Unveiling Multilinguality in Transformer Models: Exploring Language Specificity in Feed-Forward Networks

Sunit Bhattacharya and Ondřej Bojar
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
(bhattacharya,bojar)@ufal.mff.cuni.cz

Abstract

Recent research suggests that the feed-forward module within Transformers can be viewed as a collection of key-value memories, where the keys learn to capture specific patterns from the input based on the training examples. The values then combine the output from the 'memories' of the keys to generate predictions about the next token. This leads to an incremental process of prediction that gradually converges towards the final token choice near the output layers.

This interesting perspective raises questions about how multilingual models might leverage this mechanism. Specifically, for autoregressive models trained on two or more languages, do all neurons (across layers) respond equally to all languages? No! Our hypothesis centers around the notion that during pretraining, certain model parameters learn strong language-specific features, while others learn more language-agnostic (shared across languages) features. To validate this, we conduct experiments utilizing parallel corpora of two languages that the model was initially pretrained on. Our findings reveal that the layers closest to the network's input or output tend to exhibit more language-specific behaviour compared to the layers in the middle.

1 Introduction

One of the least studied aspects of the Transformer (Vaswani et al., 2017) models in general and Large Language Models (LLMs) in particular is the feed-forward layers (FFNs). Although they contain almost two-thirds of the parameters, it is only recently[1] that their role in the working of the models is being seriously studied.

Geva et al. (2021, 2022) have earlier demonstrated that FFNs could be seen as "key-value memories" where each neuron (key)[2] in the lower sub-layer of the FFN gets triggered by specific patterns in the input data and the higher sub-layer (values) produces a distribution over the output vocabulary. This leads us to a perspective (Figure 1) where the FFN first captures certain patterns or concepts[3] in the input (conceptualization), selects the important aspects (using the activation function i.e. selection) and then combines them to emit an output which can be interpreted as a prediction of the possible next-word token for that layer, i.e. synthesis. To highlight this view throughout the rest of the paper, we will use the term *'detectors'* instead of the rather generic 'keys' to refer to the neurons in the earlier layer and *'combinators'* instead of 'values' to refer to the later layer. Repeating this across layers leads to a process of incremental prediction of the next token, with the prediction from previous layers being refined in the next layers (Belrose et al., 2023). This perspective however raises an important question. For models trained with a causal-language modeling objective in multilingual settings, what sort of patterns do the detectors encode across layers? More precisely, are some detectors triggered by input only from specific languages?

In this paper, we investigate this phenomenon of language specificity of the detectors in a multilingual model, pretrained on 30 languages from 16 language

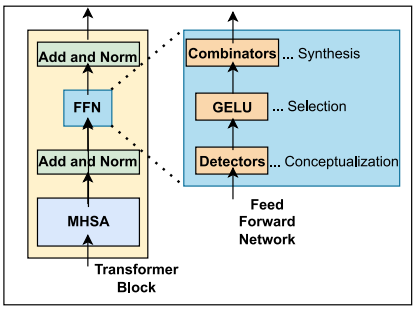

Figure 1: Transformer block and the structure of FFN

[1] Although the work by (Wang and Tu, 2020) is relevant in this regard, their analysis was done for all the components of the Transformer and not just the FFNs.

[2] While Geva et al. (2021) use the word *'keys'*, some other authors use the word *neuron* in this context.

[3] Shallow processing would require them to be good at capturing certain syntax patterns while semantic processing would require them to be good at capturing more thematic/conceptual patterns.

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 120–126

families. Earlier work has shown that Transformer models encode more shallow features in the earlier layers[4] while encoding more semantic features in the later layers[5] (Tenney et al., 2019). We hypothesise that the shallow processing would require more language-specific detectors than the semantic aspects of the input. And hence, we posit that during pretraining of the multilingual models, two kinds of neurons would emerge: **language-specific** and **language-agnostic**.

Thorough investigations into the role of the FFN layers in Transformer is an interesting research direction, and to our best knowledge, this is the first work that tries to look at the FFN[6] from the perspective of multilinguality. The rest of the paper is structured as follows: a brief discussion of the related works (Section 2) is followed by the description of the models and data (Section 3) and models (Section 4). This is followed by the presentation (Section 5) and simultaneous discussion of the results (Sections 6 and 7).

2 Related Work

Exploring the role and capabilities of the FFN sub-layer in Transformer models is a still nascent field of research with only a few papers exploring their working. As mentioned earlier, Geva et al. (2021, 2022) have proposed an interesting perspective of looking at how the FFN layer of the Transformer contributes during language generation. Recent work (Meng et al., 2022; Yao et al., 2022) exploring the capabilities of the FFN has also looked into how the activations of FFNs could be used for understanding how autoregressive models deal with facts. Other works (Li et al., 2022; Zhang et al., 2022) have analysed activation patterns in FFNs to study sparsity in Transformers. In other words, they show that only a few neurons in the FFNs are activated corresponding to inputs to Transformers.

On the front of studying multilingual models, Libovický et al. (2019) demonstrated that representations in encoder-only models can be split into language specific and language-neutral components. But to our best knowledge, no equivalent study has been done for autoregressive language models. Additionally, Deshpande et al. (2022); Blevins et al. (2022); Lauscher et al. (2020); Choudhury and Deshpande (2021); Kudugunta et al. (2019) have studied the pretraining behaviour and capabilities of various encoder-only multilingual models. More recently, Pfeiffer et al. (2022) demonstrated how separating parameters into language-specific modules during training can help improve the performance across languages.

From the perspective of studying multilinguality in the human brain, neuroimaging studies (Crinion et al., 2006; Videsott et al., 2010; Miozzo et al., 2010) have shown that although neural circuits for different languages are highly overlapping, there are distinct brain areas for language-specific processing and areas that are language-agnostic.

3 Model and testing data

We use a pretrained XGLM model (Lin et al., 2021) with 1.7 billion parameters, available on the Hugging Face (Wolf et al., 2019) repository[7] for our experiments.

We use sentences from the training data of the CzEng 2.0 corpus[8] (Kocmi et al., 2020) for our experiments. The model description of the XGLM model states that the model was trained on CommonCrawl data of various languages. CzEng heavily relies on various freely accessible web sources and a part of the data included in CzEng is also drawn from CommonCrawl among other sources. Thus, we expect that the sentences used for the experiments are of the same domain/style as the model was originally trained on, and they can even overlap. We do not consider such a possible overlap a serious problem for our analysis, because we are not measuring any processing performance or generalization capability.

4 Experiment

We first extract a sample of sentences from the CzEng corpus, giving us a set of Czech and English parallel sentences. We only select sentences with lengths between 20 and 50. We then feed the model with all 'prefixes' of the sampled sentences from both languages. In other words, for each sentence, we incrementally feed the model one subword at a time and record our observations. For instance, for a Czech sentence like "Tenhle úkol je obtížný" (This task is difficult), the prefixes fed to the model would be "Tenhle", "Tenhle úkol", "Tenhle úkol je" and "Tenhle úkol je obtížný". The parallel sentences ensure that the semantic contents of the sentences for the two languages are similar. We go on to collect the data about the model state corresponding to each prefix.

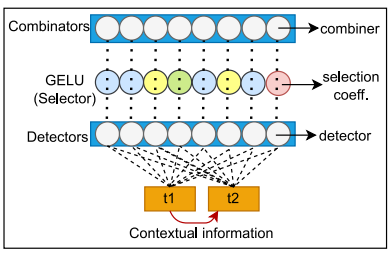

Figure 2: FFN in close detail

[4] close to the input
[5] near the output
[6] in a decoder-only Transformer model
[7] https://huggingface.co/facebook/xglm-1.7B
[8] https://ufal.mff.cuni.cz/czeng

From the collected data[9], we extract the "selection coefficients" corresponding to each prefix for all detectors across the layers of the model. Specifically, for detector d_i in layer L_j, we define the selection coefficient for a prefix p_k as:

$$C_{p_k}^{(L_j,d_i)} = GeLU\{d_i(p_k)\} \quad (1)$$

Thus, for each prefix we obtain layer-wise selection coefficients for the detectors (an example can be visualised in Table 1). We then sort the detectors based on the values of their corresponding selection coefficients. We posit that for a layer, certain detectors are triggered by specific prefix templates or languages. The selection coefficient is the indicator of the extent to which a particular detector is triggered by a prefix. Thus, observing the selection coefficients of the detectors across prefixes of different languages should indicate which (and how many) detectors are relevant bilingually and which (and how many) are relevant only for one of the two examined languages. We do this by analysing the top-k detectors after sorting the detectors by decreasing selection coefficients.

Table 1: Selection coefficients of m detectors in layer L for a total of n prefixes

Lang1, sent1, prefix_1	$C_{11}C_{12}C_{13}\ldots C_{1m}$
Lang1, sent1, prefix_2	$C_{21}C_{22}C_{23}\ldots C_{2m}$
⋮	⋮
Lang2, sentN, prefix_xx	$C_{k1}C_{k2}C_{k3}\ldots C_{km}$
Lang2, sentN, prefix_xy	$C_{n1}C_{n2}C_{n3}\ldots C_{nm}$

5 Observations

As an example, Table 2 shows the top-1 detector (detector with maximum selection coefficient) for the prefixes of an English and Czech sentence.

In the following sections, we present the results from our observations of the selection coefficients of detectors across the layers of the model.

5.1 Distribution of active detectors across layers

We collect the indices of the top-10 and top-100[10] detectors for each prefix. For a prefix P_i of all the considered prefixes $P_0, P_1, ..., P_n$, we denote the set of the top detectors D_i where $|D_i| = t$ (i.e. the set cardinality of $|D_i|$ is t). This way, we collect the list of the top t detectors for all prefixes in a layer. For each layer L_k, we obtain $L_k = D_0 \cup D_1 \cup ... \cup D_n$ and we plot the $|L_k|$

[9]from all sentences across Czech and English

[10]The top-10 list implies that we extract the list of the 10 detectors that had the maximum selection coefficients for a prefix. Similarly, for the top-100 list, we extract 100 detectors with the maximum selection coefficients.

Prefix	Detector
Europol	2149
Europol zpracovává	2149
Europol zpracovává a	3942
Europol zpracovává a předává	200
Europol zpracovává a předává údaje	200
Europol	2149
Europol shall	2149
Europol shall process	2149
Europol shall process and	3424
Europol shall process and transfer	2149

Table 2: Prefixes from an example Czech-English sentence pair, listing the most active detector ID (according to selection coefficients) from layer 1.

across the layers (e.g. Figure 3). In other words, we are checking how many unique detectors across prefixes belong to the list of 10 or 100 most active detectors for that layer. The fewer detectors in this set, the more "compact" the representation of these sentences are. The more detectors is in this set, the more "network capacity" is used when processing the given sentences. We make the plots for each of the two languages. Hence, using the example in Table 2: for layer 1, we have $L_1^{en} = (2149, 3424)$ and $L_1^{cs} = (2149, 3942, 200)$ and so $|L_1^{en}| = 2$ and $|L_1^{cs}| = 3$.

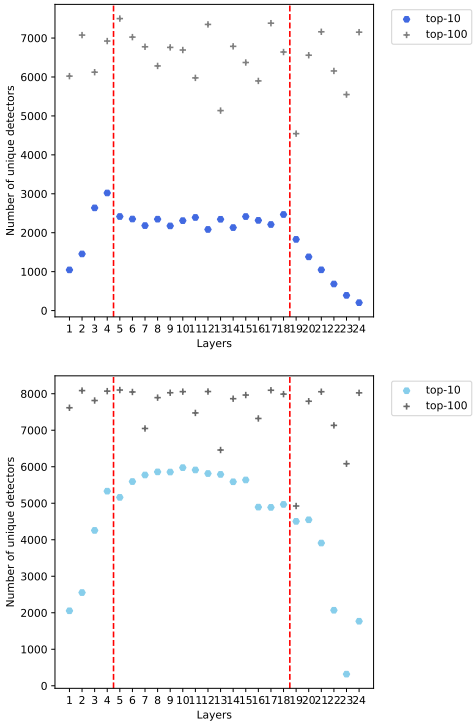

Figure 3: Number of top detectors ($|L_i|$) used across layers when processing Czech (top plot) and English (bottom plot) sentences.

Figure 3 shows that the top-100 list does not seem to show any pattern, unlike the top-10 list. We observe that for each prefix, only certain detectors exhibit high values of selection coefficient. Selecting the top-100 leads to the inclusion of many detectors that repeatedly appear across many prefixes with tiny values of selection coefficient. We reason that, this leads to the pattern seen with the top-10 list. We also posit that this is a callback to the previous research that has indicated that FFNs exhibit patterns of sparse activation.

The top-10 list shows that the number of detectors for both languages increases between layers 1 to 4 (near the input) and then decrease between layers 19 to 24 (near the output). Since this observation also includes detectors that get triggered for both languages[11], we analyse the number of detectors that are intersecting between the two languages (Czech and English). That is, for each layer L_k, we identify the intersecting detectors $I_k = L_i^{cs} \cap L_i^{en}$. In other words, we examine how the number of keys getting triggered by both English and Czech prefixes (multilingual detectors) vary across the layers.

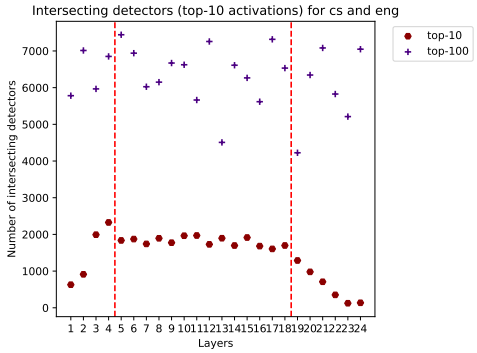

Figure 4: Distribution of multilingual detectors (intersecting detectors)

As Figure 4 shows, the number of intersecting detectors also follows the same pattern as observed in Figure 3. The number starts increasing in the layers near the input and decrease near the output. It may be argued that the spike in the number of unique detectors (for individual languages) in the middle layers might imply that the number of intersecting detectors would also increase in the middle layers. However, we argue that it might not always be the case. We validate our argument in the following sections.

To look at the language specific responses of the detectors across the layers, we look at the set difference of the detectors seen in, Figure 3 i.e. the language-specific detectors. So, for some layer k, we analyse $en_k = L_k^{en} \setminus L_i^{cs}$ and[12] $cs_k = L_i^{cs} \setminus L_i^{en}$. From the results in Figure 5, we see that there is a steady drop in the number of Czech-specific detectors in the middle layers. No such effect is seen for English. Also, across all the results presented here, we note that the observed number of detectors getting triggered by English prefixes is considerably higher than that of Czech prefixes.

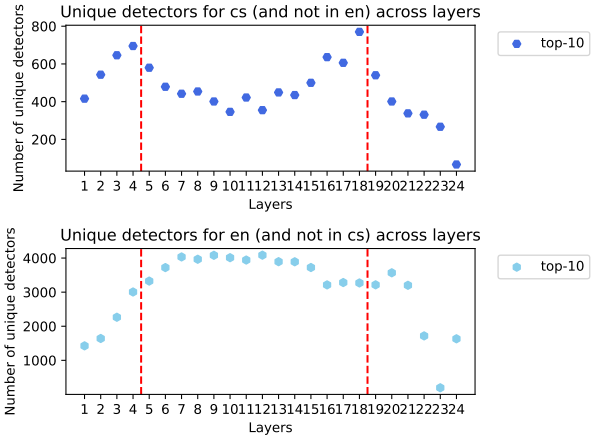

Figure 5: Distribution of language specific detectors

Next, we determine to what extent the actual language can be identified from the detector activity.

5.2 Layers close to the input and output are language specific

To confirm the existence of language-specific detectors, we train a linear classifier over all the detectors for each layer. The task of the classifier is to use the selection coefficients to determine if the given prefix was in English or Czech. The results from the experiment are shown in Figure 6. In the plot, we show the number of detectors across different performance brackets. Each series shows the number of detectors classifying with an accuracy of $>= k\%$.

We see that for performance brackets $< 80\%$, the layer closer to the input shows the highest accuracy in predicting the language. Again for slabs, $> 70\%$ we see that the accuracy increases in the last few layers. Thus, we conclude that layers closer to the input and output are more language-specific than the others.

6 Discussion

We started with the hypothesis that language-specific detectors would be more common in the layers closer to the input and output. We analysed the detectors across the layers using sentences from a Czech-English parallel corpus. We note that in the underlying XGLM model, English (with 803,527 million training tokens) was much more dominant than Czech (with 8,616 million training tokens) (Lin et al., 2021). We thus consider the model to be a primarily English model that

[11] for example, detector 2149 in the example shown in Table 2

[12] From the example in Table 2, $en_k = 3424$ and $cs_k = 3942, 200$

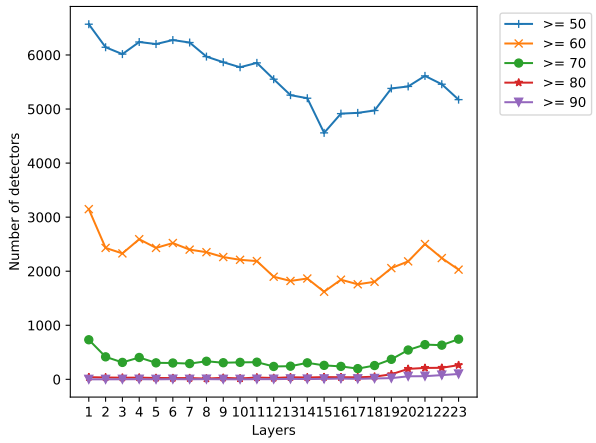

Figure 6: Classification percentages across layers. The colour indicates the reached accuracy level of the prediction.

saw some Czech sentences during pretraining. From the results, we observe that the layers closer to the input and output indeed perform more language specific processing than others. We also see that considerably lower number of detectors are triggered by the Czech prefixes than English prefixes, probably reflecting the data imbalance during training. While looking at the behaviour of Czech-specific detectors, we find that their numbers drop near the middle layers (8-15). We know that the model is primarily English centric. And since it is well known that higher-layers of Transformers are involved in more semantic processing, it is likely that the model uses more language-agnostic detectors and only a few Czech-specific detectors for processing semantic aspects of the input. Studies with humans have previously shown that semantic processing in humans is often language-agnostic. We thus see a possible way to connect these observations in the future.

From a different perspective, the analysis of the selection coefficients also agrees with the recent theories and observations about the sparse nature of FFN modules. We hypothesise that the sparsity (lesser numbers of unique detectors) might be an indicator of shallow processing and density might be an indicator of semantic processing. The sparsity argument might also be extended to claim that only a subset of detectors are required for language specific processing while greater numbers of detectors are required for more language-agnostic (i.e. semantic) However, such claims warrant extensive experimentation that we wish to conduct as a followup to this work.

7 Conclusion

In this study, we focused on the analysis of the Feed Forward Layers (FFNs) of a pretrained multilingual Transformer model. We look at the FFNs as a system that first identifies patterns in the input representations (detector), selects the relevant information (selector), and then combines it to make a guess of the next token (combiner). We assess the degree of language specificity of the detectors in this multilingual model with two experiments. We observe that there are greater number of language specific detectors near the input and output of the model. Additionally, we observe how data imbalance during training is reflected in the behaviour of the multilingual detectors. We also try to link our observations with recent studies on the sparse activations in FFNs. Overall, our findings shed light on the language specificity of FFNs in multilingual models.

Limitations

While our analysis provides valuable insights into the behaviour of "detectors" in a multilingual Transformer model's Feed Forward Layers (FFNs), there is an important limitation to consider. Our analysis is limited to only the XGLM model. This work does not consider the multilingual dynamics of other models. Also, our study is centred on the Czech-English language pair. Different languages exhibit diverse linguistic characteristics and complexities, and the behaviour of detectors could vary significantly across various language pairs. Extrapolating our findings to multilingual behaviour involving other languages requires caution and further investigation. Further, while we categorize detectors as language-specific or multilingual based on their activation patterns, the specific linguistic cues that trigger their activation remain complex and challenging to interpret. Our study focuses on the quantitative aspects of detector behaviour, and a deeper qualitative analysis of the linguistic information captured by these detectors could provide additional insights.

Ethics Statement

As the work is dedicated to evaluating existing models on publicly available datasets, we are not aware of any potential ethical issues or negative impacts.

Future Work

We wish to extend this work and test the generalizability of our hypothesis across more language pairs and other multilingual autoregressive language models.

8 Acknowledgements

This work has been funded from the 19-26934X (NEUREM3) grant of the Czech Science Foundation and the grant 205-09/260698 (SVV) of Charles University. The work has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 (LINDAT/CLARIAH-CZ).

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models. *arXiv preprint arXiv:2205.11758*.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12710–12718.

Jenny Crinion, Robert Turner, Alice Grogan, Takashi Hanakawa, Uta Noppeney, Joseph T Devlin, Toshihiko Aso, Shinichi Urayama, Hidenao Fukuyama, Katharine Stockton, et al. 2006. Language control in the bilingual brain. *Science*, 312(5779):1537–1540.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is bert multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, et al. 2022. Large models are parsimonious learners: Activation sparsity in trained transformers. *arXiv preprint arXiv:2210.06313*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Michele Miozzo, Albert Costa, Mireia Hernandez, and Brenda Rapp. 2010. Lexical processing in the bilingual brain: Evidence from grammatical/morphological deficits. *Aphasiology*, 24(2):262–287.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Gerda Videsott, Bärbel Herrnberger, Klaus Hoenig, Edgar Schilly, Jo Grothe, Werner Wiater, Manfred Spitzer, and Markus Kiefer. 2010. Speaking in multiple languages: Neural correlates of language proficiency in multilingual word production. *Brain and language*, 113(3):103–112.

Wenxuan Wang and Zhaopeng Tu. 2020. Rethinking the value of transformer components. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6019–6029.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yunzhi Yao, Shaohan Huang, Li Dong, Furu Wei, Huajun Chen, and Ningyu Zhang. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pages 131–143. Springer.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Moefication: Transformer feed-forward layers are mixtures of experts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 877–890.