WhyNet - Hebrew style transfer, text generation and classification

Nir Weingarten, Tamir Tshuva
IDC, Israel

Abstract

Hebrew is a Morphological rich language, making it's modeling harder than simpler language. Recent developments such as Transformers in general and Bert in particular opened a path for Hebrew models that reach SOTA results, not falling short from other non-MRL languages. We explore the cutting edge in this field performing style transfer, text generation and classification over news articles collected from online archives. Furthermore, the news portals that feed our collective consciousness are an interesting corpus to study, as their analysis and tracing might reveal insights about our society and discourse.

The Problem in More Details

Hebrew was classified as a Morphological Rich Language (MRL) by Tsarfaty et al. (2010). In MRLs prefixes and suffixes are appended to words to change their grammatical meaning. This structure has been shown to results in inherent morphological ambiguity in the language. For this reason, amongst others, Hebrew NLP algorithms lag behind other non MRL languages. The use of powerful new tools and the right tokenization techniques allow us to perform different supervised and unsupervised NLP tasks with good results in a short amount of time. Learning on more than 200K labeled news articles scraped from online archive we manage to: 1. Generate coherent news articles from any one of 11 different news sections and any month since January 2000; 2. Classify news articles to their correct news section with high accuracy and minimal learning time; 3. Cast the representation of an article's title or any other sequence down to 2D and examine it's proximity to classified clusters.

Method

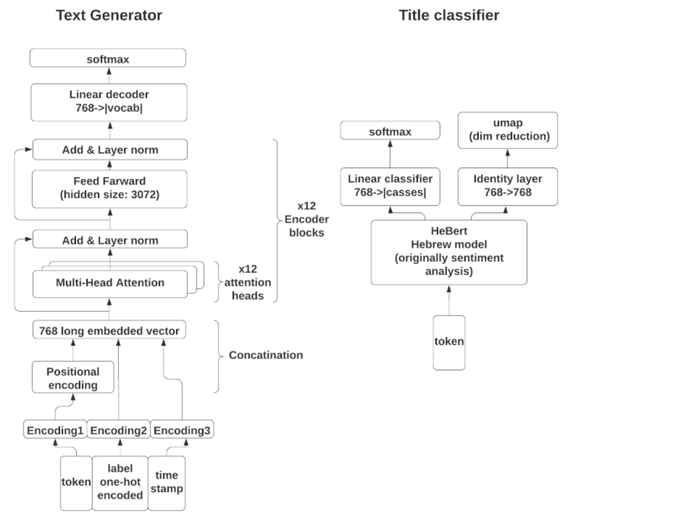
Data collection: Data was collected from online news archive using a specially built python web scraper. The data was later split 0.9-0.1 to train and validation sets.
Data Article title classification: The classifier is based on a newly released pre-trained Hebrew Bert model called HeBert. A sentiment analysis version of HeBert is available online and was put to use. It's final linear classification layer was replaced with a new layer of the relevant dimensions. A YAP based tokenizer built for the HeBert model was used to tokenize article titles and new new model was fine tuned to classify article titles to one of 11 different news sections.
Text generation and style transfer: As HeBert is based on a YAP tokenizer it couldn't be fine tuned for text generation. A new Bert like encoder only model was constructed using Smiling-face's Transformers library. A char-based tokenizer was used as it was shown to yield the best results in unsupervised Hebrew tasks. News articles were processed to fit a Bert like structure s.t. the model recognizes title, sub-title and article body sequences. Style transfer was achieved by concatenating an embedding of the news section and timestamp to the embedded vector of each token. Upon generation a start string, news section and timestamp is selected by the user and inserted into the model.

Results

The classification model reached an accuracy of 0.83 on validation set after three training Epochs using SGD and wAdam optimizer, each epoch taking less than 15 minutes on a Tesla p100-pcie-16gb GPU. Further training did not improve the results. The generator model trained for 20 epochs before starting to overfit. Text generation is an unsupervised task, making it's performance difficult to measure. A perplexity of 2.55 on validation set was measured on the last epoch and we provide qualitative observations and examples. The model seems to generate coherent sentences that relate to the starting vector and in most cases to the news section and time provided to it. Some outputs have minor grammatical flaws and some tend to recourse into a loop of the same few word. Most outputs are clear, logical, exhibit the required style and in some cases seem genuine.

Correct classifications and network architecture

News section classification	Article title (Hebrew)
General (כללי)	בת 60 נהרגה בתאונת דרכים בעוטף עזה
Law and Justice (משפט ופלילי)	אדם אבוטבול נעצר : גנב אופנוע במטרה לפגע
World economics (כלכלה עולמית)	האצה באינפלציה בגוש האירו בספטמבר
Military (צבאי)	חיילים התבקשו לספר : "משהו מצחיק מצוק איתן"



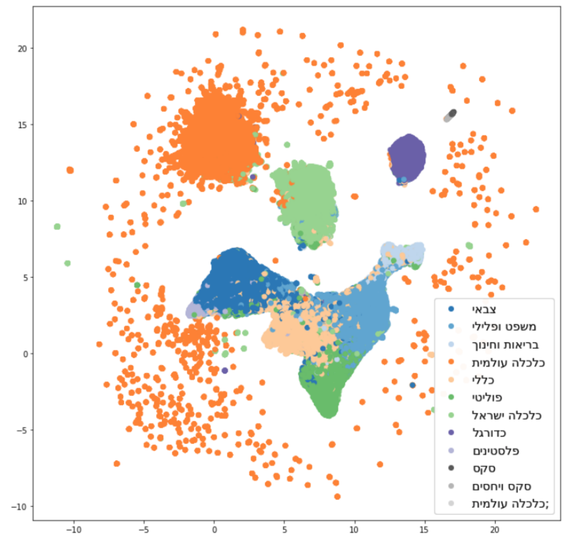
A web site was created to host the model at `www.why-net.net`

Conclusions

Despite the ambiguity and morphological richness of Hebrew, it is evident that transformers in general and Bert in particular are good tools for quick and transferable Hebrew models. Using the right tokenization and architecture and sufficient data Transformers can be used to generate compelling text and accurate classification. Furthermore, we find that rich style transfer is easily attained in transformers using simple concatination in the embedding layer. Assuming $H \supseteq C we can project that the latent representations of titles in our model is accurate to the assuming all of the above, it's location will reveal it's place in the discourse of the corp$

References

Tsarfaty R, Seddah D, Goldberg Y, K¨ubler S, Versley Y, Candito M, Foster J, Rehbein I, Tounsi L (2010) Statistical parsing of morphologically rich languages (spmrl) what, how and whither. Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, 1–12. Jacob Devlin. 2018. Multilingual bert readme document Reut Tsarfaty, Amit Seker, Shoval Sadde, Stav Klein. 2019. What is wrong with Hebrew NLP? And how to make it right Amir More and Reut Tsarfaty. 2016. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In Proceedings of COLING, pages 337–348. The COLING 2016 Organizing Committee. Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. Computational Linguistics, 37(1):105–151. Avihay Chriqui and Inbal Yahav. 2020. HeBERT HebEMO: a Hebrew BERT model and a tool for polarity analysis and emotion recognition Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. Computational Linguistics, 37(1):105–151. Avihay Chriqui and Inbal Yahav. 2020. HeBERT HebEMO: a Hebrew BERT model and a tool for polarity analysis and emotion recognition Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In Proceedings of the ACL, HLT '11, pages 188–193, Stroudsburg, PA, USA. ACL. Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin. 2017. Attention is All you Need. Devlin, Chang, Lee, Toutanova. 2018. BERT: Pre-training of Deep Bidirectional transformers for Language Understanding.



Scatter plot of dimension reduced latent representation of article titles