AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level

Anonymous ACL submission

Abstract

Large Pre-trained Language Models (PLMs) have become ubiquitous in the development of language understanding technology and lie at the heart of many artificial intelligence advances. While advances reported for English using PLMs are unprecedented, reported advances using PLMs for Hebrew are few and far between. The problem is twofold. First, so far, Hebrew resources for training large language models are not of the same magnitude as their English counterparts. Second, there are no accepted benchmarks to evaluate the progress of Hebrew PLMs on, and in particular, sub-word (morphological) tasks. We aim to remedy both aspects. We present *AlephBERT*, a large PLM for Modern Hebrew, trained on larger vocabulary and a larger dataset than any Hebrew PLM before. Moreover, we introduce a *novel language-agnostic* architecture that can recover all of the sub-word morphological segments encoded in contextualized word embedding vectors. Based on this new morphological component we offer a new PLM evaluation suite consisting of multiple tasks and benchmarks, that cover *sentence level word-level* and *sub-word level* analyses. On all tasks, *AlephBERT* obtains state-of-the-art results beyond contemporary Hebrew baselines. We make our *AlephBERT* model, the morphological extraction mode, and the Hebrew evaluation suite publicly available, providing a single point of entry for assessing Hebrew PLMs.

1 Introduction

We presents a case study of PLM development for a *morphologically-rich* and *medium-resourced* language. Specifically, we address Modern Hebrew, a Semitic language, long known to be notoriously hard to process (Tsarfaty et al., 2019). The challenges posed to automatically processing Hebrew and obtaining good accuracy on downstream tasks stem from (at least) two main factors. The first is the internal-complexity of word-tokens, resulting from the rich morphology, complex orthography, and lack of diacritization in Hebrew written texts. Space-delimited tokens have non-transparent decomposition and are highly ambiguous, making even the simplest of the tasks in the pipeline very challenging (Tsarfaty et al., 2019). The second factor is the fact that Modern Hebrew, with only a few dozens of millions of native speakers, is often studied in resource-scarce settings.

Contextualized word representations, provided by models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), were shown in recent years to be critical for obtaining state-of-the-art performance on a wide range of Natural Language Processing (NLP) tasks — such as tagging and parsing, question answering, natural language inference, text summarization, natural language generation, and many more. These contextualized word representations are obtained by pre-training a large language model on massive quantities of unlabeled textual data, aiming to optimize simple yet effective objectives such as *masked word prediction* and *next sentence prediction*.

While advances reported for English using such models are unprecedented, in Modern Hebrew, previously reported results using PLMs are far from satisfactory. Specifically, the BERT-based Hebrew section of multilingual-BERT (Devlin et al., 2019) (henceforth, mBERT), did not provide a similar boost in performance as observed by the English section of mBERT. In fact, for several reported tasks, the mBERT model results are on a par with pre-neural models, or neural models based on non-contextualized embedding (Tsarfaty et al., 2020; Klein and Tsarfaty, 2020). An additional Hebrew BERT-based model, HeBERT (Chriqui and Yahav, 2021), has been recently released, yet without empirical evidence of performance improvements on key components of the Hebrew NLP pipeline.

The deficiency in Hebrew resources is problematic for PLM development in at least two ways.

First, the amount of raw text published and *available* for training PLMs is relatively small. To wit, the Hebrew Wikipedia used for training mBERT is of orders of magnitude smaller than the English Wikipedia (See Table 1). Secondly, there are no commonly accepted benchmarks for evaluating the performance of Hebrew PLMs on NL processing and understanding tasks. Translation of the English NLU benchmarks into Hebrew is a feasible solution for initial PLM evaluation. However no such effort has been undertaken to date, and, more importantly, such tasks do not address morphological-level evaluation, which is critical for *Morphologically Rich Languages* (MLRs).

Evaluating BERT-based models on morpheme-level tasks is non trivial. PLMs employ sub-word tokenization mechanisms such as WordPiece and Byte-Pair Encoding, for the purposes of minimizing Out-Of-Vocabulary words. These sub-word tokens are generated in a pre-processing step and passed as input to the PLM. In particular they are generated in a statistical manner without utilization of linguistic information, and consequently these sub-word tokens are assigned contextualized vectors by PLMs but they do not reflect morphological segments in any way. Extracting morphological units from contextualized vectors provided by PLMs is thus challenging, yet necessary in order to enable morphological level evaluation. To address this we introduce a *novel language-agnostic* architecture that recovers the *morphological* sub-word segments encoded in the contextualized embeddings output by PLMs.

We propose an evaluation setup for PLMs covering various processing levels tailored to fit MRLs, i.e. test on sentence, word and most importantly sub-word morphological tasks. These tasks include: **Segmentation, Part-of-Speech Tagging, full Morphological Tagging, Dependency Parsing, Named Entity Recognition** and **Sentiment Analysis**.

We present *AlephBERT*, a Hebrew pre-trained language model, larger and trained on more data than any Hebrew PLM before, and confirm SOTA results on *all* existing Hebrew benchmarks and scheme variants. We make our PLM and online demo publicly available[1] allowing to qualitatively assess present and future Hebrew PLMs.

[1] www.anonymous.org

2 Previous Work

Contextualized word embedding vectors are a major driver for improved performance of deep learning models on many NLU tasks. Initially, ELMo (Peters et al., 2018) and ULMFit (Howard and Ruder, 2018) introduced contextualized word embedding frameworks by training LSTM-based models on massive amounts of texts. The linguistic quality encoded in these models was demonstrated over 6 NLU tasks: Question Answering, Textual Entailment, Semantic Role labeling, Coreference Resolution, Name Entity Extraction, and Sentiment Analysis. The next big leap was obtained with the introduction of the GPT-1 framework by Radford and Sutskever (2018). Instead of using LSTM layers, GPT is based on 12 layers of Transformer decoders with each decoder layer is composed of a 768-dimensional feed-forward layer and 12 self-attention heads. Devlin et al. (2019) followed along the same lines as GPT and implemented Bidirectional Encoder Representations from Transformers, or BERT in short. BERT attends to the input tokens in both forward and backward directions while optimizing a *Masked Language Model* and a *Next Sentence Prediction* objective objectives.

BERT Benchmarks An integral part involved in developing various PLMs is providing NLU multi-task benchmarks used to demonstrate the linguistic abilities of new models and approaches. English BERT models are evaluated on 3 standard major benchmarks. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) is used to test paragraph level reading comprehension abilities. Wang et al. (2018) selected a diverse and relatively hard set of sentence and sentence-pair tasks which comprise the General Language Understanding Evaluation (GLUE) benchmark. The SWAG (Situations With Adversarial Generations) dataset (Zellers et al., 2018) presents models with partial description of grounded situations to see if they can consistently predict relevant scenarios that come next thus indicating the ability for commonsense reasoning. When evaluating Hebrew PLMs, one of the key pitfalls is that there are no Hebrew versions for these benchmarks. Furthermore, none of the suggested benchmarks account for examining the capacity of PLMs. In particular, currently there is no standard accepted way for evaluating the word-internal morphological structures which are inherent for MRLs and for the Hebrew language.

2.1 Multilingual vs Monolingual BERT

Devlin et al. (2019) produced 2 BERT models for English and Chinese. To support other languages they trained a multilingual BERT (mBERT) model combining texts covering over 100 languages. They hoped to benefit low resourced languages with the linguistic information obtained from other languages with large dataset sizes. In reality however mBERT performance on specific languages have not been as successful as English.

Consequently several research efforts focused on building monolingual BERT models as well as providing language specific evaluation benchmarks. Liu et al. (2019) trained CamemBERT, a French BERT model evaluated on syntactic and semantic tasks in addition to natural language inference tasks. Rybak et al. (2020) trained HerBERT, a BERT PLM for Polish. They evaluated it on a diverse set of existing NLU benchmarks as well as a new dataset for sentiment analysis for the e-commerce domain. Polignano et al. (2019) created Alberto, a BERT model for Italian, using a massive tweet collection. They tested it on NLU tasks - subjectivity, polarity (sentiment) and irony detection in tweets. In order to obtain a large enough training corpus in low-resources languages such as Finnish (Virtanen et al., 2019) and Persian (Farahani et al., 2020) a great deal of effort went into filtering and cleaning text samples obtained from web crawls.

Languages with rich morphology introduce another challenge involving identification and extraction of sub-word morphological information. Nguyen and Tuan Nguyen (2020) applied a specialized segmenter on the training data and normalized all the syllables and words before training their Vietnamese PheBERT model. In Arabic, like in Hebrew, words are composed of sub-word morphological units with each morpheme acting as a single syntactic unit (the way words are in English). Antoun et al. (2020) acknowledged this by pre-processing the training data using a morphological segmenter producing segments that were used instead of the actual words to train AraBERT. Doing so they were able to produce output vectors that correspond to morphological segments as opposed to the original words. On the other hand, this approach requires the application of the same segmenter at inference time as well.

Like any pipeline approach, this setup is susceptible to error propagation stemming from the fact that words can be morphologically ambiguous and the predicted segments in fact might not represent the correct interpretation of the words. As a result, the quality of the PLM depends on the accuracy achieved by the segmenting component. We, on the other hand, do not make any changes to the input, letting the PLM encode relevant morphological information associated with *complete* Hebrew words. Rather, we post-process the output by transforming contextualized word vectors into morphological-level segments to be used by the downstream tasks.

Across all of the above-mentioned language-specific PLMs, evaluation was performed on the token-,sentence- or paragraph-level. Non of these benchmarks examine the capacity of PLMs to encode sub-word morphological-level information which we focus on in this work.

Language	Oscar Size	Wikipedia Articles
English	2.3T	6,282,774
Russian	1.2T	1,713,164
Chinese	508G	1,188,715
French	282G	2,316,002
Arabic	82G	1,109,879
Hebrew	**20G**	**292,201**

Table 1: Corpora Size Comparison: High-resource (and Medium-resourced) languages vs. Hebrew.

Corpus	File Size	Sentences	Words
Oscar (deduped)	9.8GB	20.9M	1,043M
Twitter	6.9GB	71.5M	774M
Wikipedia	1.1GB	6.3M	127M
Total	**17.9GB**	**98.7M**	**1.9B**

Table 2: Data Statistics for AlephBERT's training sets.

3 AlephBERT Pre-Training

Data The PLM termed here *AlephBERT* is trained on a larger dataset and a larger vocabulary than any Hebrew BERT instantiation before. The Hebrew portions of Oscar and Wikipedia provides us with a training set size order of magnitude smaller compared with resource-savvy languages, as shown in Table 1. In order to build a strong PLM we need a considerable boost in the amount of sentences the PLM can learn from, which in our case comes form massive amounts of tweets added to the training set. We acknowledge the potential inherent concerns associated with this data source (population bias, behavior patterns, bot masquerading as humans etc.) and note that we have not made any explicit attempt to identify these cases.

Honoring ethical and legal constraints we have not manually analyzed nor published this data source. While the free form language expressed in tweets might differ significantly from the text found in Oscar and Wikipedia, the sheer volume of tweets helps us close the resource gap substantially with minimal effort. Data statistics are provided in Table 2.

Specifically, we employ the following datasets for pre-training:

- **Oscar:** A deduplicated Hebrew portion of the OSCARcorpus which is "extracted from Common Crawl via language classification, filtering and cleaning" (Ortiz Suárez et al., 2020).
- **Twitter:** Texts of Hebrew tweets collected between 2014-09-28 and 2018-03-07. We manually cleaned up the texts by removing markers (such "*RT:*", user mentions (e.g. "*@username*"), and URLs), and eliminating duplicates.
- **Wikipedia:** The texts in all of Hebrew Wikipedia, extracted using Attardi (2015)[2]

Configuration We used the Transformers training framework of Huggingface (Wolf et al., 2020) and trained two different models — a small model with 6 hidden layers learned from the Oscar portion of our dataset, and a base model with 12 hidden layers which was trained on the entire dataset. The processing units used are wordpieces generated by training BERT tokenizers over the respective datasets with a vocabulary size of 52K in both cases. Following the work on RoBERTa (Liu et al., 2019) we optimize AlephBERT with a masked-token prediction loss. We deploy the default masking configuration - 15% of word piece tokens are masked, In 80% of the cases, they are replaced by [MASK], in 10% of the cases, they are replaced by a random token and in the remaining cases, the masked tokens are left as is.

Operation To optimize GPU utilization and decrease training time we split the dataset into 4 chunks based on the number of tokens in a sentence and consequently we are able to increase batch sizes, resulting in dramatically shorter training times.

We trained for 5 epochs with learning rate set to 1e-4 followed by an additional 5 epochs with learning rate set to 5e-5 for a total of 10 epochs. We trained AlephBERT$_{\text{base}}$ over the entire dataset on an NVidia DGX server with 8 V100 GPUs which took us 8 days. AlephBERT$_{\text{small}}$ was trained over the Oscar portion only using 4 GTX 2080ti GPUs taking 5 days in total.

	chunk1	chunk2	chunk3	chunk4
max tokens	0>32	32>64	64>128	128>512
num sentences	70M	20M	5M	2M

4 Experimental Setup

Our two AlephBERT variants allow us to empirically gauge the effect of model size and data size on the quality of the language model. In addition, we compared the performance of all Hebrew BERT instantiations on various Hebrew NLP tasks using the following benchmarks:

- **Word Segmentation, Part-of-Speech Tagging, Full Morphological Tagging, Dependency Parsing:**
  - The Hebrew Section of the SPMRL Task (Seddah et al., 2013)
  - The Hebrew Section of the UD[3] treebanks collection (Sadde et al., 2018)
- **Named Entity Recognition:**
  - Token-based NER evaluation based on the Ben-Mordecai (henceforth BMC) corpus (Ben Mordecai and Elhadad, 2005)
  - Token-based and Morpheme-based NER evaluation based on the Named Entities and MOrphology (henceforth NEMO) corpus (Bareket and Tsarfaty, 2020)
- **Sentiment Analysis:**
  - Sentiment Analysis evaluation based on a fixed version of the Facebook (henceforth FB) corpus of Amram et al. (2018).

4.1 Sentence-Based Modeling

Sentiment Analysis We first report on a classification task, assigning a sentence with one of three values: negative, positive, neutral. By appending a classification head we turn a BERT model into a sentence level classifier (utilizing sentence level

[2]We make the corpus available on www.anonymous.com .

[3]https://universaldependencies.org

Raw input	לבית הלבן				
Space-delimited tokens	הלבן		לבית		
Index	5	4	3	2	1
Segmentation	לבן	ה	בית	ה	ל
POS	ADJ	DET	NOUN	DET	ADP
Morphology	Gender=Masc|Number=Sing	PronType=Art	Gender=Masc|Number=Sing	PronType=Art	-
Dependencies	3/amod	5/det	1/obj	3/def	0/ROOT
Token-level NER	E-ORG		B-ORG		
Morpheme-level NER	E-ORG	I-ORG	I-ORG	B-ORG	O

Table 3: Illustration of Evaluated Token and Morpheme-Based Downstream Tasks. The input is the two-word input phrase "לבית הלבן" (*to the White House*). Sequence and Hebrew text goes from right to left.

embedded vector representation associated with the special [CLS] BERT token).

We used a version of the Hebrew Sentiment dataset which we corrected by removing the leaked samples and re-partitioned to add a development set. This version has a total of 8,465 samples.[4]. We fine-tuned all models for 15 epochs with 5 different seeds and report the mean accuracy.

4.2 Token-Based Modeling

Named Entity Recognition Here we assume a token-based sequence labeling model. The input comprises of the sequence of tokens in the sentence, and the output contains BIOSE tags indicating entity spans. By appending a token-classification head we predict NER class labels for each word vector provided by the PLM (in cases of multiple word pieces we use the first one).

We evaluate this model on two corpora. We first evaluate on the BMC corpus which provides token-level annotations. It contains 3294 sentences and 4600 entities, and has seven different entity categories (DATE, LOC, MONEY, ORG, PER, PERCENT, TIME). To remain compatible with the original work we train and test the models on the 3 different splits as in Bareket and Tsarfaty (2020).[5] We then move to evaluate on the NEMO corpus which is an extension of the SPMRL dataset with Named Entities, marked by BIOSE tags. This corpus provides both token and morpheme based entity annotations, where the latter contains the accurate (token-internal) entity boundaries. The NEMO corpus has nine categories (ANG, DUC, EVE, FAC, GPE, LOC, ORG, PER, WOA). It contains 6220 sentences and 7713 entities, and we used the standard SPMRL train-dev-test. All sequence labeling models were trained for 15 epochs.

[4] www.anonymous.org

[5] www.anonymous.org

4.3 Morpheme-Based Modeling

Modern Hebrew is a Semitic language with rich morphology and complex orthography. As a result, the basic processing units in the language are typically smaller than a given token's span. To probe AlephBERT's capacity to accurately predict such token-internal linguistic structure, we test our models on five tasks that require knowledge of the internal morphology of the raw tokens. The input to all these tasks is a Hebrew sentence containing raw space-delimited tokens:

- **Segmentation**
  Output: A sequence of morphological segments representing basic processing units.[6]

- **Part-of-Speech Tagging**
  Output: Segmentation of the tokens to basic processing units as above, where each segment is tagged with its single disambiguated part-of-speech tag.

- **Morphological Tagging**
  Output: Segmentation of the tokens to basic processing units as above, where each segment is tagged with a single POS tag and a set of morphological features.[7]

- **Dependency Parsing**
  Output: Segmentation of the tokens to basic processing units as above, where each segment is tagged with a single POS tag and a set of morphological features and assigned with labeled dependency relations.

[6] These units comply with the 2-level representation of tokens defined by UD, where each basic unit corresponds to a single POS tag. https://universaldependencies.org/u/overview/tokenization.html

[7] Equivalent to the AllTags evaluation metric defined in the CoNLL18 shared task. https://universaldependencies.org/conll18/results-alltags.html

- **Morpheme-Based NER**
  Output: Segmentation of the tokens to basic processing as above, where each segment is tagged with a BIOSE tag indicating entity spans, along with the entity-type label.

An illustration of these tasks is given in Table 3.

In order to provide proper segmentation and labeling for the aforementioned tasks we developed a model designated to produce the morphological segments of each word in context. This morphological segmentation model consumes words and their associated contextualized embedded vectors (produced by a PLM), feeds them into a char-based seq2seq module and produces sub-token morphological segments as output. The seq2seq module is composed of an encoder implemented as a simple char-based BiLSTM, and a decoder implemented as a char-based LSTM generating the output character symbols, or a space symbol signalling the end of a morphological segment. We train the model for 15 epochs, optimized with next-character prediction loss. For tasks involving both segmentation and labeling (POS, Features, NER) we deploy an MTL (multi-task learning) setup. That is, when generating an end-of-segment symbol, the morphological model then predicts task labels which can be one or more of the following: POS-tag, NER-tag, morphological features. In order to guide the training we optimize the combined segmentation and label prediction loss values.

For the NER task, we design another setup in which we first segment the text, and feed the morphological segments into the PLM to produce contextualized embedded vectors for the segments. We are then able to perform fine-tuning with a token classification attention head directly applied to the PLM output (similar to the way we fine-tune the PLM for the token-based NER task described in the previous section). We acknowledge the fact that we are fine-tuning the PLM using morphological segments even though it was originally pre-trained without any morphological knowledge but, as we shall see shortly, this seemingly unintuitive strategy performs surprisingly well.

Finally, we set up a dependency parsing evaluation pipeline. For this purpose we choose the standalone Hebrew parser offered by More et al. (2019) (a.k.a YAP) which was trained and produces SPMRL dependency labels. The morphological information encoded in the PLMs is recovered for each word by our morphological extraction model and used as input features to the YAP standalone parser.

4.3.1 Morpheme Level Evaluation

Aligned Segment The CoNLL18 Shared Task evaluation campaign[8] reports scores for segmentation and POS tagging[9] for all participating languages. For multi-segment words, the gold and predicted segments are aligned by their Longest Common Sub-sequence, and only matching segments are counted as true positives. We use the script to compare aligned segment and tagging scores between oracle (gold) segmentation and realistic (predicted) segmentation.

Aligned Multi-Set In addition we compute F1 scores similar to the aforementioned with a slight but important difference as defined by More et al. (2019) and Seker and Tsarfaty (2020). For each word, counts are based on multi-set intersections of the gold and predicted labels ignoring the order of the segments while accounting for the number of each segment. *Aligned mset* is based on set difference which acknowledges the possible undercover of covert morphemes which is an appropriate measure of morphological accuracy.

Discussion To illustrate the difference between *aligned segment* and *aligned mset*, let us take for example the gold segmented tag sequence: *b/IN, h/DET, bit/NOUN* and the predicted segmented tag sequence *b/IN, bit/NOUN*. According to *aligned segment*, the first segment (*b/IN*) is aligned and counted as a true positive, the second segment however is considered as a false positive (*bit/NOUN*) and false negative (*h/DET*) while the third gold segment is also counted as a false negative (*bit/NOUN*). On the other hand with aligned mulit-set both *b/IN* and *bit/NOUN* exist in the gold and predicted sets and counted as true positives, while *h/DET* is mismatched and counted as a false negative. In both cased the total counts across words in the entire datasets are incremented accordingly and finally used for computing Precision, Recall and F1.

5 Results

Sentence-Based Tasks Sentiment analysis accuracy results are provided in Table 4. All BERT-based models substantially outperform the original

[8] https://universaldependencies.org/conll18/results.html

[9] respectively referred to as 'Segmented Words' and 'UPOS' in the CoNLL18 evaluation script

Task	NER (Token)		Sentiment
Corpus	NEMO	BMC	FB
Prev. SOTA	77.75	85.22	NA
mBERT	79.07	87.77	79.07
HeBERT	81.48	89.41	81.48
AlephBERT$_{small}$	78.69	89.07	78.69
AlephBERT$_{base}$	**84.91**	**91.12**	**84.91**

Table 4: Token-based NER F1. Previous SOTA on both corpora reported by the NEMO models of Bareket and Tsarfaty (2020). Sentiment Analysis accuracy on the corrected version of the Facebook corpus.

Task	Segment	POS	Features	UAS	LAS
Prev. SOTA	NA	90.49	85.98	75.73	69.41
mBERT	97.36	93.37	89.36	80.17	74.9
HeBERT	97.97	94.61	90.93	81.86	76.54
AlephBERT$_{small}$	97.71	94.11	90.56	81.5	76.07
AlephBERT$_{base}$	**98.10**	**94.90**	**91.41**	**82.07**	**76.9**

Table 5: Morpheme-Based results on the SPMRL corpus. Aligned MultiSet (mset) F1 for Segmentation, POS tags and Morphological Features - previous SOTA reported by (Seker and Tsarfaty, 2020) (POS) and (More et al., 2019) (features). Labeled and Unlabeled Accuracy Scores for morphological-level Dependency Parsing - previous SOTA reported by (More et al., 2019) (uninfused/realistic scenario)

CNN Baseline reported by Amram et al. (2018). AlephBERT$_{base}$ is setting new SOTA.

Token-Based Tasks On our two NER benchmarks, we report F1 scores on the token-based fine-tuned model in Table 4. Although we see noticeable improvements for the mBERT and HeBert variants over the current SOTA, the most significant increase is achieved by AlephBERT$_{base}$.

Morpheme-Based Tasks As a particular novelty of this work, we report BERT-based results on sub-token (segment-level) information. Specifically, we evaluate segmentation, POS, Morphological Features, NER and dependencies compared against morphologically-labeled test sets. In all cases we use raw space-delimited tokens as input and produce morphological segments with our new morphological extraction model which uses BERT-based output as features.

Table 5 presents evaluation results for the SPRML dataset as done in previous work on Hebrew (More et al., 2019). We report aligned multiset F1 scores for 3 tasks: segmentation, POS tagging, and morphological features extraction. In addition we report labeled and unlabeled accuracy

Task	Segment	POS	Features
Prev. SOTA	NA	94.02	NA
mBERT	97.70	94.76	90.98
HeBERT	98.05	96.07	92.53
AlephBERT$_{small}$	97.86	95.58	92.06
AlephBERT$_{base}$	**98.20**	**96.20**	**93.05**

Table 6: Morpheme-Based Aligned MultiSet (mset) F1 results on the UD corpus. Previous SOTA reported by (Seker and Tsarfaty, 2020) (POS)

Task	Segment	POS	Features
Prev. SOTA	96.03	93.75	91.24
mBERT	97.17	94.27	90.51
HeBERT	97.54	95.60	92.15
AlephBERT$_{small}$	97.31	95.13	91.65
AlephBERT$_{base}$	**97.70**	**95.84**	**92.71**

Table 7: Morpheme-Based Aligned (CoNLL shared task) F1 on the UD corpus. Previous SOTA reported by Minh Van Nguyen and Nguyen (2021)

scores of the dependency trees produced by our dependency parsing pipeline setup. We see that segmentation results for all BERT-based models are similar, in the high range of 97-98 F1 scores, which are hard to improve further.[10] For POS tagging and morphological features, all BERT-based models considerably outperform previous SOTA.

The most impressive improvement is observed in dependency parsing attachment scores where we observe a large gain compared to the previous SOTA joint morpho-syntactic framework. It confirms the impact that morphological errors early in the pipeline have on downstream tasks, and highlight the importance of morphologically-driven benchmark as part of any PLM evaluation.

In all tasks of the SPMRL dataset, we notice a repeating trend placing AlephBERT$_{base}$ as the best model for all morphological tasks, indicating that the improvement provided by the depth of the model and a larger dataset does improve the ability to capture token-internal structure.

These trends are replicated on the UD Hebrew corpus, for two different evaluation metrics — the Aligned MultiSet F1 Scores as in previous work on Hebrew (More et al., 2019), (Seker and Tsarfaty, 2020), and the Aligned Segment F1 scores metrics as described in the UD shared task (Zeman et al., 2018) — reported in Tables 6 and 7 respectively.

[10] Some of these errors are due to annotation errors, or truly ambiguous cases.

Architecture	Pipeline (Oracle)		Pipeline (Predicted)		MultiTask	
Segmentation Task	Seg	NER	Seg	NER	Seg	NER
Prev. SOTA	100.00	79.10	95.15	69.52	97.05	77.11
mBERT	100.00	77.92	97.68	72.72	97.24	72.97
HeBERT	100.00	82	98.15	76.74	97.92	74.86
AlephBERT$_{\text{small}}$	100.00	79.44	97.78	73.08	97.74	72.46
AlephBERT$_{\text{base}}$	100.00	83.94	**98.29**	**80.15**	98.19	79.15

Table 8: Morpheme-Based NER F1 on the NEMO corpus. Previous SOTA reported by Bareket and Tsarfaty (2020) for the Pipeline (Oracle), Pipeline (Predicted) and a Hybrid (almost-joint) scenarios, respectively.

Morpheme-Based NER Earlier in this section we considered NER as a token-based task that simply requires fine-tuning on the token level. However, this setup is not accurate enough and less useful for downstream tasks, since the exact entity boundaries are often token internal (Bareket and Tsarfaty, 2020). We hence report morpheme-based NER evaluation, respecting exact boundaries of entity mentions. To obtain morpheme-based labeled-span of Named Entities we could either employ a pipeline, first predicting segmentation and then applying a fine tuned labeling model *directly on the segments*, or employ multi-task model and predict NER labels *while* performing segmentation.

Table 8 presents segmentation and NER results for three different scenarios: (i) pipeline assuming gold segmentation (ii) pipeline assuming predicted segmentation (iii) segmentation and NER labels obtained jointly in multi-task setup. AlephBERT$_{\text{base}}$ consistently scores highest in all 3 setups.

Looking at the Pipeline-Predicted scores, there is a clear correlation between a higher segmentation quality of a PLM and its ability to produce better NER results. Moreover, the differences in NER scores are considerable (unlike the subtle differences in segmentation, POS and morphological features scores) and draw our attention to the relationship between the size of the PLM, the size of the pre-training data and the quality of the final NER models. Specifically, HeBERT and AlephBERT$_{\text{small}}$ were both pre-trained on similar datasets and comparable vocabulary sizes (heBERT with 30K and AlephBERT-small with 52K) but HeBERT, with its 12 hidden layers, performs better compared to AlephBERT$_{\text{small}}$ which is composed of only 6 hidden layers. It thus appears that semantic information is learned in those deeper layers helping in both discriminating entities and improving the overall morphological segmentation capacity. In addition, comparing HeBERT to AlephBERT$_{\text{base}}$ we point to the fact that they are both modeled with the same 12 hidden layer architecture, the only differences between them are in the size of their vocabularies (30K vs 52K respectively) and the size of the training data (Oscar-Wikipedia vs Oscar-Wikipedia-Tweets). The improvements exhibited by AlephBERT$_{\text{base}}$, compared to HeBERT, suggests that it is a result of the large amounts of training data and larger vocabulary. By exposing AlephBERT$_{\text{base}}$ to a substantially larger amount of text we increased the ability of the PLM to encode syntactic and semantic signals associated with Named Entities. Finally, our NER experiments suggest that a pipeline composed of our near-perfect morphological segmentation model followed by AlephBERT$_{\text{base}}$ augmented with a token classification head is the best strategy for generating morphologically-aware NER labels.

6 Conclusion

Modern Hebrew, a morphologically-rich and medium-resource language, has for long suffered from a gap in the resources available for NLP applications, and lower level of empirical results than observed in other, resource-rich languages. This work provides the first step in remedying the situation, by making available a large Hebrew PLM, nicknamed AlephBERT, with larger vocabulary and larger training set than any Hebrew PLM before, and with clear evidence as to its empirical advantages. Crucially, we propose a language-agnostic pipeline with a morphological disambiguation component that does not require any particular (possibly noisy) pre-processing. This opens the door for developing an entire suite of morphological benchmarks for testing PLMs for MRLs. AlephBERT$_{\text{base}}$ obtains state-of-the-art results on the tasks of morphological segmentation, part-of-speech tagging, morphological feature extraction, dependency parsing, named-entity recognition, and sentiment analysis outperforming both multilingual (mBERT) and language-specific (HeBERT) PLMs. Our proposed morphologically-driven test benchmarks serve as a solid foundation for future development and evaluation of Hebrew and MRLs in general.

References

Adam Amram, Anat Ben-David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from modern hebrew. In *Proceedings of the 27th International Conference on*

Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 2242–2252.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Dan Bareket and Reut Tsarfaty. 2020. Neural modeling for named entities and morphology (nemo^2). *CoRR*, abs/2007.15620.

Naama Ben Mordecai and Michael Elhadad. 2005. Hebrew named entity recognition.

Avihay Chriqui and Inbal Yahav. 2021. Hebert \& hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, SIGMORPHON 2020, Online, July 10, 2020*, pages 204–209.

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

Amir Pouran Ben Veyseh Minh Van Nguyen, Viet Lai and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Amir More, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint transition-based models for morpho-syntactic parsing: Parsing strategies for mrls and a case study from modern hebrew. *Trans. Assoc. Comput. Linguistics*, 7:33–48.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets.

Alec Radford and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *arxiv*.

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

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Shoval Sadde, Amit Seker, and Reut Tsarfaty. 2018. The hebrew universal dependency treebank: Past present and future. In *Proceedings of the Second Workshop on Universal Dependencies, UDW@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 133–143.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Wolinski, Alina Wróblewska, and Éric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, SPMRL@EMNLP 2013, Seattle, Washington, USA, October 18, 2013*, pages 146–182.

Amit Seker and Reut Tsarfaty. 2020. A pointer network architecture for joint morphological segmentation and tagging. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4368–4378, Online. Association for Computational Linguistics.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: what did we learn (and unlearn) in a decade of parsing morphologically-rich languages (mrls)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7396–7408.

Reut Tsarfaty, Shoval Sadde, Stav Klein, and Amit Seker. 2019. What's wrong with hebrew nlp? and how to make it right. 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 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 259–264.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.