How to Translate SQuAD to German? A Comparative Study of Answer Span Retrieval Methods for Question Answering Dataset Creation

Jens Kaiser[1] and **Agnieszka Falenska**[1,2]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
firstname.lastname@ims.uni-stuttgart.de

Abstract

This paper investigates the effectiveness of automatic span retrieval methods for translating SQuAD to German through a comparative analysis across two scenarios. First, we assume no gold-standard target data and find that TAR, a method using an alignment model, results in the highest QA scores. Secondly, we switch to a scenario with a small target data and assess the impact of retrieval methods on fine-tuned models. Our results indicate that while fine-tuning generally enhances model performance, its effectiveness is dependent on the alignment of training and testing datasets.

1 Introduction

Extractive question answering (QA) is an NLP task in which a model receives a question and a context and needs to identify a context span that best answers this question. Figure 1 shows an example from a well-known extractive QA dataset **SQuAD** (the Stanford Question Answering Dataset, Rajpurkar et al. (2016, 2018)): For a given question, "What happened in 1971 and 1972?" the model should find the span of "two more launch failures" within the given context text.

To achieve high-performance in QA, one requires a robust training dataset with gold-standard annotations. However, such resources exist only for a few languages (Rogers et al., 2023). Therefore, to perform QA in a new language or domain, one must choose from: (1) manually curating a new dataset (d'Hoffschmidt et al., 2020; Heinrich et al., 2022; Efimov et al., 2020; Lim et al., 2019; Kazemi et al., 2022), (2) automatically translating a well-established dataset such as SQuAD into the target language (Mozannar et al., 2019; Kazi and Khoja, 2021; Vemula et al., 2022), or (3) using a hybrid approach and combining translation with a small manually annotated data (Möller et al., 2021). Given the varying costs associated with each option, it is crucial that researchers not only share their datasets but also insights learned during their creation, thereby aiding future similar initiatives.

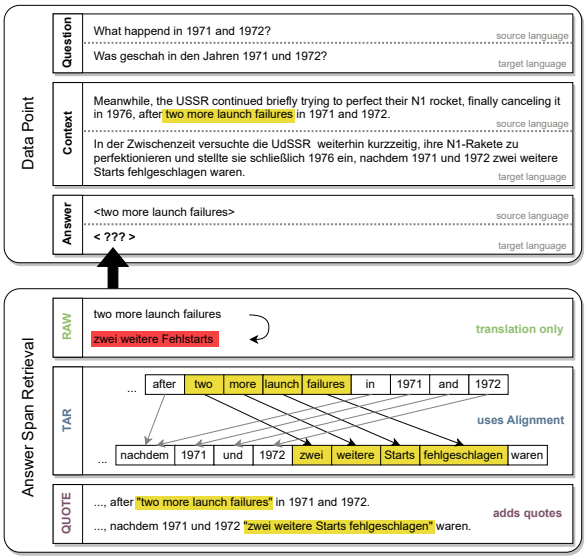

Figure 1: Context-based QA is the task of extracting answer span based on the question and context. The top depicts challenges in converting an English QA pair to German. The bottom shows different approaches for retrieving the answer span from the translated context.

For German, such valuable observations were provided by Möller et al. (2021). The authors not only introduced a new, manually annotated dataset, a state-of-the-art QA model, but also shared lessons learned during its creation, such as successful strategies for hybrid QA approaches and their generalization capabilities in out-of-dataset scenarios. However, the authors skipped a crucial aspect – the selection of the method for *answer span retrieval*. Translating SQuAD to a new language introduces challenges, such as answers that do not match the translated context. Figure 1 illustrates such a common issue. After translating the gold-standard English question and context pair to German, the translated answer "zwei weitere Fehlstarts" does not appear in the translated context anymore, making the datapoint unusable in the QA system. To deal with

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 134–140

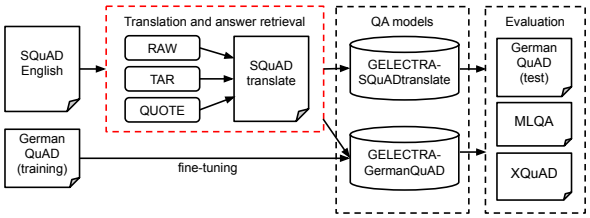

Figure 2: Experimental setup.

such issues, it is necessary to use additional answer retrieval methods (see the bottom of Figure 1 and the details of the methods in Section 2.2). However, there is a notable research gap regarding the comparative effectiveness of these heuristics and their influence on the German QA systems.

In this work, we aim to facilitate future approaches to QA dataset creation. Based on the premise that robust and high-quality training datasets lead to higher QA scores, we seek to answer two methodological research questions:

RQ1 Which answer span retrieval method yields the best-performing German SQuAD translation?

RQ2 Do span retrieval methods influence hybrid, fine-tuned models?

To address these questions, we replicate the experimental setup from Möller et al. (2021) using various answer retrieval methods (§2). We find that the effectiveness of these methods significantly depends on the type of existing data. In scenarios where only translated SQuAD is available, retrieving answers with an alignment model yields the best QA results (§3). However, for the hybrid QA models that additionally use small target data, the impact of span retrieval methods is dependent on the application and origin of the evaluation data (§4). While our analyses focus only on German, the results presented here can serve as guidelines for the future creation of SQuAD-based datasets in other languages.

2 Experimental Setup

Figure 2 illustrates the experimental setup from Möller et al. (2021) expanded by various answer retrieval methods (marked in red box). Below, we provide details of all the setup steps, beginning with an overview of the data used.

2.1 Data

Our experimental setup includes four different QA datasets (one English and three German). Based on the survey by Rogers et al. (2023), these are all SQuAD-like datasets that exist for German.

SQuAD 1.1 (Rajpurkar et al., 2016) is our source English QA dataset. It contains 107.785 QA pairs for 536 paragraphs taken from Wikipedia articles. For simplicity reasons, we use version 1.1 and not SQuAD 2.0 (Rajpurkar et al., 2018), which additionally includes over 50k unanswerable questions.[1] Moreover, we employ only the training part, with 87k QA pairs.

GermanQUAD is the German recreation of SQuAD from Möller et al. (2021). We use it for fine-tuning hybrid QA models and evaluation (see Figure 2). It comprises 13,722 manually created QA pairs. The original dataset comes only with training and test parts, so we leave out 20% of the training data as a development set.

XQuAD and MLQA (German parts) are used only for evaluation. XQuAD contains 1190 QA pairs from SQuAD translated by professionals to ten languages (Artetxe et al., 2020). MLQA (5027 pairs) was created from scratch following the SQuAD methodology (Lewis et al., 2020). Möller et al. (2021) call these two datasets out-of-domain for GermanQuAD. However, the main difference between them and GermanQuAD lies in the details of their creation, and not domains – all three resources are based on Wikipedia articles and the SQuAD framework. Therefore, we use the term *cross-dataset* to refer to the experiments where models are trained on GermanQuAD and applied to XQuAD and MLQA.

2.2 Translation and span retrieval

The first step in Figure 2 consists of translating SQuAD to German. Originally, Möller et al. (2021) used data translated with Facebook's commercial model (Lewis et al., 2020). We replace it with an open-source model called FAIRSEQ (Ott et al., 2019). Moreover, we differ the answer span retrieval method to one of the three approaches identified in the literature:

RAW simply filters out cases where the translated answer does not appear exactly once in the context.

TAR (Translate Align Retrieve) was introduced by Carrino et al. (2020) to translate SQuAD to Spanish. The method addresses the complex cases that

[1] Unanswerable questions have an empty answer span and are, therefore, exempt from the issue at hand.

	Dataset Size	GermanQUAD		MLQA		XQuAD	
		F1	EM	F1	EM	F1	EM
RAW	42.3k	65.3	51.2	63.1	47.7	76.5	60.8
TAR	83.3k	73.2	**55.5**	**66.9**	**50.9**	**77.9**	**62.5**
QUOTE	76.5k	**73.4**	51.3	66.8	47.6	77.7	56.1

Table 1: Performance of the QA systems trained only on the automatically translated SQuAD. The size of the datasets is measured in the number of individual QA pairs. The highest numbers in each column are in bold.

the RAW approach typically discards. It uses an alignment model to extract answer spans by mapping tokens between the source and target contexts (cf., Figure 1). We re-implement TAR with XML-Align (Chi et al., 2021), a better-performing aligner than the originally used efmaral (Östling and Tiedemann, 2016).[2]

QUOTE was first used by Lee et al. (2018) for translating SQuAD to Korean. The heuristic takes advantage of translation models frequently overlooking certain symbols, like quotation marks, and directly copying them to the outputs. It involves surrounding the answer span with such symbols before translation to then easily identify the corresponding span in the translated context. We tested three different symbols – *"*, *'*, and *()* – and found that FAIRSEQ preserves quotation marks the best.

2.3 QA Training and fine-tuning

As the next step from Figure 2, we implement two QA models following Möller et al.'s (2021) best-performing systems. They are based on GELECTRA large (Chan et al., 2020) and have two versions: SQuADtranslate, trained only on the translated data, and the hybrid model, fine-tuned on GermanQuAD (see hyperparameters in Appendix A).

2.4 Evaluation

We use two evaluation metrics: averaged F1 and exact match (EM) scores. F1 measures the similarity between the predicted and gold-standard answers, where the score is above zero as long as there is some word overlap between the two. EM, on the other hand, is a binary measure, giving 1 only if the predicted answer is equal to the gold-standard answer and 0 otherwise.

3 QA with No Target Data

We begin by addressing **RQ1** and evaluating which answer retrieval method gives the best QA results.

[2]All the developed code is publicly available at https://github.com/JensKaiser96/HowToTranslateSQuAD.

	GermanQUAD			
	F1	ΔF1	EM	ΔEM
RAW	65.3	–	51.2	–
$\text{TAR}_{\text{REDUCED}}$	70.1	-3.1	52.0	-3.5
$\text{QUOTE}_{\text{REDUCED}}$	72.7	-0.7	51.3	0.0

Table 2: Performance of the QA systems trained on 42.3k randomly selected, automatically translated SQuAD instances. Δs report losses from the data reduction (cf. Table 1).

3.1 Results

Table 1 shows the results for the three QA models using different answer retrieval methods. Firstly, we observe the influence of retrieval approaches on the training data size. With RAW, which excludes all data points where the translated answer does not appear exactly once in the translated context, roughly half of the training data is lost (training part of SQuAD has 87k pairs). In contrast, TAR allows for keeping almost 100% of the dataset. Finally, QUOTE preserves approximately 90% of the data, filtering out for example pairs where the translation did not keep the quotation marks.

Next, we move to the accuracy of the QA systems.[3] While the evaluation datasets clearly vary in difficulty, with MLQA being the most challenging, the relative performance of the models remains consistent across them. Interestingly, the two metrics – F1 and EM – prioritize different methods. Under F1, which allows for partial matches, RAW significantly underperforms compared to the other two methods, which achieve very similar results. In contrast, under the EM metric, TAR emerges as the clear leader, outperforming QUOTE by as much as 6.4 EM points on XQuAD.

[3]Differences to the results reported by Möller et al. (2021) most likely stem from the translation method and hyperparameters. However, since our goal is to observe differences between the models, we do not aim at SOTA performance.

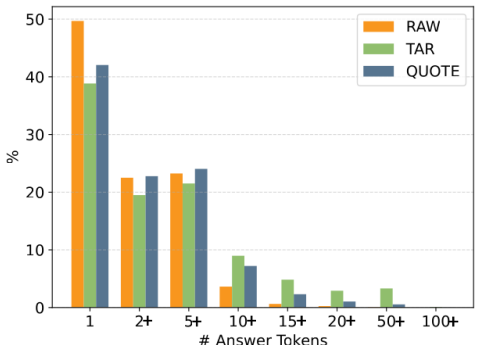

Figure 3: Percentages of answer lengths in the datasets.

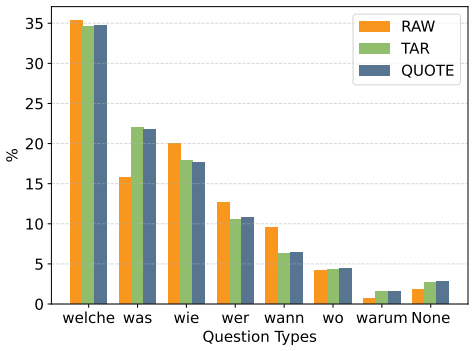

Figure 4: Percentages of question types in the datasets.

3.2 Analysis

So far, TAR resulted in the best QA system. However, it is unclear if its advantage stems solely from the larger dataset or the quality of the generated QA pairs. To analyze the influence of the dataset size on the model performance, we randomly subsample the TAR and QUOTE datasets to match the size of RAW (42.3k) and train two new reduced QA systems (see Table 2). As expected, the performance of both models decreases compared to Table 1 (reported in Δ columns). However, even with equivalent training sizes, they still exceed the performance of RAW. With dataset size ruled out as the only contributing factor, we analyze what other differences we can find.

Answer lengths The first factor that potentially varies within the datasets is the answer length. RAW, which keeps only QA pairs where the translated answer directly appears in the context, might be affected by the answer's length and perform better for typically short answers, such as numbers, dates, and names. Similarly, TAR might encounter more issues when setting the answer span at extreme context points following token mapping.[4] To evaluate this hypothesis, Figure 3 presents the distribution of answer length across all datasets.[5] For RAW, there is approximately 10% more single-token answers compared to TAR and QUOTE. Additionally, only about 6% of RAW's answers extend beyond five tokens, and none exceed 21. In contrast, TAR and QUOTE exhibit more similar distributions. TAR has fewer answers than QUOTE up to five tokens, but the situation reverses afterwards.

[4]Consider the example *Emma*[1] *bought*[2] *ice*[3] *at the new store in town* translated to German *Emma*[1] *hat*[2] *Eis*[3] *bei dem neuen Laden in der Stadt gekauft*[2] and with the retrieved span including all tokens in between the aligned words.

[5]Answers in-between are counted towards higher buckets.

Question types The observed variations in answer lengths may indirectly influence the distribution of question types. Typically, questions, such as *who* (*wer*) or *when* (*wann*) are associated with shorter answers, while *what* (*was*) or *why* (*warum*) require more elaborate responses. To test if this is the case in our datasets, we categorize questions based on their initial words and present results in Figure 4. We find that distributions for TAR and QUOTE are very similar. However, RAW exhibits a notably different pattern with fewer questions requiring complex answers, such as *what* (*was*) and *why* (*warum*) and more necessitating shorter responses, such as *who* (*wer*) and *when* (*wann*)

4 QA with Small Target Data

So far, we have assumed no gold-standard data in the target language. Now, we switch to **RQ2** and analyze the influence of span retrieval methods on the hybrid models. We prepare four versions of the GELECTRA-GermanQuAD model from Figure 2: **ONLY_FT**, which uses only GermanQuAD, and $\textbf{RAW}_{\textbf{FT}}$, $\textbf{TAR}_{\textbf{FT}}$, and $\textbf{QUOTE}_{\textbf{FT}}$, models that are first trained on translated SQuAD and then fine-tuned. Table 3 presents the results of all the models and their respective gains/losses from fine-tuning (i.e., differences to Table 1). For comparison, we also report **NO_FT** numbers – the highest results achieved by models that did not use fine-tuning (i.e., best results from Table 1). As all results span two distinct scenarios, we discuss each separately.

In-dataset evaluation When models are fine-tuned and evaluated with data from the same source – GermanQuAD – already ONLY_FT outperforms NO_FT, i.e., models with no additional training signals (see Table 3a). Further boosts can be observed from fine-tuning, which strongly reduces performance differences between FT approaches.

	GermanQUAD			
	F1	ΔF1	EM	ΔEM
NO_FT	73.4	–	55.5	–
ONLY_FT	77.5	–	63.0	–
RAW_{FT}	**84.1**	+18.8	**70.5**	+19.3
TAR_{FT}	82.2	+9.0	66.7	+11.2
$QUOTE_{FT}$	83.0	+9.6	68.4	+17.1

(a) In-dataset results

	MLQA				XQuAD			
	F1	ΔF1	EM	ΔEM	F1	ΔF1	EM	ΔEM
NO_FT	**66.9**	–	**50.9**	–	**77.9**	–	**62.5**	–
ONLY_FT	50.4	–	28.2	–	64.9	–	38.4	–
RAW_{FT}	60.2	-2.9	37.1	-10.6	71.4	-5.1	46.3	-14.5
TAR_{FT}	60.4	-6.5	35.2	-15.7	69.9	-8.0	42.3	-20.2
$QUOTE_{FT}$	62.2	-4.6	37.7	-9.9	71.3	-6.4	44.8	-11.3

(b) Cross-dataset results

Table 3: Performance of the fine-tuned QA systems; Δs reports gains/losses from fine-tuning (cf. Table 1).

Interestingly, their magnitude varies considerably among the models, ranging from 9 F1 points for TAR_{FT} to 18.8 points for RAW_{FT}. Surprisingly, RAW_{FT}, which previously was the weakest method, achieves the best results.

Cross-dataset evaluation Similarly to Möller et al. (2021), we find that fine-tuning in the cross-dataset setting degrades performance of the QA models (see Table 3b). ONLY_FT and all hybrid systems, irrespective of the answer span retrieval method, achieve significantly lower scores compared to the models trained only on the translated SQuAD. Interestingly, bigger drops in performance are observed for EM than for F1, suggesting that tuning leads to overfitting to the specific dataset characteristics.

5 Conclusion

In this paper, we explored best approaches to automatically translating SQuAD to German, highlighting the crucial role of the span retrieval methods in this process. We performed a comparative study of the three most-commonly used in the literature methods in two settings – with and without fine-tuning. Addressing **RQ1**, we found that when no fine-tuning is possible, TAR is the best practical choice, yielding more training data and higher (EM) or comparable (F1) results than QUOTE. RAW performs the worst – its strict filtering not only reduces the dataset size by half, but also skews question-answer distributions toward shorter queries about *who*, *when*, and *how*.

Responding to **RQ2**, the effectiveness of span retrieval methods varies when small target data is available. If this data comes from the same dataset as the evaluation set, automatically translated SQuAD is ideally used as a preliminary step before fine-tuning. In such cases, the differences between span retrieval methods are minor. However, if training data comes from a different origin, fine-tuning can lead to large drops in performance. In such cases, a well-translated, high-quality SQuAD dataset emerges as a more reliable source, again underscoring the importance of a carefully chosen method for the answer span retrieval.

6 Limitations

This work provides methodological insights into the creation of SQuAD-based datasets in German. Therefore, our experiments are limited to a single language. However, we believe that presented results, particularly the importance of careful selection of the answer span retrieval method, can be beneficial for researchers aiming to create new datasets also in other languages.

Secondly, we evaluate QA models using only three manually-curated datasets and fine-tune with just one. While a broader selection of datasets would enhance the generalizability of our results, to the best of our knowledge, we have used all the data that is currently available in German.

Finally, to ensure a fair comparison between approaches, the only variables we altered in the experimental setup were the span retrieval methods and the datasets used for training and fine-tuning. We did not experiment with other language models or QA systems. This decision was based on the findings of Möller et al. (2021), who evaluated various approaches and determined that models based on GELECTRA performed the best.

7 Acknowledgements

We acknowledge the support of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9-19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and

the support by the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. 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.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. Sberquad–russian reading comprehension dataset: Description and analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 3–15. Springer.

Quentin Heinrich, Gautier Viaud, and Wacim Belblidia. 2022. FQuAD2.0: French question answering and learning when you don't know. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2205–2214, Marseille, France. European Language Resources Association.

Arefeh Kazemi, Jamshid Mozafari, and Mohammad Ali Nematbakhsh. 2022. Persianquad: The native question answering dataset for the persian language. *IEEE Access*, 10:26045–26057.

Samreen Kazi and Shakeel Khoja. 2021. Uquad1. 0: Development of an urdu question answering training data for machine reading comprehension. *arXiv preprint arXiv:2111.01543*.

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. 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*.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. 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.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP

resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Rakesh Vemula, Mani Nuthi, and Manish Srivastava. 2022. TeQuAD:Telugu question answering dataset. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 300–307, New Delhi, India. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106.

A Appendix

A.1 Hyperparameters

We based the selection of the hyperparameters for training QA models and fine-tuning on two different sources. For training QA models, Möller et al. (2021) point to the default settings of a legacy framework which is no longer public. Therefore, we choose the parameters based on `https://huggingface.co/docs/transformers/tasks/question_answering` and `https://github.com/google-research/electra/blob/master/configure_finetuning.py` and used a batch size of 4, a learning rate of e-5, and 6 epochs. After each epoch, the model is evaluated using the development set, and the checkpoint with the lowest loss is saved.

For fine-tuning, we follow the recommendations from Möller et al. (2021): learning rate of 3e-5 and two epochs.