Saved You A Click In Hebrew: Fine-tuning A multilingual LLM for generalized abstractive QA

Daria Lioubashevski
daria.lioubashevsky@mail.huji.ac.il

Yuval Shalev
Yuval.Shalev2@mail.huji.ac.il

Noam Dahan
Noam.dahan1@mail.huji.ac.il

1 Introduction

The term "clickbait" refers to a common practice of presenting a title of an article next to a link, with the sole purpose of luring the user into clicking and visiting the article's website. To achieve these, titles often intentionally leave out the most interesting part of the story to spark the curiosity of the reader. The main goal of this project is to train a LLM that can take a clickbait title and a news article body in Hebrew as input, and generate the missing information as output, "saving" the user a click.

This project has two main contributions: a new labeled real-world dataset in Hebrew and a fine-tuned version of a pre-trained multilingual LLM for a variant of question answering task. This is especially notable, since labeled datasets in Hebrew are scarce compared to English, as shown by Joshi et al. (2020). In 2021, Keren and Levy (2021) presented the first QA dataset in Hebrew, but it was extractive and used synthetic data. Moreover, we argue that collecting news domain data could be beneficial for advancing Hebrew capabilities as similar sets inspired advancements in the field of NLP in English (eg. Hermann et al. (2015)'s CNN/Daily Mail).

Besides the possible practical applications of this project, this task can be used to explore a more general ability of natural language processing: identifying missing information in a given sentence and retrieving it from a provided context. This ability can be referred to as a generalized question-answering because the input is not phrased as a question, but only implies one.

Johnson et al. (2022) presented a parallel task in English, and demonstrated improvement in performance by fine-tuning RoBERTa and T5 models on extractive and abstractive variants of the task respectively. Our work also builds upon a recent study in the Hebrew domain, where Eyal et al. (2022) showed that fine-tuning powerful, multilingual, pretrained sequence-to-sequence models such as mT5 on Hebrew data led to substantial improvements in the QA tasks.

2 Data (n = 2625 data points)

In order to achieve our goal, we have created a dataset based on the Facebook page "this.is.amlk". This page contains posts with three components: a clickbait title of a news article, a link to said article, and the text of the post itself. The post's text provides the missing information that is not included in the clickbait title. In keeping with the QA format, the "question" is the clickbait title, the "context" is the article body and the "answer" is the text of the post.

We scraped **9,392** posts from the Facebook page mentioned above. These posts contain links to various Israeli news websites, but we decided to focus on five: Israel Hayom, Mako, Walla News, Maariv and Ynet. This resulted in **4,374** data samples. To ensure high-quality clean data, we applied multiple automatic filters that removed data samples unsuitable for our purposes, such as posts that contained text unrelated to the article content or sarcastic comments that could confuse our model during training. The filters were based on manual analysis to identify patterns. After the filtering stage, our final dataset contains **n=2625** data samples. Fig. 1 presents the domain distribution. We split this dataset into training, validation, and test sets, at the ratio of approximately 8:1:1, respectively. The dataset is accessible through the Hugging Face platform here.

3 Methods

We finetuned mT5 (Xue et al. (2021)), a pre-trained sequence-to-sequence multilingual model, in variants small (300M), base (580M) and large (1.2B). We chose to focus on this model both because it was found to deal best with the rich morphology in Hebrew (Eyal et al. (2022)), and it is a variant of a model that achieved good results in the parallel task in English Johnson et al. (2022). We pre-processed our data to fit the text-to-text model by using the template "question: [clickbait title] context: [article body]" for the input.

In addition to the mT5 model, we also tested a

Hebrew text generation model based on Black et al. (2021), EleutherAI's gpt-neo, that was trained on Hebrew corpra. We fine-tuned the variants tiny, small and XL. We found that the model tends to quickly overfit to the train data, and even when experimenting with different combinations of hyperparameters (batch size, learning rate, etc.) and regularization methods (such as dropout and weight decay), we couldn't achieve relevant predictions. This may suggest that the model required more data, as due to the fact that it's a masked LM it essentially needs to learn the distribution of all the clickbait titles and news articles, as well as that of the posts' text, resulting in a much harder task.

The evaluation process included both automatic scoring methods and human evaluation. To compare the different models and hyper-parameter choices, we calculated BLEU (Papineni et al. (2002)), ROUGE (Lin (2004)) and BERTscore (Zhang et al. (2020)) on the validation set's predictions. The best model achieved the highest scores across the board. We included ROUGE since this metric is frequently used for evaluation of QA and summarization tasks. As in many cases the answers span only a few words, we used ROUGE-1 and ROUGE-2, and ROUGE-L that measures the longest common sub-sequence for overall similarity. As this method calculates exact overlap between words or n-grams, predictions may get a misleading low score even when they are semantically very close to the reference due to paraphrasing. Hence, we also used BERTscore which measures similarity between contextual word embeddings. As there is contradictory empirical evidence to support whether or not BERTscore has a higher correlation with human judgment than ROUGE-L scores specifically for QA tasks, as shown by Chen et al. (2019), we report both as well as human annotation.

4 Results

The model chosen was mT5-large trained on a context length of 300 tokens and in batches of 4. The results of the model on the test set are shown in Table 1. We achieved improvements in all metrics in comparison to the baseline: the same model before fine-tuning. Significantly, we report an improvement of 12 points in BERTscore F-score and 11 in ROUGE-L F-score. Examples of predictions are presented in Table 2. We note that the results achieved are lower than the baseline in the parallel English work, which might imply that there are challenges unique to Hebrew. Other reasons could be that mT5's abilities in Hebrew are not as strong as T5's abilities in English, or that the parallel English dataset is cleaner and easier to learn.

During the experimentation process, we concluded that the context length has the most impact on model performance. By using longer contexts, we managed to achieve similar results with the base model to those of the large model despite the latter having more than double the parameters. Interestingly, although up to a certain threshold the model performance improves monotonically as the context length increases, above it the performance actually starts decreasing (Fig. 2). We suggest that this is due to the fact that on the one hand the model needs a certain amount of information from the news articles, which is cut off by the shorter contexts, while on the other, contexts that are too long might include irrelevant details misleading the model.

Two annotators scored 50 test set samples by 1-5 according to a guide. The average was 2.72 (Cohen's kappa 0.23). Best predictions had over 0.73 BERTscore. Error analysis on 100 examples found two interesting categories: evaluation errors and question problems. In 8% of the worst performing samples BERTscore-wise, the model was correct but phrased the answers very differently than the reference, in 7% the title was not a clickbait at all (examples in Table 3 and Table 4 respectively). We also note that in some cases the model predicted a non-relevant country name which might suggest a spurious correlation, as many of the information missing from the titles concerns a location.

Further analysis on the results demonstrated that the model achieved better predictions on data from the news website "Israel Hayom", the main source for articles in our dataset (see Fig. 3). In future research it may be valuable to balance the data ahead of training to enable better generalization.

5 Conclusion

We created a Hebrew dataset for a variant of the abstractive QA task and used it to fine-tuned a multilingual LLM to answer clickbait titles in Hebrew. Our model achieved great improvements upon baseline, yet the quality of the results is still far from English benchmark. Ways to improve our results could be further cleaning the data and experimenting with larger models, possibly with quantization to bypass the resources barrier. All relevant code for this project is availible on GitHub.

	Bleu	ROUGE_1			ROUGE_2			ROUGE_L			BERTscore		
		P	R	F	P	R	F	P	R	F	P	R	F
mT5-large	2.19	4.57	7.84	4.91	1.08	2.39	1.37	4.51	7.80	4.87	54.58	63.11	58.36
fmT5-large	5.03	16.13	11.61	11.55	5.13	3.03	3.32	11.45	11.53	**16.00**	72.61	68.48	**70.32**

Table 1: The fine-tuned model *fmT5-large* results compare to the baseline *mT5-large*. *P* is Precision, *R* is the Recall, *F* is the F score.

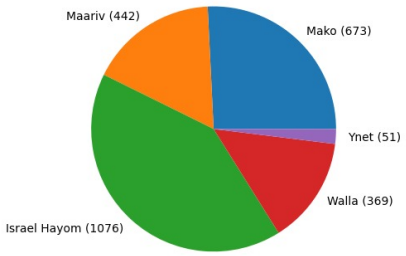

Figure 1: The distribution of articles after filtering

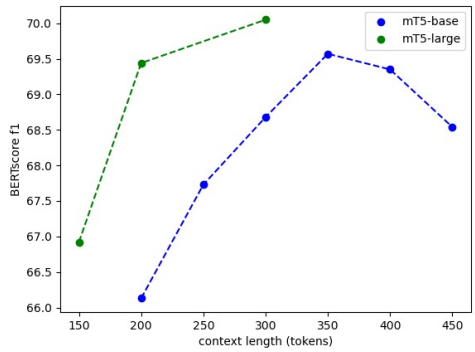

Figure 2: Increasing the context's size improved predictions on the validation set, until reaching a threshold

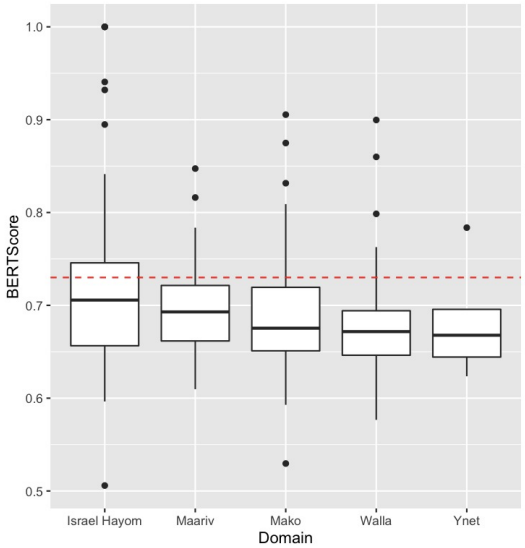

Figure 3: The distribution of the predictions' BERT scores grouped by domain. The red dashed line represents a threshold determined by human evaluation.

Clickbait	Reference	Prediction
הסיוט של כולנו חוזר לחיים	פאוץ	תיקי פאוץ
גאנס אנד רוזס: ההחלטה המפתיעה שהטריפה את הדתיים	הקדימו את תחילת המופע לשעה 20:00	מקדימים את ההופעה ל 20:00
מפתיע: לאחר 28 שנות קריירה של משחק זה מה שאקי אבני עושה היום	הפיק סרטון תדמית לפרוייקט דירות בתל אביב	סרט תדמית לפרויקט הדירות Port-TLV
גאווה ישראלית שמצילה חיים - כך ניצל הלוחם שנדקר בשער שכם	לבש מגן צוואר	עי מגן הצוואר

Table 2: Some of the models predictions

Clickbait	Reference	Prediction
אחרי המון ספקולציות מצדדים שונים - צהל טוען זו הסיבה האמיתית לפיצוצים העזים בשמי מרכז	בום על-קולי	מטיסות בחיל האוויר
האם החיסון יעיל נגד המוטציות מבריטניה ודרום אפריקה	כן על פי המחקר של פייזר	כן

Table 3: Correct answers that are hard to evaluate

Clickbait	Reference	Prediction
בן גביר חוזר בו ניצב אשד יישאר בתפקידו לפחות עד הרמדאן כל העדכונים	היועמשית עצרה את ההחלטה אז בינתיים הניצב בתפקיד	מטיסות בחיל האוויר
הכלה התחתנה בשמלת פיצה וחטפה מהגולשים	אנשים רוצים לאכול פיצה לא להסתכל עליה	היא כתבה בפוסט אחר

Table 4: Errors caused by the title not being a clickbait

References

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Matan Eyal, Hila Noga, Roee Aharoni, Idan Szpektor, and Reut Tsarfaty. 2022. Multilingual sequence-to-sequence models for hebrew nlp.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Oliver Johnson, Beicheng Lou, Janet Zhong, and Andrey Kurenkov. 2022. Saved you a click: Automatically answering clickbait titles.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Omri Keren and Omer Levy. 2021. Parashoot: A hebrew question answering dataset. *arXiv preprint arXiv:2109.11314*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.