HeAna - Hebrew Analogies

Maxim Ifergan maxim.ifergan@mail.huji.ac.il
Esther Shizgal esther.shizgal@mail.huji.ac.il
Mor Turgeman mor.turgeman2@mail.huji.ac.il

1 Introduction

The ability to identify abstract relations between words in the form of 'Analogy Quizzes' has been used to evaluate language ability in humans until today. Since it received attention in Mikolov et al. (2013), not much work has been done in the domain in the LM era, especially with more abstract forms of analogies, and certainly not in Hebrew. Previous work as Mikolov et al. (2013) focused on a defined simple set of relations, such as 'Capital of' rather than complex semantic relations like the questions found in the Israeli Psychometric exams, and the American SATs. A Hebrew example:

קרב אגרוף : תגרה ∼ מדורה : שריפה

The first attempt to solve this type of analogies using Language modeling (Ushio et al., 2021), showed that transformer-based language models struggle with abstract and complex relations and receive similar results to older generation models.

We present HeAna - a Hebrew Analogies dataset. To our knowledge, this is the first Hebrew dataset for this assignment. Very few datasets for this assignment are available in English as well. Our contributions in HeAna are: (1) A dataset with over 500 psychometric questions and 2.5K context sentences (2) Dataset analysis (3) Multiple experimental setups attempting to solve the questions. The project code and dataset is available at GitHub.

2 HeAna

HeAna is a complex analogy dataset that contains 552 relations structured as four-choice cloze-questions. Each sample includes five context sentences; two of them describe the correct relation and the other three demonstrate why the other relations are false.

2.1 Data Collection

The data was collected from The National Institute for Testing and Evaluation (NITE) and from the Niv-Revach solutions archive (Revach). The questions were collected using a scrapping script, and the context sentences were typed manually. We organize the data in the form illustrated in Table 1. The metadata that we include for each sample contains the question's difficulty as appears in NITE, whether it was originally written with Niqqud (ניקוד) and does it include proverbs or phrases.

We divide HeAna into three splits: train, test, and validation, by the fractions 90%, 5%, and 5% respectively, while maintaining the original proportion of the difficulty level and uniform distribution of the correct answer.

2.2 Dataset Analysis

This subsection refers to the data in the cloze-question part of the dataset, without the context sentences.

The samples in HeAna consist of six difficulty levels, with 92 questions of each level; each one has four options for the correct answer. In order to minimize possible artifacts in the dataset, we reorganize the data to assure that the correct answer is distributed equally.

We explore the lexical overlap between the dataset splits as presented in Table 4.

The samples that originally appeared in the psychometric with Niqqud, and ones containing phrases, are likely to be more challenging to solve; as the models available for this assignment do not recognize Niqqud, and the likelihood of these phrases in Hebrew texts in lower than that of other words. In Table 3 we analyze the portion of each split containing these samples.

We use Yap (Adler and Elhadad, 2006) and Trankit (Nguyen et al., 2021) to analyze the morphology of the words by their context sentences in HeAna. We divide our data into four main parts of speech (POS): Nouns, Verbs, Adjectives and Phrases, covering 83% of our data. The rest of the data consists of other POS or is not recognized by our parsers. The final analysis is a combination of both parsers and is presented in figures 1 - 2. It can be seen that the majority of the words are nouns and verbs. In addition, we examine the different POS of word-pairs. For this purpose, for each ques-

tion, we look only at POS relation of the base pair of words, since the options of the answer consist of the same POS relation as the base pair.

3 Experimental Results

We examine several baselines as described below. The results are shown in Table 6 and discussed in the results section 3.3.

3.1 Classification Models

We use AlephBertGimmel (ABG) (Guetta et al., 2022) for the classification task. For this task, we consider two main input formats: Multiclass Classification and Semi-binary classification.

For the multiclass case, we create input sentences containing the base relation and the four optional word-pair relations. We fine-tune ABG on various input formats (Table 5) in order to achieve the model with the best results.

We construct the train set of the semi-binary setup. For every question in HeAna train split, we create two samples; one for the base relation with the right relation, and one for the base relation with one of the wrong relations, which was chosen randomly. We form the test and validation sets by the following strategy: For every question we create four samples, one that is correct and three that are wrong. The predicted answer is the sample with the highest score.

3.2 Generative LMs

Inspired by the way humans are guided to solve analogies, we explore the idea of finding the correct analogies word-pair by first determining the accurate relation in context sentences. This approach requires working with models that can describe the relation, hence Generative LMs.

We use *hebrew-gpt_neo-xl* (Adler) an open-source Hebrew generative LM base on *GPT-neo XL* (Gao et al., 2020). The three different setups we performed were: (1) Multiclass classification - generating the correct class label. (2) Finetune the model to generate the correct two context sentences and choose the correct answer. (3) A two step approach, first finetuning the model to generate the context sentence for a given word-pair; then performing the procedure described in setup 2. For the pretraining data, we only used the false answers of the training set context sentences.

3.3 Results

The results suggest that HeAna is a challenging dataset. The different setups barely overpass the random guess. This is reasonable as the dataset was originally meant to test human lingual capabilities. Our results align with previous work in English (Ushio et al., 2021) showing that encoder-based models receive better results than generative models. Moreover, The simpler semi-binary classification overcomes the multiclass classification setting in ABG. Even though the binary case leaves much of the data outside, we believe it simplifies the problem.

As to the generative approach, as expected the model that generated the context sentences leverages the guidance and performs better on the same examples. Furthermore, in the pretraining setup, the model acquires the ability to describe relations and to properly detect the correct relation in the quizzes similar to the human guidelines. Although *hebrew-gpt_neo-xl* underperforms ABG. The model's ability to describe the relation sheds light on the output and assures the answer is not a simple guess. This is demonstrated in Table 2.

4 Future directions

Due to the data origin as a human test, comparing ML performance to human success can be insightful. Does the model struggle with the same difficult questions?

In future research, we would like to explore additional pretraining of ABG, on data from our domain, which might accelerate its performance on HeAna. Context sentences and Psychometric vocabulary lexicons are available for this cause.

We ran ChatGPT on a few exmaples of HeAna. Trying different prompts and few-shot settings requires unlimited access to commercial LLM's and is left for further research. There are generative LLMs that perform well on English text while struggle on inputs written in other languages. We hypothesize that adjusting HeAna to English either by transliteration or translation *Text2Text* models can give interesting results.

Other experiments that may give insights to this research include: Generating new relations, rewriting the examples so no Niqqud is needed to understand them ("Ktiv-Maleh"), examining the effect of the POS of the word pairs on the model performance, and cross-lingual transfer from English SAT exams.

base	base description	...	option 3	option 3 description	...	correct	vowelized	phrases
לכונן : נוסד	לכונן זה לגרום למשהו להיווסד		לכלות : אזל	לכלות זה לגרום למשהו לאזול		3	1	0
שבע רצון : סיפוק	שבע רצון הוא אדם שחש סיפוק		מנוכר : זרות	מנוכר הוא אדם שחש זרות		3	0	1

Table 1: Examples from HeAna

Example	Model's output context sentences	Gold context sentences
1	אורווה : סייח - אורווה היא מקום בו מאחסנים סייח. לול : אפרוח - לול הוא מקום בו מאחסנים אפרוח.	אורווה : סייח - אורווה היא המבנה שבו מתגורר הסייח צאצא של הסוס. לול : אפרוח - לול הוא המבנה בו גר האפרוח צאצא של התרנגולת.
2	אינסופי : גבול - דבר מה סופי הוא דבר שאין לו גבול. מושלם : פגם - דבר מה מושלם הוא דבר שאין לו פגם.	אינסופי : גבול - דבר מה אינסופי הוא דבר שאין לו גבול מושלם : פגם - דבר מושלם הוא דבר שאין לו פגם

Table 2: Examples of gpt context sentences outputs

Split	Niqqud [#, %]	Phrases [#, %]
All	123, 22	87, 16
Train	108, 22	76, 15
Test	6, 21	6, 21
Validation	9, 32	5, 18

Table 3: Portion of Vowelized and phrases samples in the data

a = train, b = test/val	Lexical Overlap [$\frac{a \cap b}{b} * 100\%$]
train + test	54.73
train + validation	46.57

Table 4: Lexical unique words overlap between the dataset splits

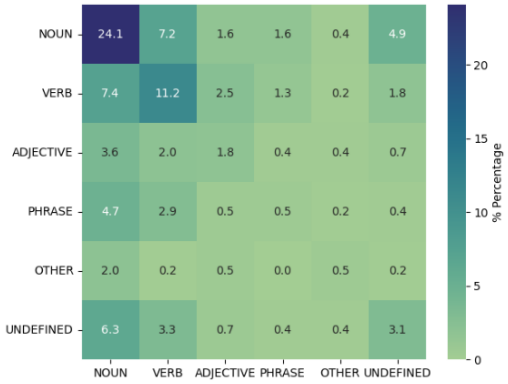

Figure 1: Distribution of the POS relation of word-pairs in the data. Rows represent the POS of the first word of the pair, columns represent the POS of the second word of the pair.

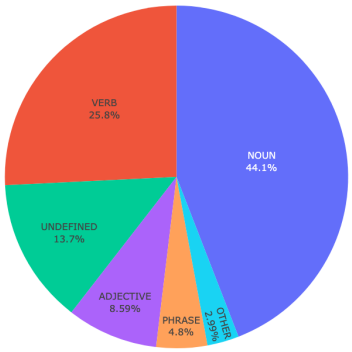

Figure 2: Distribution of the different part of speech of the data

input format	label
לדונ:דנ <sep> לקלף:קליפה; לאפות:מאפה; להקפיא:גלידה; לקטוף:פרי	4
לדונ:דנ <sep> 1 לקלף:קליפה 2 לאפות:מאפה 3 לקטוף:פרי 4 להקפיא:גלידה	4
לקטוף:פרי; לדונ:דנ	1
להקפיא:גלידה; לדונ:דנ	0

Table 5: Partial table of the input formats that were used in the classification task

Model	Accuracy [%]
AlephBertGimmel multiclass classification	32.33 ± 2.87
AlephBertGimmel semi-binary classification	**36.33 ± 2.05**
hebrew-gpt_neo-xl multiclass classification	20.64 ± 0.7
hebrew-gpt_neo-xl context sentences	27.33 ± 0.84
hebrew-gpt_neo-xl context sentences + additional relation finetuning	32.14 ± 0.87
Random guess	25

Table 6: Results- Mean and std calculate on 3 different training seeds.

References

Doron Adler. hebrew-gpt_neo-xl. https://huggingface.co/Norod78/hebrew-gpt_neo-xl.

Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based hmm for Hebrew morphological disambiguation. In *ACL*. The Association for Computer Linguistics.

Leo Gao, Stella Rose Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027.

Eylon Guetta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2022. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. *ArXiv*, abs/2211.15199.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, 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*.

NITE. The national institute for testing and evaluation, practice tests in hebrew. https://www.nite.org.il/psychometric-entrance-test/preparation/hebrew-practice-tests/?lang=en.

Nir Revach. Nir revach psychometric solutions, 2023. https://www.psychometry.co.il/nite-exams.php.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and José Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? *ArXiv*, abs/2105.04949.