AlephBERT: Pre-training and End-to-End Language Model 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 in Hebrew are few and far between. The problem is twofold. First, Hebrew resources for training large language models are not at the same order of magnitude as their English counterparts. Second, there are no accepted tasks and benchmarks to evaluate the progress of Hebrew PLMs on, and in particular, evaluation on 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 extracts all of the sub-word morphological segments encoded in contextualized word embedding vectors. Utilizing this new morphological component we offer a new PLM evaluation pipeline of multiple Hebrew tasks and benchmarks, that cover *word-level*, *sub-word level* and *sentence level* tasks. With *AlephBERT* we achieve state-of-the-art results compared against contemporary baselines. We make our *AlephBERT* model and evaluation pipeline publicly available, providing a single point of entry for evaluating and comparing Hebrew PLMs.

1 Introduction

This paper presents a case study for PLM development for a *morphologically-rich* and *resource-poor* language. Specifically, we address Modern Hebrew, a Semitic, morphologically-rich language, that is long known to be notoriously hard to process. The challenges posed to automatically processing Hebrew texts 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 impressive. 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 released, yet there is no reported evidence on performance improvements on key components of the Hebrew NLP pipeline

The scarcity in Hebrew resources is problematic for PLM development in at least two ways. First, there are insufficient amounts of free unlabeled text for pre-training. To wit, the Hebrew Wikipedia that was the source for training multilingual BERT is of orders of magnitude smaller than the English Wikipedia (See Table 1).[1] Secondly, there are no large-scale open-access commonly accepted benchmarks for evaluating the performance of Hebrew PLMs on NL processing and understanding tasks.

We present *AlephBERT*, a Hebrew pre-trained language model, larger and more effective than any Hebrew PLM before. More importantly, we show a novel pipeline covering a range of essential tasks in Hebrew NLP, tailored to fit a *morphologically-rich language*, i.e., test for sentence-level, word-level and sub-word level accuracy, including: **Segmentation, Part-of-Speech Tagging, full morphological tagging, Named Entity Recognition** and **Sentiment Analysis**. Since previous Hebrew NLP studies used varied corpora and annotation schemes, we confirm our results on *all* existing Hebrew benchmarks and scheme variants.[2]

Crucially, evaluating BERT-based models on morpheme-level tasks is non trivial. BERT outputs word-level embedded vectors, however sub-word morpheme-level vectors which are required for morpheme based tasks are not readily available. To address this we introduce a neural component that operates on contextualized word vectors and extracts morphological segments and vector representations for these segmented form.

We make our model and online demo publicly available[3] allowing to qualitatively compare the prediction capacity of different PLMs available for Hebrew. Furthermore, we release the complete pipeline we developed, as means for evaluating and comparing PLMs for MRLs, and as a starting point for developing further downstream applications and tasks that require access to sub-word morphological information. In the future we plan to showcase *AlephBERT*'s capacities on downstream tasks as: Information Extraction, Summarization, Reading Comprehension, and many more.

[1] Of course, ample Hebrew data does exist online, but most of it is closed due to copy-right issues and paywalls.

[2] For morphology and POS tagging, we test on both the Hebrew section of SPMRL (Seddah et al., 2013), and the Hebrew UD corpus (Sadde et al., 2018). For Named Entity recognition, we test on the corpora of Ben Mordecai and Elhadad (2005) and Bareket and Tsarfaty (2020). For sentiment analysis we test on the Facebook corpus Amram et al. (2018)

[3] www.anonymous.org

2 Previous Work

Using contextualized word embedding vectors improved the 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 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, don't 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 which is an 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 text that the PLM can learn from, which in our case comes form massive amounts of tweets added to the training set. 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 OSCAR corpus, 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 slightly cleaned up the texts by removing retweet signals "*RT:*", user mentions (e.g. "@*username*"), and URLs.
- **Wikipedia:** The texts in all of Hebrew Wikipedia, extracted using Attardi (2015)[4]

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 significantly increase the batch sizes, resulting in dramatically shorter training times.

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

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 $\text{AlephBERT}_{\text{base}}$ over the entire dataset on an NVidia DGX server with 8 V100 GPUs which took us 8 days. $\text{AlephBERT}_{\text{small}}$ was trained over the Oscar portion only using 4 GTX 2080ti GPUs taking 5 days in total.

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

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:**
  - The Hebrew Section of the SPMRL Task (Seddah et al., 2013)
  - The Hebrew Section of the UD[5] treebanks collection (Sadde et al., 2018)
- **Named Entity Recognition:**
  - Token-based NER evaluation based on the corpus of Ben-Mordecai and Elhadad (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 the corpus of Amram et al. (2018).
  - Since the aforementioned corpus is reported to be leaking (shared material between test and train), we provide a cleaned up version and evaluate on the updated split.

4.1 Sentence-Based Modeling

Sentiment Analysis The first task we report on is a simple 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 embedded vector representation associated with the special [CLS] BERT token). We trained and evaluated BERT-based sentence classification on two variants of the Hebrew Sentiment dataset of Amram et al. (2018) – first with an additional split to create a dev set (which is missing in the original dataset), and second with

[5]https://universaldependencies.org

Raw input	לבית הלבן				
Space-delimited tokens	הלבן		לבית		
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	-
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.

a corrected dataset without duplicate samples that leaked across splits. After removing the duplicates out of the original 12,804 sentences, the dataset is left with 8,465 samples.[6] We fine-tuned all the models for 15 epochs on 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 BIOES tags indicating entity spans. This token-based model is designed by simply appending a token-classification head predicting a NER class label for each contextualized word embedded vector provided by the PLM (in cases of multiple word pieces, the first one is used as the word representation).

We evaluate this model on two corpora. The first is the corpus by Ben Mordecai and Elhadad (2005), henceforth, the BMC corpus. The BMC corpus annotates entities at Token-level. 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).[7] For the BMC corpus we report token-based F1 scores on the detected entity mentions.

The second corpus is an extension of the SPMRL dataset with Named Entities annotation, also marked by BIOSE tags, respecting the precise (token-internal) morphological boundaries of NEs (henceforth, NEMO, standing for Named Entities and MOrphology) (Bareket and Tsarfaty, 2020). This corpus provides both a token-based and a morpheme-based annotation of the entities, 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 over the annotated datasets. For both benchmarks we report token-based F1 scores on the detected entity mentions.

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 four tasks that require knowledge of the internal morphology of the raw tokens:

- **Segmentation**
  Input: A Hebrew sentence containing raw space-delimited tokens
  Output: A sequence of morphological segments representing basic processing units.[8]

- **Part-of-Speech Tagging**
  Input: A Hebrew sentence containing raw space-delimited tokens
  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**
  Input: A Hebrew sentence containing raw space-delimited tokens
  Output: Segmentation of the tokens to basic processing units as above, where each segment is tagged with a single POS tag and a set

[6] www.anonymous.org
[7] www.anonymous.org
[8] 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

of morphological features.[9]

- **Morpheme-Based NER**
  Input: A Hebrew sentence containing raw space-delimited tokens
  Output: Segmentation of the tokens to basic processing as above where segment is tagged with a BIOSE tags 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 four aforementioned tasks we developed a model designated to produce the morphological segments of each token in context. The morphological segmentation model consumes words and their associated contextualized embedded vectors (produced by a PLM), fed into a char-based seq2seq module that extracts the output segments. 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 other tasks, involving both segmentation and labeling 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.

We design another setup for morphological NER in which we first segment the text (using the above-mentioned segmentation model), 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.

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

4.3.1 Morpheme Level Evaluation

Aligned Segment The CoNLL18 Shared Task evaluation campaign[10] reports scores for segmentation and POS tagging[11] 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 results are provided in Table 4. All BERT-based models substantially outperform the original CNN Baseline reported by Amram et al. (2018). $\text{BERT}_{\text{base}}$ is setting new SOTA results on the new (corrected) dataset.

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

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

	Old_{token}	Old_{morph}	New_{token}	New_{morph}
Prev. SOTA	89.2	87.5	NA	NA
mBERT	92.12	92.18	84.21	85.58
HeBERT	92.48	92.27	87.13	86.88
$AlephBERT_{small}$	**93.15**	**92.70**	88.3	87.38
$AlephBERT_{base}$	91.63	92.01	**89.02**	**88.71**

Table 4: Sentiment Analysis Accuracy Scores on the Old (leaked) and New (corrected) Facebook Corpus variants. Previous SOTA is reported by Amram et al. (2018)

	NEMO	BMC
Prev. SOTA	77.75	85.22
mBERT	79.07	87.77
HeBERT	81.48	89.41
$AlephBERT_{small}$	78.69	89.07
$AlephBERT_{base}$	**84.91**	**91.12**

Table 5: Token-Based NER F1 Results on the NEMO and the Ben-Mordecai Corpora. Previous SOTA on both corpora has been reported by the NEMO models of Bareket and Tsarfaty (2020).

Token-Based Tasks On our two NER benchmarks, we report F1 scores on the token-based fine-tuned model in Table 5.

Although we see noticeable improvements for the mBERT and HeBert variants over the current SOTA, the most significant increase is achieved by $AlephBERT_{base}$. $AlephBERT_{base}$ provides a new SOTA results on both datasets. Crucially, this holds for the *token-based* evaluation metrics (as defined in Bareket and Tsarfaty (2020)).

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 F1, POS F1, Morphological Features F1 and morphem-base NER F1, compared against the labeled morphological segments. In all cases we use raw space-delimited tokens as input, letting the BERT-based models perform *both* the segmentation and labeling.

Table 6 presents the segmentation, POS tags, and morphological features F1 for the SPMRL dataset, all evaluated at the granularity of morphological segments. We report the aligned multiset F1 Scores as in previous work on Hebrew (More et al., 2019).

We see that segmentation results for all BERT-based models are similar, and they are already at the higher range of 97-98 F1 scores, which are hard to improve further.[12] For POS tagging and morphological features, all BERT-based models

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

	Segment	POS	Features
Prev. SOTA	NA	90.49	85.98
mBERT-morph	97.36	93.37	89.36
HeBERT-morph	97.97	94.61	90.93
$AlephBERT_{small}$-morph	97.71	94.11	90.56
$AlephBERT_{base}$-morph	**98.10**	**94.90**	**91.41**

Table 6: Morpheme-Based Aligned MultiSet (mset) F1 Results on the SPMRL Corpus. Previous SOTA is as reported by (Seker and Tsarfaty, 2020) (POS) and (More et al., 2019) (morphological features)

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

Table 7: Morpheme-Based Aligned MultiSet (mset) F1 Results on the UD Corpus. Previous SOTA is as reprted by (Seker and Tsarfaty, 2020) (POS)

significantly outperform the previous SOTA provided by (Seker and Tsarfaty, 2020) (referred to as PtrNet) for POS tags and (More et al., 2019) (referred to as YAP) for morphological features. With respect to all BERT-based variants, we see an improvement for AlephBERT on all other alternatives, but on a smaller scale. That said, we do notice a repeating trend that places $AlephBERT_{base}$ as the best model for all of our morphological tasks, indicating that the improvement provided by the depth of the model and a larger dataset does also 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 7 and 8 respectively. $AlephBERT_{base}$ obtains the best results for all tasks, even if not by a large margin.

Morpheme-Based NER Earlier in this section we considered NER as a token-based task that simply requires fine-tuning on the token labels. 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 also report here morpheme-based NER evaluation, respecting the exact boundaries of the Entity mentions. To obtain morpheme-based labeled-span of Named En-

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

Table 8: Morpheme-Based Aligned (CoNLL shared task) F1 Results on the UD Corpus. Previous SOTA is as reported by Minh Van Nguyen and Nguyen (2021)

Architecture	Pipeline (Oracle)		Pipeline (Predicted)		MultiTask	
Segmentation Scores (aligned mset F1)	Seg	NER	Seg	NER	Seg	NER
Prev. SOTA (NEMO)	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$_{small}$	100.00	79.44	97.78	73.08	97.74	72.46
AlephBERT$_{base}$	100.00	83.94	**98.29**	**80.15**	98.19	79.15

Table 9: Morpheme-Based NER F1 Evaluation on the NEMO Corpus. Previous SOTA is as reported by Bareket and Tsarfaty (2020) for the Pipeline (Oracle), Pipeline (Predicted) and a Hybrid (almost-joint) Scenarios, respectively.

tities as discussed above we could either employ a pipeline, first predicting segmentation and then applying a fine tuned labeling model *directly on the segments*, or we can use the MTL model and predict NER labels *while* performing the segmentation.

Table 9 presents segmentation and NER results for three different scenarios: (i) a pipeline assuming gold segmentation (ii) a pipeline assuming the best predicted segmentation (as predicted above) (iii) obtaining the segmentation and NER labels jointly in the MTL setup.

AlephBERT$_{base}$ consistently scores highest in both pipeline (oracle and predicted) and multi-task 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 between the models 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$_{small}$ were pre-trained on similar datasets and comparable vocabulary sizes (heBERT with 30K and AlephBERT-small with 52K). However we notice that HeBERT, with its 12 hidden layers, performs significantly better compared to AlephBERT$_{small}$ which is composed of only 6 hidden layers. It thus appears that semantic information is learned in those deeper layers which helps in both learning to discriminate entities and improve the overall morphological segmentation capacity.

In addition, comparing HeBERT to AlephBERT$_{base}$ 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$_{base}$, compared to HeBERT, suggests that it is a result of the large amounts of training data and larger vocabulary available in our setup. By exposing AlephBERT$_{base}$ to an amount of text which is order of magnitude larger we increased the ability of the PLM to encode the 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$_{base}$ augmented with a token classification head is the best strategy for generating morphologically-aware NER labels.

6 Conclusion

Modern Hebrew, a morphologically-rich and resource-scarce 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 acts on *complete* tokens and 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. Our AlephBERT$_{base}$ model obtains state-of-the-art results on the tasks of morphological segmentation, Part-of-Speech Tagging, Named Entity Recognition, and Sentiment Analysis. outperforming both multilingual (mBERT) and language-specific (HeBERT) PLMs. AlephBERT and the proposed pipeline serve as a solid foundation for future development and evaluation of Hebrew PLMs.

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.

Giusepppe 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.