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

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky,
Refael Shaked Greenfeld, Reut Tsarfaty
Department of Computer Science, Bar Ilan University, Ramat-Gan, Israel
{aseker00,elronbandel,dbareket,bruslil,
shakedgreenfeld,reut.tsarfaty}@gmail.com

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, most benchmarks available to evaluate progress in Hebrew NLP require morphological boundaries which are not available in the output of PLMs. In this work we 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 neural architecture that recovers the morphological segments encoded in contextualized embedding vectors. Based on this new morphological component we offer an 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 state-of-the-art models. We make our *AlephBERT* model, the morphological extraction component, and the Hebrew evaluation suite publicly available, for future investigations and evaluations of Hebrew PLMs.

1 Introduction

Contextualized word representations provided by models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT3 (Brown et al., 2020), T5 (Raffel et al., 2020) and more, were shown in recent years to be a critical component for obtaining state-of-the-art performance on a wide range of Natural Language Processing (NLP) tasks, from surface syntactic tasks as tagging and parsing, to downstream semantic tasks as question answering, information extraction and text summarization.

While advances reported for English using such models are unprecedented, previously reported results using PLMs in Modern Hebrew 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 results of the mBERT model are on a par with pre-neural models or neural models based on non-contextual embeddings (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 challenge of developing PLMs for *morphologically-rich* and *medium-resourced* languages such as Modern Hebrew is twofold. First, contextualized word representations are obtained by pre-training a large language model on massive quantities of unlabeled texts. In Hebrew, the size of published texts *available* for training is relatively small. To wit, Hebrew Wikipedia (300K articles) used for training mBERT is orders of magnitude smaller compared to English Wikipedia (6M articles). Second, commonly accepted benchmarks for evaluating Hebrew models, via Morpho-Syntactic Tagging and Parsing (Sadde et al., 2018), or Named Entity Recognition (Bareket and Tsarfaty, 2020) require decomposition of words into *morphemes*,[1] which are distinct of the sub-words (a.k.a. word-pieces) provided by standard PLMs. Such *morphemes* are as of yet not readily available in the PLMs' output embeddings.

Evaluating BERT-based models on morpheme-level tasks is thus non-trivial due to the mismatch between the sub-word tokens used as sub-word

[1] These morphemes are affixes and clitics bearing their own POS. They are termed *syntactic words* in UD (Zeman et al., 2018), or *segments* in previous literature on Hebrew NLP.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 46 - 56

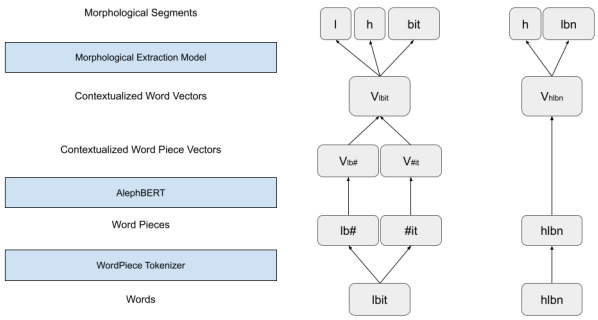

Figure 1: PLM Morphological Extraction Pipeline. The two-word phrase "לבית הלבן", transliterated as "lbit hlbn", mapped to word-pieces which are consumed by a PLM to generate contextualized vectors and extract the sub-word morphological units. In this example the WordPiece Tokenizer splits the first word, "lbit", into two pieces while leaving the second word, "hlbn", intact. Consequently, AlephBERT generates 3 embedded vectors - the vectors associated with the split word pieces are averaged to form a single contextualized vector. Finally, the resulting two word vectors are used by the Morphological Extraction Model that generates the disambiguated morphological segments.

input units used by the PLMs and the sub-word morphological units needed for evaluation. PLMs employ sub-word tokenization mechanisms such as WordPiece or Byte-Pair Encoding (BPE) for the purposes of minimizing Out-Of-Vocabulary words (Sennrich et al., 2016). These sub-word tokens are generated in a pre-processing step, without utilization of any linguistic information, and passed as input to the PLM. Crucially, such word-pieces *do not reflect morphological units*. Extracting morphological units from contextualized vectors provided by PLMs is challenging yet necessary in order to enable morphological-level evaluation of Hebrew PLMs on standard benchmarks.

In this paper we introduce *AlephBERT*, a Hebrew PLM trained on more data and a larger vocabulary than any Hebrew PLM before.[2] Moreover, we propose a novel architecture that extracts the *morphological* sub-word units *implicitly* encoded in the contextualized vectors outputted by PLMs. Using AlephBERT and the proposed morphological extraction model we enable evaluation on *all* existing Hebrew benchmarks. We thus present a processing and evaluation pipeline tailored to fit Morphologically Rich Languages (MRLs), i.e., covering sentence-level, word-level and most importantly sub-word morphological-level tasks (*Segmentation, Part-of-Speech Tagging, full Morphological Tagging, Dependency Parsing, Named Entity Recognition (NER)* and *Sentiment Analysis*), and present new and improved SOTA for Modern Hebrew on all of these tasks.

[2]We make our PLM https://huggingface.co/onlplab/alephbert-base and demo https://nlp.biu.ac.il/~amitse/alephbert/ publicly available, to qualitatively assess present and future Hebrew PLMs.

2 Previous Work

Contextualized word embedding vectors are a major driver for improved performance of deep learning models on many Natural Language Understanding (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 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 composed of a 768-dimensional feed-forward layer and 12 self-attention heads. Devlin et al. (2019) followed along the same lines 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 for testing 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 subsequent scenarios, thus indicating abilities of 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 for encoding the word-internal morphological structures which are inherent in MRLs. In this work we enable a generic morphological-level evaluation pipeline that is suited for PLMs of MRLs.

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, in the hoped to benefit low-resource languages with the linguistic information obtained from languages with larger datasets. In reality, however, mBERT performance on specific languages has 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 several 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.

BERT for MRLs Languages with rich morphology introduce another challenge involving the identification and extraction of sub-word morphological information. In many MRLs words are composed of sub-word morphological units, with each unit acting as a single syntactic unit bearing as single POS tag (mimicking 'words' in English). Antoun et al. (2020) addressed this for Arabic, a Semitic MRLs, by pre-processing the training data using a morphological segmenter, producing morphological segments to be used for training AraBERT instead of the actual words. By doing so, they were able to produce output vectors that correspond to morphological segments rather than the original space-delimited word-tokens. However, this approach requires the application of the same segmenter at inference time as well, and like any pipeline approach, this setup is susceptible to error propagation. This risk is magnified as words in MRLs may be morphologically ambiguous, and the predicted segments 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. A particular novelty of this work is *not* making any changes to the input, letting the PLM encode morphological information associated with *complete* Hebrew tokens. Instead, transforming the resulting contextualized word vectors into morphological-level segments via a novel neural architecture which we discuss shortly.

Language	Oscar (duped) 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: Resource-savvy languages vs. Hebrew.

Evaluating PLMs for MRLs Across all of the above-mentioned language-specific PLMs, evaluation was performed on the word-,sentence- or paragraph-level. Non examined the capacity of PLMs to encode sub-word morphological-level information which we focus on in this work. Şahin et al. (2019) probed various information types encoded in embedded word vectors. Similarly to us, they focused on languages with rich morphology where linguistic signals are encoded at the morphological, subword level. Their work is more about explainability — showing high positive correlation of probing tasks to the downstream tasks, especially for morphologically rich languages. Unlike us, they assume a single POS tag and set of features per word in their probing tasks. In Hebrew, Arabic and other MRLs, tokens may carry multiple POS per word, and are required to be segmented for further processing. We provide a framework that extracts subword morphological units given contextualized word vectors, that enables to evaluate PLMs on morphologically-aware datasets where words can have multiple POS tags and feature-bundles.

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: AlephBERT's Training Data.

3 AlephBERT Pre-Training

Data The PLM termed *AlephBERT* that we provide herein is trained on a larger dataset and a larger vocabulary than any Hebrew BERT instantiation before. The data we train on is listed in Table 2. Concretely, we employ the following datasets for pre-training: **(i) Oscar:** Deduplicated Hebrew portion extracted from Common Crawl via language classification, filtering and cleaning (Ortiz Suárez et al., 2020). **(ii) Wikipedia:** Texts from all of Hebrew Wikipedia, extracted using Attardi (2015). **(iii) Twitter:** Hebrew tweets collected between 2014-09-28 and 2018-03-07. We removed markers ("*RT:*", "@" user mentions and URLs), and eliminated duplicates. For data statistics, see Table 2.

The Hebrew portions of **Oscar** and **Wikipedia** provide us with a training-set size orders-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.[3]

Model 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 where 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 and dramatically shorten training time.

	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 1e-4 followed by an additional 5 epochs with learning rate at 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 8 days. $\text{AlephBERT}_{\text{small}}$ was trained over the Oscar portion only, using 4 GTX 2080ti GPUs taking 5 days in total.

4 The Morphological Extraction Model

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 raw space-delimited tokens. Subsequently, most standard evaluation tasks require knowledge of the internal morphological boundaries within the raw tokens. To accommodate this granularity requirement we developed a neural model designed to produce the *disambiguated* morphological segments for each token in context. These linguistic segmentations are distinct of the word-pieces employed by the PLM.

In the morphological extraction neural model, each input token is represented by (one or more) contextualized word-vectors produced by the PLM. Each word-piece token is associated with a vector, and for each space-delimited token, we average the word-piece vectors. We feed the resulting vector into a seq2seq model and encode the surface token as a sequence of characters using a BiLSTM, followed by a decoder that generates an output sequence of characters, using *space* as a special symbol signaling morphological boundaries.

[3]For more details and an ethical discussion, see Section 8.

Raw input	לבית הלבן (lbit hlbn)				
Space-delimited words	הלבן (hlbn)		לבית (lbit)		
Index	5	4	3	2	1
Segmentation	לבן (lbn) white	ה (h) the	בית (bit) house	ה (h) the	ל (l) to
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
Word-level NER	E-ORG		B-ORG		
Morpheme-level NER	E-ORG	I-ORG	I-ORG	B-ORG	O

Table 3: Illustration of Evaluated Word-Based and Morpheme-Based Downstream Tasks. The two-word input phrase "לבית הלבן", transliterated as "lbit hlbn" (*to the White House*), decompose into five morphological segments ('to-the-house the-white'). The Hebrew text goes from right to left.

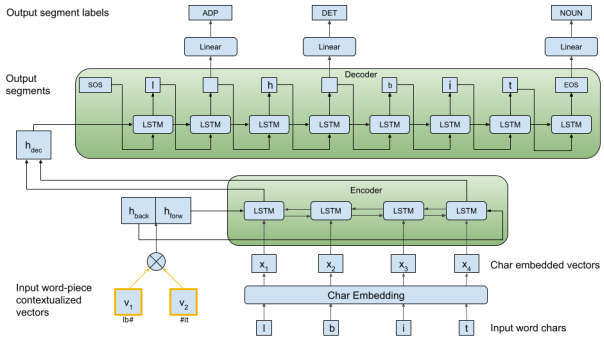

Figure 2: Illustration of the Morphological Extraction Model. The embedded vectors associated with the word-pieces (v1 and v2 representing word-piece vectors generated in Figure 1) are combined (averaged) to produce a single word context vector. This context vector initializes the hidden (forward and backward) state of a BiLSTM that encodes the characters of the origin word. The decoder LSTM outputs a sequence of characters, where a special empty symbol indicates a morphological segment boundary. In multi-task setup, a fully connected linear layer is used to predict a label whenever a segment boundary is detected.

For tasks involving both segments and labels (Part-of-Speech Tagging, Morphological-Features Tagging, Named-Entity Recognition) we expand this network in a multi-task learning setup; when generating an end-of-segment (space) symbol, the model also predicts task label, and we combine the segment-label losses. The complete morphological extraction architecture is illustrated in Figure 2.

5 Experimental Setup

Goal In order to empirically gauge the effect of model size and data quantity on the quality of the language model, we compare the performance of AlephBERT (both *small* and *base*) with all existing Hebrew BERT instantiations. In this Section, we detail the tasks and evaluation metrics. In the next Section, we present and analyze the results.

5.1 Sentence-Based Modeling

Sentiment Analysis We first report on a sentence classification task, assigning a sentence with one of three sentiment values: negative, positive, neutral. Sentence-level predictions are achieved by directly fine-tuning the PLM using an additional sentence-classification head The sentence-level embedding vector representation is the one associated with the special [CLS] BERT token.

We used a version of the Hebrew Facebook Sentiment dataset (henceforth FB) of Amram et al. (2018) which we corrected by removing leaked samples.[4] We fine-tuned all models for 15 epochs with 5 different seeds, and report mean accuracy.

5.2 Word-Based Modeling

Named Entity Recognition In this setup we assume a sequence labeling task based on space-delimited word-tokens. The input comprises of the sequence of words in the sentence, and the output contains BIOES tags indicating entity spans. Word-level NER predictions are achieved by directly fine-tuning the PLMs using an additional token-classification head In cases where a word is split into multiple word pieces by the PLM tokenizer, we employ common practice and use the first word-piece vector.

We evaluate this model on two corpora. (i) The Ben-Mordecai (BMC) corpus (Ben Mordecai and Elhadad, 2005), which contains 3294 sentences with 4600 entities and seven different entity categories (Date, Location, Money, Organization, Person, Percent, Time). To remain compatible with the original work we train and test the models on 3

[4]This version has a total of 8,465 samples and is publicly available here: https://github.com/OnlpLab/Hebrew-Sentiment-Data

different splits as in Bareket and Tsarfaty (2020). (ii) The Named Entities and MOrphology (NEMO) corpus[5] (Bareket and Tsarfaty, 2020) which is an extension of the SPMRL dataset with Named Entities. The NEMO corpus contains 6220 sentences with 7713 entities of nine entity types (Language, Product, Event, Facility, Geo-Political Entity, Location, Organization, Person, Work-Of-Art). We trained both models for 15 epochs with 5 different seeds and report mean F1 scores on entity spans.

5.3 Morpheme-Based Modeling

Finally, to probe the PLM capacity to accurately predict word-internal structure, we test all models on five tasks that require knowledge of the internal morphology of raw words. The input to all these tasks is a Hebrew sentence represented as a raw sequence of space-delimited words:

> **(i) Segmentation**: Generating a sequence of morphological segments representing the basic processing units. These units comply with the 2-level representation of tokens defined by UD, each unit with a single POS tag.[6]
> **(ii) Part-of-Speech (POS) Tagging**: Tagging each segment with a single POS.
> **(iii) Morphological Tagging**: Tagging each segment with a single POS and a set of features. Equivalent to the AllTags evaluation defined in the CoNLL18 shared task.[7]
> **(iv) Morpheme-Based NER**: Tagging each segment with a BIOES and its entity-type.
> **(v) Dependency Parsing**: Use each segment as a node in the predicted dependency tree.

We train and test all morphologically-aware models using two available morphologically-aware Hebrew resources:

- The Hebrew Section of the SPMRL Task (Seddah et al., 2013).
- The Hebrew Section of the UD treebanks collection (Sadde et al., 2018)

All models were trained for 15 epochs with 5 different seeds and we report two variants of mean F1 scores as described next.

For tasks (i)–(iv) we use the morphological extraction model (Section 4) to extract the morphological segments of each word in context and also predict the labels via Multitask training.

For task (iv) the NER task, we use the morphologically-annotated data files of the aforementioned SPMRL-based NEMO corpus (Bareket and Tsarfaty, 2020). In addition to the multi-task setup described earlier, we design another setup in which we first *only* segment the text, and then perform fine-tuning with a token classification attention head directly applied to the PLM output for the segmented tokens (similar to the way we fine-tune the PLM for the word-based NER task described in the previous section). We acknowledge that we are fine-tuning the PLM on morphological segments the model was not originally pre-trained on, however, as we shall see shortly, this seemingly unintuitive strategy performs surprisingly well.

For task (v) we set up a dependency parsing evaluation pipeline using the standalone Hebrew parser offered by More et al. (2019) (a.k.a YAP) which was trained to produce SPMRL dependency labels. The morphological information for each word (namely the segments and POS tags) is recovered by our morphological extraction model, and is used as input features for the YAP standalone dependency parser.

5.4 Morpheme-Based Evaluation Metrics

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 to the CoNLL18 metrics, we compute F1 scores, with a slight but important difference from the shared task, 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 account-

[5] Available here: https://github.com/OnlpLab/NEMO-Corpus
[6] https://universaldependencies.org/u/overview/tokenization.html
[7] https://universaldependencies.org/conll18/results-alltags.html
[8] https://universaldependencies.org/conll18/results.html
[9] respectively referred to as 'Segmented Words' and 'UPOS' in the CoNLL18 evaluation script

Task	NER (Word)		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: Word-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 corpus of Amram et al. (2018).

ing 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 multi-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.

6 Results

Sentence-Level Task Sentiment analysis accuracy results are provided in Table 4. All BERT-based models substantially outperform the original CNN Baseline reported by Amram et al. (2018). AlephBERT$_{base}$ is setting a new SOTA.

Word-Based Task On our two NER benchmarks, we report F1 scores on the word-based fine-tuned model in Table 4. While we see noticeable improvements for the mBERT and HeBert variants over the current SOTA, the most significant increase is achieved by AlephBERT$_{base}$, setting a new and improved SOTA on this task.

Morpheme-Level Tasks As a particular novelty of this work, we report BERT-based results on sub-word (segment-level) information. Specifically, we evaluate word segmentation, POS, Morphological Features, NER and dependencies compared against morphologically-labeled test sets.

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)

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)

In all cases, we use raw space-delimited tokens as input and produce morphological segments with our morphological extraction model.

Table 5 presents evaluation results for the SPRML dataset, compared against the previous SOTA of More et al. (2019). For segmentation, POS tagging, and morphological tagging we report aligned multiset F1 scores. BERT-based segmentations are similar, all scoring in the high range of 97-98 F1, which are hard to improve further.[10]

For POS tagging and morphological features, all BERT-based models considerably outperform the previous SOTA. For syntactic dependencies we report labeled and unlabeled accuracy scores of the trees generated by YAP (More et al., 2019) on our predicted segmentation. Here we see impressive improvement compared to the previous SOTA of a joint morpho-syntactic framework. It confirms that morphological errors early in the pipeline negatively impact downstream tasks, and highlight the importance of morphologically-driven benchmarks

[10] According to error analysis, most of these errors are annotation errors or truly ambiguous cases.

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)

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$_{small}$	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 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.

as an integral part of PLM evaluation for MRLs.

All in all we see a repeating trend placing AlephBERT$_{base}$ first on all morphological tasks, indicating the depth of the model and a larger pre-training dataset improve the ability of the PLM to capture word-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.

Morpheme-Level NER results Earlier in this section we considered NER a word-level task that simply requires fine-tuning on the word level. However, this setup is not accurate enough and less useful for downstream tasks, since the exact entity boundaries are often word internal (Bareket and Tsarfaty, 2020). We hence report morpheme-based NER evaluation, respecting the 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 a multi-task model and predict NER labels *while* performing segmentation.

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

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$_{small}$ 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$_{small}$ 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 morphological segmentation capacity.

In addition, comparing AlephBERT$_{base}$ and HeBERT we note 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, suggest large amounts of training data and larger vocabulary are invaluable. By exposing AlephBERT$_{base}$ to a substantially larger amount of text we increased the ability of the PLM to encode syntactic and semantic signals associated with Named Entities.

Our NER experiments further suggest that a pipeline composed of our accurate morphological segmentation model followed by AlephBERT$_{base}$ with a token classification head is the best strategy for generating morphologically-aware NER labels. Finally, we observe that while AlephBERT excels at morphosyntactic tasks, on tasks with a more semantic flavor there is room for improvement.

7 Conclusion

Modern Hebrew, a morphologically-rich and medium-resourced 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, named AlephBERT, with larger vocabulary and larger training set than any Hebrew PLM before, and with clear evidence as to its empirical advantages. Crucially, we augment the PLM with a morphological disambiguation component that matches the input granularity of the downstream tasks. Our system does not presuppose Hebrew-specific linguistic-rules, and can be transparently applied to any language for which 2-level segmentation data (i.e., the standard UD benchmarks) exists. AlephBERT$_{\text{base}}$ obtains state-of-the-art results on morphological segmentation, POS tagging, morphological feature extraction, dependency parsing, named-entity recognition, and sentiment analysis, outperforming all existing Hebrew PLMs. Our proposed morphologically-driven pipeline[11] serves as a solid foundation for future evaluation of Hebrew PLMs and of MRLs in general.

8 Ethical Statement

We follow Bender and Friedman (2018) regarding professional practice for NLP technology and address ethical issues that result from the use of data in the development of the models in our work.

Pre-Training Data. The two initial data sources we used to pre-train the language models are Oscar and Wikipedia. In using the Wikipedia and Oscar we followed standard language model training efforts, such as BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019). We use the language-specific Oscar data according to the terms specified in Ortiz Suárez et al. (2020) and we extract texts from language-specific Wikipedia dumps. On top of that, a big portion of the data used to train AlephBERT originates from the Twitter sample stream.[12] As shown in Table 2, this data set includes 70M Hebrew tweets which were collected over a period of 4 years (2014 to 2018). We acknowledge the potential concerns inherently associated with Twitter data (population bias, behavior patterns, bot masquerading as humans etc.) and note that we have not made any explicit attempt to identify these cases. We only used the text field of the tweets and *completely discard* any other information included in the stream (such as identities, followers, structure of threads, date of publication, etc). We have not made any effort to identify or filter out any samples based on user properties such as age, gender and location nor have we made any effort to identify content characteristics such as genre or topic. To reduce exposure of private information we cleaned up all user mentions and URLs from the text. 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/Wikipedia, the sheer volume of tweets helps us close the substantial resource gap.

Training and Evaluation Benchmarks. The SPMRL (Seddah et al., 2013) and UD (Sadde et al., 2018) datasets we used for evaluating segmentation, tagging and parsing, were used to both train our morphological extraction model as well as provide us with the test data to evaluate on morphological level tasks. Both datasets are publicly available and widely used in research and industry.

The NEMO corpus (Bareket and Tsarfaty, 2020) used to train and evaluate word and morpheme level NER is an extension of the SPMRL dataset augmented with entities and follows the same license terms. The BMC dataset used for training and evaluating word-level NER was created and published by Ben Mordecai and Elhadad (2005) and it is publicly available for NER evaluation.

We used the sentiment analysis dataset of Amram et al. (2018) for training and evaluating AlephBERT on a sentence level task, and we follow their terms of use. As mentioned, this dataset had some flows, and we describe carefully the steps we've taken to fix them before using this corpus in our experiments for internal evaluation purposes. We make our in-house cleaning scripts and split information publicly available.

Acknowledgements

This research was funded by the European Research Council (ERC grant agreement no. 677352) and by a research grant from the Ministry of Science and Technology (MOST) of the Israeli Government, for which we are grateful.

[11] Available at https://github.com/OnlpLab/AlephBERT

[12] https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/api-reference/get-tweets-sample-stream

References

Adam Amram, Anat Ben-David, and Reut Tsarfaty. 2018. Representations and architectures in neu-

ral 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.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the

limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2019. LINSPECTOR: multilingual probing tasks for word representations. *CoRR*, abs/1903.09442.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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.

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.