MsBERT:
A New Model for the Reconstruction of Lacunae in Hebrew Manuscripts

Avi Shmidman[1,2,†], **Ometz Shmidman**[1,‡], **Hillel Gershuni**[1,2,‡], **Moshe Koppel**[1,2,‡]
[1]DICTA / Jerusalem, Israel [2]Bar Ilan University / Ramat Gan, Israel
[†]avi.shmidman@biu.ac.il
[‡]{ometzshmidman,gershuni,moishk}@gmail.com

Abstract

Hebrew manuscripts provide thousands of textual transmissions of post-Biblical Hebrew texts. In many cases, the text in the manuscripts is not fully decipherable, whether due to deterioration, perforation, burns, or otherwise. Existing BERT models for Hebrew struggle to fill these gaps, due to the many orthographical deviations found in Hebrew manuscripts. We have pretrained a new dedicated BERT model, dubbed MsBERT (short for: Manuscript BERT), designed from the ground up to handle Hebrew manuscript text. MsBERT substantially outperforms all existing Hebrew BERT models regarding the prediction of missing words in fragmentary Hebrew manuscript transcriptions in multiple genres, as well as regarding the task of differentiating between quoted passages and exegetical elaborations. We provide MsBERT for free download and unrestricted use, and we also provide an interactive and user-friendly website to allow manuscript scholars to leverage the power of MsBERT in their scholarly work of reconstructing fragmentary Hebrew manuscripts.[1]

1 Introduction

Hebrew manuscripts preserve thousands of textual transmissions of post-Biblical Hebrew texts from the first millennium (Richler, 2014). In many cases, the text in the manuscripts is not fully decipherable, whether due to deterioration, perforation, burns, or otherwise. Hebrew Studies scholars spend hours upon hours attempting to determine these missing words, in order to reconstruct the original texts.

Prima facie, BERT models are optimally suited for this task, given their Masked Language Modeling objective (Devlin et al., 2019a). Indeed, a variety of high-performing BERT models for Hebrew texts have been released over the last few years, including AlephBERT (Seker et al., 2021), AlephBERTGimmel (Gueta et al., 2023), and BEREL (Shmidman et al., 2022). A recent study even showed that these models can be leveraged to complete Biblical verses (Fono et al., 2024). However, as we will show, these models are not adequately equipped to handle Hebrew manuscript texts. In order to address this need, we have pretrained a new BERT model specifically for Hebrew manuscript transcriptions. Our new model is dubbed MsBERT, short for: Manuscript BERT.

2 Reconstruction of Textual Lacunae via Deep Learning in Other Languages

Over the last few years, deep learning techniques have been utilitized for reconstruction of textual lacunae in a number of other languages. For instance, Assael et al. (2019) applied such techniques to Greek epigraphy; Bamman and Burns (2020) did so with Latin; and Fetaya et al. (2020) did so regarding Akkadian texts found in Mesopotamian cuneiform tablets. For a full survey of existing research regarding computational textual restoration, see Sommerschield et al. (2023, Section 4).

3 Challenges of Hebrew Manuscript Texts

Most existing Hebrew BERT models, including AlephBERT and AlephBERTGimmel, were trained on modern Hebrew alone. The historical texts found in Hebrew manuscripts admit to a very different writing style. Differences abound regarding vocabulary, morphology, syntax, semantics, and more. It is therefore not surprising that these models stumble when faced with historical Hebrew texts.

One notable exception is BEREL. This model was specifically trained on a corpus of historical Hebrew texts, and it is thus suited to handle the linguistic norms of such texts. However, although it can handle the *morphology and syntax* of these

[1]Link to model: https://huggingface.co/dicta-il/MsBERT
Link to website: https://mss--dicta-bert-demo.netlify.app/

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 13–18
August 15, 2024 ©2024 Association for Computational Linguistics

texts, it falls flat when confronted with the *orthography* of the manuscript transcriptions. Virtually all of BEREL's training data originates from printed editions of historical Hebrew texts. Although these printed editions date as far back as the cradle of printing at the end of the fifteenth century, they still conform to a narrow set of orthographic norms assumed by the Hebrew printing press.

In contrast, the scribes of the Hebrew manuscripts did not adhere to such norms. Examples of where the orthography of the manuscripts deviates from that of the printing press include:

- *Matres lectionis* (consonants representing vowels). Manuscripts use *matres lectionis* in a far more varied set of positions (e.g. מיצטרף rather than מצטרף).
- Acronyms. Manuscripts tend to use multiple apostrophes rather than a single double quote mark (e.g. ה'ק'ב'ה rather than הקב"ה).
- Truncated words. The manuscript scribes often transcribed only one or two letters of a given word, relying on the reader to fill in the rest from context. Hebrew manuscripts often contain long sequences of such minimal word subsets (e.g. ד' א' כי' ק' ה' rather than דבר אחר כי קדוש הוא).[2]
- Treating the preposition של ("of") as a proclitic rather than as an independent word (e.g. שלתרומה vs. של תרומה).

Needless to say, these orthographic discrepancies lead to a situation wherein texts of Hebrew manuscripts are not well supported in the BEREL model. Many of the words in the texts (including words noted above, such as מיצטרף, שלתרומה, and ה'ק'ב'ה), end up as sequences of word-pieces that the model was simply not trained for. The orthographic deviations noted above are not occasional but rather rampant throughout these texts, and thus they take their toll on BEREL's ability to handle the text.

Due to all of the foregoing, there is a need for a new specialized model for Hebrew manuscript texts, designed from the ground up - from the tokenization level and through all phases of training - specifically to handle the type of text found in Hebrew manuscripts.[3] The present paper does precisely this.

4 Model

4.1 Tokenizer

The first stage of our model design involves the training of a new word-piece tokenizer to build a BERT vocabulary that is optimally suited for Hebrew manuscript texts. For the training corpus for the tokenizer we start with our full set of manuscript transcriptions (section 4.3.1). Additionally, we add in a corpus of standard editions of Hebrew texts from before the printing era (section 4.3.2), to widen the vocabulary with additional words that are likely to be found in Hebrew manuscripts, even if they aren't in our particular corpus of manuscript transcriptions.

We use the Word-Piece tokenization method proposed by Song et al. (2021), with adjustments to handle the apostrophe and double-quote marks, which otherwise would have been tokenized into separate word pieces. Specifically, we avoid breaking on a double-quote between Hebrew letters (e.g., תנ"ך), or on apostrophes which succeed Hebrew letters (e.g., א'ע'ג).

Following previous work (Gueta et al., 2023), the tokenizer was trained with a vocabulary size of 128,000 tokens. In addition, in order to properly represent the fragmentary nature of Hebrew manuscripts, we add two special tokens to the vocabulary: `[GAP]` (indicating a large gap, or a gap of an unknown number of words) and `[ONEGAP]` (indicating a single missing word).

4.2 Architecture

The model's architecture is based on the BERT-base architecture (Devlin et al., 2019b), trained

[2]This particular sequence is attested in a Cairo Genizah fragment of *mekhilta de-rashbi*, a legal midrash; see Kahana (2005), p. 25.

[3]To be sure, to a certain extent, challenges of manuscript orthography can be addressed with existing models if normalization is applied during preprocessing. However, the oddities of manuscript orthography often result in ambiguous forms which must be disambiguated prior to normalization, and aggressively normalizing such forms would likely result in errors early on in the pipeline, adversely impacting the model's capabilities overall. Furthermore, the oddities of manuscript orthography are not entirely predictable, and constructing a completely comprehensive normalization routine would prove difficult. Additionally, for downstream tasks such as handwritten text recognition, it is desirable to have a model which can predict the specific orthographic forms which fits the orthographic norms of the context words; this would not be possible if everything was normalized in advance. For these reasons, we opted to produce the new model presented here, tokenized and pretrained from scratch. Nevertheless, in future work we hope to explore the preprocessing normalization approach as well, and to properly compare the results.

on a DGX-A100 with 4xA100 40GB cards. The training was done with the fused lamb optimizer combined with AMP (Automatic Mixed Precision). A polynomial warmup learning rate scheduler was used to warm up for a portion of the training steps and then decay the learning rate over the total steps.

4.3 Training Data

On the one hand, we wish to train the model specifically for Hebrew manuscript texts; yet our corpus of Hebrew manuscript texts is not sufficiently large to train a BERT model alone, and thus we need to augment it with larger corpora of Hebrew. We first describe the multiple corpora which we used as part of this process, and then describe how we combine them together during the training process.

4.3.1 Hebrew Manuscript Corpus

We collected transcriptions of Hebrew manuscripts from Hebrew Studies scholars who generously agreed to provide their transcriptions for this project. All in all, this corpus consists of over 67 million words, representing texts authored between the 3rd and 13th centuries.

4.3.2 Pre-Print Rabbinic Corpus

The Pre-Print Rabbinic Corpus is a collection of digitized Rabbinic texts authored before the age of printing (that is, before the end of the 15th century). This corpus contains a total of 49 million words.

4.3.3 Comprehensive Rabbinic Corpus

This corpus contains a maximally comprehensive set of digitized Rabbinic Hebrew texts from all available time periods, stretching from the 3rd century until today. It contains over 400 million words, including the full corpus of texts from Sefaria [4], plus many texts which we have scanned and digitized in-house.

4.4 Training Objectives

We train our model on the Masked Language Modeling objective. We implement two restrictions when selecting the random tokens to mask:

1. We don't allow masking of word-piece tokens which are not full words. The task of predicting just one part of a word given the rest of the word is too easy and does not result in significant optimization.

2. We don't allow masking of the `[GAP]` and `[ONEGAP]` tokens, since we wish to train the model to predict actual Hebrew words.

[4] sefaria.org.il

During training we chunk the texts into sequences of up to 256 tokens. To ensure we train on sentences of substance, we remove sentences with fewer than 3 words or where most of the sentence consisted of `[GAP]` tokens.

4.5 Training Phases

In order to leverage the larger Hebrew corpora, while still placing the emphasis specifically on the manuscript transcriptions, we used a three-stage procedure, as follows:

Phase 1: For the first phase of the training - when the model is most malleable - we trained only on the manuscript Corpus (4.3.1) and the Pre-Print Corpus (4.3.2). We trained for one full epoch over these corpora, using a global batch size of 2048 examples per iteration, for a total of 4200 iterations. The learning rate was initialized to 0, and was warmed up to 6e-5 by the end of this phase. Total training time was 7 hours.

Phase 2: For the second phase of the training, we continued training with all three corpora. We trained for a total of 5.5 epochs of the corpora, using a global batch size of 8192 examples, for a total of 15,400 iterations. We continued warming up the learning rate until 6e-3 and then applied a polynomial scheduler with a degree of 0.5. Total training time was 2.1 days.

Phase 3: For the third phase of the training, we confined the training corpus solely to our set of Hebrew manuscript transcriptions. We ran this corpus for 3.5 epochs with a batch size of 1024, for a total of 15,800 iterations. We used a learning rate of 5e-5, with the same scheduler as in phase 2. Total training time was 5.5 hours.

5 Experiments and Results

We evaluate the performance of MsBERT in comparison with the three BERT models discussed above. We evaluate MsBERT both in its final form (*MsBERT-Full*), as well the checkpoint upon completing phase 2 (*MsBERT-Ph2*), before the final training phase on the dedicated manuscript corpus, in order to evaluate the impact of that final training phase.

Our first test evaluates the models' ability to predict a masked word within a Hebrew manuscript transcription. We tested the models on Hebrew manuscript transcriptions from two separate genres: first, manuscripts of a homiletic text from the

5th-6th century (*shir hashirim rabba*),[5] and second, a manuscript of a Hebrew legal text from the fourth quarter of the first millennium dubbed *me'en sh'iltot* (Emanuel, 2019, 82-148). These transcriptions were not part of the training corpus of any of the BERT models.

It should be emphasized that this word prediction task is particularly difficult due to the fragmentary nature of the aforementioned manuscripts. Many words are damaged or indecipherable throughout both manuscripts, and many of the extant words are truncated. It should also be noted that although MsBERT was trained with the special GAP and ONEGAP tokens in order to provide it with optimal knowledge of the type of gaps found in Hebrew manuscript, here we avoided use of those tokens, to allow for a fair comparison with the other models in which those tokens are not available. Instead, we replace any single-word gaps with the universal MASK token, and we treat GAP tokens as paragraph separators, cutting the input samples at that points. We run the word-prediction test on all full words within the text (we don't include truncated words in the test, because they can potentially match multiple forms). In all, we test predictions for 9333 words in the first corpus, and 9475 words in the second corpus.

We report accuracy indicating how often the masked word was correctly predicted within the top 1, top 3, or top 10 (ignoring predictions of truncated words, word pieces, or punctuation). When we test for word equivalence, we ignore medial *vav* and *yod* characters, because words that differ only in their *matres lectionis* are essentially the same word. The results can be seen in Tables 1 and 2. MsBERT outperforms all models on both tests. As expected, BEREL (184M params) performs far better than both AlephBERT (120M params) and AlephBERTGimmel (184M params), due to its exposure to a large Rabbinic Hebrew corpus. Yet, at the same time, the substantial gap between BEREL and MsBERT (also 184M params) demonstrates the critical importance of our new training corpus which reflects the orthographic range of Hebrew manuscripts. Furthermore, the results demonstrate that the final phase of manuscript-only training does in fact provide a boost in the model's ability to handle these fragmentary transcriptions.

Our second test evaluates the models' ability to analyze the content of the texts, by testing whether the models can identify the words that comprise quoted citations. Our evaluation involves two genres: legal midrash and homiletic midrash. Many citations of Biblical verses are interspersed throughout such texts. Unlike modern texts, these texts do not use any form of quotation marks or braces to mark the citations; rather, the reader must figure this out from context. Thus, this test poses an ample challenge for our BERT models, to determine how well they are able to parse the context and to thus determine which words comprise the claims and discussion, and which words are source material interwoven within. The test set includes manuscripts transcriptions of *mekhilta de-rashbi* (a legal midrash),[6] and *shir hashirim rabah* (a homiletic midrash).[7] The training set includes excerpts from standard print editions of *sifre* Deuteronomy (a legal midrash) and *kohelet rabba* (a homiletic midrash). We selected training texts from printed editions in order to increase the challenge: the BERT models must apply the lessons learned from standard Hebrew texts to Hebrew manuscripts with their nonstandardized orthography. This challenge is particularly acute when it comes to identifying citations, because print editions tend to quote sources in full, whereas the manuscript scribes, painstakingly writing by hand, generally sufficed with more subtle references of only two or three words.

All of these texts were annotated by our in-house expert who marked the words that comprise the source citations. We include both full words and truncated words in the experiment. In total, the test set includes 1753 words, 288 of which are citations; the train set includes 3976 words, 1122 of which are citations.

We fine-tune each of the BERT models on the task of classifying words as "Citation" or "Not Citation". We input sequences of 64 tokens (batch size = 2, LR = 5e-5, Epochs = 30). We report the results in Table 3. Although precision is similar across the various models, MsBERT far outperforms all of the other models on the recall.

6 Conclusion

The BERT model we present here is the first of its kind: a model specifically trained to handle

[5] https://schechter.ac.il/midrash/shir-hashirim-raba/; we use the set of 16 Cairo Genizah fragments downloadable there.

[6] We test on fragment 13 from Kahana (2005), p. 161-162.

[7] We test on Cairo Genizah fragments 15 and 16 from https://schechter.ac.il/midrash/shir-hashirim-raba/.

Model	Top	Top 3	Top 10
AlephBERT	22.90	31.95	40.86
AlephBERTGimmel	25.57	34.89	43.96
BEREL	47.33	58.99	67.91
MsBERT-Ph2	56.77	69.50	77.25
MsBERT-Full	**59.99**	**71.99**	**79.10**

Table 1: Word prediction on mss of *shir hashirim rabba*

Model	Top	Top 3	Top 10
AlephBERT	26.37	37.27	46.80
AlephBERTGimmel	31.18	42.91	53.11
BEREL	56.24	68.88	76.89
MsBERT-Ph2	62.43	74.5	82.06
MsBERT-Full	**63.99**	**75.85**	**82.99**

Table 2: Word prediction on the *me'en sh'iltot* manuscript.

the orthographic oddities of Hebrew manuscript transcriptions. As we have shown, our model substantially outperforms all existing Hebrew BERT models on a variety of tests regarding Hebrew manuscript texts. We release the model for unrestricted use and free download.

We expect that this new model will aid Hebrew manuscript scholarship in a number of ways. First and foremost, this model provides a computational foundation to aid scholars in deciphering and reconstructing Hebrew manuscript text. As noted, we have in fact already developed an interactive and user-friendly website to bridge the gap between the scholar and the technology; scholars can input their text as they have deciphered it so far, and then receive predictions from the model which fit the context and any additional extant letters. Moreover, in addition to the basic word-prediction task, we have demonstrated that this model also excels beyond other models in its ability to classify parts of the text. Thus, this model provides a critical foundation for researchers who wish to build deep learning models for automatic analysis of Hebrew manuscripts. Finally, because this model is so keenly aware of the orthographic reality of Hebrew manuscripts, it provides an ideal foundation on which to build Handwritten Text Recognition systems for Hebrew manuscripts.

7 Limitations

When building the training corpus of Hebrew manuscript transcriptions, we endeavored to include as many genres as possible, to ensure maximal applicability of the model. However, we note that there is one specialized genre found in Hebrew manuscripts which is not at all covered in the present model: the genre of Hebrew liturgical poetry. These Hebrew poems draw upon all sorts of unusual and unique words which are not represented in the present model, and which really require a separate specialized model in and of itself. We don't expect this model to perform well on manuscripts containing Hebrew liturgical poetry.

Model	Precision	Recall
AlephBERT	76.99	20.21
AlephBERTGimmel	77.40	47.60
BEREL	78.67	81.94
MsBERT-Ph2	**79.31**	87.85
MsBERT-Full	78.20	**89.93**

Table 3: Evaluation on the citation identification test.

Acknowledgements

This paper has been funded by the Israel Science Foundation (Grant No. 2617/22) and by the European Union (ERC, MiDRASH, Project No. 101071829), for which we are grateful. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

We will to express our thanks to Eli Handel for his substantial help in preparing and preprocessing the input corpus.

References

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on Greek epigraphy. 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)*, pages 6368–6375, Hong Kong, China. Association for Computational Linguistics.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Simcha Emanuel. 2019. *Hidden Treasures from Europe [Hebrew]*, volume 2. Mekize Nirdamim, Jerusalem.

Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary Babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 117 (37):22743—-22751.

Niv Fono, Harel Moshayof, Eldar Karol, Itai Assraf, and Mark Last. 2024. Embible: Reconstruction of Ancient Hebrew and Aramaic texts using transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 846–852, St. Julian's, Malta. Association for Computational Linguistics.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2023. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. *Preprint*, arXiv:2211.15199.

Menachem I. Kahana. 2005. *The Genizah Fragments of the Halakhic Midrashim [Hebrew]*. Magnes Press, Jerusalem.

Benjamin Richler. 2014. *Guide to Hebrew Manuscript Collections*, second, revised edition. The Israel Academy of Sciences and Humanities, Jerusalem.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert:a hebrew large pre-trained language model to start-off your hebrew nlp application with. *Preprint*, arXiv:2104.04052.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing berel: Bert embeddings for rabbinic-encoded language. *Preprint*, arXiv:2208.01875.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, and Jonathan Prag. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast wordpiece tokenization. *Preprint*, arXiv:2012.15524.