FigureOut - Automatic Detection of Metaphors in Hebrew Across the Eras

Münz-Manor, Ophir
ophirmm@openu.ac.il
The Open University of Israel, Israel

Toker, Michael
tok@campus.technion.ac.il
Technion - Israel Institue of Technologhy

Mishali, Oren
oren.mishali@gmail.com
Technion - Israel Institue of Technologhy

Kimmelfeld, Benny
bennyk@cs.technion.ac.il
Technion - Israel Institue of Technologhy

Belinkov, Yonatan
belinkov@technion.ac.il
Technion - Israel Institue of Technologhy

Cohen, Adir
adircohen@campus.technion.ac.il
Technion - Israel Institue of Technologhy

The ability to distinguish between figurative and literal language is essential for computers to better understand text in a natural language. This ability is crucial for many tasks, such as sentiment analysis, machine translation and summarization. At the same time, the differentiation between the figurative and the literal is a major interpretative task for scholars of literature, especially of poetry. While several machine learning models for this task have been developed, most of them focus on modern non-poetic texts, and as far as we know none of them deals with Hebrew. In this project we fill this gap by developing tools for the automatic detection of figurative language in Hebrew poetry. We focus specifically on metaphors in a corpus of poems that were written in the Galilee between the fifth and eighth century of the common era. The metaphors in the poems were annotated manually by the literary experts and we use them in order to train our models and validate their results.

A metaphor is a figure of speech in which a word or phrase is used to refer to an object or action that differs from its literal meaning. A statement such as "The warrior has a heart made of stone" does not refer to the word stone in the literal sense. The word stone is used to describe the warrior's emotionlessness, which is characterized by the stone. Recent advances in NLP that were driven by the transformer architecture (Vaswani et al., 2017), show very promising results on the metaphor detection task in English and in other European languages. These models are typically based on pretrained models, which are trained on a large dataset, then fine-tuned for the specific task. Virtually every NLP task is now able to be solved with a high degree of accuracy by fine-tuning a transformer model.

In view of the fact that Hebrew is a morphologically-rich language, NLP in Hebrew is more challenging than that in languages with simpler morphologies (Tsarfaty, et al., 2019). Several transformer-based models have been pre-trained on Hebrew, but these models have been trained on a relatively small dataset compared to English (18 GB vs 160 GB). Token classification can be accomplished using a trained encoder, such as BERT (Devlin et al., 2018), that was pre-trained on masked language modeling on a large corpus and then fine-tuned on a small labeled dataset. For Hebrew, we are using AlephBert (Seker et al., 2021) and finetuning it for metaphor detection using the aforementioned labeled dataset. We consider additional alternatives, one of them is to fine-tune BEREL (Shmidman et al., 2022), which was pretrained on Rabbinic Hebrew that is more similar to ours, although its corpus was much smaller (220M compared to 1.9B in AlephBert). Our hypothesis is that the BEREL model produces better results since it was trained in a language that is more similar to Piyyut than modern Hebrew. As previously noted, the task is challenging and we obtain a score of 48.3 F1 on it.

The project introduces a new, challenging dataset in Hebrew and it seeks to extend automatic metaphor detection capabilities in pre-modern Hebrew. In the poster, we present the major literary characteristics of the corpus, the computational approaches and methods we use, and the results together with error analysis.

Finally, FigureOut is a collaborative project of literary scholars specializing in medieval Hebrew poetry and computer scientists who specializes in NLP and deep machine learning. We envision the humanities and computational connection not merely as practical one where one “side” needs a solution and the “other” provides it. We strongly believe that both teams bring to the table their background and expertise but at the same time a deep desire to understand the premises, procedures and methods of each discipline. For us, this approach is the most appropriate way to build a bridge between the two fields and to truly expand the knowledge and methods of both of them. By so doing we seek to bolster the Digital Humanities and in particular Computational Literary Studies and exemplify the great potential of real, deep interdisciplinary collaboration.

Bibliography

Devlin, Jacob / Chang , Ming-Wei / Lee, Kenton / Toutanova , Kristina (2018): “Bert: Pre-training of deep bidirectional transformers for language understanding”, https://arxiv.org/pdf/1810.04805.pdf [25.04.2023].

Vaswani, Ashish / Shazeer, Noam / Parmar, Niki / Uszkoreit, Jakob / Jones, Llion / Gomez, Aidan N / Kaiser, Łukasz / Po-losukhin, Illia (2017): “Attention is all you need”, in: *Advances in Neural Information Processing Systems* , 30.

Tsarfaty, Reut / Seker, Amit / Sadde, Shoval / Klein, Stav (2019): “What’s wrong with hebrew nlp? and how to make it right”, https://arxiv.org/pdf/1908.05453.pdf [25.04.2023].

Seker, Amit / Bandel, Elron / Bareket, Dan / Brusilovsky, Idan / Shaked, Refael Greenfeld / Tsarfaty, Reut (2021): “Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with”, https://arxiv.org/pdf/2104.04052.pdf [25.04.2023].

Shmidman, Avi / Guedalia, Joshua / Shmidman, Shaltiel / Shmidman, Cheyn Shmuel / Handel, Eli / Koppel, Moshe (2022): “Introducing berel: Bert embeddings for rabbinic-encoded language”, https://arxiv.org/pdf/2208.01875.pdf [25.04.2023].