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

Anonymous ACL submission

Abstract

Large Pre-trained Language Models (PLMs) have become ubiquitous in the development of language understanding technology. However, while advances reported for English using PLMs are unprecedented, advances using PLMs reported for Hebrew are few and far between. The problem is twofold. First, Hebrew resources for training large language models have not been 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 standard PLMs. In this work we remedy both aspects. We present *AlephBERT*, a large PLM for Modern Hebrew, trained on larger vocabulary and larger dataset than any Hebrew PLM before. Moreover, we introduce a novel neural architecture that recovers the morphological segments encoded in contextualized embeddings. 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 *all* existing Hebrew models. We make *AlephBERT*, the morphological extraction model, and the novel evaluation pipeline publicly available for evaluating future 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, in Modern Hebrew, previously reported results using PLMs are far from satisfactory. Specifically, the BERT-based Hebrew section of multilingual-BERT (Devlin et al., 2019) (henceforth, mBERT) did not provide a similar boost in performance as observed by the English section of mBERT. In fact, for several reported tasks, the mBERT model results are on par with pre-neural models, or neural models based on non-contextualized embedding (Tsarfaty et al., 2020; Klein and Tsarfaty, 2020). An additional Hebrew BERT-based model, HeBERT (Chriqui and Yahav, 2021), has been recently released, yet without empirical evidence of performance improvements on key components of the Hebrew NLP pipeline.

Development of PLMs for *morphologically-rich* and *medium-resourced* languages such as Modern Hebrew introduces two challenges. First, contextualized word representations are obtained by pre-training a large language model on massive quantities of unlabeled textual data. In Hebrew, however, 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, and are not readily available in their output embeddings.

Evaluating BERT-based models on morpheme-level tasks is in fact non trivial. PLMs employ sub-word tokenization, such as WordPiece, for minimizing Out-Of-Vocabulary cases. Word-pieces are statistically generated in a pre-processing step without utilization of any linguistic information.

[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.

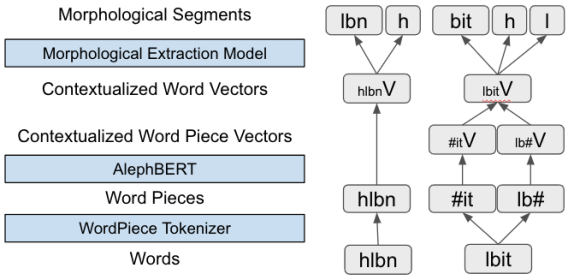

Figure 1: Illustration of the morphological extraction process. The two-word phrase "לבית הלבן", transliterated as "lbit hlbn", is mapped to word-pieces which are consumed by a PLM to generate contextualized vectors and extract the sub-word morphological units.

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 1: Data Statistics for AlephBERT's training sets.

Crucially, word-pieces *do not reflect morphological segments*. Extracting morphological units from contextualized vectors generated by PLMs is thus challenging, yet necessary in order to enable the evaluation of Hebrew PLMs on standard benchmarks. To address this we introduce a *novel* neural architecture that recovers the *morphological* sub-word segments encoded in contextualized embeddings. See Figure 1 for the relationships between the different processing units.

We present *AlephBERT*, a Hebrew pre-trained language model, trained on more data and a larger vocabulary than any Hebrew PLM before. Then, using the proposed extraction model we confirm SOTA results on *all* existing Hebrew benchmarks.[2] We thus present an evaluation suite tailored to fit MRLs, i.e., covering sentence-level, word-level and importantly sub-word morphological-level tasks (*Segmentation, Part-of-Speech Tagging, full Morphological Tagging, Dependency Parsing, Named Entity Recognition* and *Sentiment Analysis*), presenting new and improved SOTA on all tasks.

2 AlephBERT Pre-Training

Data. We acknowledge the gap in training data size compared with resource-savvy languages[3] and address it by including massive amounts of tweets.

Specifically, we employ the following datasets: **(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.[4] For data statistics, see Table 1.

Model. We used the Transformers training framework of Huggingface (Wolf et al., 2020) and trained two different models: (i) *small*, with 6 hidden layers learned from the Oscar portion of our dataset, and (ii) *base*, with 12 hidden layers trained on the entire dataset. The processing units used are the default word-pieces generated by training BERT tokenizers over the respective datasets, with a vocabulary size of 52K in both cases.[5]

3 Morphological Extraction

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 words' span. Subsequently, most standard evaluation tasks require knowledge of internal morphological boundaries within the raw words (illustrated in Table 2):
(i) Segmentation: A sequence of morphological segments representing basic processing units.[6]
(ii) Part-of-Speech (POS) Tagging: Tag each segment with a single POS.
(iii) Morphological Tagging: Tag each segment with a POS and a set of morphological features.[7]
(iv) Dependency Parsing: Use each segment as a node in the predicted dependency tree.
(v) Morpheme-Based NER: Tag each segment with a BIOES along with its entity-type label.

To accommodate the input granularity of these tasks, we developed a neural model designated to produce the *disambiguated* morphological segments for each word in context. These linguistic segmentations are thus distinct of the WordPieces.

[2] We make our PLM and online demo publicly available www.anonymous.org allowing to qualitatively assess present and future Hebrew PLMs.

[3] See Appendix B for a cross-linguistic survey & statistics.

[4] For more details and an ethical discussion, see Section 6.

[5] See Appendix C for training details, times & compute.

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

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

Raw input	לבית הלבן (lbit hlbn)				
Space-delimited words	הלבן (hlbn)		לבית (lbit)		
Index	5	4	3	2	1
Segmentation	לבן (lbn)	ה (h)	בית (bit)	ה (h)	ל (l)
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 2: Illustration of Evaluated Word and Morpheme-Based Downstream Tasks. The input is the two-word input phrase "לבית הלבן", transliterated as "lbit hlbn" (*to the White House*), which is decomposed to 5 morphological segments ('to the white the house'). The Hebrew text goes from right to left.

This morphological extraction network works as follows. Each input word is represented as (one or more) word-pieces associated with contextualized embedding vectors produced by the PLM. For each word, we average the word-pieces vectors and feed the result into a seq2seq module (The "Morphological Extraction Model" in Figure 1) that encodes the surface form as a sequence of characters using a BiLSTM, followed by a decoder that generates output sequence of characters, space used as a special symbol signalling morphological segment boundaries. We train it for 15 epochs optimized using next-char prediction loss.

For tasks involving labels (POS, Morphological Features, NER) we expand this network in a multi-task learning setup; when generating am end-of-segment symbol, the model also predicts task labels and we combine the segment-label losses.

4 Experiments

We set out to evaluate Hebrew PLMs on standard benchmarks covering sentence, word and sub-word (morphological) levels. We compare the performance of AlephBERT with all existing Hebrew BERT instantiations. First, we evaluate on **Word Segmentation, Part-of-Speech Tagging, Full Morphological Tagging, Dependency Parsing** using two available benchmarks: (i) The Hebrew Section of the SPMRL Task (Seddah et al., 2013), (ii) the Hebrew Section of the Universal Dependencies (UD) treebanks (Sadde et al., 2018). Next, we evaluate **Named Entity Recognition**. We provide Word-based NER evaluation based on the Ben-Mordecai (henceforth BMC) corpus (Ben Mordecai and Elhadad, 2005), and evaluate Word-based and Morpheme-based NER based on the Named Entities and MOrphology (NEMO) corpus (Bareket and Tsarfaty, 2020). Finally, we evaluate sentence-based **Sentiment Analysis** on a deduplicated version of Amram et al. (2018).

Task	NER (Morpheme)	NER (Word)		Sentiment
Corpus	NEMO (SPMRL)	NEMO	BMC	FB
Prev. SOTA	77.11	77.75	85.22	NA
mBERT	72.97	79.07	87.77	79.07
HeBERT	74.86	81.48	89.41	81.48
$\text{AlephBERT}_{\text{small}}$	72.46	78.69	89.07	78.69
$\text{AlephBERT}_{\text{base}}$	**79.15**	**84.91**	**91.12**	**84.91**

Table 3: Morpheme-based and Word-based NER F1. Previous SOTA is reported by Bareket and Tsarfaty (2020). Sentiment Analysis accuracy is reported on a deduplicated version of Amram et al. (2018).

To evaluate sentence-level classification we report sentence accuracy. To evaluate word-level NER performance we report F1 scores on entity spans. To evaluate morpheme-level tasks we use two variants that have been used in the literature: 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 used in the UD shared tasks (Zeman et al., 2018).[8]

Sentence-Level Tasks Sentiment analysis accuracy results are provided in Table 3. Sentence level predictions are achieved by directly fine-tuning the PLMs using an additional sentence-classification head. All BERT-based models substantially outperform the original CNN Baseline reported by Amram et al. (2018), where $\text{AlephBERT}_{\text{base}}$ is setting a new SOTA.[9]

Word-Level Tasks In Table 3 we report F1 NER scores on the two word-level test sets. Word-level NER predictions are achieved by directly fine-tuning the PLMs using an additional token-classification head. While we see noticeable improvements for the mBERT and HeBert variants

[8]For further discussion of the metrics (strengths weaknesses and comparison) we refer the reader to Appendix E

[9]For more sentence-level and word-level experimental details see Appendix D.

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$_{\text{small}}$	97.71	94.11	90.56	81.5	76.07
AlephBERT$_{\text{base}}$	**98.10**	**94.90**	**91.41**	**82.07**	**76.9**

Table 4: Morpheme-Based results on the SPMRL corpus. Aligned MultiSet (mset) F1 on Segmentation, POS tags and Morphological Features - previous SOTA reported by (Seker and Tsarfaty, 2020) (POS) and (More et al., 2019) (Features). Un/Labeled Accuracy Scores on morphological-level Dependency Parsing - previous SOTA reported by (More et al., 2019).

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$_{\text{small}}$	97.31	95.13	91.65
AlephBERT$_{\text{base}}$	**97.70**	**95.84**	**92.71**

Table 5: Morpheme-Based Aligned (CoNLL shared task) F1 on the UD corpus. Previous SOTA reported by Minh Van Nguyen and Nguyen (2021)

over the current SOTA, the most significant increase is achieved by AlephBERT$_{\text{base}}$, 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 morphological sub-words (segment-level) information. Specifically, we evaluate Word segmentation, POS, Morphological Features, NER and dependencies compared against morphologically-labeled test sets. In all cases we use raw space-delimited words as input and produce morphological segments with our new morphological extraction model.

Table 4 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 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 joint morpho-syntactic framework. It confirms how morphological errors early in the pipeline negatively impact downstream tasks, and highlight the importance of morphologically-driven benchmarks as an integral part of PLM evaluation for MRLs.

We see a repeating trend placing AlephBERT$_{\text{base}}$ 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 reported in Table 5.

Earlier in this section we considered NER as a word-based task that simply requires fine-tuning on the word level. However, this setup is not accurate enough and less useful for downstream tasks, since exact entity boundaries are often word-internal (Bareket and Tsarfaty, 2020). We hence report morpheme-based NER evaluation, respecting exact boundaries of entity mentions. To obtain morpheme-based NER labels we use the multi-task model that predicts NER labels *while* performing segmentation. The results are reported in Table 3. The differences in NER scores are substantial 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 predictions. Also, we see that while AlephBERT excels at morphosyntactic tasks, on tasks with a more semantic flavour there is room for improvement.

5 Conclusion

Modern Hebrew, a morphologically-rich and medium-resource language, has for long suffered from a gap in the resources available for NLP applications, and lower level of empirical results than observed in other, resource-rich languages. This work provides the first step in remedying the situation, by making available a large Hebrew PLM, nicknamed AlephBERT, with larger vocabulary and larger training set than any Hebrew PLM before. Crucially, we augment the PLM with a morphological disambiguation component that matches the input granularity of the downstream tasks. AlephBERT$_{\text{base}}$ obtains state-of-the-art results on the tasks of morphological segmentation, POS tagging, morphological feature extraction, dependency parsing, named-entity recognition, and sentiment analysis outperforming all existing Hebrew PLMs. Our proposed morphologically-driven suite serves as a solid foundation for future evaluation of Hebrew PLMs and of MRLs in general.

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

6 Ethical Statement

We follow the proposal of Bender and Friedman (2018) regarding professional practice for NLP technologists and address ethical issues that result from the use of data in the development of the models described 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 our AlephBERT language model originates from the Twitter sample stream.[11] As shown in Table 1 this data set includes 70M Hebrew tweets which were collected over a period of 4 years (from 2014 to 2018). We acknowledge the potential inherent concerns 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, network of 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 with minimal effort.

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 BCM dataset used for training and evaluating word-leve NER was created and published by Ben Mordecai and Elhadad (2005) and it is publicly available for NER evaluation.[12].

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 has major flows, and while we describe carefully the steps we've taken to fix them before using this corpus in our experiments, we performed this cleaning for internal evaluation purposes and we note that we have not published the fixed version of the corpus. We will make our in-hous cleaning scripts and split information publically available.

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

[12] At https://www.cs.bgu.ac.il/~elhadad/nlpproj/naama/HebrewNER.zip

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.

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.

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*.

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.

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.

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.

A Related Work

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

BERT Benchmarks An integral part involved in developing various PLMs is providing NLU multi-task benchmarks used to demonstrate the linguistic abilities of new models and approaches. English BERT models are evaluated on 3 standard major benchmarks. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) is used to test paragraph level reading comprehension abilities. Wang et al. (2018) selected a diverse and relatively hard set of sentence and sentence-pair tasks which comprise the General Language Understanding Evaluation (GLUE) benchmark. The SWAG (Situations With Adversarial Generations) dataset (Zellers et al., 2018) presents models with partial description of grounded situations to see if they can consistently predict relevant scenarios that come next thus indicating the ability for commonsense reasoning. When evaluating Hebrew PLMs, one of

the key pitfalls is that there are no Hebrew versions for these benchmarks. Furthermore, none of the suggested benchmarks account for examining the capacity of PLMs for encoding the word-internal morphological structures which are inherent for MRLs.

A.1 Multilingual vs Monolingual BERT

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

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

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

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

B PLM Training Data Size Comparison

The Hebrew portions of Oscar and Wikipedia provides us with a training set size order of magnitude smaller compared with resource-savvy languages, as shown in Table 6.

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

C AlephBERT Pre-training Details

Following the work of 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. 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.

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

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

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

D Sentence-based and Word-based Experimental Details

D.1 Sentiment Analysis

We first report on a classification task, assigning a sentence with one of three values: negative, positive, neutral. By appending a classification head we turn a BERT model into a sentence level classifier (utilizing sentence level embedded vector representation associated with the special [CLS] BERT token).

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

D.2 Word-based Named Entity Recognition

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

We evaluate this model on two corpora. We first evaluate on the BMC corpus which provides word-level annotations. It contains 3294 sentences and 4600 entities, and has 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 the 3 different splits as in Bareket and Tsarfaty (2020).[13] We then move to evaluate on the NEMO corpus which is an extension of the SPMRL dataset with Named Entities, marked by BIOES tags. This corpus provides both word and morpheme based entity annotations, where the latter contains the accurate (word-internal) entity boundaries. The NEMO corpus has nine categories (Language, Product, Event, Facility, Geo-Political, Location, Organization, Person, Work-Of-Art). It contains 6220 sentences and 7713 entities, and we used the standard SPMRL train-dev-test. Both word-based and morpheme-based models were trained for 15 epochs.

E Morpheme Level Evaluation Metrics

Aligned Segment The CoNLL18 Shared Task evaluation campaign[14] reports scores for segmentation and POS tagging[15] 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*).

[13] www.anonymous.org

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

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

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.