ParaShoot: A Hebrew Question Answering Dataset

Omri Keren **Omer Levy**
The Blavatnik School of Computer Science
Tel Aviv University
omrikeren@mail.tau.ac.il

Abstract

NLP research in Hebrew has largely focused on morphology and syntax, where rich annotated datasets in the spirit of Universal Dependencies are available. Semantic datasets, however, are in short supply, hindering crucial advances in the development of NLP technology in Hebrew. In this work, we present ParaShoot, the first question answering dataset in modern Hebrew. The dataset follows the format and crowdsourcing methodology of SQuAD, and contains approximately 3000 annotated examples, similar to other question-answering datasets in low-resource languages. We provide the first baseline results using recently-released BERT-style models for Hebrew, showing that there is significant room for improvement on this task.

1 Introduction

Natural language processing has seen a surge in the pretraining paradigm in recent years with the appearance of pretrained models in a plethora of languages, including Hebrew (Chriqui and Yahav, 2021; Seker et al., 2021). While such models have shown to perform remarkably well on a variety of tasks, most of the evaluation of the Hebrew models, however, has been focused on morphology and syntax tasks in the spirit of universal dependencies (Nivre et al., 2017), while end-user-focused evaluation has been limited to sentiment analysis (Chriqui and Yahav, 2021) and named entity recognition (Bareket and Tsarfaty, 2020).

In this paper, we try to remedy the scarcity of semantic datasets by presenting ParaShoot,[1] the first question answering dataset in Hebrew, in the style of SQuAD (Rajpurkar et al., 2016). We follow similar work in constructing non-English question answering datasets (d'Hoffschmidt et al., 2020; Mozannar et al., 2019; Lim et al., 2019, *inter alia*), and turn to Hebrew-speaking crowdsource workers, asking them to write questions given paragraphs sampled at random from Hebrew Wikipedia. Through this process, we collect approximately 3000 annotated (*paragraph*, *question*, *answer*) triplets, in a setting that may be suitable for few-shot learning, simulating the amount of data a startup or academic group can quickly collect with a limited annotation budget or a short deadline.

Statistical analysis of ParaShoot shows that the dataset is diverse in question types and complexity, and that the annotations are of decent quality. We provide baseline results based on two recently-released BERT-style models in Hebrew, showing that there is much potential in devising better pretraining and fine-tuning schemes to improve the performance of Hebrew language models on this dataset. We hope that this new dataset will pave the way for practitioners and researchers to advance natural language understanding in Hebrew.[2]

2 Dataset

We present ParaShoot, a question answering dataset in Hebrew, in a format that closely follows that of SQuAD (Rajpurkar et al., 2016). Each example in the dataset is a triplet consisting of a paragraph, a question, and a span from the paragraph text constituting the answer to the question. We scrape paragraphs from random Hebrew Wikipedia articles, and crowdsource questions and answers for each one, resulting in 3038 annotated examples. While larger datasets may facilitate better-performing models, recent work has advocated for research on smaller labeled datasets (Ram et al., 2021), which more accurately reflect the amount of data a startup or academic lab can collect in a short amount of time and resources.

[1] A portmanteau of *paragraph* and ש"ות (*shoot*), the Hebrew abbreviation of Q&A.

[2] The dataset is publicly available at https://github.com/omrikeren/ParaShoot

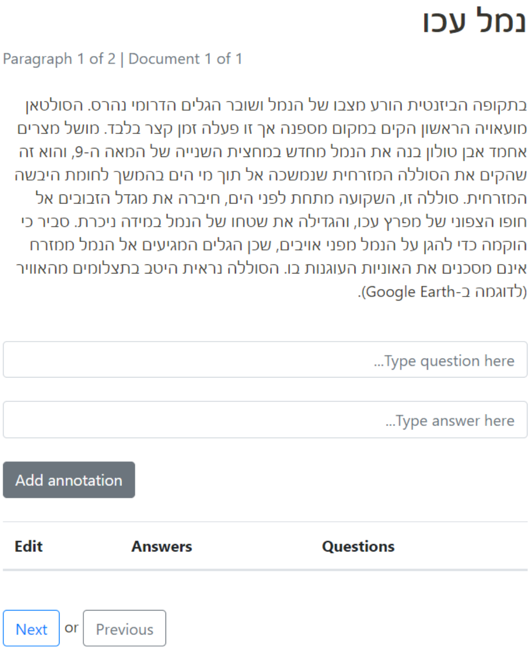

Figure 1: The annotation user interface, containing the article's title, the paragraph, a slot for entering a question, and an additional slot for entering the answer. Dragging the mouse over a span in the paragraph automatically fills the question slot, allowing for quick and accurate annotation of answer spans.

2.1 Corpus

We collect random articles from Hebrew Wikipedia, covering a wide range of domains and topics. We only sample articles containing at least two paragraphs and 500 characters.[3] Finally, for each such article, two candidate paragraphs are randomly sampled and added to the annotation corpus. These paragraphs will eventually become the passages in the question answering dataset.

2.2 Annotation

We recruit annotators by using the Prolific crowdsourcing platform.[4] Being a native Hebrew speaker is the only required qualification, allowing the participation of a few dozen annotators in the campaign. Annotators are presented with random paragraphs from the annotation set, and tasked to write 3-5 questions that are explicitly answered by the given text, for each paragraph. As in the original SQuAD annotation campaign, annotators are instructed to phrase the questions in their own words, and highlight the minimal span of characters from the paragraph that contains the answer to each question. Our implementation also provides automatic data validation heuristics that alert the annotators if, for instance, the answer span is too long or not a substring of the paragraph. Figure 1 shows a screenshot from the annotation web page.[5]

We acknowledge the fact that this data collection technique is known to encourage annotation artifacts (Gururangan et al., 2018; Kaushik and Lipton, 2018), and several newer annotation methods, such as TyDi QA (Clark et al., 2020), have been introduced to alleviate them. Nevertheless, we follow SQuAD's annotation methodology, as it necessitates considerably fewer resources. Maintaining an hourly wage of over $10,[6] we were able to collect our entire dataset, including discarded data from development runs, for under $800.

	#Articles	#Paragraphs	#Questions
Train	295	565	1792
Validation	33	63	221
Test	165	319	1025
Total	493	947	3038

Table 1: The number of unique articles, paragraphs, and questions in each split of PARASHOOT. The dataset is partitioned by articles.

2.3 Post-Processing

In total, we amass 3106 question-answer examples. Of those, we discard 68 examples (2.2%) that contained yes/no questions or extremely short/long answers. The resulting dataset contains 3038 examples, which we divide to training, validation, and test by article, preventing content overlap. Table 1 details the amount of unique articles, paragraphs, and questions of each split.

3 Analysis

We analyze the dataset in various ways to assess its quality and limitations as a benchmark.

3.1 Annotation Quality

To measure the quality of the annotated data, we randomly select 50 examples from the validation set, and manually analyze them ourselves.[7] Specifically, we check whether the annotated answer span is *correct* (answers the question) and *minimal* (contains only the answer). Table 2 shows that the

[3] We filter out images, tables, etc.
[4] www.prolific.co

[5] The platform's code is based upon https://github.com/cdqa-suite/cdQA-annotator.
[6] 7.50 GBP ≈ 10.50 USD, at the time of writing.
[7] The authors are native speakers of modern Hebrew.

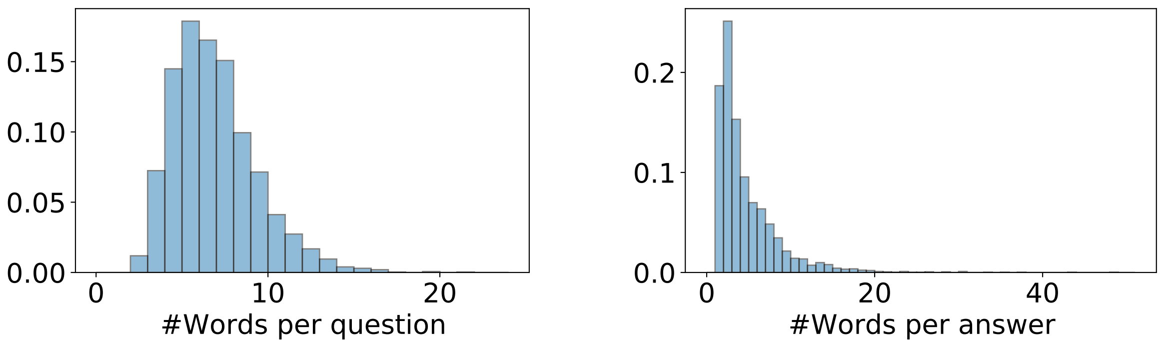

Figure 2: The length distribution of questions (left) and answers (right) in the entire dataset.

Answer Span	**Frequency**
Minimal	70%
Too Long	28%
Too Short	2%

Table 2: Distribution of annotated answer span quality, based on manual analysis of 50 examples from the validation set.

Question Word		**Frequency**
What	מה/מהו/...	16.29%
Which	איזה/איזו/...	15.84%
Who	מי/מיהו/...	14.03%
When	מתי/ממתי	13.57%
Where	איפה/היכן/...	10.86%
How	איך/כיצד	6.79%
How much/many	כמה/בכמה/...	5.43%
Why	למה/מדוע	4.52%

Table 3: Question word distribution, according to the first word of each question in the validation set. Inflected words and synonyms are clustered together to better align with English question types.

majority of the annotations are indeed valid, answering the questions with a minimal span. Yet, a significant minority contains additional supporting information, which makes the answer span longer than the desired minimal span by 2.5 times on average. We can thus expect an upper bound of 57% token F1 on those examples, setting the performance ceiling at around 84% F1 for the entire dataset. Finally, we present examples from the validation set that illustrate the annotation quality (Figure 3).

3.2 Question Diversity

To measure the dataset's diversity, we cluster questions by their question word (typically the first word in the question). Table 3 shows that *what* (מה) and *which* (איזה) questions account for a third of the sample, with other answer types being distributed in a rather balanced distribution. We also observe that about 11% of the data contains *how* (איך) and *why* (למה) questions, which may reflect more complex instances.

3.3 Sequence Length

We measure the length in words (using whitespace tokenization) of each question and each answer. Figure 2 shows the distributions of annotated questions and answers. We observe that most questions use between 4-7 words, which is typical of simple questions in Hebrew. More complicated questions constitute 27.6% of the data, for example: איך נקראת האופרה האחרונה שכתבו גילברט וסאליבן יחדיו? (*What is the last opera written jointly by Gilbert and Sullivan called?*) There are even questions with only 2 words; due to Hebrew's rich morphology, these questions are usually translated to 3-4 words in English, e.g. מהו המניכאיזם? (*What is Manichaeism?*) Answer lengths, however, can vary greatly, depending on whether the annotators wrote minimal spans (typically 1-4 words) or included supporting information in the answer spans (see Section 3.1).

3.4 Linguistic Phenomena

As a morphologically-rich language (Tsarfaty et al., 2010; Seddah et al., 2013), modern Hebrew exhibits a variety of non-trivial phenomena that are uncommon in English and could be challenging for NLP models (Tsarfaty et al., 2020). We can identify some of these phenomena in our dataset. Consider for example the following question-answer pair from the validation set:

Q: ?מה היה שטחו של כפר שמריהו כשהוקם

ma haya shitkho shel kfar shmaryahu
what was area-of-it of Kfar Shmaryahu

kshe-hukam
when-was.established

'What was Kfar Shmaryahu's area when it was established?'

A: ... הישוב הוקם על שטח של

ha-yeshuv hukam al shetakh
the-village was.established on area

shel ...
of ...

'The village was established on an area of ...'

This example illustrates a morphological variation between the question and the answer: the same entity appears as a morpheme in a compound word in the question's text: שטחו (*its area*), כשהוקם (*when it was established*), but as a standalone word (i.e. without inflection) in the answer: שטח (*area*), הוקם (*was established*). These phenomena make exact match-optimized predictions more difficult for models aimed to solve this task.

4 Baselines

We establish baseline results for PARASHOOT using BERT-style models. Results indicate the task is challenging, leaving much room for future work in Hebrew NLP to advance the state of the art.

4.1 Experiment Setup

We fine-tune three adaptations of BERT (Devlin et al., 2019): *mBERT*, trained by the original authors on a corpus consisting of the entire Wikipedia dumps of 100 languages; *HeBERT* (Chriqui and Yahav, 2021), trained on the OSCAR corpus (Ortiz Suárez et al., 2020) and Hebrew Wikipedia; *AlephBERT* (Seker et al., 2021), also trained on the OSCAR corpus, with an additional 71.5 million tweets in Hebrew. All models are equivalent in size to BERT-base, i.e. 12 layers, 768 model dimensions, and 110M parameters in total.

We fine-tune the models using the default implementation of HuggingFace Transformers (Wolf et al., 2020). We select the best model by validation set performance over the following hyperparameter grid: learning rate $\in \{3e{-}5, 5e{-}5, 1e{-}4\}$, batch size $\in \{16, 32, 64\}$, and update steps $\in \{512, 800, 1024\}$. We compare the models' predictions to the annotated answer using token-wise

Model	F1	EM
HeBERT	36.7	18.2
AlephBERT	49.6	26.0
mBERT	**56.1**	**32.0**

Table 4: Baseline performance on the test set.

F1 score and exact match (EM), as defined by Rajpurkar et al. (2016).

4.2 Results

Table 4 shows the performance of each model on PARASHOOT, with mBERT achieving the highest performance (56.1 F1). We also observe significant variance across the models, with mBERT and AlephBERT performing significantly better than HeBERT. It is not immediately clear where this discrepancy stems from; one possibility is that the introduction of noisy data via multilinguality (mBERT) or tweets (AlephBERT) makes that model more robust to potential noise in the annotated questions (e.g. typos). Comparing these results to the estimated ceiling performance of 84 F1 (see Section 3.1), we can infer that PARASHOOT poses a genuine challenge to future Hebrew models and encourages further analysis of the semantic capabilities of the current models.

4.3 Error Analysis

We analyze the error distribution by sampling 50 examples from the validation set and comparing AlephBERT's predictions to the annotated answers. Table 5 shows how the examples are distributed into five categories, accounting for every type of overlap between the model's prediction and the annotated answer. Putting aside exact matches (which account for about a quarter of examples), nearly half of the errors stem from zero overlap between the annotated answer and the model's prediction. We observe that a significant part of the sample (22%) contains cases where the annotated answer is a substring of the model's prediction, which might be, to a large extent, an artifact of the long answer annotations we observe in Section 3.1. For examples of erroneous predictions see Appendix A.

5 Conclusion

In this paper, we present PARASHOOT, the first question answering dataset in modern Hebrew, in a style and data collection methodology similar to that of SQuAD. Baseline results demonstrate the

Context: ... לאחר מכן **הוא מתארס עם חברתה הטובה שארלוט לוקאס** (קארן מורלי) ...
... He later becomes engaged to her best friend Charlotte Lucas (Karen Morely) ...
Question: למי מר קולינס מתארס?
To whom does Mr. Collins get engaged?

Context: ... מ**שנות ה-60 של המאה ה-20** התחילה תקופה של חפירות ארכאולוגיות בוואל קמוניקה שנמשכת ללא הפסקה ...
... From the 60s of the 20th century began a period of archeological excavations in Valcamonica that continues unabated ...
Question: מתי התחילה תקופה של חפירות ארכאולוגיות בוואל קמוניקה?
When did a period of archeological excavations begin in Valcamonica?

Figure 3: Examples from the validation set. The text in bold shows crowd-annotated answers. The underlined text represents the (expert-annotated) minimal answer span. The first example demonstrates a non-minimal span that has some overlap with the question's text. The second example demonstrates a valid minimal span selection.

Overlap Type	Sample Frequency	Error Frequency
Model = Annotation	26%	–
Model ⊂ Annotation	14%	19%
Model ⊃ Annotation	22%	30%
Model ∩ Annotation ≠ ∅	4%	5%
Model ∩ Annotation = ∅	34%	46%

Table 5: An error analysis of 50 random examples from the validation set, based on AlephBERT's predictions. The first reflects exact matches, and the last case accounts for zero overlap between model prediction and annotated answer. The three categories in the middle refer to partially correct answers, where the model's prediction has some overlap with the annotated answer.

potential of this dataset for researchers and practitioners alike to develop better models and datasets for natural language understanding in Hebrew.

Acknowledgements

This work was supported by the Tel Aviv University Data Science Center, Len Blavatnik and the Blavatnik Family foundation, the Alon Scholarship, Intel Corporation, and the Yandex Initiative for Machine Learning. We thank Reut Tsarfaty for her valuable feedback.

References

Dan Bareket and Reut Tsarfaty. 2020. Neural modeling for named entities and morphology (nemoˆ2). *arXiv preprint arXiv:2007.15620*.

Avihay Chriqui and Inbal Yahav. 2021. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3066–3079. Association for Computational Linguistics.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA. Association for Computational Linguistics.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12, Los Angeles, CA, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Error Examples

Context: ... לאחר התפטרות ניקסון ב-1974 בעקבות פרשת ווטרגייט, עבד צ'ייני בצוות שארגן את העברת הממשל לידי ג'רלד פורד ואחר כך המשיך בתפקיד עוזר בכיר לנשיא ...

... Following Nixon's resignation in 1974 following the Watergate affair, Cheney worked on a team that organized the transfer of administration to Gerald Ford and then continued as a senior assistant to the president ...

Question: מי היה נשיא ארצות הברית אחרי ניקסון?

Who was the president of the United States after Nixon?

Predicted Answer: צ'ייני (*Cheney*)

Context: תֵּל מִיכַל הוא אתר ארכאולוגי ובית גידול ים תיכוני הנמצא על רכס הכורכר החופי, מול מרינה הרצליה בדרום-מערבה של הרצליה ...

Tel Michal is an archeological site and Mediterranean habitat located on the coastal kurkar ridge, opposite the Herzliya Marina in the southwest of Herzliya ...

Question: איפה נמצא תל מיכל?

Where is Tel Michal located?

Predicted Answer: מול מרינה הרצליה (*opposite the Herzliya Marina*)

Context: ... בשנת 1895, בהיותו בן 17, עזב ואלזר את ביל ועבר לבזל. לאחר זמן קצר עבר לשטוטגרט שבגרמניה, עיר מגוריו של אחיו קרל ...

... In 1895, at the age of 17, Walser left Biel and moved to Basel. He soon moved to Stuttgart, Germany, the hometown of his brother Karl ...

Question: באיזה גיל ואלזר עבר לבאזל?

At what age did Walser move to Basel?

Predicted Answer: בהיותו בן 17, עזב ואלזר את ביל ועבר לבזל (*at the age of 17, Walser left Biel and moved to Basel*)

Figure A.1: Predictions made by fine-tuned AlephBERT vs. annotated answers. In the first example, the prediction produced by the model is clearly an error. In the second example, the annotated answer span is excessively long, and the model predicts a more accurate substring of this span. In the third example, the model predicts a full sentence, while the annotated answer span is shorter.