Reproducing targeted syntactic evaluations of language models

Kalle Hilsenbek[1,⋆]

[1] *Universität Trier NLP*

Abstract. This work reproduces Cross-Linguistic Assessment of Models on Syntax (CLAMS) with Hugging Face. This CLAMS metric can supplement the loss and perplexity function to measure a model when fine-tuning on benchmarks isn't possible or too costly. A faster map-reduce algorithm is contributed to utilize parallel hardware. The test on different models shows its limitations due to different vocabulary and data set sizes between languages and models. A solution is proposed to adjust the different vocabularies, yet it isn't fully analyzed. The fork of the CLAMS repository is publicly accessible at https://github.com/Bachstelze/clams

1 Introduction

A word prediction model or language model (LM) specifies a probability distribution across word sequences. Numerous LM architectures based on neural networks have exploded as a result of recent technological advancements. This is similar to the development of translation models with its seq2seq architecture, which had an attention module to connect the encoder with the decoder [36]. The architectures of LMs, which can be seen as a single encoder, developed from recurrent neural networks (RNN [4, 23]) to Long short-term memory (LSTM [39]), convolutional neural networks (CNN) and fully attention-based models (BERT [15]).

In order to generally compare the developed models, they are fine-tuned on a variety of benchmarks [14, 27, 22]. While those benchmarks are widely adopted in the scientific community to measure new architectures, not all new trained models with a common model are measured like this and sometimes fall back to the loss function or perplexity. Perplexity is a common method for assessing language models. This measure combines various factors, such as popular colloquialisms, semantics, pragmatics, syntax, and others to determine whether a word will be predicted correctly or incorrectly [33]. Since most sentences are grammatically simple and most words can be predicted from their local context, ambiguity rewards LMs primarily for semantic predictions, making the quality of the LMs syntactic predictions appear to be particularly difficult to measure.

Marvin and Linzen created a data set, which was evaluated by humans and three LMs: an RNN LM trained on an unannotated corpus, an n-gram baseline, and an RNN LM trained on a multitask goal. On straightforward situations, the RNN LMs performed admirably, but poorly on more challenging ones. The RNN performed better after multi-task training with a supervised syntactic aim, although it was still far less effective than humans [33].

In "Targeted Syntactic Evaluation of Language Models" Rebecca Marvin and Tal Linzen suggest adding a metric to perplexity that determines if the probability distribution defined by the model complies with the language's grammar [33]. Syntactically difficult sentences in corpora have been identified in previous work on focused syntactic evaluation of language models [29]. Given two minimally different phrases, one of which is grammatical and the other is not [29,

⋆e-mail: kalle@hilsenbek.de

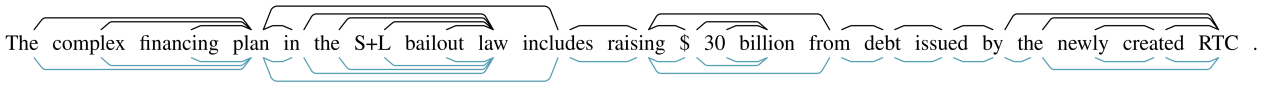


Figure 1. Minimum spanning trees resultant from predicted squared distances on BERTLARGE16. Black edges are the gold parse, above each sentence; blue are BERTLARGE16 [21].

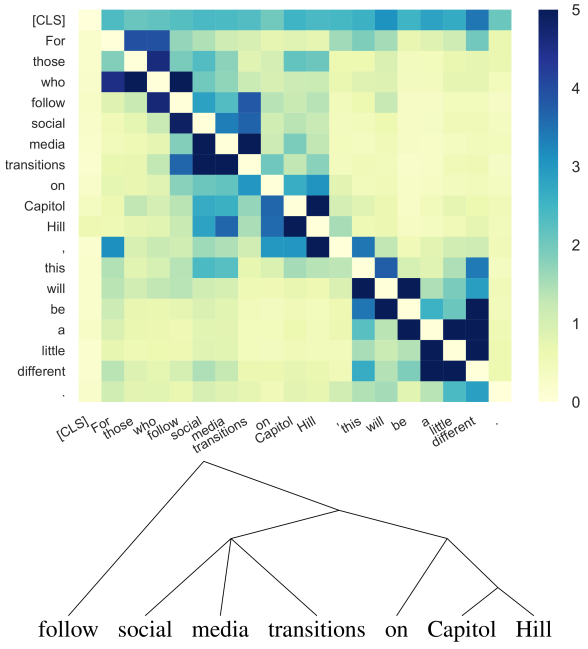


Figure 2. Words sharing syntactic subtrees have a greater impact on one another in the MLM prediction, according to a parameter-free test of syntactic knowledge by Wu et al. [48].

20], it is preferable for the model to provide a larger probability to the grammatical one.

The attempt by Tran et al. [24], which compared a LSTM language model to a fully attentional network (FAN) [36], provides an example of the value of this strategy. The attention-only model had three times as many errors as the LSTM when the models were explicitly tested on challenging subject-verb agreement dependencies, while having slightly lower perplexity on the validation set. Syntactical challenges are poorly represented in corpora. In the corpus, examples of pertinent constructions may be hard to uncover, and frequently, naturally occurring sentences contain statistical signals which allow the model to predict the right form of the verb without doing a sufficient syntactic analysis [45, 17, 43]. Rebecca Marvin and Tal Linzen constructed in "Targeted Syntactic Evaluation of Language Models" [33] a dataset that enables us to examine a considerably wider variety of particular grammatical phenomena (subject-verb agreement, reflexive anaphora, negative polarity items). The subject-verb agreement is a prototypical example. The verb must frequently agree in number (in this case, singular or plural) with the subject in many languages, including English. The following example is taken from Müller et al. [35] (asterisks denote word predictions that aren't grammatically correct): The key to the cabinets is/*are next to the coins.

The model must decide that the word "key", not "cabinets" or "coins", is the head of the sentence to properly forecast the form of the verb.

Besides the ability to cover all tokens of tokenizers like wordpiece [47], the vocabulary differs in most models due to the different vocabulary size and corpora. The probability of a target word can be calculated by the probabilities of its subtokens, though the syntactic evaluations just excluded the sentence pairs with multiple subtokens as the target. This leads to different sizes of test sets between languages and models, which lessens the comparability.

2 Related work

2.1 BERTology

Investigating the inner workings of massive transformers like BERT is a major subject of research; some refer to it as "BERTology". The model shows the typical NLP pipeline's processes in a comprehensible and localizable manner. Qualitative examination suggests that the model may dynamically modify this pipeline, altering lower-level judgments based on information from higher-level representations that can be used to resolve ambiguities [40]. BERT's attention heads display patterns like attending to delimiter tokens, certain positional offsets, or extensively attending throughout the entire phrase.

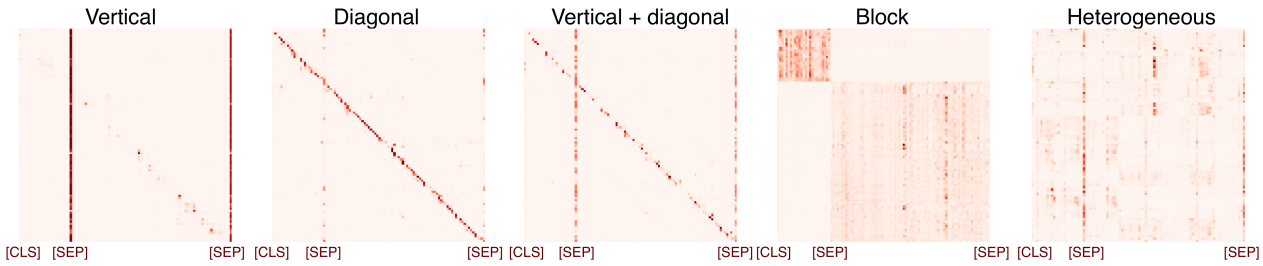


Figure 3. Attention patterns from Kovaleva et al. [25].

Heads in the same layer frequently display behaviors that are similar to one another [11].

To make it easier for users to access the internal representations, hugging face has added a few extra characteristics to the BERT models, e.g. a pipeline to fill in a masked token. Moreover, it is possible to access all of the BERT hidden states and all of the attention weights for each head to determine which model components depend more heavily on having several heads, so the obtained head values and gradients show the relevance score for the head to eventually trim the head [34].

The ability to capture syntax-sensitive phenomena is typically considered to prove the presence of a hierarchical structure. In "Open Sesame: Getting inside BERT's Linguistic Knowledge" [28] it is shown that BERT representations are mostly hierarchical. The syntactic structure can be reconstructed from the token representations [21] (see figure 1). According to the used fill probes in this work, BERT considers subject-predicate agreement while completing the mask task [38], even for phrases that lack sense or include distracting clauses in between the subject and the verb [18]. In terms of negative polarity items, BERT performs better then previous models [44, 37].

Many studies point to a shared structure between languages, if they are trained together like in the multilingual version of BERT (mBERT) [1, 5, 30, 41, 46, 32, 3, 12]. This multilingual structure exists even though the model was only trained on monolingual corpora without an interlingual signal or objective. In this limited work, we don't see an overall significant improvement in transfer learning with an optimized version of mBERT. But some models explicitly define an interlingual trainings objective, that could improve it.

Simple Agreement:
The author laughs/*laugh.

Across a Prepositional Phrase:
The farmer near the parents smiles/*smile.

Across a Subject Relative Clause:
The officers that love the skater *smiles/smile.

Short Verb Phrase Coordination:
The senator smiles and laughs/*laugh.

Long Verb Phrase Coordination:
The manager writes in a journal every day and likes/*like to watch television shows.

Across Object Relative Clause:
The farmer that the parents love swims/*swim.

Within Object Relative Clause:
The farmer that the parents *loves/love swims.

Figure 4. Syntactic constructions used in CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

The ability of BERT to capture linguistic knowledge has certain limitations and does not reach the gold standard. Several studies point out those drawbacks [16, 42, 2]. This leads us to the question, to which extent the limitations are overcome in newly developed models and how can we easily measure and compare the inner state, best already during training and fine-tuning.

2.2 Cross-Linguistic Assessment of Models on Syntax

In "Cross-Linguistic Syntactic Evaluation of Word Prediction Model" Müller et al. extend

```
vary: V[]
S[]     → je V[1,s]
V[1,s]  → pense
V[2,s]  → penses
V[1,p]  → pensons
V[2,p]  → pensez

True      je pense
False     je penses
False     je pensons
False     je pensez
```

Figure 5. An example of an attribute-varying grammar from Cross-Linguistic Assessment of Models on Syntax (CLAMS) by Müller et al. [35]. Preterminals are blue and attributes are orange. Here, the first statement is the "vary" statement. This is followed by a template, with the special S keyword in red. All remaining statements are preterminal definitions. The output of this AVG is at the bottom with "True" as grammatically correct.

Simple Agreement:

a. The surgeons laugh/*laughs.

b. Le pilote parle / *parlent.
The pilot laughs / *laugh.

c. Der Schriftsteller spricht / *sprechen.
The writer speaks / *speak.

d. Ha meltsar yashen / yeshenim.
The server sleeps / *sleep.

e. Врачи говорят / *говорит.
Doctors speak / *speaks.

Figure 6. This simple agreement involves agreeing a verb with its adjacent subject. Taken from CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

VP coordination (short):

a. The author swims and smiles/*smile.

b. Les directeurs parlent et déménagent / *déménage.
The directors talk and move / *moves.

c. Der Polizist schwimmt und lacht / *lachen.
The police.officer swims and laughs / *laugh.

d. Ha tabaxim rokdim ve soxim / *soxe.
The cooks dance and swim / *swims.

e. Профессор старый и читает / *читают.
Professor is.old and reads / *read.

Figure 7. Short verb-phrase coordination from CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

a subset from Marvin and Linzen (M+L) to the languages of German, French, Hebrew and Russian. The performance is considered to be comparable by concentrating on particular linguistic phenomena of the models across languages.

Müller et al.- built the data set using a grammatical engineering framework based on templates, they call attribute-varying grammars (AVGs). It is flexible and prevents sentences produced by a recursive context-free language from having an infinite depth [9]. The templates are made up of preterminals and terminals. A list of preterminals and the corresponding properties for each are listed in the "vary" statement (see figure 5). In most cases, only one preterminal per grammar is changed, such that each grammatical case is internally compatible with the changed syntactic feature.

2.2.1 Linguistic Examples

This section provides examples and descriptions of the syntactic structures taken as citations from the paper "Targeted Syntactic Evaluation of Language Model" [33] (only the reflexiva anaphora) and the rest is from the appendix of the CLAMS paper. For Hebrew, the original right-to-left script was transliterated into the left-to-right Latin script; this makes labeling and glossing more consistent across languages. Hebrew was not transliterated in the training/development/test corpora or in the evaluation sets. In all examples, (a) is English, (b) is French, (c) is German, (d) is Hebrew, and (e) is Russian. The first case is simple agreement. It simply involves agreeing a verb with its adjacent subject (see figure 6).

VP coordination (long):

a. The teacher knows many different foreign languages and likes/*like to watch television shows.

b. L' agriculteur écrit dans un journal tous
The farmer writes in a journal all
les jours et préfère / *préfèrent jouer au
the days and prefers / *prefer to.play at.the
tennis avec des collègues.
tennis with some colleagues.

c. Die Bauern sprechen viele verschiedene
The farmers speak many various
Sprachen und sehen / *sieht gern
languages and watch / *watches gladly
Fernsehprogramme.
TV.shows.

d. Ha tabax ohev litspot be toxniot televizya
The cook likes to.watch in shows TV
ve gar / *garim be merkaz ha ir.
and lives / *live in center the city.

e. Автор знает много иностранных
Author knows many foreign
языков и любит / *любят смотреть
languages and likes / *like to.watch
телепередачи.
TV.shows.

Figure 8. Long verb-phrase coordination from CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

Short verb-phrase coordination introduces some slight distance between the subject and verb, though the presence of the previous verb should give a model a clue as to which inflection should be more probable (see figure 7).

Long verb-phrase coordination is similar, but makes each verb phrase much longer to introduce more distance and attractors between the subject and target verb (see figure 8).

Agreement across a subject relative clause involves a subject with an attached relative clause containing a verb and object, followed by the main verb (see figure 9).

Agreement within an object relative clause requires the model to inflect the proper verb inside of an object relative clause; the object relative clause contains a noun and an associated

Across a subject relative clause:

a. The officers that love the chef are/*is old.

b. Les chirurgiens qui détestent le garde
The surgeons that hate the guard
retournent / *retourne.
return / *returns

c. Der Kunde, der die Architekten hasst, ist /
The customer that the architects hates is /
*sind klein.
*are short.

d. Ha menahel she ma'arits et ha shomer
The manager who admires ACC the guard
rats / *ratsim.
runs / *run.

e. Пилоты, которые понимают агентов,
Pilots that understand agents
говорят / *говорит.
speak / *speaks.

Figure 9. Agreement across a subject relative clause from CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

Within an object relative clause:

a. The senator that the executives love/*loves laughs.

b. Les professeurs que le chef admire /
The professors that the boss admires /
*admirent parlent.
*admire talk.

c. Die Polizisten, die der Bruder hasst, /
The police.officers that the brother hates /
*hassen, sind alt
*hate are old.

d. Ha menahel she ha nahag ma'aritz /
The manager that the driver admires /
*ma'aritsim soxe.
*admire swims.

e. Сенаторы, которых рабочие ищут, /
Senators that workers seek /
*ищет, ждали.
*seeks wait.

Figure 10. Agreement within an object relative clause from CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

Across an object relative clause:

a. The senator that the executives love laughs/*laugh.
b. Les professeurs que le chef admire parlent / The professors that the boss admires talk / *parle. *talks.
c. Der Senator, den die Tänzer mögen, spricht / The senator that the dancers like speaks / *sprechen. *speak.
d. Ha katsin she ha zamar ohev soxe / The officer that the singer likes swims / *soxim. *swim.
e. Фермеры, которых танцоры хотят, Farmers that dancers want большие / *большой. are.big / *is.big.

Figure 11. Agreement across an object relative clause from CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

Across a prepositional phrase:

a. The consultants behind the executive smile/*smiles.
b. Les clients devant l' adjoint sont / *est The clients in.front.of the deputy are / *is vieux. old.
c. Der Lehrer neben den Ministern lacht / The teacher next.to the ministers laughs / *lachen. *laugh.
d. Ha meltsarim leyad ha zamarim nos'im / The servers near the singers drive / *nose'a. *drives.
e. Режиссёры перед агентами Directors in.front.of agents маленькие / *маленький. are.small / *is.small.

Figure 12. Agreement across a prepositional phrase from CLAMS by Müller et al. [35]. Ungrammatical forms are marked with asterisks.

Simple reflexive:

a. The senators embarrassed themselves.
b. *The senators embarrassed herself.

Reflexive in a sentential complement:

a. The bankers thought the pilot embarrassed himself.
b. *The bankers thought the pilot embarrassed themselves.

Reflexive across an object relative clause:

a. The manager that the architects like doubted himself.
b. *The manager that the architects like doubted themselves.

Figure 13. Reflexive anaphora from Marvin and Linzen [33]. Ungrammatical forms are marked with asterisks.

transitive verb whose object requirement is filled by the relative pronoun. The model must choose the proper verb inflection given the noun within the relative clause as opposed to the noun outside of it. This may seem similar to simple agreement, but we now have an attractor which appears before the noun of the target verb (see figure 10).

Agreement across an object relative clause is similar, but now the model must choose the correct inflection for the noun outside of the relative clause (see figure 11).

Finally, agreement across a prepositional phrase entails placing a prepositional phrase after the subject; the prepositional phrase contains an attractor, which makes choosing the correct inflection more difficult (see figure 12).

Some of the constructions used by Marvin and Linzen could not be replicated across languages. This includes reflexive anaphora (see figure 13). A reflexive pronoun such as himself needs to have an antecedent from which it derives its interpretation. The pronoun needs to agree in number (and gender) with its antecedent [33]. None of the non-English languages use quite the same syntactic structures as English (or even to each other) when employing reflexive verbs and pronouns. Some

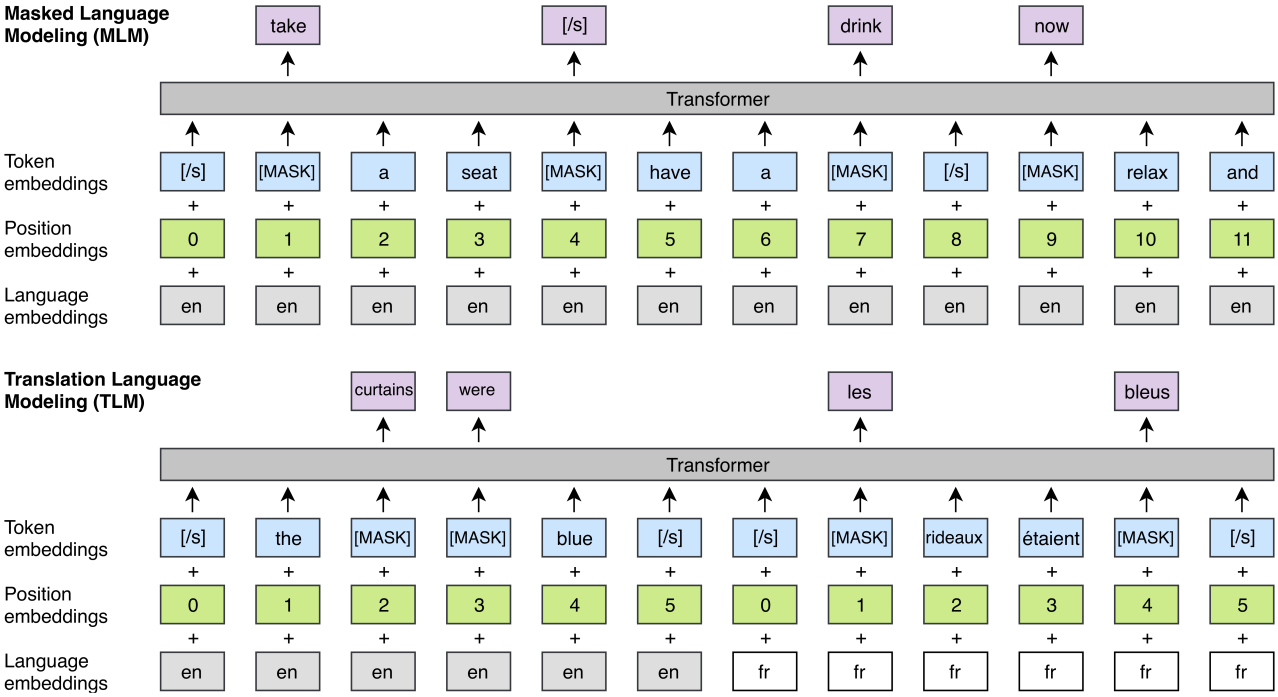


Figure 14. Cross-lingual language model pretraining. The MLM objective is similar to the one of Devlin et al. [15], but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

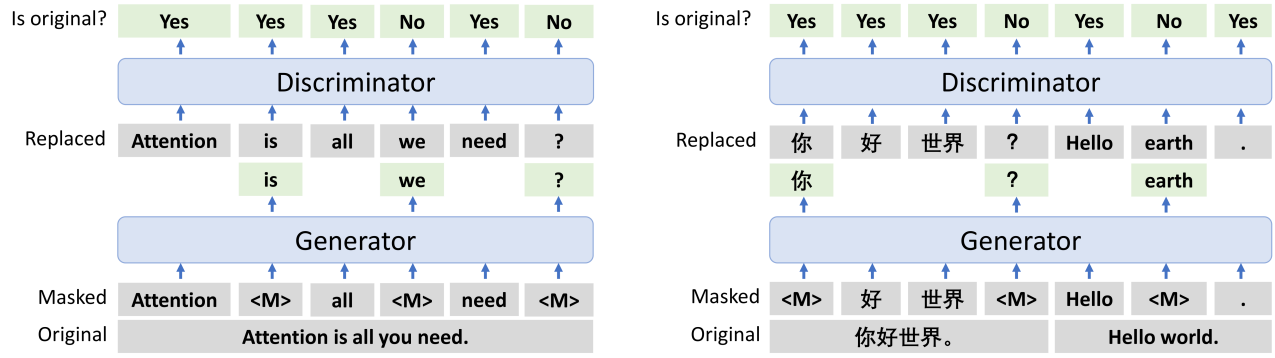
do not even have separate reflexive pronouns for third-person singular and plural distinctions (like French and German). Moreover, the English reflexive examples rely on the syncretism between past-tense verbs for any English person and number,[1] whereas other languages often have different surface forms for different person and number combinations in the past tense. This would give the model a large clue as to which reflexive is correct. Thus, any results on reflexive anaphora would not be comparable cross-linguistically. For the monolingual, English test cases, the reflexive anaphora is still valid and used in this work.

3 Multilingual attention-based models

LMs show the ability to model many languages in a shared embedding and transfer knowledge to smaller languages, similar to translation models. The well-established multilingual BERT has seen various improvements in its model like its monolingual version. Those models are good candidates to analyze with the CLAMS metric, because they aim to amplitude the multilingual signal. Although this metric also can be used for solely monolingual models. A next major model after BERT was RoBERTa (a Robustly Optimized BERT Pretraining Approach [31]). It removed the next sentence prediction objective and used a bigger corpus, vocabulary and parameter size. It was also adopted to a multilingual version with XLM-R and for the first time, a multilingual model has outperformed conventional monolingual baselines that were based on a pretrained model [13, 19]. The initial translation language modeling (TLM) was removed from the previous cross-lingual language model pretraining (XLM) [26] (see figure 14).

Other models altered the multilingual training signal [7]. XLM-E [8] uses the faster pretraining discriminator objective of ELECTRA [10] in a multilingual setting (see figure 15) and leverages

[1] For example, regardless of whether the subject is singular, plural, first- or third-person, etc., the past-tense of see is always saw.



(a) Multilingual replaced token detection (MRTD)

(b) Translation replaced token detection (TRTD)

Figure 15. Overview of two pre-training tasks of XLM-E [8], i.e., multilingual replaced token detection, and translation replaced token detection. The generator predicts the masked tokens given a masked sentence or a masked translation pair, and the discriminator distinguishes whether the tokens are replaced by the generator.

parallel data with the translation replaced token detection (TRTD). Or the denoising word alignment, a novel iterative cross-lingual pretraining objective used by XLM-Align (see figure 16) [6].

4 Experimental setup

The sequential algorithm by Müller et al. - which is itself a variation of former publications - is rewritten with the Hugging Face fill mask pipeline [2]. A mapping of the target tokens structures the sentences into dictionaries. In this way, for each target, there is a joined list of sentences with the similar target. The information, which sentence belongs to which sentence pair, is stored in separate dictionaries. After all lists have been put through the fill mask pipeline with a defined batch size, the scores of targets are compared and reduced to a single boolean value.

A good batch size of 32 was determined with the hardware from google colabs (see table 2). Only targets which are fully in the vocabulary are measured. The monolingual model doesn't share its vocabulary and has therefore the smallest amount of out-of-vocabulary tokens. The XLM-R model has due to its bigger vocabulary size a smaller of out-of-vocabulary tokens (see table 1).

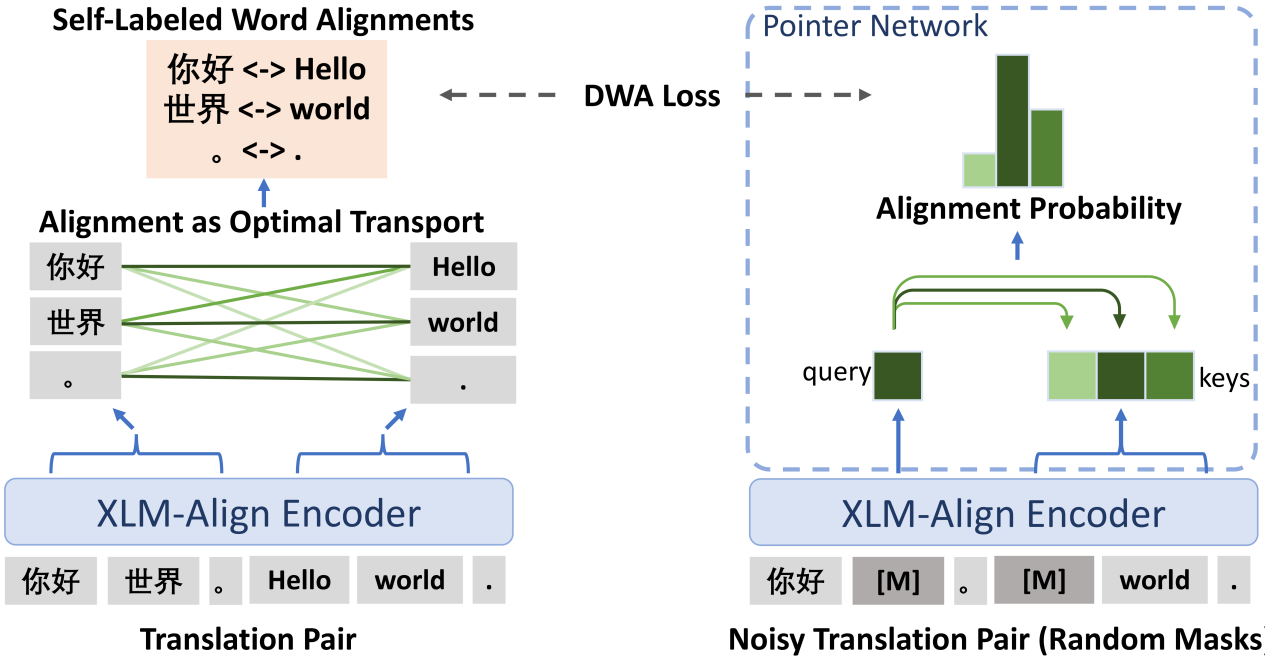
[2]https://huggingface.co/docs/transformers/main_classes/pipelines transformers.FillMaskPipeline

5 Conclusions

Both the monolingual English test and the multilingual test were carried out with the described setup. The monolingual subject-verb-agreement cases (see table 3) prove that the implemented algorithm mostly reproduces the previously reported values. The only difference was produced by the BERT with Hugging Face in the sentential complement, where all cases are stated as correct. The newly tested XLM-R has a slightly better average than the RNN, but the case numbers aren't the same (see table 4) and ultimately not directly comparable.

Much more comparable are the cases of the reflexiva anaphor, all numbers of the same type align to each other (see table 6). Humans score the highest accuracy in those cases, followed by BERT and its multilingual version. XLM-R performs slightly worse than mBERT, besides having a better performance in the simple reflexiva anaphora cases.

The multilingual tests between mBERT from the CLAMS paper, mBERT and XLM-R from Hugging Face show identical results for mBERT except the prepositional phrase in Hebrew (see table 9). XLM-R performs better in the lower resource languages Hebrew and Russian, though only the Russian are comparable. On the other hand, the XLM-R model performs worse at English, French and German In the comparable cases.



(a) Word alignment self-labeling

(b) Denoising word alignment

Figure 16. XLM-ALIGN is pretrained in an expectation-maximization manner with two alternating steps. (a) Word alignment self-labeling: we formulate word alignment as an optimal transport problem, and self-labels word alignments of the input translation pair on-the-fly; (b) Denoising word alignment: update of the model parameters with the denoising word alignment task, where the model uses a pointer network to predict the aligned tokens from the perturbed translation pair [6].

	English	French	German	Hebrew	Russian
mBERT	12/22	15/24	12/20	32/36	23/30
XLM-R	10/22	14/24	9/20	22/36	11/30
en BERT	2/33	-	-	-	-

Table 1. Ratio between out-of-vocabulary tokens to overall amount of target tokens in the CLAMS data set.

A complete comparison is prevented by different vocabularies, which lead to different set sizes and multiple levels of difficulty between languages and models. The possible solution is to calculate all the scores of the targets by iterative scoring the subtokens[3]. This would also complicate the level of difficulty, as there would be words tested, which were rare in the training corpus. For ease of use, the linked implementation should be reimplemented into the evaluation library of Hugging Face[4].

[3] https://github.com/Bachstelze/clams/blob/master/bert-syntax/eval_missing_vocab.py

[4] https://github.com/huggingface/evaluate

References

[1] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. "On the Cross-lingual Transferability of Monolingual Representations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4623–4637. DOI: 10.18653/v1/2020.acl-main.421. URL: https://aclanthology.org/2020.acl-main.421.

[2] Sriram Balasubramanian et al. "What's in a Name? Are BERT Named Entity Representations just as Good for any other Name?" In: *Proceedings of the 5th Workshop on Representation Learning for NLP*.

batch size	GPU seconds	CPU seconds
0	73.1	294
1	15.1	156.6
2	9.5	110.8
4	6.9	90.93
8	5.5	86.3
16	5.5	80.4
32	5.2	73.9
64	5.4	73.4
128	5.3	80.6
256	5.7	85.1
512	12.8	83.1

Table 2. Speed comparison of GPU and CPU due to the batch size with 1024 sentences

Online: Association for Computational Linguistics, July 2020, pp. 205–214. DOI: 10.18653/v1/2020.repl4nlp-1.24. URL: https://aclanthology.org/2020.repl4nlp-1.24.

[3] Johannes Bjerva and Isabelle Augenstein. "Does Typological Blinding Impede Cross-Lingual Sharing?" In: *CoRR* abs/2101.11888 (2021). arXiv: 2101.11888. URL: https://arxiv.org/abs/2101.11888.

[4] Tomas Mikolov; Martin Karafiat; Lukas Burget; Jan Cernocky and Sanjeev Khudanpur. "Recurrent neural network based language model". In: *In Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), pages 1045–1048* (2010).

[5] Ethan A. Chi, John Hewitt, and Christopher D. Manning. "Finding Universal Grammatical Relations in Multilingual BERT". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5564–5577. DOI: 10.18653/v1/2020.acl-main.493. URL: https://aclanthology.org/2020.acl-main.493.

[6] Zewen Chi et al. "Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3418–3430. DOI: 10.18653/v1/2021.acl-long.265. URL: https://aclanthology.org/2021.acl-long.265.

[7] Zewen Chi et al. "InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 3576–3588. DOI: 10.18653/v1/2021.naacl-main.280. URL: https://aclanthology.org/2021.naacl-main.280.

[8] Zewen Chi et al. "XLM-E: Cross-lingual Language Model Pre-training via ELECTRA". In: *CoRR* abs/2106.16138 (2021). arXiv: 2106.16138. URL: https://arxiv.org/abs/2106.16138.

[9] N. Chomsky. "Three models for the description of language". In: *IRE Transactions on Information Theory* 2.3 (1956), pp. 113–124. DOI: 10.1109/TIT.1956.1056813.

[10] Kevin Clark et al. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *CoRR* abs/2003.10555 (2020). arXiv: 2003.10555. URL: https://arxiv.org/abs/2003.10555.

[11] Kevin Clark et al. "What Does BERT Look At? An Analysis of BERT's Attention". In: *CoRR* abs/1906.04341 (2019). arXiv: 1906.04341. URL: http://arxiv.org/abs/1906.04341.

[12] Alexis Conneau et al. "Emerging Cross-lingual Structure in Pretrained Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 6022–6034. DOI: 10.18653/v1/2020.acl-main.536. URL: https://aclanthology.org/2020.acl-main.536.

Subject-verb-agreement cases	Humans	RNN	BERT	BERT pipe	mBERT	mBERT pipe	XLM-R pipe
Simple	0.96	1.00	1.00	1.00	1.00	1.00	1.00
In a sentential complement	0.93	0.93	0.83	1.00	1.00	1.00	1.00
VP coordination (short)	0.94	0.90	0.89	0.89	1.00	1.00	0.99
VP coordination (long)	0.82	0.81	0.98	0.98	0.92	0.92	0.97
Across subject rel. clause	0.88	0.74	0.84	0.84	0.88	0.88	0.59
Within object rel. clause	0.78	0.89	0.95	0.95	0.83	0.83	-
Within object rel. clause (no that)	0.79	0.81	0.79	0.79	0.61	0.61	-
Across object rel. clause	0.85	0.57	0.89	0.89	0.87	0.87	0.79
Across object rel. clause (no that)	0.82	0.52	0.86	0.86	0.64	0.64	0.71
Across prepositional phrase	0.85	0.69	0.85	0.85	0.9	0.92	0.63
Average accuracy	0.86	0.79	0.89	0.89	0.87	0.87	0,835

Table 3. Subject-verb-agreement accuracies with English BERT (base), multilingual BERT and XLM-R on the English stimuli are from Marvin and Linzen [33]. Also the results for humans and the multitask RNN are from Marvin and Linzen. Monolingual BERT results are taken from Goldberg [18]. Multilingual results are taken from CLAMS. The pipe versions are the results produced with the hugging face pipeline.

Subject-verb-agreement quantity	Humans (M+L)	RNN (M+L)	M+L pairs	BERT	mBERT	XLM-R
Simple	140	140	140	120	80	80
In a sentential complement	1680	1680	1680	1440	960	960
VP coordination (short)	840	840	840	720	480	480
VP coordination (long)	400	400	400	400	240	160
Across subject rel. clause	11200	11200	11200	9600	6400	6400
Within object rel. clause	22400	22400	22400	15960	5320	-
Within object rel. clause (no that)	22400	22400	22400	15960	5320	-
Across object rel. clause	22400	22400	22400	19680	16480	13600
Across object rel. clause (no that)	22400	22400	22400	19680	16480	13600
Across prepositional phrase	22400	22400	22400	19440	14640	13200

Table 4. The quantity of subject-verb-agreement sentence pairs, most of the numbers don't align.

Reflexiva anaphora cases	Humans	RNN	BERT	BERT pipe	mBERT	mBERT pipe	XLM-R pipe
Simple	**0.96**	0.86	0.94	0.94	0.87	0.87	0.93
In a sentential complement	**0.91**	0.83	0.89	0.89	0.89	0.89	0.78
Across a relative clause	**0.87**	0.56	0.80	0.80	0.74	0.74	0.74
Average accuracy	**0.91**	0.75	0.88	0.88	0.83	0.83	0.82

Table 5. Overall accuracies for the reflexiva anaphora with the same load as in table 1.

Reflexiva anaphora quantity	Humans (M+L)	RNN (M+L)	M+L pairs	BERT	mBERT	XLM-R	BERT pairs
Simple	280	280	280	280	280	280	280
In a sentential complement	3360	3360	3360	3360	3360	3360	3360
Across a relative clause	22400	22400	22400	22400	22400	22400	22400

Table 6. The quantity of reflexiva anaphora sentence pairs, all numbers of the same type align to each other.

Negative polarity items	Humans	RNN	BERT	BERT pipe	mBERT	mBERT pipe	XLM-R pipe
Simple	0.98	0.48	1.0	1.0	0.0	0.0	0.89
Across a relative clause	0.81	0.73	-	-	-	-	-
Average accuracy	0.9	0.6	1.0	1.0	0.0	0.0	0.89

Table 7. Overall accuracies for the negative polarity items with the same load as in table 1.

	Humans (M+L)	RNN (M+L)	all M+L pairs	BERT	mBERT	XLM-R	BERT pairs
Simple	792	792	792	246	246	246	246
Across a relative clause	31680	31680	31680	-	-	-	-

Table 8. The quantity of negative polarity items

	English	French	German	Hebrew	Russian
Simple agreement	**1.00 | 1.00 | 1.00**	**1.00 | 1.00 | 1.00**	**0.95 | 0.95 | 0.95**	**0.70 | 0.70** | 0.78	**0.65 | 0.65 | 0.78**
VP coordination (short)	**1.00 | 1.00 | 0.99**	**1.00 | 1.00 | 1.00**	**0.97 | 0.97 | 1.00**	**0.91 | 0.90** | 0.98	**0.80 | 0.80 | 0.98**
VP coordination (long)	**0.92 | 0.92** | 0.97	**0.98 | 0.98 | 1.00**	**1.00 | 1.00** | 1.00	**0.73 | 0.73** | 1.00	**- | -** | 1.00
Across subject relative clause	**0.88 | 0.88 | 0.59**	**0.57 | 0.57 | 0.52**	0.73 | **1.00 | 0.98**	**0.61 | 0.61** | 0.93	**0.70 | 0.70 | 0.93**
Within object relative clause	0.83 | 0.81 | -	**- | - | -**	**- | - | -**	**- | -** | 0.96	**- | -** | 0.96
Across object relative clause	0.87 | **0.86 | 0.67**	**0.86 | 0.86 | 0.71**	0.93 | **1.00 | 0.95**	**0.55 | 0.55** | 0.97	**0.67 | 0.67 | 0.97**
Across prepositional phrase	0.92 | **0.92 | 0.57**	**0.57 | 0.57 | 0.56**	**0.95 | 0.95 | 0.76**	**0.62 | 0.55** | 0.55	**0.56 | 0.56 | 0.55**

Table 9. Multilingual BERT accuracies on CLAMS. The three items in the triplets are standing for the original mBERT value, the reproduced mBERT result and the XLM-R result. Bold items have the same amount of cases and are comparable.

[13] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.02116 (2019). arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

[14] Alexis Conneau et al. "XNLI: Evaluating Cross-lingual Sentence Representations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2475–2485. DOI: 10.18653/v1/D18-1269. URL: https://aclanthology.org/D18-1269.

[15] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[16] Allyson Ettinger. "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models". In: *CoRR* abs/1907.13528 (2019). arXiv: 1907.13528. URL: http://arxiv.org/abs/1907.13528.

[17] Richard Futrell et al. "Neural language models as psycholinguistic subjects: Representations of syntactic state". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 32–42. DOI: 10.18653/v1/N19-1004. URL: https://aclanthology.org/N19-1004.

[18] Yoav Goldberg. "Assessing BERT's Syntactic Abilities". In: *CoRR* abs/1901.05287 (2019). arXiv: 1901.05287. URL: http://arxiv.org/abs/1901.05287.

[19] Naman Goyal et al. "Larger-Scale Transformers for Multilingual Masked Language Modeling". In: *CoRR* abs/2105.00572 (2021). arXiv: 2105.00572. URL: https://arxiv.org/abs/2105.00572.

	English	French	German	Hebrew	Russian
Simple agreement	**80|80|80**	**40|40|40**	**100|100|100**	**20|20**|80	**80|80|80**
VP coordination short	**480|480|480**	**140|140|140**	**700|700|700**	**140|140**|280	**280|280|280**
VP coordination long	**240|240**|160	**100|100|100**	**300|300**|400	**100|100**|300	**0|0**|300
Across subject relative cl.	**6400|6400|6400**	**1600|1600|1600**	5406|**8000|8000**	**1600|1600**|2880	**2880|2880|2880**
Within object rel. cl.	5320|2800|0	**0|0|0**	**0|0|0**	**0|0**|8400	**0|0**|8400
Across object rel. cl.	16480|**6400|6400**	**1600|1600|1600**	5620|**8000|8000**	**1600|1600**|3200	**3200|3200|3200**
Across prep. phrase	14640|**9600|9600**	**2000|2000|2000**	**9000|9000|9000**	**800|800**|1680	**1680|1680|1680**

Table 10. Quantities of CLAMS sentences pairs. The three items in the triplets are standing for the original mBERT quantity, the reproduced mBERT quantity and the XLM-R quantity. Bold items have the same amount and are comparable.

[20] Kristina Gulordava et al. "Colorless Green Recurrent Networks Dream Hierarchically". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1195–1205. DOI: 10.18653/v1/N18-1108. URL: https://aclanthology.org/N18-1108.

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

[22] Junjie Hu et al. "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization". In: *CoRR* abs/2003.11080 (2020). arXiv: 2003.11080. URL: https://arxiv.org/abs/2003.11080.

[23] Rafal Józefowicz et al. "Exploring the Limits of Language Modeling". In: *CoRR* abs/1602.02410 (2016). arXiv: 1602.02410. URL: http://arxiv.org/abs/1602.02410.

[24] Christof Monz Ke Tran Arianna Bisazza. "The Importance of Being Recurrent for Modeling Hierarchical Structure". In: *Arxiv* (2018). repository: https://github.com/ketranm/fan_vs_rnn. URL: https://arxiv.org/pdf/1803.03585.pdf.

[25] Olga Kovaleva et al. "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4365–4374. DOI: 10.18653/v1/D19-1445. URL: https://aclanthology.org/D19-1445.

[26] Guillaume Lample and Alexis Conneau. "Cross-lingual Language Model Pretraining". In: *CoRR* abs/1901.07291 (2019). arXiv: 1901.07291. URL: http://arxiv.org/abs/1901.07291.

[27] Patrick Lewis et al. "MLQA: Evaluating Cross-lingual Extractive Question Answering". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7315–7330. DOI: 10.18653/v1/2020.acl-main.653. URL: https://aclanthology.org/2020.acl-main.653.

[28] Yongjie Lin, Yi Chern Tan, and Robert Frank. "Open Sesame: Getting inside BERT's Linguistic Knowledge". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 241–253. DOI: 10.

18653 / v1 / W19 - 4825. URL: https : / / aclanthology.org/W19-4825.

[29] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535. DOI: 10.1162/tacl_a_00115. URL: https://aclanthology.org/Q16-1037.

[30] Qi Liu, Matt J. Kusner, and Phil Blunsom. "A Survey on Contextual Embeddings". In: *CoRR* abs/2003.07278 (2020). arXiv: 2003.07278. URL: https://arxiv.org/abs/2003.07278.

[31] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[32] José Antonio Hernández López et al. *AST-Probe: Recovering abstract syntax trees from hidden representations of pre-trained language models*. 2022. DOI: 10.48550/ARXIV.2206.11719. URL: https://arxiv.org/abs/2206.11719.

[33] Rebecca Marvin and Tal Linzen. "Targeted Syntactic Evaluation of Language Models". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1192–1202. DOI: 10.18653/v1/D18-1151. URL: https://aclanthology.org/D18-1151.

[34] Paul Michel, Omer Levy, and Graham Neubig. "Are Sixteen Heads Really Better than One?" In: *CoRR* abs/1905.10650 (2019). arXiv: 1905.10650. URL: http://arxiv.org/abs/1905.10650.

[35] Aaron Mueller et al. "Cross-Linguistic Syntactic Evaluation of Word Prediction Models". In: *CoRR* abs/2005.00187 (2020). arXiv: 2005.00187. URL: https://arxiv.org/abs/2005.00187.

[36] Ashish Vaswani; Noam Shazeer Niki Parmar; Jakob Uszkoreit; Aidan N. Gomez; Łukasz Kaiser; Illia Polosukhin. "Attention Is All You Need". In: *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*. (2017). URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[37] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A Primer in BERTology: What we know about how BERT works". In: *CoRR* abs/2002.12327 (2020). arXiv: 2002.12327. URL: https://arxiv.org/abs/2002.12327.

[38] Marten van Schijndel, Aaron Mueller, and Tal Linzen. "Quantity doesn't buy quality syntax with neural language models". 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5831–5837. DOI: 10.18653/v1/D19-1592. URL: https://aclanthology.org/D19-1592.

[39] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. "LSTM Neural Networks for Language Modeling". In: *INTERSPEECH*. 2012.

[40] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline". In: *CoRR* abs/1905.05950 (2019). arXiv: 1905.05950. URL: http://arxiv.org/abs/1905.05950.

[41] Ivan Vulic et al. "Multi-SimLex: A Large-Scale Evaluation of Multilingual and Cross-Lingual Lexical Semantic Similarity". In: *CoRR* abs/2003.04866 (2020). arXiv: 2003.04866. URL: https://arxiv.org/abs/2003.04866.

[42] Eric Wallace et al. "Do NLP Models Know Numbers? Probing Numeracy in Embeddings". In: *CoRR* abs/1909.07940 (2019). arXiv: 1909.07940. URL: http://arxiv.org/abs/1909.07940.

[43] Alex Warstadt et al. "BLiMP: A Benchmark of Linguistic Minimal Pairs for English". In: *CoRR* abs/1912.00582 (2019). arXiv: 1912.00582. URL: http://arxiv.org/abs/1912.00582.

[44] Alex Warstadt et al. “Investigating BERT’s Knowledge of Language: Five Analysis Methods with NPIs”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2877–2887. DOI: 10.18653/v1/D19-1286. URL: https://aclanthology.org/D19-1286.

[45] Ethan Wilcox et al. “What do RNN Language Models Learn about Filler–Gap Dependencies?” In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 211–221. DOI: 10.18653/v1/W18-5423. URL: https://aclanthology.org/W18-5423.

[46] Shijie Wu and Mark Dredze. “Are All Languages Created Equal in Multilingual BERT?” In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. Online: Association for Computational Linguistics, July 2020, pp. 120–130. DOI: 10.18653/v1/2020.repl4nlp-1.16. URL: https://aclanthology.org/2020.repl4nlp-1.16.

[47] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

[48] Zhiyong Wu et al. “Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4166–4176. DOI: 10.18653/v1/2020.acl-main.383. URL: https://aclanthology.org/2020.acl-main.383.