Attention Can Reflect Syntactic Structure
(If You Let It)

***Vinit Ravishankar**[†] ***Artur Kulmizev**[‡] **Mostafa Abdou**[§]
Anders Søgaard[§] **Joakim Nivre**[‡]

[†] Language Technology Group, Department of Informatics, University of Oslo
[‡] Department of Linguistics and Philology, Uppsala University
[§] Department of Computer Science, University of Copenhagen

[†] vinitr@ifi.uio.no
[‡] {artur.kulmizev,joakim.nivre}@lingfil.uu.se
[§] {abdou,soegaard}@di.ku.dk

Abstract

Since the popularization of the Transformer as a general-purpose feature encoder for NLP, many studies have attempted to decode linguistic structure from its novel multi-head attention mechanism. However, much of such work focused almost exclusively on English — a language with rigid word order and a lack of inflectional morphology. In this study, we present decoding experiments for multilingual BERT across 18 languages in order to test the generalizability of the claim that dependency syntax is reflected in attention patterns. We show that full trees can be decoded above baseline accuracy from single attention heads, and that individual relations are often tracked by the same heads across languages. Furthermore, in an attempt to address recent debates about the status of attention as an explanatory mechanism, we experiment with fine-tuning mBERT on a supervised parsing objective while freezing different series of parameters. Interestingly, in steering the objective to learn explicit linguistic structure, we find much of the same structure represented in the resulting attention patterns, with interesting differences with respect to which parameters are frozen.

1 Introduction

In recent years, the attention mechanism proposed by Bahdanau et al. (2014) has become an indispensable component of many NLP systems. Its widespread adoption was, in part, heralded by the introduction of the Transformer architecture (Vaswani et al., 2017a), which constrains a soft alignment to be learned across discrete states in the input (self-attention), rather than across input and output (e.g., Xu et al., 2015; Rocktäschel et al., 2015). The Transformer has, by now, supplanted the popular LSTM (Hochreiter and Schmidhuber, 1997) as NLP's feature-encoder-of-choice, largely due to its compatibility with parallelized training regimes and ability to handle long-distance dependencies.

Certainly, the nature of attention as a distribution over tokens lends itself to a straightforward interpretation of a model's inner workings. Bahdanau et al. (2014) illustrate this nicely in the context of `seq2seq` machine translation, showing that the attention learned by their models reflects expected cross-lingual idiosyncrasies between English and French, e.g., concerning word order. With self-attentive Transformers, interpretation becomes slightly more difficult, as attention is distributed across words within the input itself. This is further compounded by the use of multiple layers and heads, each combination of which yields its own alignment, representing a different (possibly redundant) view of the data. Given the similarity of such attention matrices to the score matrices employed in arc-factored dependency parsing (McDonald et al., 2005a,b), a salient question concerning interpretability becomes: Can we expect some combination of these parameters to capture linguistic structure in the form of a dependency tree, especially if the model performs well on NLP tasks? If not, can we relax the expectation and examine the extent to which subcomponents of the linguistic structure, such as subject-verb relations, are represented? This prospect was first posed by Raganato et al. (2018) for MT encoders, and later explored by Clark et al. (2019) for BERT. Ultimately, the consensus of these and other studies (Voita et al., 2019; Htut et al., 2019; Limisiewicz et al., 2020) was that, while there appears to exist no "generalist" head responsible for extracting full dependency structures, standalone heads often specialize in capturing individual grammatical relations.

Unfortunately, most of such studies focused their

*Equal contribution. Order was decided by a coin toss.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3031–3045
April 19 - 23, 2021.

experiments entirely on English, which is typologically favored to succeed in such scenarios due to its rigid word order and lack of inflectional morphology. It remains to be seen whether the attention patterns of such models can capture structural features across typologically diverse languages, or if the reported experiments on English are a misrepresentation of local positional heuristics as such. Furthermore, though previous work has investigated how attention patterns might change after fine-tuning on different tasks (Htut et al., 2019), a recent debate about attention as an explanatory mechanism (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) has cast the entire enterprise in doubt. Indeed, it remains to be seen whether fine-tuning on an explicit structured prediction task, e.g. dependency parsing, can force attention to represent the structure being learned, or if the patterns observed in pretrained models are not altered in any meaningful way.

To address these issues, we investigate the prospect of extracting linguistic structure from the attention weights of multilingual Transformer-based language models. In light of the surveyed literature, our research questions are as follows:

1. Can we decode dependency trees for some languages better than others?
2. Do the same layer–head combinations track the same relations across languages?
3. How do attention patterns change after fine-tuning with explicit syntactic annotation?
4. Which components of the model are involved in these changes?

In answering these questions, we believe we can shed further light on the (cross-)linguistic properties of Transformer-based language models, as well as address the question of attention patterns being a reliable representation of linguistic structure.

2 Attention as Structure

Transformers The focus of the present study is mBERT, a multilingual variant of the exceedingly popular language model (Devlin et al., 2019). BERT is built upon the Transformer architecture (Vaswani et al., 2017b), which is a self-attention-based encoder-decoder model (though only the encoder is relevant to our purposes). A Transformer takes a sequence of vectors $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_n}]$ as input and applies a positional encoding to them, in order to retain the order of words in a sentence. These inputs are then transformed into query (Q), key (K), and value (V) vectors via three separate linear transformations and passed to an attention mechanism. A single attention head computes scaled dot-product attention between K and Q, outputting a weighted sum of V:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

For multihead attention (MHA), the same process is repeated for k heads, allowing the model to jointly attend to information from different representation subspaces at different positions (Vaswani et al., 2017b). Ultimately, the output of all heads is concatenated and passed through a linear projection W^O:

$$H_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

$$\text{MHA}(Q, K, V) = \text{concat}(H_1, H_2, ..., H_k)W^O \quad (3)$$

Every layer also consists of a feed-forward network (FFN), consisting of two Dense layers with ReLU activation functions. For each layer, therefore, the output of MHA is passed through a LayerNorm with residual connections, passed through FFN, and then through another LayerNorm with residual connections.

Searching for structure Often, the line of inquiry regarding interpretability in NLP has been concerned with extracting and analyzing linguistic information from neural network models of language (Belinkov and Glass, 2019). Recently, such investigations have targeted Transformer models (Hewitt and Manning, 2019; Rosa and Mareček, 2019; Tenney et al., 2019), at least in part because the self-attention mechanism employed by these models offers a possible window into their inner workings. With large-scale machine translation and language models being openly distributed for experimentation, several researchers have wondered if self-attention is capable of representing syntactic structure, despite not being trained with any overt parsing objective.

In pursuit of this question, Raganato et al. (2018) applied a maximum-spanning-tree algorithm over the attention weights of several trained MT models, comparing them with gold trees from Universal Dependencies (Nivre et al., 2016, 2020). They found that, while the accuracy was not comparable to that of a supervised parser, it was nonetheless higher than several strong baselines, implying that some

structure was consistently represented. Clark et al. (2019) corroborated the same findings for BERT when decoding full trees, but observed that individual dependency relations were often tracked by specialized heads and were decodable with much higher accuracy than some fixed-offset baselines. Concurrently, Voita et al. (2019) made a similar observation about heads specializing in specific dependency relations, proposing a coarse taxonomy of head attention functions: *positional*, where heads attend to adjacent tokens; *syntactic*, where heads attend to specific syntactic relations; and *rare words*, where heads point to the least frequent tokens in the sentence. Htut et al. (2019) followed Raganato et al. (2018) in decoding dependency trees from BERT-based models, finding that fine-tuning on two classification tasks did not produce syntactically plausible attention patterns. Lastly, Limisiewicz et al. (2020) modified UD annotation to better represent attention patterns and introduced a supervised head-ensembling method for consolidating shared syntactic information across heads.

Does attention have explanatory value? Though many studies have yielded insight about how attention behaves in a variety of models, the question of whether it can be seen as a "faithful" explanation of model predictions has been subject to much recent debate. For example, Jain and Wallace (2019) present compelling arguments that attention does not offer a faithful explanation of predictions. Primarily, they demonstrate that there is little correlation between standard feature importance measures and attention weights. Furthermore, they contend that there exist *counterfactual* attention distributions, which are substantially different from learned attention weights but that do not alter a model's predictions. Using a similar methodology, Serrano and Smith (2019) corroborate that attention does not provide an adequate account of an input component's importance.

In response to these findings, Wiegreffe and Pinter (2019) question the assumptions underlying such claims. Attention, they argue, is not a *primitive*, i.e., it cannot be detached from the rest of a model's components as is done in the experiments of Jain and Wallace (2019). They propose a set of four analyses to test whether a given model's attention mechanism can provide meaningful explanation and demonstrate that the alternative attention distributions found via adversarial training methods do, in fact, perform poorly compared to standard attention mechanisms. On a theoretical level, they argue that, although attention weights do not give an *exclusive* "faithful" explanation, they do provide a meaningful *plausible* explanation.

This discussion is relevant to our study because it remains unclear whether or not attending to syntactic structure serves, in practice, as plausible explanation for model behavior, or whether or not it is even capable of serving as such. Indeed, the studies of Raganato et al. (2018) and Clark et al. (2019) relate a convincing but incomplete picture — tree decoding accuracy just marginally exceeds baselines and various relations tend to be tracked across varying heads and layers. Thus, our fine-tuning experiments (detailed in the following section) serve to enable an "easy" setting wherein we explicitly inform our models of the same structure that we are trying to extract. We posit that, if, after fine-tuning, syntactic structures were still *not* decodable from the attention weights, one could safely conclude that these structures are being stored via a non-transparent mechanism that may not even involve attention weights. Such an insight would allow us to conclude that attention weights cannot provide even a plausible explanation for models relying on syntax.

3 Experimental Design

To examine the extent to which we can decode dependency trees from attention patterns, we run a tree decoding algorithm over mBERT's attention heads — before and after fine-tuning via a parsing objective. We surmise that doing so will enable us to determine if attention can be interpreted as a reliable mechanism for capturing linguistic structure.

3.1 Model

We employ mBERT[1] in our experiments, which has been shown to perform well across a variety of NLP tasks (Hu et al., 2020; Kondratyuk and Straka, 2019a) and capture aspects of syntactic structure cross-lingually (Pires et al., 2019; Chi et al., 2020). mBERT features 12 layers with 768 hidden units and 12 attention heads, with a joint WordPiece sub-word vocabulary across languages. The model was trained on the concatenation of WikiDumps for the top 104 languages with the largest Wikipedias,where principled sampling was employed to enforce a balance between high- and

[1] https://github.com/google-research/bert

low-resource languages.

3.2 Decoding Algorithm

For decoding dependency trees, we follow Raganato et al. (2018) in applying the Chu-Liu-Edmonds maximum spanning tree algorithm (Chu, 1965) to every layer/head combination available in mBERT ($12 \times 12 = 144$ in total). In order for the matrices to correspond to gold treebank tokenization, we remove the cells corresponding to the BERT delimiter tokens (`[CLS]` and `[SEP]`). In addition to this, we sum the columns and average the rows corresponding to the constituent subwords of gold tokens, respectively (Clark et al., 2019). Lastly, since attention patterns across heads may differ in whether they represent heads attending to their dependents or vice versa, we take our input to be the element-wise product of a given attention matrix and its transpose ($A \circ A^\top$). We liken this to the joint probability of a head attending to its dependent and a dependent attending to its head, similarly to Limisiewicz et al. (2020). Per this point, we also follow Htut et al. (2019) in evaluating the decoded trees via Undirected Unlabeled Attachment Score (UUAS) — the percentage of undirected edges recovered correctly. Since we discount directionality, this is effectively a less strict measure than UAS, but one that has a long tradition in unsupervised dependency parsing since Klein and Manning (2004).

3.3 Data

For our data, we employ the Parallel Universal Dependencies (PUD) treebanks, as collected in UD v2.4 (Nivre et al., 2019). PUD was first released as part of the CONLL 2017 shared task (Zeman et al., 2018), containing 1000 parallel sentences, which were (professionally) translated from English, German, French, Italian, and Spanish to 14 other languages. The sentences are taken from two domains, **news** and **wikipedia**, the latter implying some overlap with mBERT's training data (though we did not investigate this). We include all PUD treebanks except Thai.[2]

3.4 Fine-Tuning Details

In addition to exploring pretrained mBERT's attention weights, we are also interested in how attention might be guided by a training objective that learns the exact tree structure we aim to decode. To this end, we employ the graph-based decoding algorithm of the biaffine parser introduced by Dozat and Manning (2016). We replace the standard BiLSTM encoder for this parser with the entire mBERT network, which we fine-tune with the parsing loss. The full parser decoder consists of four dense layers, two for head/child representations for dependency arcs (dim. 500) and two for head/child representations for dependency labels (dim. 100). These are transformed into the label space via a bilinear transform.

After training the parser, we can decode the fine-tuned mBERT parameters in the same fashion as described in Section 3.2. We surmise that, if attention heads are capable of tracking hierarchical relations between words in any capacity, it is precisely in this setting that this ability would be attested. In addition to this, we are interested in what individual *components* of the mBERT network are capable of steering attention patterns towards syntactic structure. We believe that addressing this question will help us not only in interpreting decisions made by BERT-based neural parsers, but also in aiding us developing syntax-aware models in general (Strubell et al., 2018; Swayamdipta et al., 2018). As such — beyond fine-tuning all parameters of the mBERT network (our basic setting) — we perform a series of ablation experiments wherein we update only one set of parameters per training cycle, e.g. the Query weights W_i^Q, and leave everything else frozen. This gives us a set of 6 models, which are described below. For each model, all non-BERT parser components are always left unfrozen.

- KEY: only the K components of the transformer are unfrozen; these are the representations of tokens that are paying attention *to* other tokens.
- QUERY: only the Q components are unfrozen; these, conversely, are the representations of tokens being paid attention to.
- KQ: both keys and queries are unfrozen.
- VALUE: semantic value vectors per token (V) are unfrozen; they are composed after being weighted with attention scores obtained from the K/Q matrices.
- DENSE: the dense feed-forward networks in the attention mechanism; all three per layer are unfrozen.
- NONE: The basic setting with nothing frozen; all parameters are updated with the parsing

[2] Thai is the only treebank that does not have a non-PUD treebank available in UD, which we need for our fine-tuning experiments.

	AR	CS	DE	EN	ES	FI	FR	HI	ID	IT	JA	KO	PL	PT	RU	SV	TR	ZH
BASELINE	50	40	36	36	40	42	40	46	47	40	43	55	45	41	42	39	52	41
PRE	53 7-6	53 10-8	49 10-8	47 10-8	50 9-5	48 10-8	41 2-3	48 2-3	50 9-5	41 6-4	**45** 2-3	64 9-2	52 10-8	50 9-5	51 10-8	51 10-8	55 3-8	42 2-3
NONE	**76** 11-10	**78** 11-10	**76** 11-10	**71** 10-11	**77** 10-11	**66** 10-11	**45** 11-10	**72** 11-10	**75** 11-10	**58** 11-10	42 11-10	64 11-10	**75** 11-10	**76** 11-10	**75** 10-8	**74** 10-8	55 3-8	38 2-3
KEY	62 10-8	64 10-8	58 11-12	53 10-8	59 11-12	56 10-8	41 7-12	54 10-8	59 10-8	47 9-2	44 2-3	62 10-8	64 10-8	58 11-12	61 10-8	59 12-10	55 3-12	41 2-3
QUERY	69 11-4	74 10-8	70 11-4	66 11-4	73 11-4	63 10-8	42 11-4	62 11-4	67 11-4	54 11-4	**45** 2-3	65 10-8	72 11-4	70 11-4	70 10-8	68 11-4	56 10-8	42 2-3
KQ	71 11-4	76 11-4	70 11-4	65 11-4	74 11-4	62 11-4	43 10-11	64 11-4	69 11-4	55 11-4	44 2-3	64 11-4	73 11-4	73 11-4	69 11-4	69 11-4	55 11-4	41 2-3
VALUE	75 12-5	72 12-5	72 12-5	64 12-5	76 12-5	59 12-5	**45** 12-5	63 12-5	73 12-5	55 12-5	**45** 2-3	**66** 10-8	73 12-5	74 12-5	69 12-5	65 12-5	**57** 12-5	**42** 3-8
DENSE	68 11-10	71 11-10	65 11-10	60 10-8	67 12-10	61 11-10	42 10-8	65 11-10	66 11-10	49 9-5	44 3-12	64 11-10	70 11-10	64 12-5	67 11-10	64 11-10	55 11-10	40 3-12

Table 1: Adjacent-branching baseline and maximum UUAS decoding accuracy per PUD treebank, expressed as best score and best layer/head combination for UUAS decoding. PRE refers to basic mBERT model before fine-tuning, while all cells below correspond different fine-tuned models described in Section 3.4. Best score indicated in **bold**.

loss.

We fine-tune each of these models on a concatenation of all PUD treebanks for 20 epochs, which effectively makes our model multilingual. We do so in order to 1) control for domain and annotation confounds, since all PUD sentences are parallel and are natively annotated (unlike converted UD treebanks, for instance); 2) increase the number of training samples for fine-tuning, as each PUD treebank features only 1000 sentences; and 3) induce a better parser through multilinguality, as in Kondratyuk and Straka (2019b). Furthermore, in order to gauge the overall performance of our parser across all ablated settings, we evaluate on the test set of the largest non-PUD treebank available for each language, since PUD only features test partitions. When training, we employ a combined dense/sparse Adam optimiser, at a learning rate of $3 * 10^{-5}$. We rescale gradients to have a maximum norm of 5.

4 Decoding mBERT Attention

The second row of Table 1 (PRE) depicts the UUAS after running our decoding algorithm over mBERT attention matrices, per language. We see a familiar pattern to that in Clark et al. (2019) among others — namely that attention patterns extracted directly from mBERT appear to be incapable of decoding dependency trees beyond a threshold of 50–60% UUAS accuracy. However, we also note that, in all languages, the attention-decoding algorithm outperforms a BASELINE (row 1) that draws an (undirected) edge between any two adjacent words in linear order, which implies that some non-

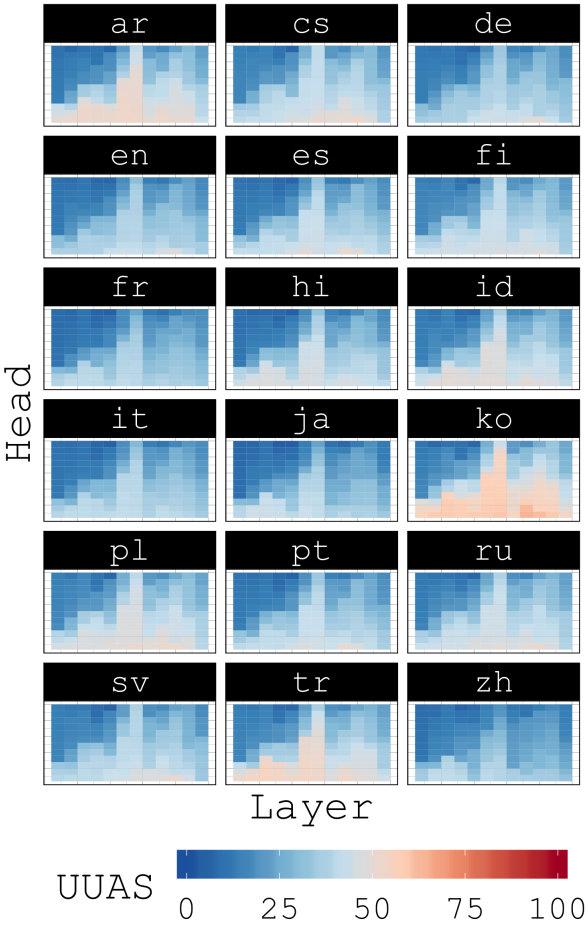

Figure 1: UUAS of MST decoding per layer and head, across languages. Heads (y-axis) are sorted by accuracy for easier visualization.

linear structures are captured with regularity. Indeed, head 8 in layer 10 appears to be particularly strong in this regard, returning the highest UUAS for 7 languages. Interestingly, the accuracy patterns

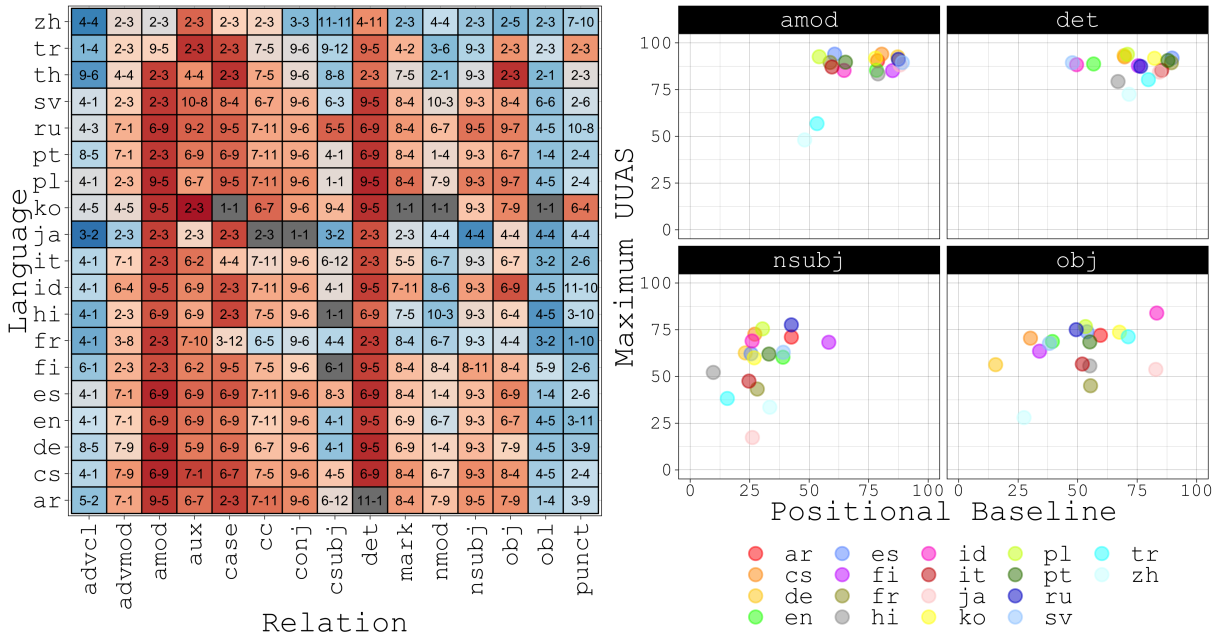

Figure 2: Left: UUAS per relation across languages (best layer/head combination indicated in cell). Right: Best UUAS as a function of best positional baseline (derived from the treebank), selected relations.

across layers depicted in Figure 1 tend to follow an identical trend for all languages, with nearly all heads in layer 7 returning high within-language accuracies.

It appears that attention for some languages (Arabic, Czech, Korean, Turkish) is comparatively easier to decode than others (French, Italian, Japanese, Chinese). A possible explanation for this result is that dependency relations between content words, which are favored by the UD annotation, are more likely to be adjacent in the morphologically rich languages of the first group (without intervening function words). This assumption seems to be corroborated by the high baseline scores for Arabic, Korean and Turkish (but not Czech). Conversely, the low baselines scores and the likewise low decoding accuracies for the latter four languages are difficult to characterize. Indeed, we could not identify what factors — typological, annotation, tokenization or otherwise — would set French and Italian apart from the remaining languages in terms of score. However, we hypothesize that the tokenization and our treatment of subword tokens plays a part in attempting to decode attention from Chinese and Japanese representations. Per the mBERT documentation,[3] Chinese and Japanese Kanji character spans within the CJK Unicode range are character-tokenized. This lies in contrast with all other languages (Korean Hangul and Japanese Hiragana and Katakana included), which rely on whitespace and WordPiece (Wu et al., 2016). It is thus possible that the attention distributions for these two languages (at least where CJK characters are relevant) are devoted to composing words, rather than structural relations, which will distort the attention matrices that we compute to correspond with gold tokenization (e.g. by maxing rows and averaging columns).

Relation analysis We can disambiguate what sort of structures are captured with regularity by looking at the UUAS returned per dependency relation. Figure 2 (left) shows that adjectival modifiers (`amod`, mean UUAS = 85 $\pm$12) and determiners (`det`, 88 $\pm$ 6) are among the easiest relations to decode across languages. Indeed, words that are connected by these relations are often adjacent to each other and may be simple to decode if a head is primarily concerned with tracking linear order. To verify the extent to which this might be happening, we plot the aforementioned decoding accuracy as a function of select relations' positional baselines in Figure 2 (right). The positional baselines, in this case, are calculated by picking the most frequent offset at which a dependent occurs with respect to its head, e.g., -1 for `det` in English, meaning one position to the left of the head. Interestingly, while we observe significant variation across the positional baselines for `amod` and `det`, the decoding

[3] https://github.com/google-research/bert/blob/master/multilingual.md

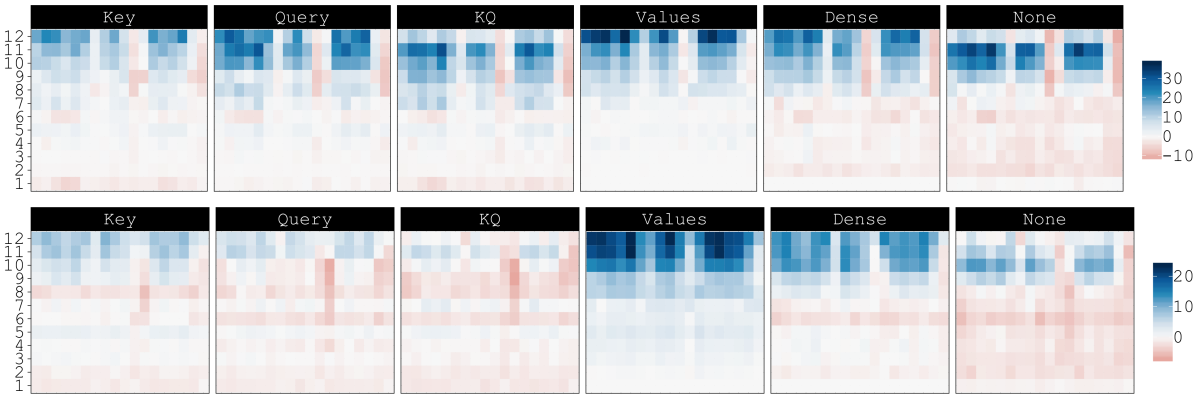

Figure 3: (Top) best scores across all heads, per language; (bottom) mean scores across all heads, per language. The languages (hidden from the X-axis for brevity) are, in order, *ar, cs, de, en, es, fi, fr, hi, id, it, ja, ko, pl, pt, ru, sv, tr, zh*

accuracy remains quite high.

In slight contrast to this, the core subject (`nsubj`, 58 ± 16 SD) and object (`obj`, 64 ± 13) relations prove to be more difficult to decode. Unlike the aforementioned relations, `nsubj` and `obj` are much more sensitive to the word order properties of the language at hand. For example, while a language like English, with Subject-Verb-Object (SVO) order, might have the subject frequently appear to the left of the verb, an SOV language like Hindi might have it several positions further away, with an object and its potential modifiers intervening. Indeed, the best positional baseline for English `nsubj` is 39 UUAS, while it is only 10 for Hindi. Despite this variation, the relation seems to be tracked with some regularity by the same head (layer 3, head 9), returning 60 UUAS for English and 52 for Hindi. The same can largely be said for `obj`, where the positional baselines return 51 ± 18. In this latter case, however, the heads tend to be much differently distributed across languages. Finally, he results for the `obj` relation provides some support for our earlier explanation concerning morphologically rich languages, as Arabic, Czech, Korean and Turkish all have among the highest accuracies (as well as positional baselines).

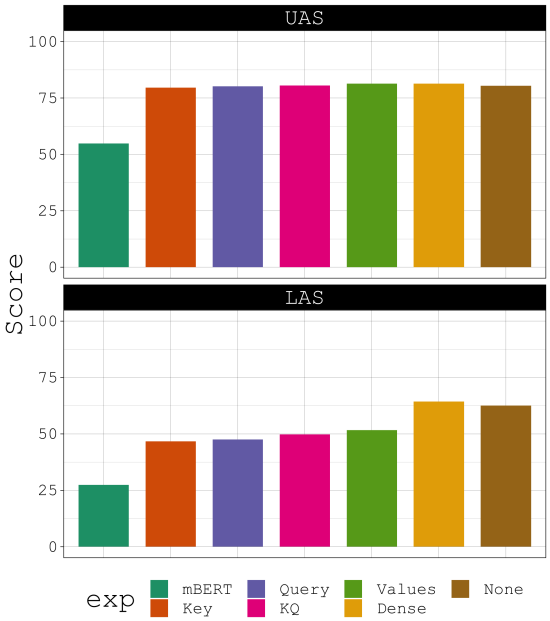

Figure 4: Mean UAS and LAS when evaluating different models on language-specific treebanks (Korean excluded due to annotation differences). MBERT refers to models where the entire mBERT network is frozen as input to the parser.

5 Fine-Tuning Experiments

Next, we investigate the effect fine-tuning has on UUAS decoding. Row 3 in Table 1 (NONE) indicates that fine-tuning does result in large improvements to UUAS decoding across most languages, often by margins as high as $\sim 30\%$. This shows that with an explicit parsing objective, attention heads are capable of serving as explanatory mechanisms for syntax; syntactic structure can be made to be transparently stored in the heads, in a manner that does not require additional probe fitting or parameterized transformation to extract.

Given that we do manage to decode reasonable syntactic trees, we can then refine our question — what components are capable of learning these trees? One obvious candidate is the key/query component pair, given that attention weights are a scaled softmax of a composition of the two. Figure 3 (top) shows the difference between pretrained

UUAS and fine-tuned UUAS per layer, across models and languages. Interestingly, the best parsing accuracies do not appear to vary much depending on what component is frozen. We do see a clear trend, however, in that decoding the attention patterns of the fine-tuned model typically yields better UUAS than the pretrained model, particularly in the highest layers. Indeed, the lowest layer at which fine-tuning appears to improve decoding is layer 7. This implies that, regardless of which component remains frozen, the parameters facing any sort of significant and positive update tend to be those appearing towards the higher-end of the network, closer to the output.

For the frozen components, the best improvements in UUAS are seen at the final layer in VALUE, which is also the only model that shows consistent improvement, as well as the highest average improvement in mean scores[4] for the last few layers. Perhaps most interestingly, the mean UUAS (Figure 3 (bottom)) for our "attentive" components – keys, queries, and their combination – does not appear to have improved by much after fine-tuning. In contrast, the maximum does show considerable improvement; this seems to imply that although all components appear to be more or less equally capable of learning decodable heads, the attentive components, when fine-tuned, appear to sharpen fewer heads.

Note that the only difference between keys and queries in an attention mechanism is that keys are transposed to index attention from/to appropriately. Surprisingly, KEY and QUERY appear to act somewhat differently, with QUERY being almost uniformly better than KEY with the best heads, whilst KEY is slightly better with averages, implying distinctions in how both store information. Furthermore, allowing both keys and queries seems to result in an interesting contradiction – the ultimate layer, which has reasonable maximums and averages for both KEY and QUERY, now seems to show a UUAS drop almost uniformly. This is also true for the completely unfrozen encoder.

Supervised Parsing In addition to decoding trees from attention matrices, we also measure supervised UAS/LAS on a held-out test set.[5] Based on Figure 4, it is apparent that all settings result in generally the same UAS. This is somewhat expected; Lauscher et al. (2020) see better results on parsing with the entire encoder frozen, implying that the task is easy enough for a biaffine parser to learn, given frozen mBERT representations.[6] The LAS distinction is, however, rather interesting: there is a marked difference between how important the dense layers are, as opposed to the attentive components. This is likely not reflected in our UUAS probe as, strictly speaking, labelling arcs is not equivalent to searching for structure in sentences, but more akin to classifying pre-identified structures. We also note that DENSE appears to be better than NONE on average, implying that non-dense components might actually be hurting labelling capacity.

In brief, consolidating the two sets of results above, we can draw three interesting conclusions about the components:

1. **Value** vectors are best aligned with syntactic dependencies; this is reflected both in the best head at the upper layers, and the average score across all heads.
2. **Dense** layers appear to have moderate informative capacity, but appear to have the best learning capacity for the task of arc labelling.
3. Perhaps most surprisingly, **Key** and **Query** vectors do not appear to make any outstanding contributions, save for sharpening a smaller subset of heads.

Our last result is especially surprising for UUAS decoding. Keys and queries, fundamentally, combine to form the attention weight matrix, which is precisely what we use to decode trees. One would expect that allowing these components to learn from labelled syntax would result in the best improvements to decoding, but all three have surprisingly negligible mean improvements. This indicates that we need to further improve our understanding of how attentive structure and weighting really works.

Cross-linguistic observations We notice no clear cross-linguistic trends here across different component sets; however, certain languages do stand out as being particularly hard to decode from the fine-tuned parser. These include Japanese, Korean, Chinese, French and Turkish. For the first three, we hypothesise that tokenization clashes with

[4]The inner average is over all heads; the outer is over all languages.

[5]Note that the test set in our scenario is from the actual, non-parallel language treebank; as such, we left Korean out of this comparison due to annotation differences.

[6]Due to training on concatenated PUD sets, however, our results are not directly comparable/

mBERT's internal representations may play a role. Indeed, as we hypothesized in Section 3.2, it could be the case that the composition of CJK characters into gold tokens for Chinese and Japanese may degrade the representations (and their corresponding attention) therein. Furthermore, for Japanese and Korean specifically, it has been observed that tokenization strategies employed by different treebanks could drastically influence the conclusions one may draw about their inherent hierarchical structure (Kulmizev et al., 2020). Turkish and French are admittedly more difficult to diagnose. Note, however, that we fine-tuned our model on a concatenation of all PUD treebanks. As such, any deviation from PUD's annotation norms is therefore likely to be heavily penalised, by virtue of signal from other languages drowning out these differences.

6 Conclusion

In this study, we revisited the prospect of decoding dependency trees from the self-attention patterns of Transformer-based language models. We elected to extend our experiments to 18 languages in order to gain better insight about how tree decoding accuracy might be affected in the face of (modest) typological diversity. Surprisingly, across all languages, we were able to decode dependency trees from attention patterns more accurately than an adjacent-linking baseline, implying that some structure was indeed being tracked by the mechanism. In looking at specific relation types, we corroborated previous studies in showing that particular layer-head combinations tracked the same relation with regularity across languages, despite typological differences concerning word order, etc.

In investigating the extent to which attention can be guided to properly capture structural relations between input words, we fine-tuned mBERT as input to a dependency parser. This, we found, yielded large improvements over the pretrained attention patterns in terms of decoding accuracy, demonstrating that the attention mechanism was learning to represent the structural objective of the parser. In addition to fine-tuning the entire mBERT network, we conducted a series of experiments, wherein we updated only select components of model and left the remainder frozen. Most surprisingly, we observed that the Transformer parameters designed for composing the attention matrix, K and Q, were only modestly capable of guiding the attention towards resembling the dependency structure. In contrast, it was the Value (V) parameters, which are used for computing a weighted sum over the KQ-produced attention, that yielded the most faithful representations of the linguistic structure via attention.

Though prior work (Kovaleva et al., 2019; Zhao and Bethard, 2020) seems to indicate that there is a lack of a substantial change in attention patterns after fine-tuning on syntax- and semantics-oriented classification tasks, the opposite effect has been observed with fine-tuning on negation scope resolution, where a more explanatory attention mechanism can be induced (Htut et al., 2019). Our results are similar to the latter, and we demonstrate that given explicit syntactic annotation, attention weights do end up storing more transparently decodable structure. It is, however, still unclear which sets of transformer parameters are best suited for learning this information and storing it in the form of attention.

Acknowledgements

Our experiments were run on resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway, under the NeIC-NLPL umbrella. Mostafa and Anders were funded by a Google Focused Research Award. We would like to thank Daniel Dakota and Ali Basirat for some fruitful discussions and the anonymous reviewers for their excellent feedback.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

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.

Yoeng-Jin Chu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

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.

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.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *arXiv:2003.11080 [cs]*. ArXiv: 2003.11080.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 479–486.

Dan Kondratyuk and Milan Straka. 2019a. 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.

Dan Kondratyuk and Milan Straka. 2019b. 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.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. 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 4364–4373, Hong Kong, China. Association for Computational Linguistics.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vuli, and Goran Glava. 2020. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. *arXiv:2005.00633 [cs]*. ArXiv: 2005.00633.

Tomasz Limisiewicz, Rudolf Rosa, and David Mareček. 2020. Universal dependencies according to bert: both more specific and more general. *arXiv preprint arXiv:2004.14620*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005a. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 91–98.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005b. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 523–530.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane

Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Kyung-Tae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.4. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*.

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.

Alessandro Raganato, Jörg Tiedemann, et al. 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*. The Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Rudolf Rosa and David Mareček. 2019. Inducing syntactic trees from bert representations. *arXiv preprint arXiv:1906.11511*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.

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. 2017a. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

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. *arXiv preprint arXiv:1905.09418*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope. page 19.

A Positional Scores Per Offset

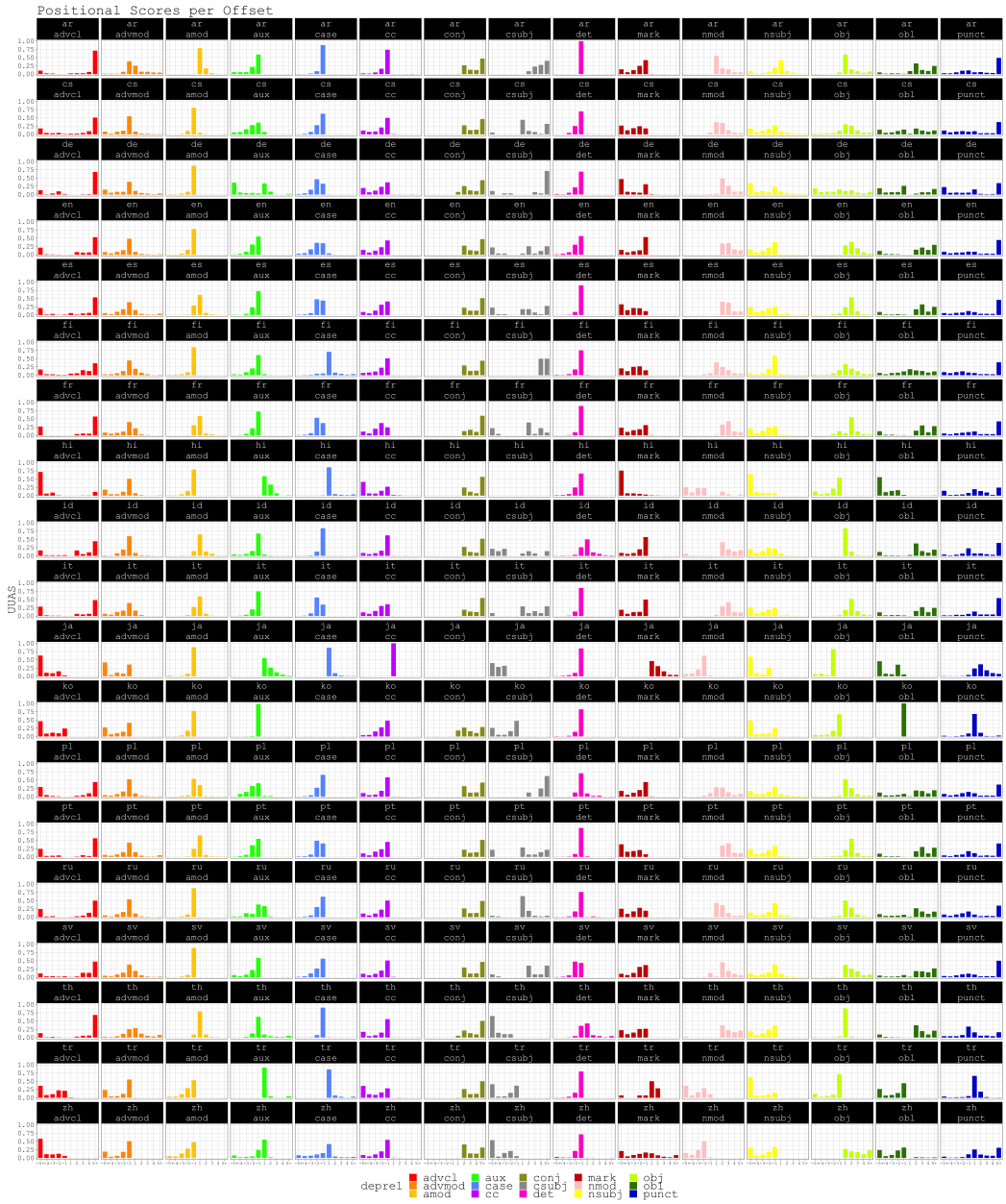

Figure 5: Positional scores across relations for all languages.

B Decoding UUAS Across Relations

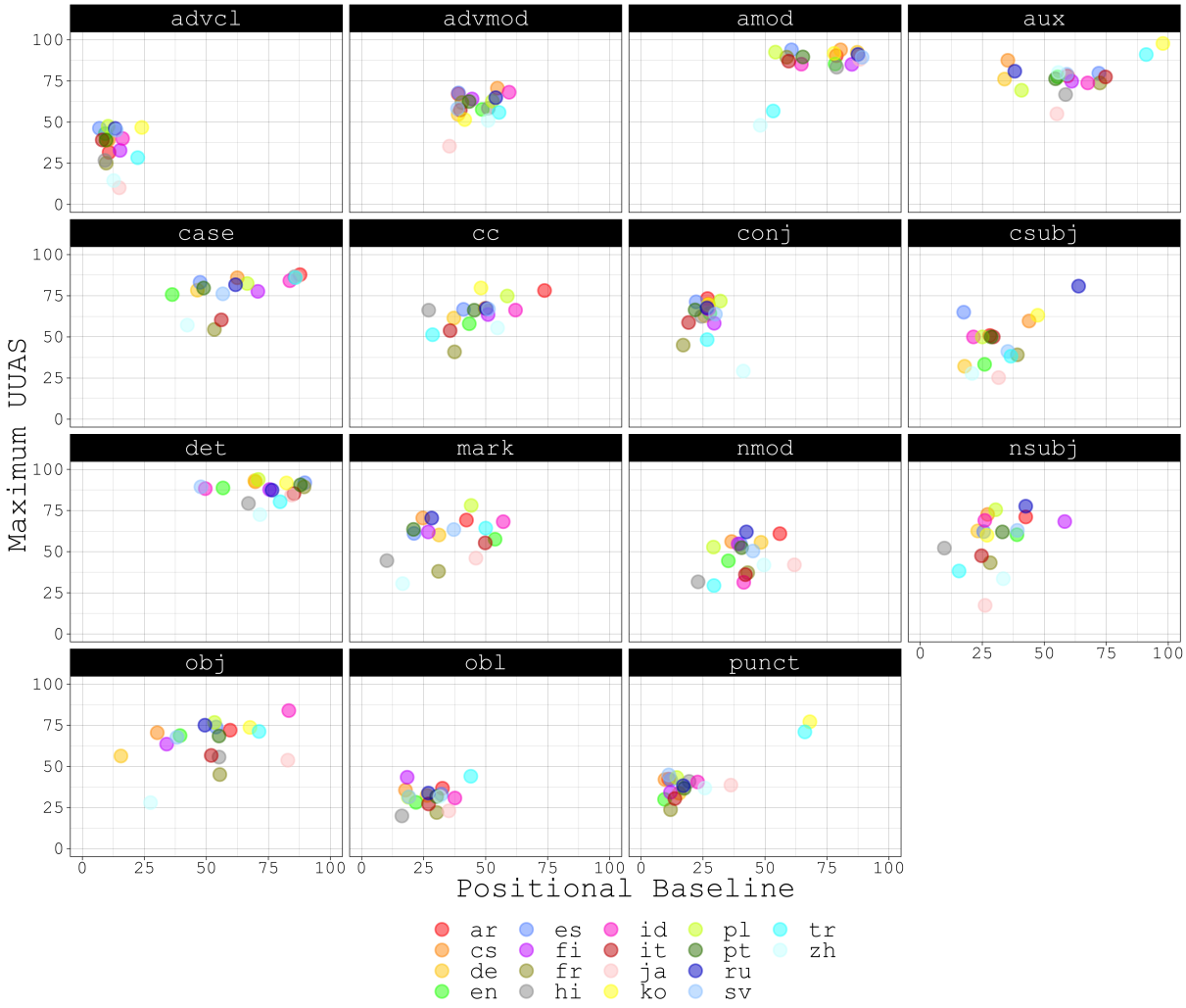

Figure 6: Decoding UUAS as a function of best positional baselines.

C Full Parsing Scores

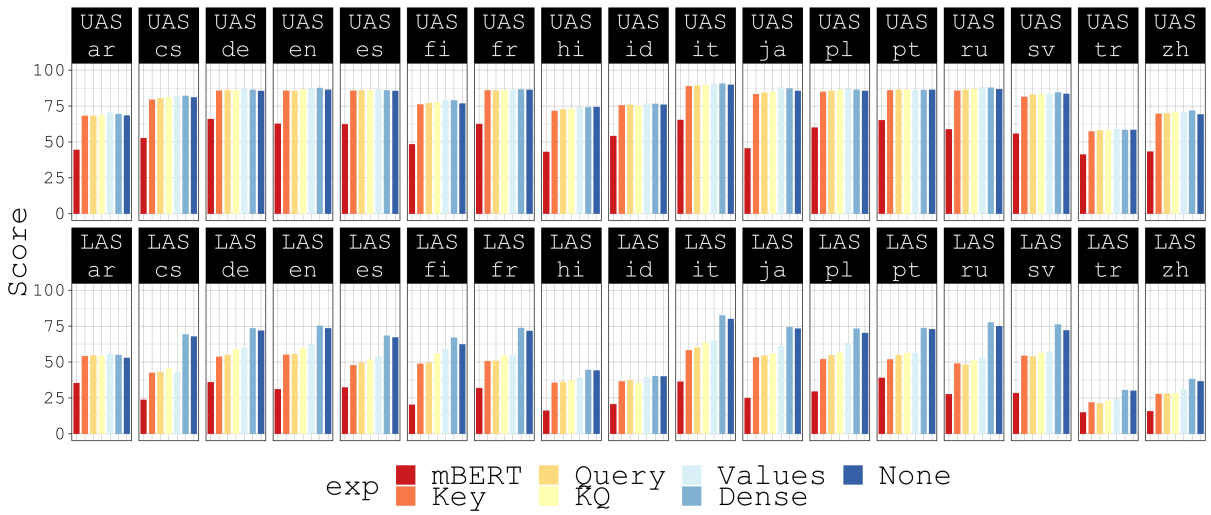

Figure 7: Parsing scores across components and languages.