Cross-Lingual QA as a Stepping Stone for Monolingual Open QA in Icelandic

Vésteinn Snæbjarnarson
Miðeind ehf
University of Iceland
vesteinn@mideind.is

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Abstract

It can be challenging to build effective open question answering (open QA) systems for languages other than English, mainly due to a lack of labeled data for training. We present a data efficient method to bootstrap such a system for languages other than English. Our approach requires only limited QA resources in the given language, along with machine-translated data, and at least a bilingual language model. To evaluate our approach, we build such a system for the Icelandic language and evaluate performance over trivia style datasets. The corpora used for training are English in origin but machine translated into Icelandic. We train a bilingual Icelandic/English language model to embed English context and Icelandic questions following methodology introduced with DensePhrases (Lee et al., 2021). The resulting system is an open domain cross-lingual QA system between Icelandic and English. Finally, the system is adapted for Icelandic only open QA, demonstrating how it is possible to efficiently create an open QA system with limited access to curated datasets in the language of interest.

1 Introduction

Open QA systems are question-answering systems that suggest answers to questions by searching through a text corpus. Such systems have improved significantly in recent years, which can, to a large extent, be attributed to transformer-based vector representations of text that are well suited for the task (Vaswani et al., 2017). The most successful systems have been trained with a focus on English using large datasets such as Natural Questions (Kwiatkowski et al., 2019) (>320k questions), and SQuAD (Okazawa, 2021) (>150k questions). In some cases, questions have been generated from text using large generative neural networks (Alberti et al., 2019). For most languages, such large datasets do not exist, and the generative models do not perform as well as for English which constitutes

the bulk of the training data. For this reason, we investigate what performance can be reached in QA for Icelandic, a language with low QA resources. In that investigation, we study the question of whether English QA data can aid QA system development through the use of machine translation.

In this paper, we present a method to bootstrap an Open QA system for Icelandic where just a few thousand labelled data entries are available. We adapt the DensePhrases (Lee et al., 2021) method by applying a bilingual language model, and machine-translated data, in a cross-lingual manner to create a monolingual Open QA system for Icelandic, the first of its kind built exclusively for the language. An overview of the build process is shown step by step in Figure 1.

2 Related work

2.1 Reading comprehension and Open QA

Open-domain question answering methods look for answers to a given question in a given text corpus (for a recent survey, see (Zhu et al., 2021)). These methods can be contrasted with reading comprehension (RC) style methods that identify an answer to a question within a single document. The RC methods are useful when an answer is sought in a given text, often referred to as the context. Open QA methods are open in the sense that the questions they can handle are open ended given a large enough underlying corpus. Open QA can be thought of as a generalization of reading comprehension since the answer is typically retrieved from a large collection of text instead of a single document. We note that most open QA methods are extractive, meaning that the suggested answer is found verbatim within a given document. There are also QA methods that provide an answer without explicitly searching through a corpus. For example, the answer can be generated based on knowledge embedded in learned parameters of a system such

as GPT-3 (Brown et al., 2020). While promising, the non-extractive methods are not considered in this paper.

Open QA methods solve a common issue in information retrieval where it is not known in what document an answer lies. The simpler reading comprehension methods can be used as components in open QA systems by combining them with a *retriever* component. BM25 (Robertson et al., 1995), a TF-IDF variant, is an example of a commonly used retriever that ranks context based on term frequencies that are shared with the question and their overall commonality. The top documents found by the retriever can then be fed to the reading comprehension component along with the question. The reading comprehension component can, for example, be a fine-tuned variant of a neural language model such as BERT (Devlin et al., 2019). The reading comprehension component is trained to predict the start and end location of an answer span or report whether an answer is not found within the given context by training on a dataset of context and question pairs.

2.2 Fast retrieval and DensePhrases

In recent years, efforts in improving open QA have focused on speeding up the lookup of documents, for example, by taking advantage of neural methods. Such a speedup has been realized by embedding documents and questions as dense vector representations such that lookup can be based on fast similarity search where the inner product of the question vector and document vector is used as a proxy for their similarity (Karpukhin et al., 2020; Lee et al., 2019; Lin et al., 2021). The embedding function can be trained such that a given question will, with a good chance, lead to the correct document being the highest ranked in the similarity search. The embedding function can also be trained end-to-end by basing the loss function on the performance of looking up the answer. A downside of these methods, in particular the end-to-end systems, is that they can be expensive to train since the document embeddings need to be updated often as a result of updates to the embedding function (Guu et al., 2020), which can be particularly expensive when many documents need to be embedded repeatedly throughout the training process. Some mitigations have been suggested; as is the case in DensePhrases (Lee et al., 2021), which is the foundation of our approach.

In DensePhrases, segments from documents are first embedded using a phrase model (and fixed), then a query model is trained to embed questions such that the inner product of question embeddings and correct context embeddings are maximized. For an incorrect pairing, the model is trained such that the inner product is minimized instead.

Fast databases intended for lookup with maximum inner product search (MIPS) (Johnson et al., 2019) enable systems such as DensePhrases to provide answers from massive datasets in subsecond time, making them excellent candidates for production-grade QA systems where an answer and its source can be reported.

2.3 Multilingual and cross-lingual QA

In cross-lingual QA, the question and answer are not required to be in the same language, and in multilingual QA the aim is to search for answers in a multilingual corpus. Multilingual QA is not necessarily cross-lingual since the answer can be generated in the same language as the query.

Interest in cross-lingual QA is likely reflected in the growing number of QA datasets in foreign languages (Rogers et al., 2021). For reading comprehension, it has been shown that multilingual LMs such as mBERT fine-tuned in an English reading comprehension task are capable of zero-shot transfer to other languages such as Japanese, French, and Hindi (Siblini et al., 2021; Gupta and Khade, 2020). Multilingual QA has been performed by extending models for English by using machine translation (MT) on the query and answer (Asai et al., 2021a), MT has also been used to adapt an English semantic parsing model for other languages (Sherborne et al., 2020; Moradshahi et al., 2020). Multilingual QA was recently implemented without explicit use of MT by extending the Dense Passage Retriever model from Karpukhin et al. (2020) with a fine-tuned mT5 model as an answer generator (Asai et al., 2021b). The answer generator receives top-scoring multilingual passages along with the question and desired answer language to generate the answer. This flexible approach even generalizes to languages not seen in the QA training process thanks to the diverse training set for crosslingual retrieval. A similar approach with an answer generator has also been applied where passage candidates come from different monolingual corpora, and the question is translated and embedded with several monolingual language mod-

els (Muller et al., 2021).

2.4 Icelandic QA data

Currently, a single extractive dataset exists for Icelandic, NQiI (Snæbjarnarson and Einarsson, 2022). It is a small Icelandic dataset containing only ∼5k question-context pairs, half of which have no answer. The dataset is sourced from the Icelandic Wikipedia following the methodology introduced in TyDi-QA (Clark et al., 2020). This limited amount of Icelandic QA data is the main reason we translate English QA datasets.

3 Methods

3.1 Translating QA data

In the first step of the process, we use an English-Icelandic translation system (Símonarson et al., 2021) for translating NewsQA (Trischler et al., 2017), SQuAD and Natural Questions (NQ) (Kwiatkowski et al., 2019). We reviewed the translated questions from SQuAD and out of 100 randomly sampled questions we found that 80 were properly translated such that the meaning was fully preserved.

We translate questions, answers and contexts independently and use a fuzzy matching algorithm (see Appendix A) to map translated answers to spans in the translated context. We refer to the fully translated versions of the datasets as NewsQA-IS, SQuAD-IS, and NQ-IS. For the translated versions of the datasets where only the questions are answered as we use NewsQA-ISQ, SQuAD-ISQ, and NQ-ISQ (for an overview, see Table 1).

In DensePhrases, questions are generated for all spans of length 0–20 words in the English Wikipedia using a fine-tuned T5 (Raffel et al., 2020) model. As no such model currently exist that can reliably generate Icelandic, we also translate the generated questions. The spans themselves can not be easily translated as the available models are mostly good at translating well-formed sentences. We refer to this dataset as DP-ISQ. For an overview of all QA datasets used see Table 1.

3.2 Pre-training an Icelandic–English language model

A bilingual language model for Icelandic and English was trained following the base XLM-RoBERTa implementation (Conneau et al., 2020). We refer to this model as LM EN-IS. The Icelandic training data is the same as the one used for Ice-BERT (Snæbjarnarson et al., 2022). The Books 3 corpus[1] is used as source for English data, it contains around 100GB of data text from a variety of books. The model was trained for 220k updates using a batch size of 8k completing 27 epochs over the data.

3.2.1 Training RC models

After pre-training, the bilingual model (LM EN-IS) is fine-tuned for cross-lingual RC where questions are asked in Icelandic and answered in English (step 3 in Figure 1). We fine-tune using SQuAD-ISQ, NewsQA-ISQ, and NQ-ISQ. We refer to this model as the IS-EN RC model.

The bilingual model (LM EN-IS) is also fine-tuned for an Icelandic only reading comprehension task (step 4 in Figure 1) using the fully translated datasets, NQ-IS, SQuAD-IS and NewsQA-IS along with NQiI. We refer to this model as the IS-IS RC model.

These RC models are later used as a teacher models (Hinton et al., 2015). The IS-EN RC model is distilled in the fifth step and the IS-IS RC model in the sixth step of the build process when fine-tuning the Open QA system. Note that to be compatible with the training of the DensePhrases model, these models do not predict missing answers.

3.2.2 Training cross-lingual DensePhrases

We also fine-tune the bilingual model (LM EN-IS) to train a DensePhrases setup [2]. We use the partially translated DP-ISQ dataset to train the cross-lingual DensePhrases model. The result is a phrase encoder that accepts English and a query encoder that accepts Icelandic. Following the DensePhrases approach, we distil the IS-EN RC model at training time. This distillation step can be beneficial since the comparison in the DensePhrases setup is based on an inner product operation, whereas the RC model was trained in a cross-attention setting. This distillation step improved the EM score by 2 points for the original DensePhrases paper and could be validated through ablation in our low-resource setting as well. We refer to the crosslingual DensePhrases model as DensePhrases-IS-EN

---

[1]This is similar to (Kobayashi, 2018) and was made available in the issue section of the GitHub repository https://github.com/soskek/bookcorpus/issues/27.

[2]With minor adjustments to work with the SentencePiece (Kudo and Richardson, 2018) tokenization used by the bilingual model

Data

Icelandic & English

English QA data

Translated IS-Q, EN-A

Translated IS-Q, IS-A

Icelandic QA data

1. Translate QA data

English QA data → EN→IS translation → Translated IS-Q, EN-A, Translated IS-Q, IS-A

2. Pretrain a bilingual language model

Icelandic & English → Pretraining → Language Model (LM) EN-IS

3. RC model with IS Q and EN A

LM EN-IS → Fine-tuning → IS-EN RC model

4. RC model with IS Q and IS A

LM EN-IS → Fine-tuning → IS-IS RC model

5. DP model with IS Q and EN A

LM EN-IS → Fine-tuning → DensePhrases IS-EN (English Phrase Encoder, Icelandic Query Encoder)

IS-EN RC model → Distillation

6. DP model with IS Q and IS A

DensePhrases IS-EN (English Phrase Encoder, Icelandic Query Encoder) → Fine-tuning → DensePhrases IS-IS (Icelandic Phrase Encoder, Icelandic Query Encoder)
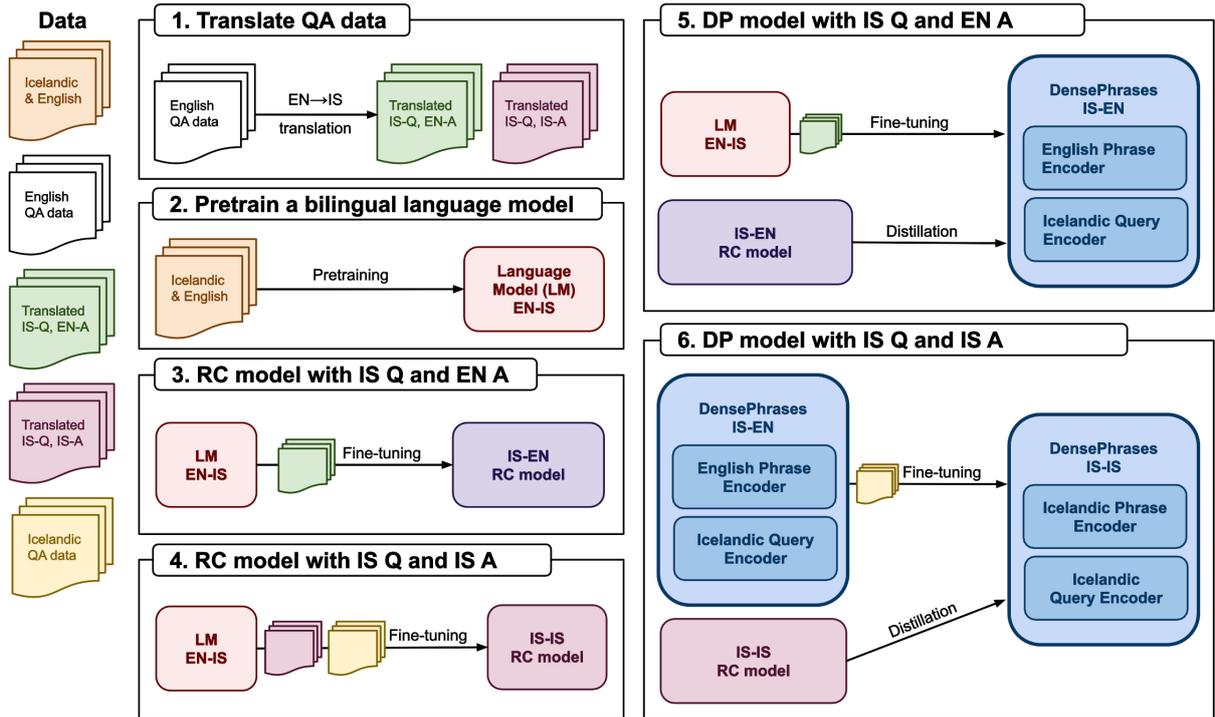
IS-IS RC model → Distillation

Figure 1: Diagram showing how to train an Icelandic DensePhrases model in six steps.

3.2.3 Training Icelandic only DensePhrases

In the last step of our process, we take the cross-lingual model DensePhrases-IS-EN and fine-tune it on NQiI to develop a fully Icelandic Open QA system. In this final step, we also distil the IS-IS RC model from the fourth step of the build process. We refer to the final Icelandic only model as DensePhrases-IS.

4 Results

4.1 Reading comprehension model performance

A comparison of RC performance is shown in 2. The table includes performance for the English model RoBERTa and untranslated SQuAD data (for the subset of the data that was successfully translated). Using the bilingual model only leads to a slight drop in performance (-1.7 F1). Translating the data further decreases the performance (-2.7 F1, row 6 in the table) but not catastrophically in any sense. In comparison, fine-tuning on an Icelandic only model (IceBERT) improves performance slightly (+0.6 F1, row 7 in the table). These models are not used in any of the steps shown in Figure 1 but the results validate not only the adequacy of the translation method applied, they also demonstrate that the bilingual model is suitable to be adapted for QA in both Icelandic and English.

All models were trained for 4 epochs, using a learning rate of 3e-5, maximum sequence length of 512 tokens and a document stride of 128.

Performance of the IS-EN RC model is measured on the development set of NQ with translated questions. We fine-tune on NQ-ISQ and SQuAD-ISQ, which refer to the Natural questions and SQuAD datasets with only the questions machine translated into Icelandic (step 3, row 5 in the table). Another RC model was fine-tuned on fully translated QA data along with NQiI (step 4, row 8 in the table). We chose that model for use in the fourth step since it was trained on more data than the models in rows 6 and 7 with a small sacrifice in performance on SQuAD-IS, 70.80 F1 and 69.51 EM. With ∼2/3 questions answered exactly, we conclude that the RC models serve well as a teacher models for the DensePhrases training (steps 5 and 6).

4.2 Open QA performance

Performance for the cross-lingual Open QA system (DensePhrases-IS-EN, from step 5) is shown in Table 3 where results are evaluated for the Natural Questions test-dataset, both for the version with machine-translated questions (Is–En) and the original one (En–En). The system still performs well on the English only data. For reference, we note

Original dataset	Transl. dataset	Question transl.	Context transl.	Step
SQuAD	SQuAD-ISQ	✓	✗	3
NewsQA	NewsQA-ISQ	✓	✗	3
NQ	NQ-ISQ	✓	✗	3
SQuAD	SQuAD-IS	✓	✓	4
NewsQA	NewsQA-IS	✓	✓	4
NQ	NQ-IS	✓	✓	4
DensePhrases (generated)	DP-ISQ	✓	✗	5
NQiI	–	✗	✗	6

Table 1: Overview of QA-datasets used in training and how they were translated. The last column refers to steps where the data is used for fine-tuning in Figure 1.

Step	Task	Model	Fine-tuning dataset	F1	EM
-	RC-EN-EN	RoBERTa (EN)	SQuAD	**75.9**	74.3
-	RC-EN-EN	LM EN-IS	SQuAD	74.2	73.0
-	RC-IS-EN	LM EN-IS	NQ-ISQ	74.9	67.1
-	RC-IS-EN	LM EN-IS	SQuAD-ISQ	59.9	50.6
3	RC-IS-EN	LM EN-IS	NQ-ISQ + SQuAD-ISQ	75.8	67.9
-	RC-IS-IS	LM EN-IS	SQuAD-IS	71.5	70.1
-	RC-IS-IS	IceBERT (IS)	SQuAD-IS	**72.1**	70.6
4	RC-IS-IS	LM EN-IS	NewsQA-IS + SQuAD-IS + NQiI	*67.4	64.8

Table 2: Performance in reading comprehension for a mono- and crosslingual setting. RC-X-Y denotes reading comprehension where the question language is X and the answer language is Y. For fine-tuning in the crosslingual setting (RC-IS-EN) in rows 3, 4 and 5, questions have been translated into Icelandic while the context and answers are in English (step 3) whereas the last three rows correspond to fine-tuning on Icelandic only (step 4). The evaluation data in the last row marked with a (*) is from a combination of the datasets used. The best performance on each task is shown in bold.

that the original DensePhrases model (Lee et al., 2021) had an exact match score of 40.9 on NQ and 39.4 on SQuAD when the query-side encoder was fine-tuned for those datasets, respectively.

The Icelandic open QA system (DensePhrases-IS, from step 6) is evaluated on NQiI as well as datasets suitable for open QA in Icelandic, the Gettu betur corpus (4,569 questions with answer) (Ólafur Páll Geirsson, 2013) and Icelandic Trivia Questions[3] (11,610 questions with answers). We note that these datasets are not guaranteed to contain answers that are present in the Icelandic Wikipedia, but serve as a future baseline for Open QA in Icelandic.

Performance results for the model in the sixth step are shown in Table 3. For comparison, a BM25 + IceBERT-QA result is included. The results are

not as good as reported for English systems in, e.g. (Karpukhin et al., 2020), which we currently attribute to the small size of the NQiI dataset.

Finally, we embed the Icelandic Wikipedia for use with CORA (Asai et al., 2021b) using the models released with the paper. The NQiI test dataset is used for evaluation. This method significantly outperforms the one presented in this paper as shown in the last row of Table 3 with F1 28.6 and EM 15.0.

5 Discussion and future work

As noted in the literature review, good results have been achieved in multilingual QA using an answer generator to generate an answer in a selected language (Asai et al., 2021b). For a monolingual setting, our approach provides a way to create an Open QA system without an answer generator as in the original DensePhrases approach.

---

[3]Available online at `https://github.com/sveinn-steinarsson/is-trivia-questions`

Step	Task	Method	Data	EM	F1	EM top 10	F1 top 10
5	Open QA IS-EN	XL-DensePhr.	NQ-ISQ	11.3	15.2	29.6	38.5
5	Open QA EN-EN	XL-DensePhr.	NQ	14.0	18.9	34.7	45.0
6	Open QA IS-IS	XL-DensePhr.	NQiI	9.7	18.8	26.8	44.6
6	Open QA IS-IS	XL-DensePhr.	G.betur	6.0	8.3	14.8	20.6
6	Open QA IS-IS	XL-DensePhr.	Trivia	5.4	6.9	14.6	18.4
-	Open QA IS-IS	BM25 + IB-QA	NQiI	2.4	17.9	2.4	18.1
-	Open QA IS-IS	CORA	NQiI	15.0	**28.6**	-	-

Table 3: Performance for open QA in a cross-lingual Icelandic and English (DensePhrases-IS-EN) setting and in a monolingual IS-IS setting (DensePhrases-IS). In the cross-lingual setting, the performance on NQ is included for reference. All the models are based on the bilingual model (LM EN-IS) except for the last one, which corresponds to using the IceBERT model along with BM25. We highlight in bold the best performance in Open QA on the NQiI dataset.

The model used in the original DensePhrases is SpanBERT (Joshi et al., 2020) whereas we trained a bilingual RoBERTa (Liu et al., 2019) model that has been proven to be successful for Icelandic. For future work, a bilingual SpanBERT model is likely to improve performance as reported in the original paper.

We also evaluated the CORA method on NQiI and it surpassed our method by a significant margin, highlighting the value of training models in a multilingual manner and using a generative model. CORA was not trained specifically on Icelandic QA although it is based on mT5 which was pre-trained on corpora that includes some Icelandic. The result highlights the potential of crosslingual transfer for QA in low-resource languages.

Finally, we emphasize that the quality of the resulting model of the process presented in this paper is affected by multiple factors. For example, it is related to the performance of the translation method but possibly also to language intricacies. A greater amount of training data for Icelandic QA, along with human translated pairs of questions and contexts would cast of light of the penalty incurred from using MT data. We believe the results can be much better with a larger and higher quality target language QA dataset, noting that, e.g. the answer span labelling in the NQiI is somewhat inconsistent. However, we also believe that QA for Icelandic is challenging, and we encourage others to try it out.

6 Conclusion

We have shown how to build an Open QA system from scratch for Icelandic, a language with very limited original QA resources. We first develop a cross-lingual QA system by taking advantage of English QA-data, a well performing translation model, a bilingual language model and the DensePhrases approach. This system is then adapted for monolingual Open QA. The method is not perfect but shows some promising results.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual Open-Retrieval Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics. 00022.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021b. One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. *Advances in Neural Information Processing Systems*, 34. 00002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Somil Gupta and Nilesh Khade. 2020. BERT Based Multilingual Machine Comprehension in English and Hindi. *arXiv:2006.01432 [cs]*. 00006 arXiv: 2006.01432.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938. PMLR. 00053 ISSN: 2640-3498.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Homemade bookcorpus. https://github.com/BIGBALLON/cifar-10-cnn.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. Learning dense representations of phrases at scale. 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 6634–6647, Online. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325. 00103 Publisher: Morgan & Claypool Publishers.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology qa semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983.

Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, and Alessandro Moschitti. 2021. Cross-Lingual GenQA: A Language-Agnostic Generative Question Answering Approach for Open-Domain Question Answering. *arXiv:2110.07150 [cs]*. 00000 arXiv: 2110.07150.

Susumu Okazawa. 2021. Swedish translation of squad2.0. https://github.com/susumu2357/SQuAD_v2_sv.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford.

1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *arXiv:2107.12708 [cs]*. 00008 arXiv: 2107.12708.

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic parser. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 499–517.

Wissam Siblini, Charlotte Pasqual, Axel Lavielle, Mohamed Challal, and Cyril Cauchois. 2021. Multilingual Question Answering from Formatted Text applied to Conversational Agents. *arXiv:1910.04659 [cs]*. 00007 arXiv: 1910.04659.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Bergur Tareq Tamimi Einarsson, Ingibjörg Iða Auðunardóttir, Unnar Ingi Sæmundsson, Hildur Bjarnadóttir, Helgi Valur Gunnarsson, and Hafsteinn Einarsson. 2021. NQiI - natural questions in icelandic - v1.0. CLARIN-IS.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022. Natural questions in icelandic. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Páll Jónsson, Vilhjálmur Þorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus – a recipe for good language models. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *CoRR*, abs/2101.00774.

Ólafur Páll Geirsson. 2013. Iceqa: Developing a Question Answering System for Icelandic.

A Answer span alignment

We apply a heuristic matching method to align the translated questions with spans in the translated context. The method does not rely on more complex word alignment methods between the source text and the translated text but is based on translating the answer and looking up the translated answer in the translated answer context using a fuzzy Levenshtein distance.

In our matching method, we search for the translated answer and then the original answer in the translated context. If either is found, we label the matched string as the answer. Otherwise, we apply a fuzzy matching approach. Denote by w_t the number of words in the translated answer. We perform a sliding window search over all contiguous sequences of words in the translated context that contain w_t, $w_t - 1$, $w_t + 1$ many words. We label and return a sequence as the answer in the translated setting if the Levenshtein distance between the translated answer and the sequence exceeds 0.9. If no sequence is sufficiently similar to the translated answer, we repeat this sliding window search using the original answer instead of the translated answer. If neither search was successful, we would discard the translated question-context pair from training in the fourth step.

Only 6,893 questions, 4.8% of the total data, were discarded from the SQuAD dataset using the matching method since an answer span could not be labelled. 11,478 questions, 9.6% of the total, were discarded from the NewsQA dataset. The only publicly released reading-comprehension style Icelandic dataset for QA, Natural Questions in Icelandic (NQiI) (Snæbjarnarson et al., 2021), is also used for training.