OtoBERT: Identifying Suffixed Verbal Forms in Modern Hebrew Literature

Avi Shmidman and Shaltiel Shmidman

Funded by the European Union
European Research Council
Established by the European Commission

What is a suffixed verbal form?

"צייר גדול צייר את דמות תבניתה של בלומה **וקבעה** בלבו של הירשל"

"A great artist painted the image of Bluma ***and he set it*** in the heart of Hershel"

(S. Y. Agnon, A Simple Story)

The Challenges

Suffixed verbal forms are exceedingly rare

Corpus	Corpus Size (Words)	Suffixed Verbs	Freq (Per 10K Words)
News1	185K	7	0.38
News2	43K	2	0.47
Lit	135K	222	17.76

Often a homograph with a frequent alternative

- ★ וקבעה ("and she set" and "and he set it")
- ★ הזמינו ("they ordered" and "he ordered it")
- ★ לימדו ("they taught" and "he taught him")
- ★ הגישה ("she offered" and "he offered her")

Our Proposal: New BERT pretraining

We pretrain a new BERT with modified tokenization

- All cases of direct-object pronouns are combined together with the preceding word as a single token.

Before: לעצמו | אותה | וקבע | חניה | תפס | השכן

After: לעצמו | וקבע_אותה | חניה | תפס | השכן

- New expressive power: the MLM head can now predict multiword tokens consisting of a verb + suffix

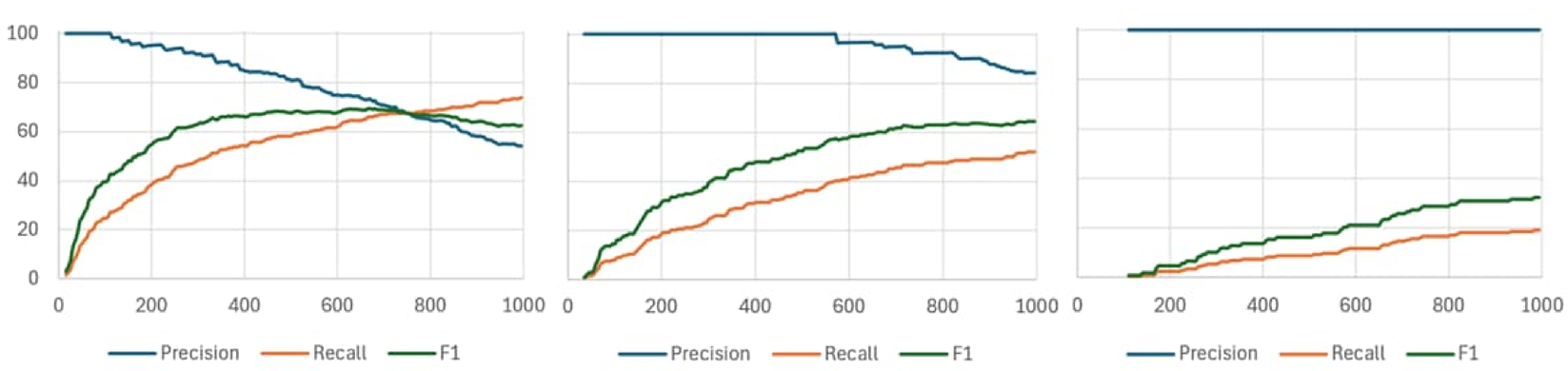
"and he set it"

A two word sequence (verb followed by direct-object pronoun) is *semantically equivalent* to the suffixed verbal from, and far more frequent

Hypothesis

MLM prediction of these multiword tokens indicates the presence of a suffixed verbal form.

Results



(a) N=5 (b) N=10 (c) N=25

Model	Precision	Recall	F1
OtoBERT, K=1000, N=1	15.75	**95.57**	.270
OtoBERT, K=1000, N=5	54.15	73.89	.625
OtoBERT, K=1000, N=10	84.13	52.22	**.644**
OtoBERT, K=1000, N=25	**100**	19.21	.322
mBERT w/ Classifier	28.23	62.12	.388
AlephBERT w/ Classifier	38.92	62.50	.480
DictaBERT w/ Classifier	48.27	73.86	.584
DictaBERT Morph Tagger	88.73	23.86	.376

Table 2: Precision, recall, and F1 vis-a-vis the class of suffixed verbal forms for each of the evaluated methods.

Download the new Model & Test Set

Model: https://huggingface.co/dicta-il/otobert

Test Set: https://huggingface.co/datasets/dicta-il/hebrew_suffix_verbal_forms

Contact info: Avi.Shmidman@biu.ac.il