Spanning the Gap: Boosting Question-Answering by Standardizing Answer Spans

Njáll Skarphéðinsson, Eysteinn Örn Jónsson, Logi Sigurðarson, Hrafn Loftsson
Reykjavik University
Iceland
{njalls, eysteinnj19, logis21, hrafn}@ru.is

Abstract

The size of training datasets plays a pivotal role in enhancing the performance of Natural Language Processing models, particularly evident in the development of state-of-the-art models for English. However, acquiring large-scale datasets for many low-resource languages is a challenging task. In response, we investigate an alternative approach to boost the performance of Question-Answering (QA) models for low-resource languages, by focusing on data quality and annotator consistency. We standardize RUQuAD, an Icelandic span-prediction QA dataset comprising approximately 23,000 questions and 12,800 answers. The answer spans in RUQuAD are noisy due to its collection by a group of roughly 1,000 crowd workers and a lack of annotation guidelines. Moreover, we present results from fine-tuning a BERT model on both standardized and unstandardized RUQuAD versions. Our model trained on the standardized data achieves an F1-score of 79.5%, substantially outperforming the 62.5% F1-score obtained by the model trained on unstandardized data. Furthermore, we observe that training on a subset of 1,000 standardized examples surpasses the performance of training on the entire unstandardized dataset. These findings highlight the significance of data quality and consistency in enhancing QA model performance for low-resource languages.

Keywords: Question Answering, Answer Spans, Standardization

1. Introduction

The size of a dataset used to train a Natural Language Processing (NLP) model is seen an important factor for increasing model performance. Historically, the creation of large NLP datasets for English, which has been enabled by abundant language resources, has driven the field forward with new state-of-the-art models. The English Stanford Question-Answering (QA) datasets (Rajpurkar et al., 2016) is one such example.

In the last few years there has been an explosion in the construction and availability of QA datasets (Rogers et al., 2023). Even though the majority of research papers in the field of QA still focus on models trained on English datasets, QA for other languages is receiving increased attention. Recently, QA datasets have been published, for example, for French (d'Hoffschmidt et al., 2020), German (Möller et al., 2021), Norwegian (Ivanova et al., 2023), Estonian (Kuulmets and Fishel, 2023), Finnish (Kylliäinen and Yangarber, 2023), and Icelandic (Skarphedinsson et al., 2023).

However, gathering large QA datasets for many of the worlds languages can be a challenging endeavour due to lack of language resources (Clark et al., 2020; Skarphedinsson et al., 2023). As a result, we aim to explore another avenue for elevating the performance of QA models for low-resource languages, i.e. by focusing on data quality and annotator consistency.

For this purpose, we standardize the answer spans in RUQuAD, an Icelandic span-prediction QA dataset consisting of about 23,000 questions and 12,800 answers (Njall Skarphedinsson et. al, 2022). Annotations in RUQuAD are noisy due to two reasons. First, the data was collected by roughly 1,000 crowd workers (Skarphedinsson et al., 2023). Second, a lack of an enforced annotation standard when marking answer spans resulted in inconsistent answer spans and a high variance in answer span lengths. Our standardization involves creating an annotation standard for the answer spans for the purpose of removing the naturally occurring inconsistencies.

RUQuAD is suitable for assessing the impact of data quality because it is of considerable size and its natural irregularities in data annotations reflect real-world data collection artifacts. As a result, we can weigh the importance on gathering more training examples naturally compared to standardizing the existing data.

The main contribution of this paper is twofold. First, we provide information on the way RUQuAD was standardized and release the standardized version (Anonymous et. al, 2023). Second, we report the results on fine-tuning IceBERT (Snæbjarnarson et al., 2022) on both the standardized and unstandardized versions of RUQuAD. We ob-

This is a draft of a paper which will be submitted to the LREC-COLING 2024 conference.

tain an F1-score of 79.5% when evaluating a model trained on the standardized data, compared to 62.5% for a model trained on the unstandarized data. Moreover, we observe that training on 1,000 standardized examples yields a higher F1-score than training on the entire unstandardized dataset.
The rest of this paper is organized as follows. We discuss related work in Section 2, present the format of RUQuAD in Section 3, and standardization of answer spans in Section 4. We describe the training and evaluation method in Section 5 and present our results in Section 6. Finally, we conclude in Section 7.

2. Related Work

QA datasets often either contain *information-seeking* questions or *probing* questions (Rogers et al., 2023). In the former case, the dataset consists of questions for which the annotators did not now the answer to, whereas, in the latter case, the questions were written by annotators who already knew the correct answer.
The Stanford Question Answering Dataset (SQuAD) is the largest QA dataset for English, containing more than 100,000 questions. SQuAD was constructed by crowd workers who were presented with a paragraph from Wikipedia. Their task was to write up to five questions about the content of the paragraph and annotate the segment of text (a span) containing the answer for each question (Rajpurkar et al., 2016). The answer spans were not edited or standardized afterwards. Since the answers to the questions constructed by the annotators were known, SQuAD falls into the category of probing questions.
The TyDi QA dataset covers 11 typologically diverse languages and comprises 204,000 question-answer pairs. Human annotators were given short prompts consisting of the first 100 characters of Wikipedia articles. They were asked to generate questions about anything interesting that came to mind and that they did not now the answer to (and that were not answered by the prompt). The top-ranked Wikipedia article from a Google search, based on the question text, was then paired with the question. Finally, the question-article pair was presented to an annotator with the task of selecting a paragraph containing an answer and marking the minimal length answer span in the paragraph (Clark et al., 2020). As with SQuAD, the answer spans in TyDi were not edited or standardized afterwards. Since the answers to the questions asked by the annotators were not known, TyDi belongs to the information-seeking category.
The Reykjavik University Question-Answering Dataset (RUQuAD) contains information-seeking questions and answers in Icelandic. RUQuAD was constructed using GameQA, a gamified mobile app platform (Skarphedinsson et al., 2023), and comprises approximately 23,000 questions and 12,800 answers. The method used for constructing RuQuAD resembles the one used for constructing TyDi, i.e., the use of Google search, article pairing, paragraph selection, and answer span marking. The main difference is that when gathering data for RUQuAD, the crowd workers were playing a game, and the search for articles was carried out in multiple answer sources, not solely using a single answer source (Wikipedia) as was the case for TyDi. About 1,000 (unpaid) crowd workers participated in compiling RUQuAD.
To reduce noise in crowdsourced QA data, a validation mechanism is often included in the data collection process. Each question is commonly assigned to more than one worker to mark an answer to and another group of workers may then validate the answers (Trischler et al., 2017; Skarphedinsson et al., 2023). Zhu et al. (2022) propose a framework for automatic aggregation of different answers spans to the same questions, but we have not been able to find work in the literature on standarizing answer spans to different questions in existing QA data sets.

3. The RUQuAD Data

RUQuAD contains about 12,800 records of answered questions. Each record has a question, a paragraph, and an answer span. Thus, each record can be described with the following tuple:

$$X_j = (Q_j, P_j, i_j^{start}, i_j^{end})$$

Where X_j is the j-th record in the dataset, Q_j is the question, P_j is a paragraph containing the answer, i_j^{start} is the index of the first character in the answer span and i_j^{end} is the index of the last one.
During the gathering of RUQuAD, two crowd workers were asked to verify that every question was understandable. Similarly, two crowd workers were asked to verify that P_j contains the answer to Q_j. Moreover, when marking answers, the users were asked to select the minimum answer span (Skarphedinsson et al., 2023). However, with the data collection being open to anyone willing to contribute and it receiving contributions from about 1,000 crowd workers, it is not surprising that the variance in answer span lengths is high. The standard deviation is 91.52 characters, compared to 20.73 and 46.12 in SQuAD and TyDi, respectively.

4. Standardization

To the best of our knowledge, no open annotation standard exists for the creation of span prediction datasets in Icelandic. Drawing from the definition

put forward by Clark et al. (2020) when constructing TyDi, we defined the minimum answer span as the "minimal answer span that completely answers the question". In order to stay consistent, we defined rules as appropriate for common patterns we routinely observed in the answer paragraphs. These patterns are documented in our annotation standard, which is presented in Appendix A.

Here, we give the reader an example of a rule from the standard. For answers of the form "*X* er [staður] í *A* í/á *B*" ('*X* is a [place] in *A* in/on *B*'), mark the minimum length answer span (MLAS) as "A í/á B". For example, for the question "Hvar er borgin Arezzo?" ('Where is the city Arezzo') and the answer passage "Arezzo er borg í Toskanahéraði á Ítalíu og höfuðstaður samnefndrar sýslu." ('Arezzo is a city in county of Tuscany in Italy and the capital of the same county.'), the MLAS should be "Toskanahéraði á Ítalíu" ('County of Tuscany in Italy').

In order to compare the value of standardizing previously annotated records to gathering more records, it is useful to compare the cost per record, in terms of records per hour. The standardization was performed by three student researchers over the course of approximately 400 working hours. We standardized 10,000 questions (the other 2,800 were either yes/no questions or duplicates), resulting in 25 questions per hour.

As we do not have a precise estimate of the time it takes to manually elicit questions and answer them, we can turn to the literature for such an estimate. Natural Questions in Icelandic (NQiL) (Snæbjarnarson and Einarsson, 2022) is comparable to RUQuAD since it is an information-seeking span-prediction QA dataset. Five students worked over the span of three months to elicit and answer 5,568 records suitable for training. This comes out to 2.47 training records per hour, assuming 150 working hours per month.

There are other factors that impact both the cost of standardizing and creating new records. However, we take this as evidence for the claim that standardizing annotations is significantly cheaper and faster than creating new ones.

5. Training and Evaluation

While preparing the data for training, we took measures to avoid data and label leakage when creating the train-test split. This was necessary due to the fact that RUQuAD contains answers from multiple sources. First, we grouped the questions on the article which contains their answer and ensured that a question that is used for training does not belong to the same article as a question used for testing. Second, we ensured that there are no duplicate questions in the dataset by doing string comparison. Lastly, we removed all Yes/No questions, since we are training the model for span prediction.

In order to evaluate the impact of standardizing the dataset, we fine-tuned IceBERT (Snæbjarnarson et al., 2022), an Icelandic BERT model pre-trained on the Icelandic Gigaword Corpus (Barkarson et al., 2022). We evaluated models trained on varying number of training examples, both for the standardized and unstandardized data. We start by sampling $k = 1,000$ training records from the train split using a simple random sample (SRS). We then fine-tune IceBERT for $\mathcal{E}$ epochs (we chose $\mathcal{E} = 6$). At every epoch we calculate the F1-score for the test split by calculating the recall and precision for the tokens in all the test records. After training and evaluating the model, we repeat the same steps but increase our SRS sample size by 1,000 records. We do this while $k < N$ where N is the number of training records. This process allows us to see how the F1-score changes with increasing number of training examples. Algorithm 1 describes the training process.

Algorithm 1 Training and Evaluation

Inputs: Training data $\mathcal{D}_{train} = \{\mathcal{X}_j, \mathcal{Y}_j\}_{j=1}^{N}$, test data $\mathcal{D}_{test} = \{\mathcal{X}_j, \mathcal{Y}_j\}_{h=1}^{M}$ where N and M are the sizes of the training and test splits, respectively, $\mathcal{E}$ is the number of epochs and $\mathcal{H}$ the pre-trained model.

Initialize a mapping $V(k) = 0, \forall k$.
$k \leftarrow 1000$
while $k < N$ **do**
    initialize model $\mathcal{H}$
    $\mathcal{X}_k, \mathcal{Y}_k \leftarrow sample(\mathcal{D}_{train}, k)$
    $\epsilon \leftarrow 0$
    while $\epsilon < \mathcal{E}$ **do**
        $\hat{\mathcal{Y}}_k \leftarrow \mathcal{H}(\mathcal{X}_k)$
        $loss \leftarrow \mathcal{L}(\hat{\mathcal{Y}}_k, \mathcal{Y}_k)$
        back-propagate loss and update $\mathcal{H}$
        $v \leftarrow eval(\mathcal{H}, \mathcal{D}_{test})$
        $V(k) \leftarrow max(V(k), v)$
        $\epsilon \leftarrow \epsilon + 1$
    end while
    $k \leftarrow k + 1000$
end while
Output V

This algorithm is performed twice. Once for the unstandardized data and again for the standardized data. Running training this way takes $O(N^2)$ time.

6. Results

After fine-tuning IceBERT on both the standardized and unstandardized data, we observed a considerable improvement after standardizing (see

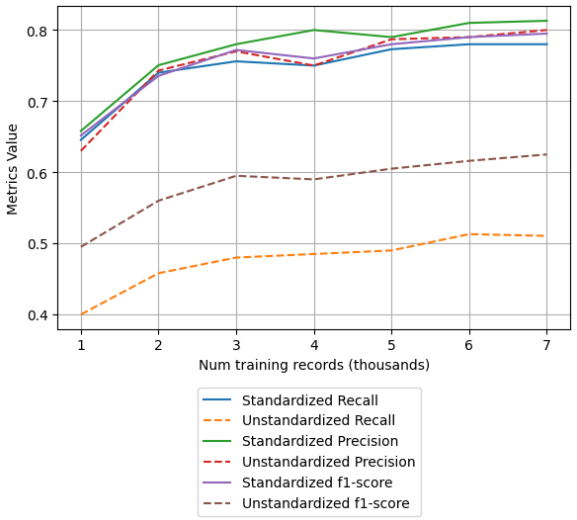

Figure 1: The figure shows how the precision, recall, and F1-score changes with varying sizes of training data, for both standardized and unstandarized versions of RUQuAD.

Figure 1). The highest F1-score is 79.5% for the standardized data, compared to 62.5% for the noisy, unstandardized data. An interesting observation is that the F1-score obtained by a model trained on 1,000 standardized records outperforms a model trained on the entire unstandardized training set.

Interestingly, noisy answer span annotations seem to only impact recall while the precision is not impacted in any noticeable way. A model trained on standardized answer spans learns from answers which in general contain fewer tokens than answers in unstandardized answer spans. Therefore such a model is able to predict with more accuracy the tokens that should be in the answer span, thus resulting in higher recall. In contrast, a model trained on unstandardized answer spans can obtain high precision, because the tokens it predicts are most often part of the answer span, whereas the recall is low because the model often does not predict all the correct tokens in the span. We hypothesise that the training objective incentivizes the model to be conservative when classifying tokens in the answer passage.

We observe that the validation metrics improve with more data as is expected. However, it is not clear how much more unstandardized data would be necessary to even match 2,000 standardized training records.

7. Conclusion

In machine learning research, we often hear the two old adages saying *garbage in, garbage out*, and *quality over quantity*. In this paper, we set out to explore the impact of data quality and annotation consistency on the performance of QA models, using RUQuAD as our testbed. Our results show that standardizing answer spans significantly improves model performance. The results further suggest that consistent annotations can even outperform larger datasets, which is particularly important for low-resource languages.

The importance of data quality is evident from the fact that even with a much smaller training dataset of 1,000 records of standardized data, a fine-tuned BERT model outperformed one that was fine-tuned on over 7,000 unstandardized answer spans.

While it is common to seek more data and larger datasets, our results demonstrate that standardizing data and remove inconsistencies that are often natural artifacts of data collection can sometimes yield more fruitful results. This is especially important for low-resource languages where more data might not be an available option.

Going forward, it would be interesting to explore how general these findings are across other NLP tasks and datasets.

8. Acknowledgements

This work described in this paper was supported by the Icelandic Student Innovation Fund, https://en.rannis.is/funding/research/icelandic-student-innovation-fund/.

9. Bibliographical References

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. Evolving large text corpora: Four versions of the Icelandic Gigaword corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

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.

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.

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023.

NorQuAD: Norwegian question answering dataset. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.

Hele-Andra Kuulmets and Mark Fishel. 2023. Translated benchmarks can be misleading: the case of Estonian question answering. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 710–716, Tórshavn, Faroe Islands. University of Tartu Library.

Ilmari Kylliäinen and Roman Yangarber. 2023. Question answering and question generation for Finnish. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 529–540, Tórshavn, Faroe Islands. University of Tartu Library.

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.

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 Comput. Surv.*, 55(10).

Njall Skarphedinsson, Breki Gudmundsson, Steinar Smari, Marta Kristin Larusdottir, Hafsteinn Einarsson, Abuzar Khan, Eric Nyberg, and Hrafn Loftsson. 2023. GameQA: Gamified mobile app platform for building multiple-domain question-answering datasets. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 152–160, Dubrovnik, Croatia. Association for Computational Linguistics.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022. Natural questions in Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4488–4496, Marseille, France. European Language Resources Association.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Peide Zhu, Zhen Wang, Claudia Hauff, Jie Yang, and Avishek Anand. 2022. Answer quality aware aggregation for extractive QA crowdsourcing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6147–6159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

10. Language Resource References

Anonymous et. al. 2023. *Reykjavik University Question-Answering Dataset (RUQuAD) 29.09*. Distributed via CLARIN: http://somepath.

Njall Skarphedinsson et. al. 2022. *Reykjavik University Question-Answering Dataset (RUQuAD) 22.02*. Distributed via CLARIN: http://hdl.handle.net/20.500.12537/310.

A. Annotation Standard

When standardizing the answer spans in RUQuAD, the main criterion was to mark the minimal length answer span (MLAS) that completely answers to associated question. We used the following main rules:

X is in [place] in/on [place]
For answers of the form "*X* er [staður] í *A* í/á *B*" ('*X* is a [place] in *A* in/on *B*'), mark the MLAS as "A í/á B". For example, for the question "Hvar er borgin Arezzo?" ('Where is the city Arezzo') and the answer passage "Arezzo er borg í Toskanahéraði á Ítalíu og höfuðstaður samnefndrar sýslu." ('Arezzo is a city in county of Tuscany in Italy and the capital of the same county.'), the MLAS should be "Toskanahéraði á Ítalíu" ('County of Tuscany in

Italy').

X is [adjectives] in A B

For answers of the form "*X* er [lýsingarorð] í *A B*" ('*X* is a [adjectives] in *A B*'), mark the MLAS as "A B". For example, for the question "Hvar er Tjadvatn?" ('Where is Lake Chad?') and the answer passage "Tjadvatn er stórt, grunnt stöðuvatn í miðri Afríku.' ('Lake Chad is a large, shallow lake in central Africa.'), the MLAS should be "í miðri Afríku" ('In central Africa').

X is A which is B

For answers of the form "*X* er *A* sem er *B*" ('*X* is *A* which is *B*'), where B is a subset of A or a more accurate description, mark the MLAS as "A". For example, for the question "Hvað eru sólstafir?" ('What are crepuscular rays?') and the answer passage "Sólstafir er veðurfyrirbrigði sem gerist þegar sólarljós skín gegnum rof í skýjum eða fjallaskörð." ('Crepuscular rays is a weather phenomena that happens when sun rays shine through a hole in the clouds or through mountain passes.'), the MLAS would be "Veðurfyrirbrigði" ('Weather phenomena').

Period of time

Include the word "Árið" ('The year') in the MLAS if the question is referring to some particular year, but not when the answer is a range of years. For example, for the question "Hvenær var Decode Genetics stofnað" ('When was Decode Genetics founded?'), the MLAS should be "Árið 1996" ('The year 1996'). In contrast, for the question "Hvenær var seinni heimsstyrjöldin?" ('When was the second world war?'), the MLAS sould be "1939–1945", but not "árin 1939–1945".

Metrics and amounts.

If the answer includes some metrics, the metric should always be included in the MLAS. For example, for the question "Hversu margir metrar eru í einni mílu" ('How many meters are in one mile'), the MLAS should be "1,609,344 metrar". Nouns following numbers/amounts should not be part of the MLAS. For example, for the question "Hvað eru mörg bein í líkamanum" ('How many bones are in the body?'), the MAS should just be "206", but not "206 bein".

Subjects

Do not include the subject in the MLAS. For example, for the question "Hvaða hlutverki þjóna nýrun?" ('What is the purpose of the kidneys?') and the answer passage "Nýrun stýra efnasamsetningu blóðs, rúmmáli og fjarlægja úrgangsefni úr því." ('The kidneys regulate blood chemical composition, volume and remove waste products from it'), the MLAS should not include parts of the answer that directly refers to the subject such as "Þau" ('They') or "Nýrun" ('The kidneys'). Thus, the MLAS would be "stýra efnasamsetningu blóðs, rúmmáli og fjarlægja úrgangsefni úr því."

Punctuation

Don not include punctuation in the MLAS if it is in the end of the answer. For example, for the question "Hver er höfuðborg Íslands?" ('What is the capital of Iceland?') and the answer passage "Höfuðborg Íslands er Reykjavík." ('The capital of Iceland is Reykjavík.'), the MLAS should be "Reykjavík", but not "Reykjavík.".

Reykjavík, 25. september, 2023

Njáll Skarphéðinsson

Njáll Skarphéðinsson

Eysteinn Örn J.

Eysteinn Örn Jónsson

Logi Sigurðarson

Logi Sigurðarson

Hrafn Loftsson

Hrafn Loftsson