date encompasses the motivations A, B, C, and partially D introduced in Section 1.1.

5.1 Universal Dependencies according to BERT

In Limisiewicz et al. (2020), we have focused on analyzing BERT's attention heads that were targeted on uncovering annotation of dependency syntax. For that purpose, we have used Universal Dependency annotation available in multiple languages (Nivre et al., 2020).

Confirming the previous observations, (Voita et al., 2019; Clark et al., 2019) we showed that attention in some heads is aligned with the places where dependency edges appear. The heads tend to be specialized in uncovering a specific type of dependency edges, as shown in Fig. 3. The novelty in our work was the method of grouping together the heads (called *head ensembles*) and averaging their attention matrices to find better alignment for specific dependency relation types. For that purpose, we required only small supervision to identify syntactic heads.

Furthermore, we extracted dependency trees from *head ensembles* following the method proposed by (Raganato and Tiedemann, 2018). We have surpassed original result (which extracted

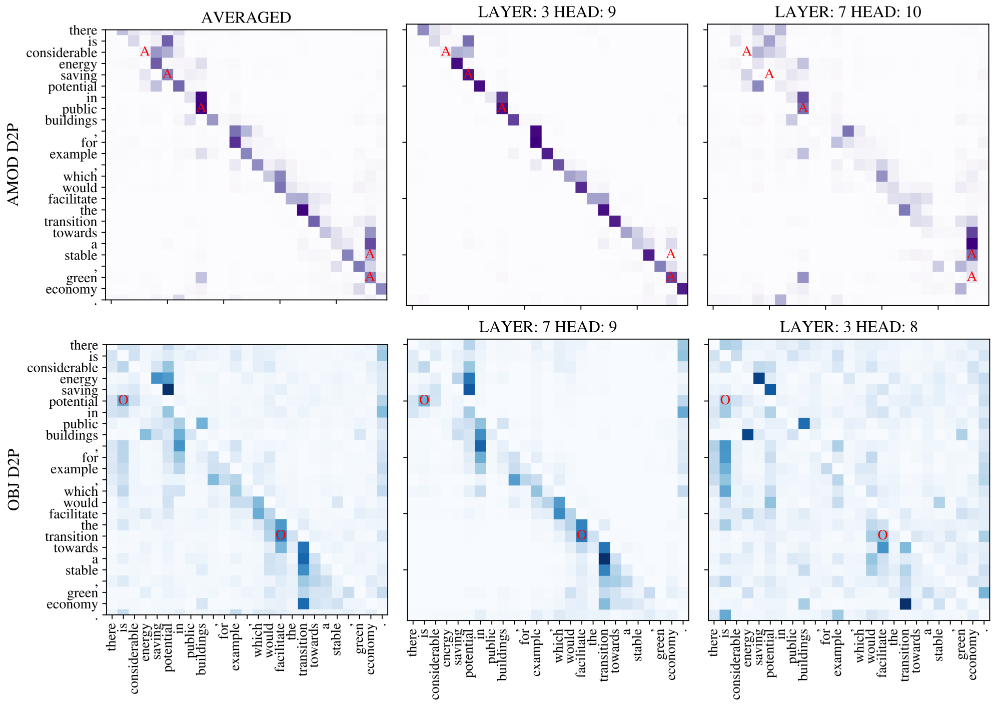

Figure 3: Examples of two BERT's attention heads covering the same relation label and their average. Gold relations are marked by red letters. In the top row (purple), both heads identify the parent noun for an adjectival modifier: Head 9 in Layer 3 if their distance is two positions or less, Head 10 in Layer 7 if they are further away (as in "a stable , green economy"). Similarly, for an object to predicate relation (blue bottom row), Head 9 in Layer 7 and Head 8 in Layer 3 capture pairs with shorter and longer positional distances, respectively. Reprint from Limisiewicz et al. (2020).

trees from a specific heads). We have repeated our analysis for mBERT in nine diverse languages.

We found out that there is no one-to-one correspondence between dependency labels and heads. Specific heads can capture multiple relation types, while many heads can also partially encode one dependency relation type. In a multilingual setting, we have observed that the same head can encode the same syntactic information even for typologically diverse languages.

The article was published in *The Findings of EMNLP 2020* and was presented at a co-located workshop: *BlackBoxNLP 2021*. It is aligned with motivations **A** and **B**.

5.2 Orthogonal Structural Probe

The work Limisiewicz and Mareček (2021b) is based on the *structural probe* method introduced by Hewitt and Manning (2019) described in Section 3.2. In our modification, we replace linear transformation with orthogonal transformation and then dimension-wise scaling of elements. The formulation is mathematically equivalent to the original one, but using the scaling coefficient allows us to analyze how important each dimension of the representation is to the task.

Moreover, we have proposed new structural probing tasks: hypernymy distance and depth in the WordNet tree (Miller, 1995), word position in the sentence. We have also evaluated how prone structural probes are to memorizing randomly generated structures.

The main outcome of the work was the possibility to identify where specific information is encoded in the network. Not only can we differentiate between different layers, but thanks to scaling, we could identify in each layer parts of the representation (or subspaces) relevant to particular tasks. The surprising finding was the fact that in most layers, subspaces encoding syntactic and lexical information were disjoint Fig. 5.

The article was published in *The Proceeding of ACL-IJCNLP 2021* and presented at the conference. The scope of work is related to the motivations **A** and **C**.

5.3 Multilingual Orthogonal Structural Probe

In the following work (Limisiewicz and Mareček, 2021a), we apply orthogonal structural probes to

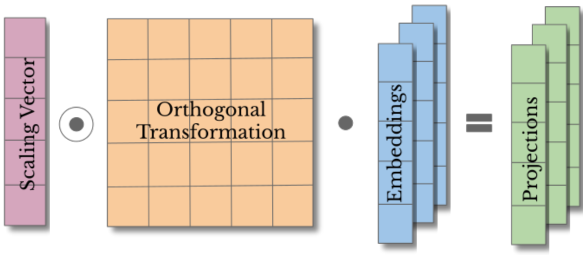

Figure 4: Schema of *orthogonal structural probes*. Embeddings are at first rotated by *Orthogonal Transformation* and element-wise multiplied by *Scaling Vector*. The matrix and the vector are gradient optimized.Reprint from Limisiewicz and Mareček (2021b).

MBERT's nine diverse language. We continue with analyzing syntactic dependency from UD (Nivre et al., 2020) and lexical hypernymy from WordNets (Bond and Foster, 2013) in many languages. The method uses orthogonal transformation that can be employed to aligning cross-lingual embeddings. The previous research showed that under some conditions when the embeddings are learned jointly, the embedding spaces are isomorphic across languages, hence we could align the representations with orthogonal transformation (Mikolov et al., 2013b; Vulić et al., 2020). Our experiments suggest that an orthogonal transformation can indeed map the representation. Furthermore, for English and languages typologically close to it, analyzed information is encoded in the same subspace, i.e., the rotation is not needed.

Moreover, we show that the *orthogonal structural probe* can be used in zero-shot dependency parsing with transfer learning from different languages.

The article is under review of *EMNLP 2021* and co-located workshop *BlackBoxNLP*.The research is aligned with the motivation **B** and **C**.

5.4 Gender Bias

In Kocmi et al. (2020), we have extended the gender bias evaluation of (Stanovsky et al., 2019) to new languages: Czech and Polish. The analysis showed that the symptoms of language bias are present in all of the analyzed translation systems submitted for the Workshop of Machine Translation. The models that perform well in automatic measures of translation proficiency - BLEU tend to rely more on the bias in the translation of profession. These findings underscore the importance of evaluating the biases in the models.

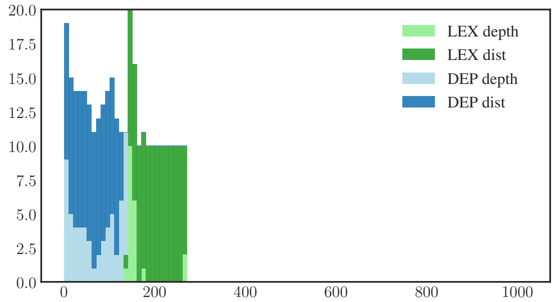

Figure 5: Histograms of dimensions selected by dependency and lexical *orthogonal structural probe* in the 16th layer of BERT. Each bin of the histogram corresponds to 10 coordinates. The height of a bar (in one color) represents how many were selected for a specific task. Reprint from Limisiewicz and Mareček (2021b).

The work is orthogonal to the previous ones, as it is not related to the interpretation and processing of the neural network's representation. The article was published in *The Proceedings of WMT 2020* and presented at the workshop. It is an initial approach to the motivation **D**.

5.5 Other Work

Besides the experimental results described in the previous chapters, I have thoroughly studied the area's literature. As a result, we have compiled the survey on the syntax representation: Limisiewicz and Mareček (2020). The article was published in *The Proceedings of ITAT 2020*. The excerpts from this work were used in the overview of background literature in Section 3.

I have also contributed to the book on interpretations of neural models Mareček et al. (2020), especially to chapters 5 and 6.

6 Future Work

The further work is a continuation of the research done so far. We will focus on the following aspects:

6.1 Uncovering and Modifying Information in Hidden Layers

Our aim is to identify and separate different kinds of linguistic information encoded in the form of subspaces of the representations. We will base our experiments on the new probing techniques (Torroba Hennigen et al., 2020; Limisiewicz and Mareček, 2021b) and work on improving them.

We will focus on semantic features in complex downstream tasks included in GLUE (Wang et al.,

2018) and SuperGLUE Wang et al. (2019a), such as natural language inference (NLI), question answering (QA), and paraphrasing. Such generalization will likely require further development of mentioned probing methods because the "semantic" tasks operate on the whole sentences instead of individual words, so averaging over the word vectors or more sophisticated techniques will be needed.

Representation of linguistic features is convoluted. One possible method for disentangling these word vectors is to find an orthogonal transformation (rotation of the vector space) that would separate the features from each other (Limisiewicz and Mareček, 2021b). This method was shown to separate some lexical and syntactic features (as described in Section 5.2). In further work, we plan to extend and adjust this approach to work with many other features.

Once we identify a subspace encoding a specific feature, we can easily regulate or suppress its activation and analyze how it affects system output. For example, we conjecture that filtering out the items related to semantics would result in generating sentences grammatically correct but nonsensical, similar to a famous example of Chomsky (1957) "Colorless green ideas sleep furiously.' Motivations: **A**, **C**.

6.2 Debasing and Filtering Out Unwanted Correlations in the Networks

We want to propose a real-world application of purely theoretical methods described in the previous part. We hypothesize that the transformation and the analysis of contextual word vectors in language models can reveal which components of these vectors are responsible for unwanted biases emerging from pre-training data. Filtering out those components could result in a new structural de-biasing technique.

We will mainly focus on effects in *autoregressive* language models and aim to eliminate the model's inclinations to generate sexist, racist, aggressive, vulgar, or in any other way toxic texts. We will approach the machine translation system analogically to examine the potential of reducing gender bias.

The evaluation of our methods will be conducted on REALTOXICITYPROMPTS (Gehman et al., 2020) for language modeling and WinoMT (Stanovsky et al., 2019) for machine translation or similar datasets. Motivation: **C**.

6.3 Analysis Beyond English

Our motivation is to show that analysis generalizes well to many languages. We will evaluate low-resource languages that were generally out of the scope of the previous studies. We hypothesize that a better understanding of the embeddings would facilitate the cross-lingual transfer. Our initial results show that the observations for English can be generalized for multiple languages (Limisiewicz and Mareček, 2021a). Therefore we will continue considering languages other than English in our further work. Motivation: **B**

7 Conclusion

We have summarised the related work and our work on the interpretation of neural networks. The field of research is rapidly growing. However, still many questions are left unanswered. Our work up to date gave initial answers to our question posed in the introduction:

A The linguistic features related to syntax, morphology, semantics are encoded in the pre-trained neural networks, even though these sources of information were not revealed during training.

B The representations of multilingual models vary across languages. However, for topologically close languages, the differences are minimal, and we can apply the same interpretation tool for their analysis.

C We have proposed *orthogonal structural probes* that is capable to divide the subspaces of the embedding spaces relevant to specific linguistic information.

D We worked on an evaluation metric that measures gender bias in machine translation systems. We have shown that the NMT systems are gender-biased.

We will continue the work on the mentioned topics. Specifically, we will improve the methods of disentangling the representation and extend their scope to other sources of information. We aim to find a way to enhance and diminish particular information in the embeddings. We hypothesize that such an approach will allow filtering out unwanted biases in the networks and benefit the practical applications of the networks.

References

Senthil Kumar B, Aravindan Chandrabose, and Bharathi Raja Chakravarthi. 2021. An overview of fairness in data – illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, Kyiv. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

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

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic tagging with deep residual networks. *arXiv preprint arXiv:1609.07053*.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *ArXiv*, abs/1611.01734.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *arXiv:1901.07291 [cs]*. ArXiv: 1901.07291.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*. ArXiv: 1909.11942.

Tomasz Limisiewicz, David Mareček, and Rudolf Rosa. 2020. Universal Dependencies According to BERT: Both More Specific and More General. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2710–2722, Online. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2020. Syntax representation in word embeddings and neural networks – a survey. In *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020)*, pages 38–48, Košice, Slovakia. Tomáš Horváth.

Tomasz Limisiewicz and David Mareček. 2021a. Examining cross-lingual contextual embeddings with orthogonal structural probes.

Tomasz Limisiewicz and David Mareček. 2021b. Introducing orthogonal constraint in structural probes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue*, 16(3):31–57.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *NAACL-HLT*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1901.11692 [cs]*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

David Mareček, Jindřich Libovický, Tomáš Musil, Rudolf Rosa, and Tomasz Limisiewicz. 2020. *Hidden in the Layers: Interpretation of Neural Networks for Natural Language Processing*, volume 20 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Prague, Czechia.

David Mareček and Rudolf Rosa. 2019. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Christoph Molnar. 2020. Interpretable Machine Learning.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2020. Facebook fair's wmt19 news translation task submission. In *Proc. of WMT*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alessandro Raganato and Jörg Tiedemann. 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*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. *arXiv preprint arXiv:1804.02472*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

Timo Schick, S. Udupa, and Hinrich Schutze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *ArXiv*, abs/2103.00453.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proc. of the First Conference on Machine Translation (WMT16)*, Berlin, Germany.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Adam Teichert, Adam Poliak, Benjamin Van Durme, and Matthew R. Gormley. 2017. Semantic protorole labeling. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 6000–6010, Long Beach, CA, USA. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*.

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

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019b. Cross-lingual bert transformation for zero-shot dependency parsing. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.