Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6674–6690
Marseille, 20-25 June 2022

BERTology for Machine Translation: What BERT Knows about Linguistic Difficulties for Translation

Yuqian Dai, Marc de Kamps, Serge Sharoff
Centre for Translation Studies, University of Leeds, LS2 9JT, United Kingdom
{mlyd, m.dekamps, s.sharoff}@leeds.ac.uk

Abstract

Pre-trained transformer-based models, such as BERT, have shown excellent performance in most natural language processing benchmark tests, but we still lack a good understanding of the linguistic knowledge of BERT in Neural Machine Translation (NMT). Our work uses syntactic probes and Quality Estimation (QE) models to analyze the performance of BERT's syntactic dependencies and their impact on machine translation quality, exploring what kind of syntactic dependencies are difficult for NMT engines based on BERT. While our probing experiments confirm that pre-trained BERT "knows" about syntactic dependencies, its ability to recognize them often decreases after fine-tuning for NMT tasks. We also detect a relationship between syntactic dependencies in three languages and the quality of their translations, which shows which specific syntactic dependencies are likely to be a significant cause of low-quality translations.

Keywords: Neural Machine Translation, BERT, Syntactic Dependencies, Probing Experiments, Quality Estimation

1. Introduction

The encoder-decoder used in Neural Machine Translation (NMT) is a distinctive representation architecture whose encoder and decoder are the product of representation learning for both source and target languages. The Transformer model (Vaswani et al., 2017) proposed in recent years is even more efficient in describing long-term links between words through a self-attention mechanism. It has become one of the most widely used models in machine translation, and the joint training of encoder and decoder makes it possible to add pre-training with better generalization ability to the machine translation task. Compared with the LSTM-based pre-trained model ELMo (Peters et al., 2018), the pre-trained model BERT (Devlin et al., 2019) takes the role of pre-training with word representation to a new level. The auto-encoding approach proposed by BERT allows the model to be modeled with self-attention in the pre-training phase, which further enhances the representation capability of the model. Inspired by BERT, more pre-trained models are proposed, such as XLM (Conneau and Lample, 2019) and RoBERTa (Liu et al., 2019). The large-scale data learning that pre-trained models possess makes it possible to acquire more general linguistic knowledge and input representations and provides better initialization parameters and generalization capabilities for downstream tasks (Edunov et al., 2019). Given the success of BERT in language understanding tasks, much work has been done on how to incorporate BERT to improve the translation quality of machine translation (Clinchant et al., 2019; Weng et al., 2020). Although BERT can be added to either the encoder or the decoder part, it is more common to incorporate it into a translation task as an encoder or as part of an encoder (Imamura and Sumita, 2019; Yang et al., 2020; Zhu et al., 2020). The language representation and feature extraction of BERT can further improve the efficiency of the decoder, and the pre-training mechanism allows the translation engine to obtain an effective initial model parameter to alleviate the challenges of low-resource languages in the translation task (Vu et al., 2021). However, BERT can also be less effective in improving some high-resource languages and can even bring about performance degradation (Zhu et al., 2020). Many studies have explored the linguistic knowledge of BERT, such as semantic knowledge (Ettinger, 2020) and syntactic knowledge (Tenney et al., 2019), and focus on how to introduce syntactic knowledge in machine translation models (Sundararaman et al., 2019). There is still a lack of discussion on the performance and impact of BERT applied to machine translation tasks from a syntactic knowledge perspective, however. In machine translation, syntactic knowledge may not be as crucial as semantic understanding and other knowledge, but past work reveals that the incorporation of syntactic knowledge is helpful for translation task (Sundararaman et al., 2019). Improving translation systems with pre-trained models and translation quality through syntactic knowledge in an encoder-decoder framework is a potential research point to be added to the field of translation. Therefore, this study investigates how a BERT-based NMT model is affected by its syntactic knowledge and how the quality of machine translation is affected by the syntactic information in different source languages.

In this work, we build NMT engines for three different languages, where the encoder is the pre-trained model BERT. We investigate the performance of BERT in machine translation tasks concerning syntactic knowledge in two ways. The first is to consider BERT as a standalone model after fine-tuning the machine translation task and explore how it predicts and knows syntactic components in sentences and thus detects its performance in syntactic knowledge through probing experiments. The second is to consider the translation en-

gines with BERT as a whole and use the Quality Estimation (QE) model to score the output translations, detecting the connection between syntactic information in the source language and the translation quality in the target language.

Our main contributions are as follows:

- We test BERT for the NMT task on a large scale on syntactic dependencies in three languages and detect changes in syntactic knowledge before and after fine-tuning. In most cases, BERT has syntactic patterns that are not affected by fine-tuning.
- We test a method to detect the link between translation quality and syntactic dependencies with the ability to recognise certain types of syntactic dependencies linked to low-quality translations.

2. Methodology

2.1. Construction of the NMT Engines

We use the pre-trained model BERT-base as the encoder, and the decoder of the vanilla Transformer model (Vaswani et al., 2017) to build our NMT engines. The NMT engines include three versions of different languages, which are Chinese to English (Zh→En), Russian to English (Ru→En), and German to English (De→En), respectively. In contrast with the vanilla transformer model, the pre-trained model BERT-base is the encoder in our NMT engines as shown in Figure 1 and is fine-tuned by the NMT task, where the architecture of our NMT engines is similar to the existing work and discussions (Imamura and Sumita, 2019; Weng et al., 2020). In detail, there are three different types of BERT-base, each acting as an encoder for different languages. We use the BERT-wwm-ext version for Chinese (Cui et al., 2021), the RuBERT version for Russian (Kuratov and Arkhipov, 2019) and the Google base German version for German (Devlin et al., 2019). There are differences in their pre-training strategies, but we are more concerned with the model's understanding of syntactic knowledge under the BERT architecture. With the same model architecture, how the syntactic knowledge varies in the machine translation task. These BERTs used for all experiments have the same basic specifications, where the number of layers = 12, attention heads = 12, embedding dimension = 768. When fine-tuning, all the internal parameters of the pre-trained BERTs will be updated.

Language	Dataset	BLEU
Zh → En	UNPC	56.34
Ru → En	UNPC	55.85
De → En	Europarl	38.06

Table 1: BLEU of three NMT engines, all engines can output understandable good translations.

The Chinese and Russian NMT engines are trained with the parallel data from the United Nations Parallel Corpus (UNPC) (Ziemski et al., 2016), while the German engine is trained with Europarl (Koehn, 2005), and

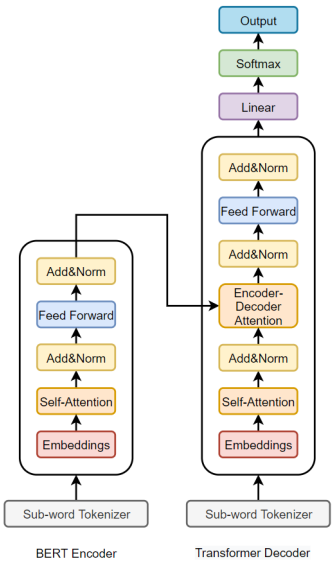


Figure 1: Construction of NMT engine with pre-trained model BERT.

all of them are from the OPUS collection (Tiedemann, 2012) [1]. All sentence pairs in the training set, validation set, and test set are randomly selected subsets from those corpora. To ensure sufficient training samples and a uniform training environment for BERTs, we randomly select approximately 1.2 million (1.2M) parallel sentence pairs as the training set for each language, and the validation set and the test set have about 6,000 parallel sentence pairs verifying the performance of the systems. The performance of the NMT engines is reported in Table 1, differences in BLEU may be influenced by the type of corpora or the pre-trained BERT-base released by different publishers.

2.2. Syntactic Probes and NMT Fine-tuning

The syntactic probing experiments aim to investigate BERT's syntactic ability changes after machine translation fine-tuning and further understand the possibility of analyzing BERT from syntactic dependency. We use two syntactic annotation corpora called Parallel Universal Dependencies (PUD) and Universal Dependencies (GSD) from Universal Dependencies (UD)[2], which contain gold annotating syntactic dependencies as the primary testing morphosyntactic features for all probing experiments.

PUD treebanks are created under the CoNLL 2017 shared task, each language contains the raw text and its linguistic annotations. PUD for each language (UD Chinese PUD[3], UD Russian PUD[4], UD German

[1] https://opus.nlpl.eu/

[2] https://github.com/UniversalDependencies

[3] https://github.com/UniversalDependencies/UD_Chinese-PUD

[4] https://github.com/UniversalDependencies/UD_Russian-PUD

PUD[5]) contains 1,000 annotated sentences respectively and always in the same order. GSD treebanks are converted from Universal Dependency treebank. We use Chinese GSD[6], Russian GSD[7], and German GSD[8] as another experimental corpora of the probing experiments. There are some differences in the syntactic types of the PUD and GSD corpora. GSD corpora has more syntactic-annotated sentences, with about 5,000 annotated sentences in Chinese GSD and Russian GSD, and 15,000 sentences in German GSD. We manually select the same syntactic dependencies between PUD and GSD corpora to conduct probing experiments to ensure the accuracy of the experiment. We only record and compare the syntactic dependencies common to both corpora, and those with a small number are excluded from the experiment.

A syntactic dependency indicates the relationship between two words. We want to know whether BERT can assign the correct syntactic dependency to the current word without specifying the target word and whether BERT can be aware of the structure of the current word in the sentence. Therefore, we treat the syntactic probing experiments as a sequence labeling task in which BERT needs to predict the syntactic dependency labels for each token in a sentence from syntactic-annotated corpora as shown in Figure 2. Inspired by previous work (Papadimitriou et al., 2021), the probing approach is straightforward in that one linear classifier is added above BERT as shown in Figure 3. The reason for this is that we need to ensure as much as possible that the results of syntactic knowledge classification are mainly from BERT. A more advanced encoder or decoder can also achieve the same effect in a complicated NMT engine if a superficial linear classification layer can capture that information. We separate the BERTs of the NMT engines for these three languages and then apply the probes to each layer of the BERTs, where BERTs are trained with a limited number of layers. The probes examine the performance of syntactic dependencies for each layer of BERT, and the results are presented as F1-score. When conducting syntactic probing experiments on BERT, BERT for all languages is divided into two groups, before and after fine-tuning. We build the training set, validation set, and test set in all experiments, all the parameters of BERT are frozen, and only the superficial classification layer is trained.

[5] https://github.com/UniversalDependencies/UD_German-PUD

[6] https://github.com/UniversalDependencies/UD_Chinese-GSDSimp

[7] https://github.com/UniversalDependencies/UD_Russian-GSD

[8] https://github.com/UniversalDependencies/UD_German-GSD

2.3. Syntactic Knowledge and Quality Estimation

To continue investigating whether syntactic knowledge impacts the quality of NMT, we use our NMT engines to translate sentences from the syntax-annotated PUD corpora. Since most input scenarios are that the source language does not have a golden reference translation, we prefer to know whether the translation engine produces a more reasonable and fluent translation rather than a standardized sentence that exactly favors the professional human translation. We use Quality Estimation (QE) model called TransQuest[9] (Ranasinghe et al., 2020) which is a state-of-the-art QE model to score machine translation quality for a number of languages. It predicts a Direct Assessment (DA) to score the adequacy and fluency of the machine-translated sentences. The score range of DA is 0-1. The higher the score, the higher the quality of the NMT engine output.

	Zh→En	Ru→En	De→En
High-quality range	0.808-0.891	0.845-0.917	0.822-0.896
Low-quality range	0.061-0.519	0.325-0.742	0.226-0.533

Table 2: The QE model scores the translations in the three languages. A range of scores is distinguished between high-quality and low-quality translations for each language.

As shown in Table 2, we consider the translation with the highest 20% score as a high-quality translation and the translation with the lowest 20% score as a low-quality translation in three different languages. A manual evaluation of a sample of translations agrees with the automatic assessment. Then we extract the golden syntactic annotations of their source sentences and count the syntactic composition and number corresponding to the high-quality and low-quality translations, respectively. We use the chi-square test to investigate the association between syntactic dependencies and translation quality also the quantitative differences between the two quality groups, with larger values of χ^2 indicating a more significant quantitative difference between high and low-quality translations for that syntactic dependency and vice versa. We report that a standard threshold p-value of 0.05 is used in the experiments, and the confidence is 95%.
(a) We detect whether there is a correlation between syntactic dependencies and machine translation quality. All syntactic dependencies are taken into account in the chi-square test. Under the assumption of independence under the hypothesis, the expected values can be obtained from the total number of observed values. The sum of the expected numbers for each sample must be equal to the sum of the observed numbers for each sample. (b) The gap in syntactic dependencies between high-quality and low-quality translations

[9] https://github.com/TharinduDR/TransQuest

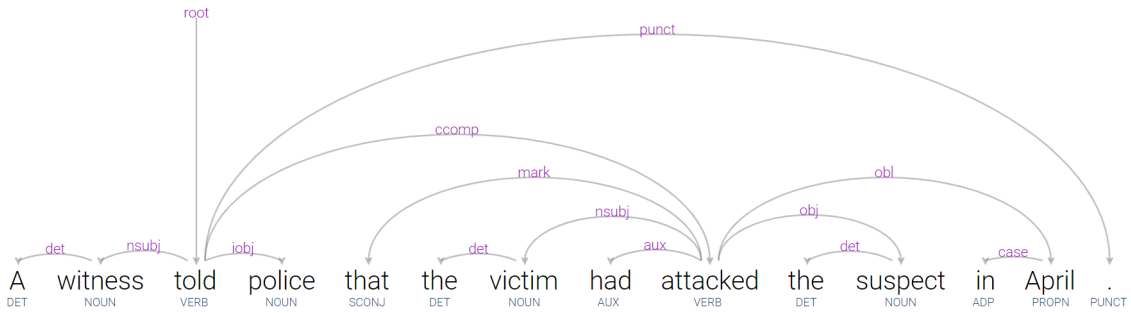


Figure 2: Example of syntactic dependencies in one English sentence. For example, the central word (root) of this sentence is a verb called "told", a subject called "witness" depends on this central word, and the syntactic dependency between them is "nsubj". There may be differences in the inventory of syntactic dependencies between different languages, and Universal Dependencies can help minimize such discrepancies.

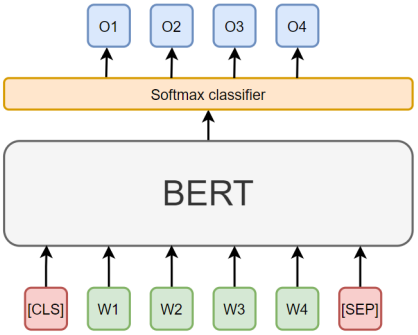


Figure 3: BERT needs to predict syntactic dependencies in sentences with a simple classifier.

is detected. To highlight the differences in specific syntactic dependencies of different quality, we define the observed value as the number of syntactic dependencies in low-quality translations and the expected value as the number of syntactic dependencies in high-quality translations. Following (Ranasinghe et al., 2020), we use DA as an evaluation metric for machine-translated sentences.

3. Results

3.1. Dependencies Probes with NMT

We observe that BERT has different mastery of different dependencies of syntactic knowledge via probes as shown in Figure 4*. Three syntactic phenomena can be classified.

- The syntactic dependencies that BERT is good at can maintain high performance, either by fine-tuning or changing the number of layers.
- Changing the number of layers does not substantially improve syntactic dependencies that BERT is not good at.
- Some syntactic dependencies in BERT are very sensitive to changes in the number of layers, which may cause their performance to fluctuate.

Most of the PUD and GSD corpora results are similar in the syntactic probing experiments. After the fine-tuning of BERT in different languages by machine translation, the performance of most of the syntactic dependencies has been reduced to varying degrees, and only a small part of the syntactic dependencies has been maintained or increased. Based on the performance curves of the F1-score, we find that BERTs have different trends of syntactic dependencies for different languages. To distinguish it from syntactic phenomena, we call it a syntactic pattern. Common syntactic patterns of syntactic dependencies for three languages in BERT are indicated with black dots below and shown in Table 3 to Table 5, where syntactic patterns of the PUD and GSD corpora are put together.

- Smooth: The performance of most layers is relatively stable, with no significant performance fluctuations.
- Climb + Decline: As the number of layers increases, the performance rises and then decreases gradually, and the performance fluctuates more smoothly from layer to layer.
- Fluctuate: Despite the overall trend, there are significant differences in performance between the layers.

The probing experiments reflect that syntactic dependencies are related to layers but are more likely determined by the working mechanism of BERT itself during the pre-training. These patterns may reveal that BERT tries to learn and process different syntactic knowledge by using different layers. Previous work (Jawahar et al., 2019) suggests that the intermediate layers perform the best for syntactic knowledge. However, our probing experiments show that the type of syntactic dependencies determines the syntactic performance of BERT, and the number of layers is not the main factor in determining the performance of syntactic dependencies. We find that syntactic patterns are similar before and after NMT fine-tuning in most cases. For example, in Figure 4, either PUD or GSD as the data set, the curves of F1-score of "appos" before and after the fine-tuning

*All experimental results for the three languages are included in the appendix.

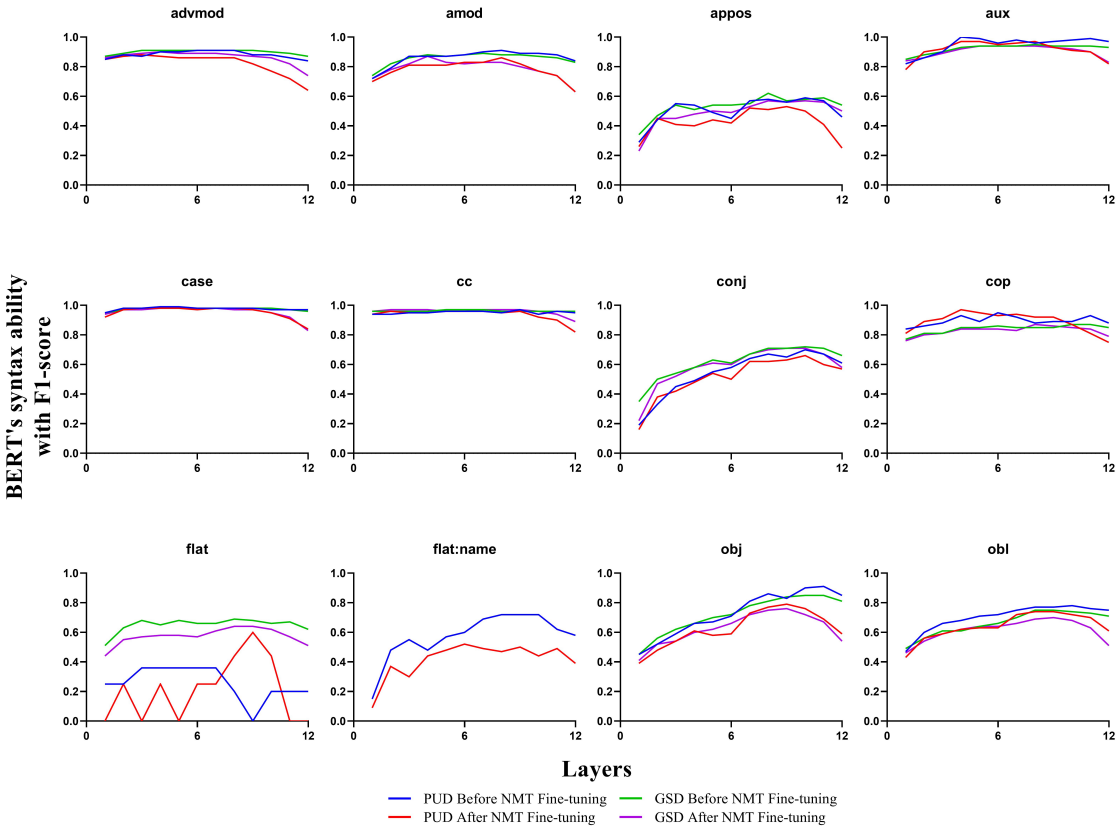


Figure 4: Some German testing results of syntactic dependencies come from the probes. The syntactic dependencies of "case", "flat" and "obl" are typical syntactic patterns.

	advmod	amod	aux:pass	case	cc	dep	det	flat	flat:foreign	discourse	mark:advb	mark	nummod
Zh					●				●	●	●	●	●
Ru	●	●	●	●	●		●	●					
De	●			●	●	●	●						

Table 3: Syntactic patterns called "Smooth" of dependencies in three languages.

	acl:relcl	acl	advcl	advmod	amod	appos	aux	aux:pass	case	case:loc
Zh	●	●	●	●	●	●	●	●	●	●
Ru										
De	●	●			●			●		
	ccomp	**clf**	**compound**	**conj**	**cop**	**det**	**flat**	**mark**	**mark:prt**	**mark:relcl**
Zh	●	●	●	●	●	●	●	●	●	●
Ru							●	●		
De	●			●	●		●		●	
	nmod	**nmod:tmod**	**nummod**	**nsubj**	**obj**	**obl**	**obl:tmod**	**root**	**xcomp**	**expl**
Zh	●	●	●	●	●	●	●	●	●	●
Ru	●			●	●	●		●	●	
De	●		●	●	●	●	●	●		●

Table 4: Syntactic patterns called "Climb + Decline" of dependencies in three languages.

of machine translation still show similar performance trends. It may mean that BERT has already formed a syntactic pattern during the pre-training phase and the fine-tuning of machine translation only changes the performance of the task instead of BERT's reconsideration of syntactic dependencies.

3.2. Quality Estimation

(a) As shown in Table 6, we count the syntactic dependencies of high-quality and low-quality translations

	acl:relcl	acl	advcl	appos	ccomp	compound	compound:prt	conj	cop
Zh					•				
Ru	•	•	•	•	•	•	•	•	•
De			•	•		•	•		
	csubj	dep	dislocated	flat	flat:name	flat:foreign	fixed	nsubj:pass	nummod
Zh	•	•	•	•	•			•	
Ru	•	•	•	•	•	•	•	•	•
De	•			•	•			•	
	nummod:gov	nmod:poss	iobj	obl:agent	obl:patient	parataxis	root	xcomp	expl
Zh					•				
Ru	•		•	•	•	•	•	•	
De		•	•			•		•	•

Table 5: Syntactic patterns called "Fluctuate" of dependencies in three languages.

for each of the three languages. The results show that the chi-square values of all three languages greatly exceed their test statistic for specific syntactic dependencies associated with high-quality and low-quality translations in each language. The null hypothesis (h_0) that translation quality and syntactic dependency are unrelated is not valid. Instead, the alternative hypothesis (h_1) is accepted that translation quality is associated with syntactic dependency.

(b) We find that syntactic dependencies occurred more frequently in low-quality translations than high-quality ones. A more significant chi-square value indicates a large difference between the number of high-quality and low-quality translations for a particular syntactic dependency, as shown in Table 7. Taking into account the differences in syntactic dependency between languages, we record common syntactic dependencies and compare them in three languages. "**appos**", "**case**", "**flat**", "**flat:name** ", "**obl**" are notable. They occur more frequently in low-quality translations of these three languages. We conjecture that BERT will have a different syntactic dependency performance as a standalone monad than an NMT engine. The quality of the translations is not precisely equivalent to the syntactic dependency performance of BERT in the probing experiments. When BERT is used as an encoder for NMT engines, the translations in all three languages show common problems caused by specific syntactic dependencies, which are possible causes of low-quality translations. However, we do not find such a significant problem in the BERT individual probing experiments. The reason may be that there are multiple neural networks involved in the work of the NMT engines, and the importance of linguistic knowledge is constantly being selected. The results of the probing experiments are not equivalent to the results that act on the downstream tasks, although they may be linked.

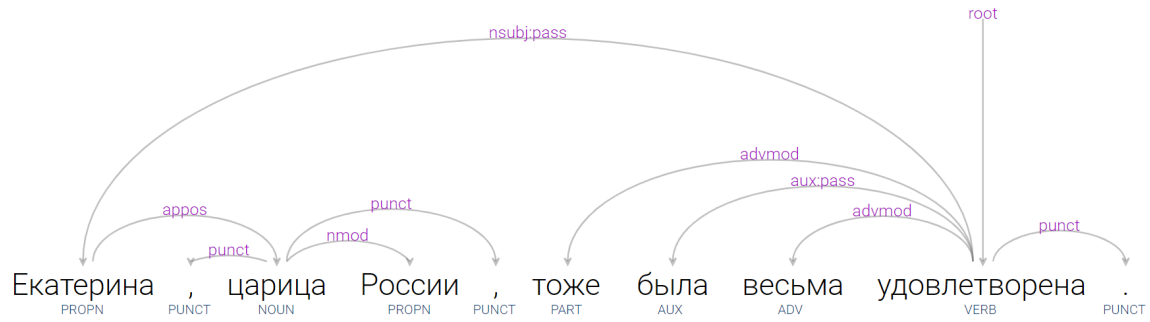
3.3. Error Analysis

We take a closer look at the specific relations which commonly cause errors. For example, while the second half of the Russian example in Figure 5 results in a satisfactory translation, its start which contains the "appos" relation does not make any sense since the BERT model is not able to predict the relationship between the two noun phrases. Also, by comparing translation quality and F1-score, we find that the F1-score of syntactic dependencies is mostly associated with translation quality, but this association is not absolute.

Languages	Dependencies	df	p-value	Test statistic	χ^2
Zh	32	31		43.77	171.4
Ru	29	28	0.05	41.34	154.9
De	30	29		42.56	182

Table 6: Dependencies show the number of syntactic dependency types in the annotated source sentences. The values of χ^2 are much larger than the test statistic, showing that the observed and expected values are significant. There is a correlation between syntactic dependencies and the quality of the translation.

- In syntactic dependency, "appos" is an appositional modifier used to modify, describe or define the noun. "flat" and "flat:name" are used to indicate the date and the syntactic structure within the proper noun. The common feature of all three is the construction of relationships between nouns in a sentence. The F1-score of the top layer of "appos" and "flat:name" in Chinese and Russian is higher than the middle layer and is one of the main syntactic dependencies that cause low-quality translations. In German, they have significantly better F1-score in the middle layer but still dominate the low-quality translations. We believe that BERT can be fine-tuned and thus take advantage of nouns' new knowledge to understand syntactic structures better. However, this knowledge may be affected by differences in the training set, e.g., UNPC may contain more noun-like information than Europarl. Although we use sub-word tokenization, the vocabularies are still based on data sets from news and conference domains such as UNPC or Europarl. The sentences contained in the PUD corpora are from different domains and contain specific and complex names of people and places. There is still a probability that they are not



MT herat , her country ' s russian federation , was also very pleased .
Reference Catherine of Russia was also very satisfied.

Figure 5: Dependencies for Russian example. BERT fails to interpret the nouns linked by "appos" in the translation *Catherine of Russia* (appos) *was also very satisfied.*

fully covered by the vocabularies, which means that BERT has never been exposed to them at the time of training and cannot understand them well when translating. Assuming that the vocabulary fully covers them but never appears in the training set, they still leave BERT with the challenge of sentence comprehension during translation.

- Although syntactically, "obl" and "case" play the function of adverb and the function of uniform analysis of different lexical forms, they have the same results. Not only are they common syntactic dependencies in low-quality translations in all three languages, but also the F1-score in the middle layer is higher than those in the top layer. It suggests that they are also likely to be one of the syntactic dependencies that contribute to low-quality translations, and the probes can reveal the effect of specific syntactic dependencies on low-quality translations. Besides, most of the syntactic dependencies in low-quality translations show that the middle layer has a better F1-score in BERT, which implies that BERT's understanding of syntactic dependencies will be reflected not only on the probes but also on the translation quality to a certain extent.

4. Related Work

The powerful non-linear and pre-training mechanisms endow BERT with extremely high representation and knowledge learning capabilities for textual inputs. However, the contribution of these capabilities to downstream tasks does not effectively reveal the role of deep knowledge in BERT. A growing body of work aims to extract the probes detected knowledge by designing probing experiments, thus understanding the type of linguistic knowledge within BERT and getting better interpretability, such as how this knowledge is learned, how it is represented, and suggested improvement methods.

A typical type of linguistic knowledge is syntactic knowledge. (Lin et al., 2019) find that BERT has hierarchical encoding and specific syntactic tree knowledge through probing experiments. (Tenney et al., 2019) investigate the problem of embedding syntactic information through the contextual representation of BERT and consider and detect specific words of a sentence (Goldberg, 2019). A similar way of working includes the exploration of semantic knowledge. The probes can detect information such as BERT encoding entity types and relationships (Tenney et al., 2019), and BERT can show sensitivity in role switching and differentiate between the same categories (Ettinger, 2020), although this ability is not as good as humans. Probing experiments can also focus on self-attention reasoning. Attention heads are probed and visualized in BERT to capture the interpretability of linguistic knowledge through attention weight distributions (Kovaleva et al., 2019; Lin et al., 2019). The possibility of injecting syntactic knowledge into pre-trained models has also been investigated (Sachan et al., 2020), and these studies show that the availability of manually labeled data essentially limits performance improvement. In downstream tasks, the addition of syntactic knowledge can be used as a complement to the sequence structure. The current mainstream Transformer models can be used for different tasks, but some work has shown that adding additional syntactic knowledge can improve performance. For example, with syntactic knowledge, the model achieves improved F1-score and accuracy in slot filling and intent detection tasks (Wang et al., 2020). The BLEU is greatly improved in the text generation task by adding syntactic knowledge to the attention mechanism (Li et al., 2020). In machine translation, (Duan et al., 2019) add syntactic knowledge to positional encoding can improve the BLEU score while maintaining the model's efficiency. (Belinkov et al., 2017) have been proposed to study NMT and syntactic knowledge, but it does not discuss BERT. In the case of long sentences with very complicated syntactic structure samples, the addition of syntactic information can still be of great use. However, testing of translation scenarios still needs to be performed, especially for syntactic dependencies in pre-trained models. If

	Dependency	Quality			F1-score	
		High	Low	χ^2	Layer-6	Layer-12
Zh	**flat:name**	**1**	**53**	**2704**	0.68	0.78
	appos	**10**	**73**	**396.9**	0.40	0.48
	flat	**2**	**24**	**242**	0.70	0.74
	dep	42	99	77.3	0.28	0.29
	advcl	62	113	41.9	0.43	0.33
	mark	32	66	36.1	0.72	0.54
	nsubj	293	380	25.8	0.63	0.60
	obl	**86**	**132**	**24.6**	0.39	0.31
	case	**214**	**286**	**24.2**	0.82	0.76
	obl:tmod	28	54	24.1	0.68	0.46
	compound	267	333	16.3	0.70	0.60
Ru	flat:foreign	1	55	2916	0.69	0.18
	flat	**1**	**31**	**900**	0.22	0.18
	flat:name	**12**	**57**	**168.7**	0.82	0.86
	appos	**8**	**31**	**48.3**	0.36	0.47
	obl	**207**	**307**	**66.1**	0.71	0.66
	parataxis	21	50	48.3	0.48	0.48
	case	**306**	**406**	**40**	0.98	0.91
	conj	93	147	32.6	0.63	0.72
	cc	82	123	31.3	0.98	0.93
	amod	274	347	20.5	0.90	0.89
	nummod:gov	10	20	19.4	0.67	0.44
De	**flat**	**0**	**9**	**-**	0.25	0
	appos	**10**	**96**	**739.6**	0.42	0.25
	flat:name	**6**	**61**	**504.1**	0.52	0.39
	compound	25	72	88.3	0.47	0.51
	obl	**212**	**327**	**62.3**	0.63	0.61
	compound:prt	10	34	57.6	0.92	0.48
	case	**324**	**459**	**56.25**	0.97	0.84
	obl:tmod	14	34	28.5	0.55	0.62
	nmod:poss	39	67	20.1	0.96	0.85
	nsubj	241	308	18.6	0.73	0.68
	advcl	27	47	14.8	0.34	0.36

Table 7: Syntactic dependencies are ordered according to the value of χ^2, the complete table is in the Appendix. χ^2 reflects the difference of the number of syntactic dependencies in two different translation qualities. Bold syntactic dependencies are a common syntactic feature of low-quality translations in all three languages. flat (De) does not appear in the high-quality translation, the result can not be calculated. F1-score is derived from BERT after fine-tuning and PUD as the data set.

the syntactic knowledge in BERT can be implicitly applied to translation tasks, it may be effective in improving the problem of low-quality translations. There is still a lack of sufficient discussion and exploration as to which syntactic knowledge in BERT is relied upon by translators and how much the syntactic knowledge in BERT affects the translation quality in translation tasks.

5. Conclusions

This work discusses how knowledge about syntactic dependency relations in BERT changes when it is fine-tuned as an encoder for an NMT engine. Testing with syntactic probes demonstrates that the F1-score for detecting most Universal Dependencies by BERT decreases after NMT fine-tuning. For example, in Chinese, Russian, and German, "advmod" and "det" show a significant downward trend after NMT fine-tuning. In addition, BERT's ability to recognize the syntactic dependency patterns does not change substantially as a result of fine-tuning for the NMT task, implying that BERT's perception of syntactic dependency may have been formed in the pre-training stage. Also we find a correlation between translation quality and syntactic dependency through a chi-square test, suggesting that lack of recognition of some syntactic dependencies can be one of the causes of low-quality translation when BERT is used as the encoder in the NMT engines. By comparing the quality of probing experiments and actual translations, we find that probing experiments on BERT alone can provide knowledge interpretability. However, this interpretability is not entirely equivalent to the performance of BERT jointly with other neural networks involved as an engine in machine translation tasks. We want to determine whether it is possible to optimize the NMT engines with BERT participation through syntactic knowledge and thus improve the translation quality. Future work will continue to focus on applying BERT in machine translation tasks to bring more interpretability from the perspective of syntactic knowledge.

6. Bibliographical References

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017). 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, July. Association for Computational Linguistics.

Clinchant, S., Jung, K. W., and Nikoulina, V. (2019). On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong, November. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Duan, S., Zhao, H., Zhou, J., and Wang, R. (2019). Syntax-aware transformer encoder for neural machine translation. In *2019 International Conference on Asian Language Processing (IALP)*, pages 396–401. IEEE.

Edunov, S., Baevski, A., and Auli, M. (2019). Pre-trained language model representations for language generation. 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 4052–4059, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Goldberg, Y. (2019). Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Imamura, K. and Sumita, E. (2019). Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong, November. Association for Computational Linguistics.

Jawahar, G., Sagot, B., and Seddah, D. (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, July. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (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 4365–4374, Hong Kong, China, November. Association for Computational Linguistics.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Li, Y., Feng, R., Rehg, I., and Zhang, C. (2020). Transformer-based neural text generation with syntactic guidance. *arXiv preprint arXiv:2010.01737*.

Lin, Y., Tan, Y. C., and Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, August. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Papadimitriou, I., Chi, E. A., Futrell, R., and Mahowald, K. (2021). Deep subjecthood: Higher-order grammatical features in multilingual bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online, April. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (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)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ranasinghe, T., Orasan, C., and Mitkov, R. (2020). Transquest: Translation quality estimation with cross-lingual transformers. *arXiv preprint arXiv:2011.01536*.

Sachan, D. S., Zhang, Y., Qi, P., and Hamilton, W. (2020). Do syntax trees help pre-trained transformers extract information? *arXiv preprint arXiv:2008.09084*.

Sundararaman, D., Subramanian, V., Wang, G., Si, S., Shen, D., Wang, D., and Carin, L. (2019). Syntax-infused transformer and bert models for machine translation and natural language understanding. *arXiv preprint arXiv:1911.06156*.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vu, V.-H., Nguyen, Q.-P., Tunyan, E. V., and Ock, C.-Y. (2021). Improving the performance of vietnamese–korean neural machine translation with contextual embedding. *Applied Sciences*, 11(23):11119.

Wang, J., Wei, K., Radfar, M., Zhang, W., and Chung, C. (2020). Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. *arXiv preprint arXiv:2012.11689*.

Weng, R., Yu, H., Huang, S., Cheng, S., and Luo, W. (2020). Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 34, pages 9266–9273.

Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., and Li, L. (2020). Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. (2020). Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May. European Language Resources Association (ELRA).

7. Appendices

7.1. Results of Probing Experiments

Whole probing experiments of syntactic dependencies in Chinese, Russian and German are shown in Figure 6, Figure 7, Figure 8, and Figure 9. Some syntactic dependencies lack complete tests because they are only contained in the PUD or GSD corpora. Usually, they include tests on the PUD and GSD corpora before and after fine-tuning the NMT task.

7.2. More Syntactic Dependencies with Quality Estimation

Syntactic dependencies with χ^2 in Chinese, Russian, and German in the high and low-quality translations are shown in Table 8, Table 9, and Table 10. Since there are some syntactic dependencies with tiny numbers in the PUD corpora, they are excluded from this table to obtain more accurate results. χ^2 reflects the difference in the number of syntactic dependencies in two different translation qualities. Layer 6 and Layer 12 show the performance of syntactic dependencies on F1-score in BERT. "-" means that this syntactic dependency does not include in the probing experiments.

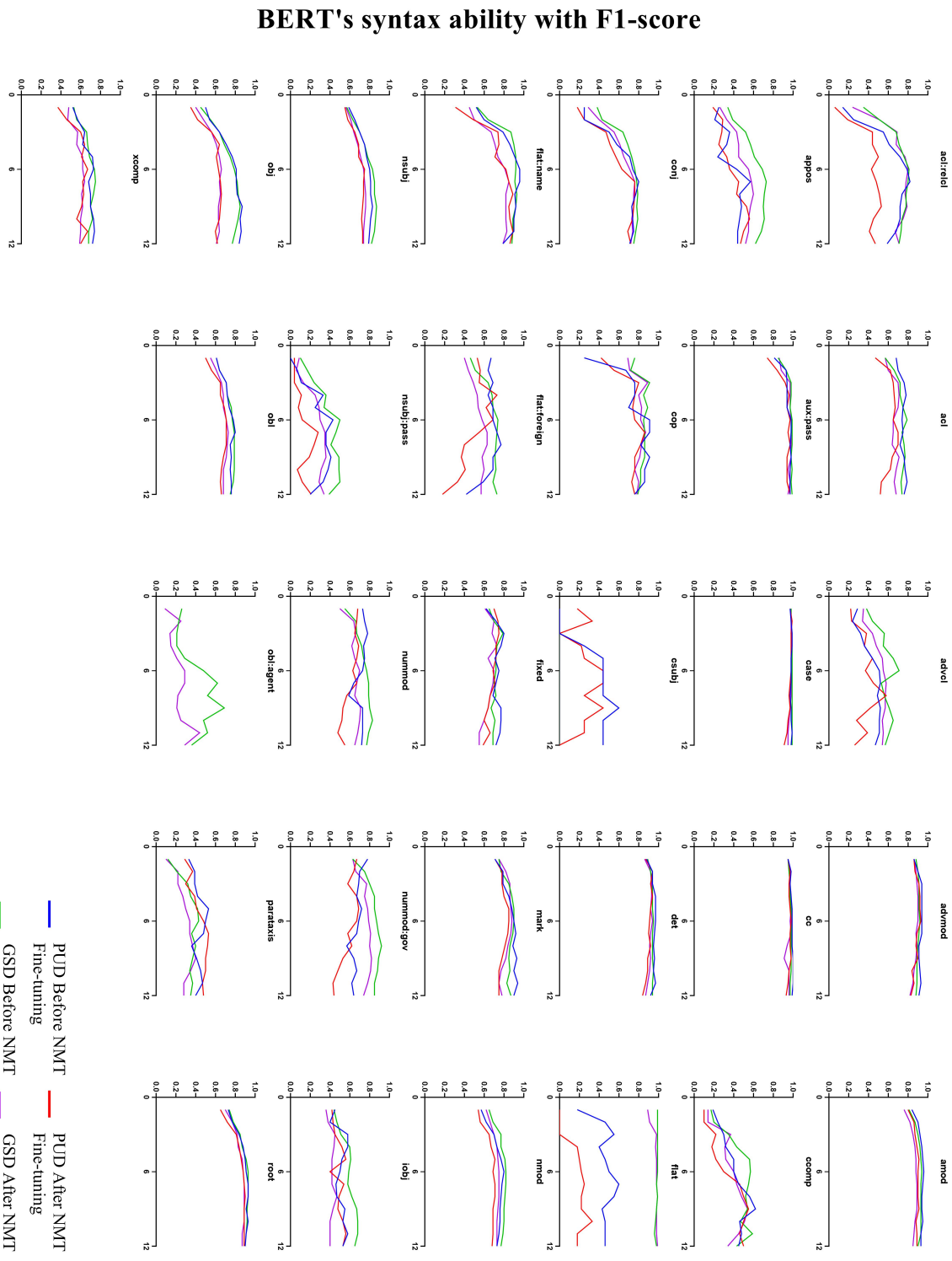


Figure 6: Full results of probing experiment for syntactic dependencies in Russian.

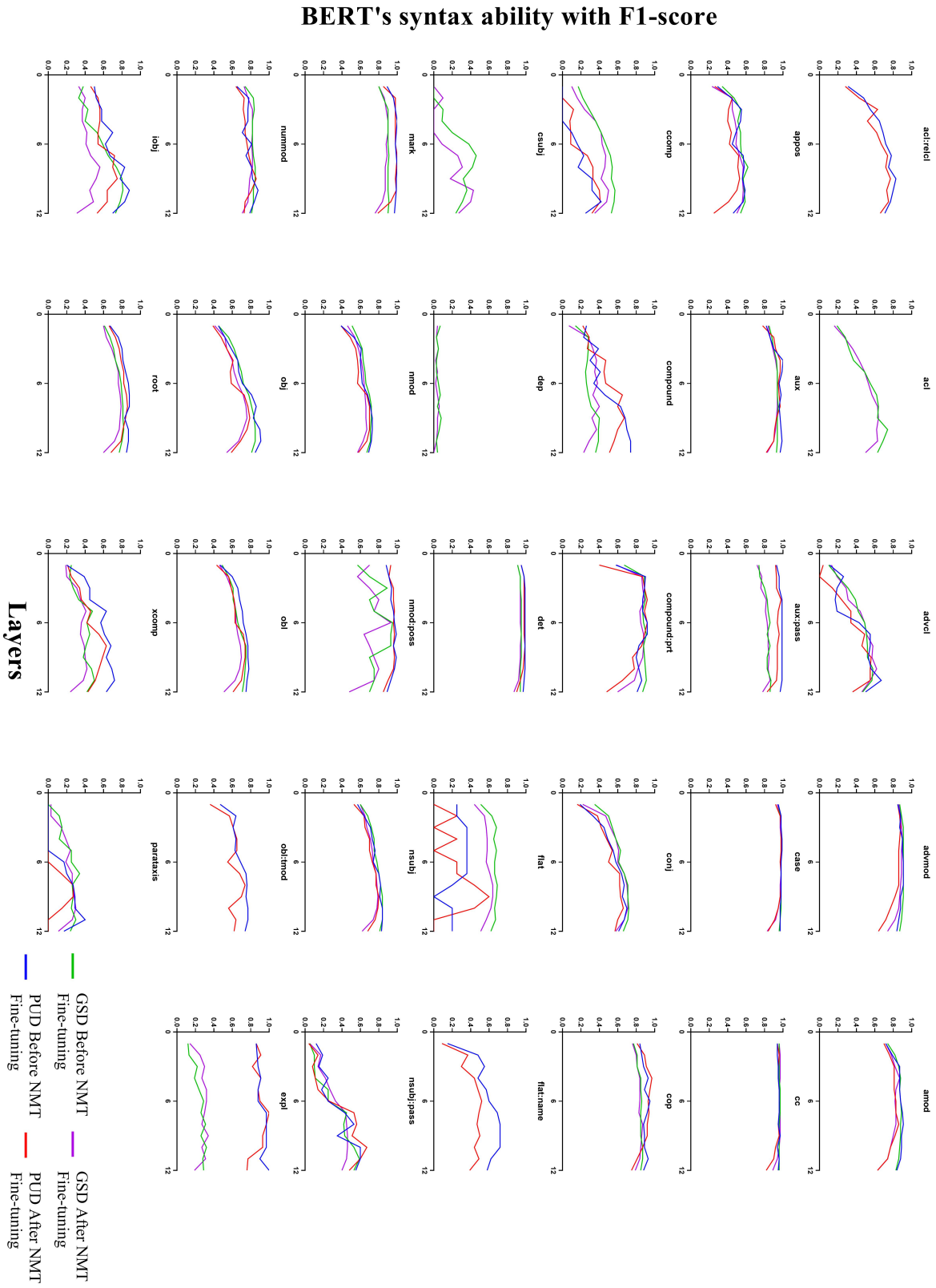


Figure 7: Full results of probing experiment for syntactic dependencies in German.

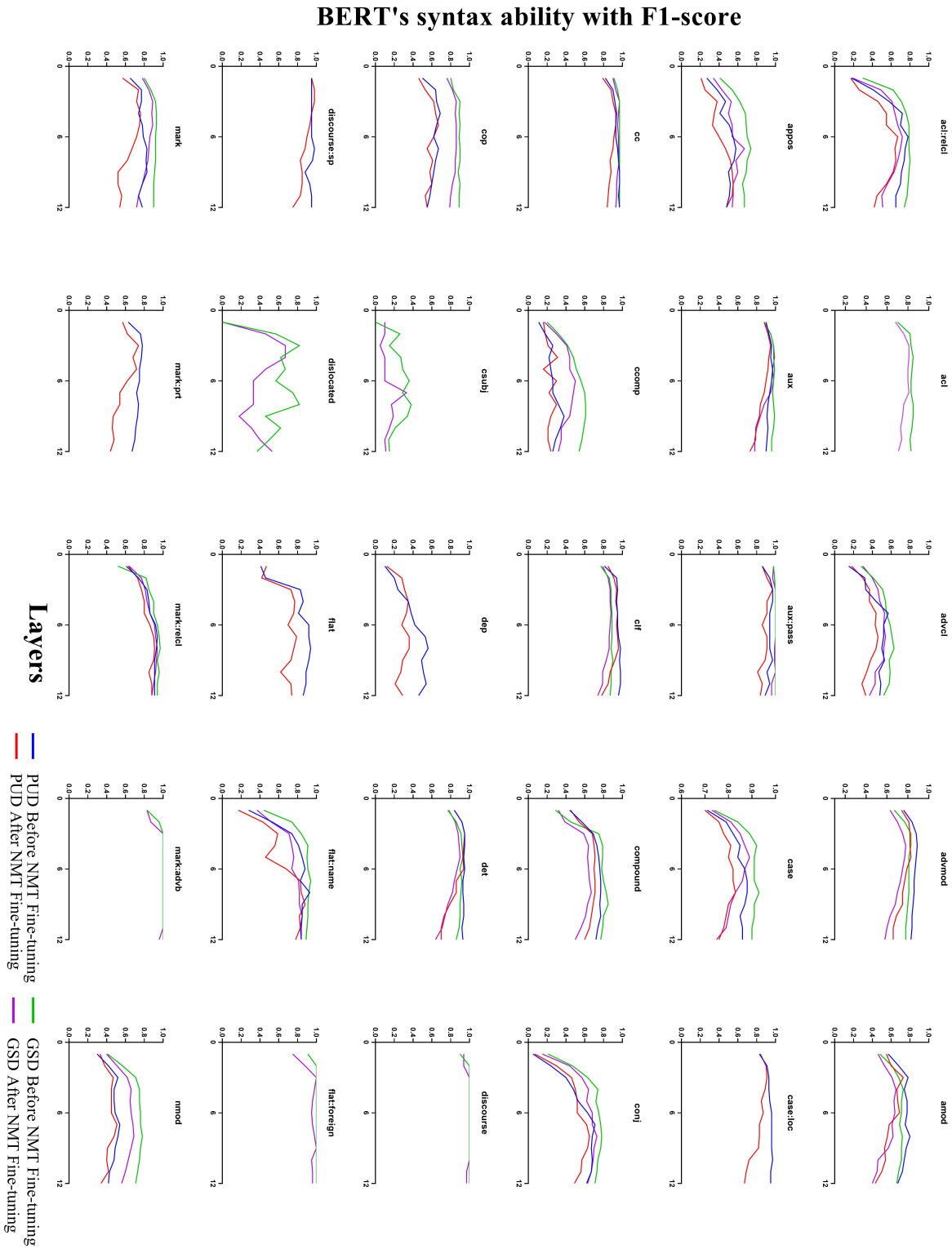


Figure 8: Full results of probing experiment for syntactic dependencies in Chinese.

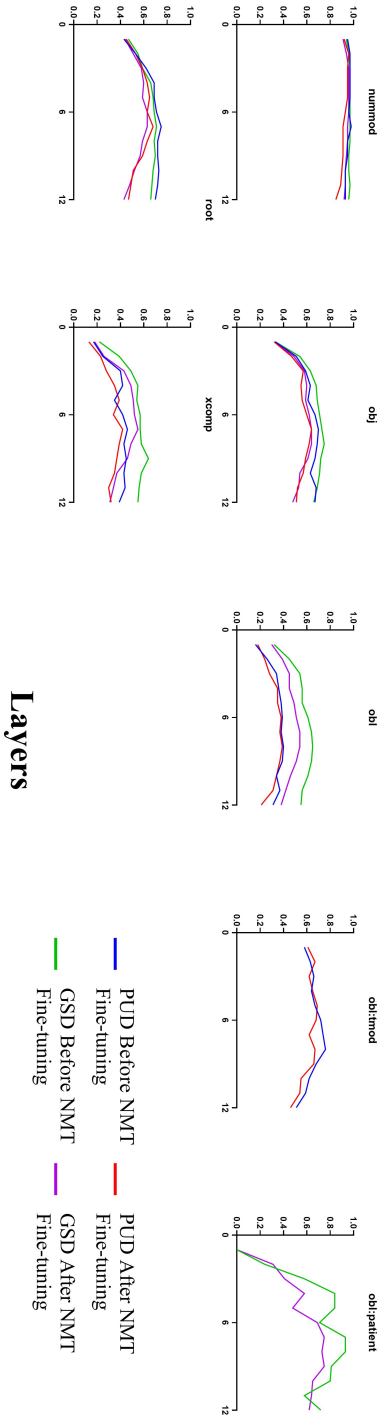


Figure 9: Full results of probing experiment for syntactic dependencies in Chinese.

Languages	Dependencies	High-quality	Low-quality	χ^2	Layer-6	Layer-12
	acl:relcl	43	56	3.93	0.67	0.66
	advcl	27	47	14.8	0.34	0.36
	advmod	222	208	0.88	0.86	0.64
	amod	197	217	2.03	0.83	0.63
	appos	10	96	739.6	0.42	0.25
	aux	70	52	4.62	0.95	0.82
	aux:pass	47	44	0.19	0.94	0.83
	case	324	459	56.25	0.97	0.84
	cc	129	142	1.31	0.96	0.82
	ccomp	20	26	1.8	0.09	0.32
	compound	25	72	88.36	0.47	0.51
	compound:prt	10	34	57.6	0.92	0.48
	conj	140	172	7.31	0.5	0.57
	cop	47	61	4.17	0.93	0.75
De	det	469	531	8.19	0.98	0.90
	flat	0	9	-	0.25	0
	flat:name	6	61	504.16	0.52	0.39
	mark	73	91	4.43	0.98	0.79
	nmod	193	177	1.32	0.57	0.58
	nmod:poss	39	67	20.1	0.96	0.85
	nsubj	241	308	18.62	0.73	0.68
	nsubj:pass	44	34	2.27	0.26	0.46
	nummod	40	45	0.62	0.76	0.73
	obj	154	179	4.05	0.59	0.59
	obl	212	327	62.38	0.63	0.61
	obl:tmod	14	34	28.57	0.55	0.62
	expl	22	11	5.5	0.90	0.76
	iobj	15	10	1.66	0.54	0.53
	xcomp	33	38	0.75	0.42	0.43
	parataxis	11	15	1.45	0	0

Table 8: Syntactic dependencies with the value of χ^2 in German.

Languages	Dependencies	High-quality	Low-quality	χ^2	Layer-6	Layer-12
	acl:relcl	67	99	15.28	0.67	0.42
	acl	3	1	1.33	-	-
	advcl	62	113	41.95	0.43	0.33
	advmod	231	250	1.56	0.79	0.66
	amod	95	75	4.21	0.69	0.43
	appos	10	73	396.9	0.4	0.48
	aux	130	132	0.03	0.9	0.73
	aux:pass	20	10	5	0.86	0.84
	case	214	286	24.22	0.82	0.76
	case:loc	63	77	3.11	0.87	0.67
	cc	52	48	0.3	0.91	0.84
	ccomp	70	72	0.05	0.30	0.24
	clf	65	77	2.21	0.94	0.84
	compound	267	333	16.31	0.70	0.60
	conj	61	71	1.63	0.52	0.49
	cop	40	65	15.62	0.62	0.55
Zh	dep	42	99	77.35	0.28	0.29
	discourse:sp	16	21	1.56	0.90	0.75
	flat	2	21	242	0.70	0.74
	flat:name	1	53	2704	0.68	0.78
	mark	32	66	36.12	0.72	0.54
	mark:prt	40	45	0.625	0.62	0.44
	mark:relcl	51	69	6.35	0.86	0.88
	nmod	123	145	3.96	0.45	0.34
	nsubj	293	380	25.83	0.63	0.60
	nsubj:pass	17	9	3.76	0	0.12
	nummod	137	169	7.47	0.93	0.85
	obj	285	297	5.89	0.60	0.51
	obl	86	132	24.6	0.38	0.21
	obl:tmod	28	54	24.14	0.68	0.46
	xcomp	76	111	16.11	0.34	0.32

Table 9: Syntactic dependencies with the value of χ^2 in Chinese.

Languages	Dependencies	High-quality	Low-quality	χ^2	Layer-6	Layer-12
	acl:relcl	21	28	2.33	0.43	0.47
	acl	35	41	1.02	0.65	0.52
	advcl	44	26	7.36	0.37	0.26
	advmod	189	162	3.85	0.93	0.83
	amod	274	347	19.44	0.90	0.89
	appos	8	31	66.12	0.36	0.47
	aux:pass	24	27	0.375	0.94	0.94
	case	306	406	32.67	0.98	0.91
	cc	82	123	20.5	0.98	0.93
	ccomp	28	22	1.28	0.30	0.50
	conj	93	147	31.35	0.63	0.72
	cop	13	17	1.23	0.76	0.76
	csubj	5	7	0.8	0.44	0
	det	75	98	7.05	0.91	0.84
	flat	1	31	900	0.22	0.18
Ru	flat:name	12	57	168.75	0.82	0.86
	flat:foreign	1	55	2916	0.69	0.18
	fixed	41	39	0.09	0.69	0.59
	mark	53	45	1.2	0.85	0.75
	nmod	309	330	1.42	0.69	0.68
	nsubj	243	273	3.7	0.75	0.73
	nsubj:pass	34	35	0.02	0.12	0.21
	nummod	34	29	0.73	0.63	0.55
	nummod:gov	10	20	10	0.67	0.44
	obj	21	35	9.33	0.63	0.62
	obl	115	137	4.2	0.71	0.66
	iobj	207	307	48.3	0.4	0.53
	xcomp	65	55	1.53	0.67	0.60
	parataxis	21	50	40.04	0.48	0.48

Table 10: Syntactic dependencies with the value of χ^2 in Russian.