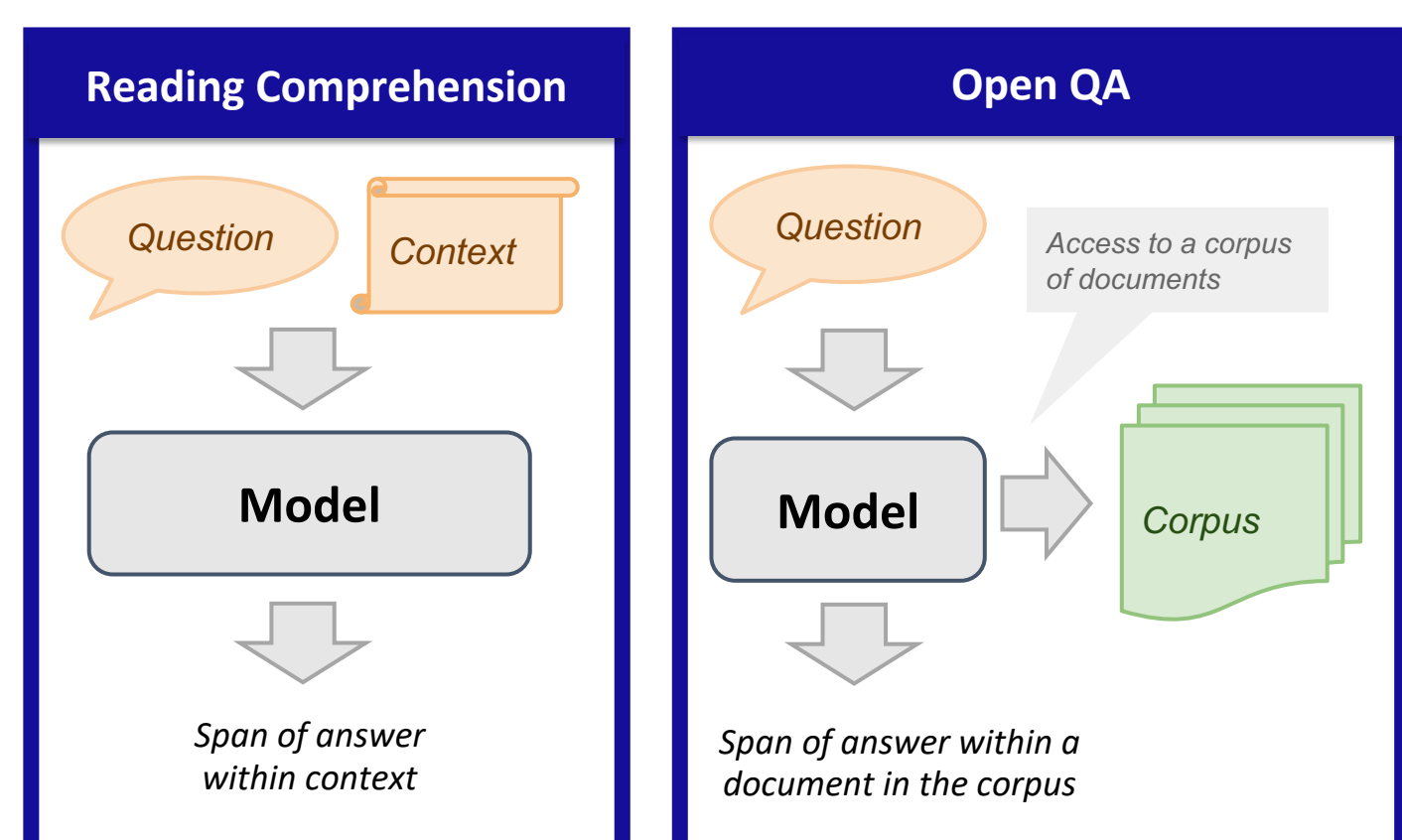


Natural Questions in Icelandic

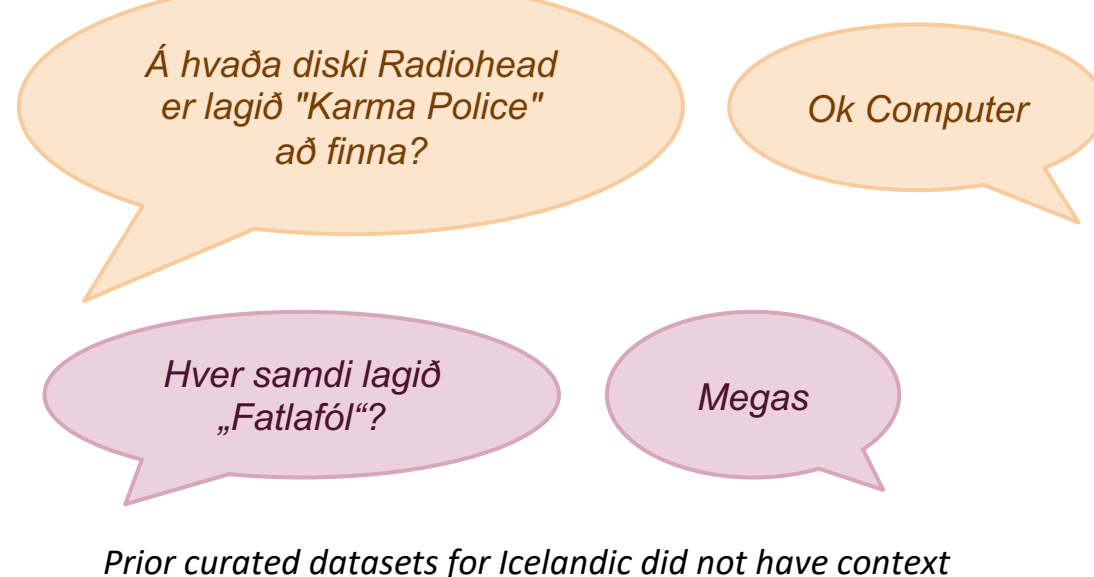
Vésteinn Snæbjarnarson & Hafsteinn Einarsson



Background



Modern QA approaches require context



Datasets

Active development of datasets for English

- WikiQA (Yang et al., 2015)
- Natural Questions (Kwiatkowski et al., 2019)
- Typologically Diverse QA (TyDi QA, Clark et al., 2020)

Other languages and formats

See QA dataset explosion (Rogers, et al., 2021)

Question Answer pair datasets for Icelandic (without context)

- Gettu betur corpus (Geirsson, 2013) - trivia style
- Community collected trivia style questions

Multilingual datasets for Open QA that include Icelandic

MQA (De Bruyn et al., 2021)

- Many questions in English labelled as Icelandic
- 94% of Icelandic questions are scraped from two hotel websites
- Thus, does not reflect information seeking behaviour

Why not translate?

Machine and human translation can introduce artifacts

- Preserved word order when not required
- Constrained language
- Might not reflect all question types
- Might not reflect language specific topics
- Translationese

Machine translation can still be helpful in the development of automated QA methods but having a good dataset is necessary to measure performance.

These observations are amongst the ones driving QA dataset development.

Typological diversity of Icelandic

Comparison with the 11 languages from the TyDi QA dataset (Clark et al., 2020).

Language	Latin script	White space tokens	Sentence boundaries	Word formation	Gender	Pro-drop
English	+	+	+	+	+	-
Arabic	-	+	+	++	+	+
Bengali	-	+	+	+	-	+
Finnish	+	+	+	+++	-	-
Indonesian	+	+	+	+	-	+
Japanese	-	-	+	+	-	+
Kiswahili	+	+	+	+++	-	+
Korean	-	+	+	+++	-	+
Russian	-	+	+	++	+	+
Telugu	-	+	+	+++	+	+
Thai	-	-	-	+	+	+
Icelandic	+	+	+	+++	+	+

Dataset

The dataset was built by five undergraduate students supported by the Icelandic student innovation fund

- Bergur Tareq Tamimi Einarsson,
- Helgi Valur Gunnarsson,
- Hildur Bjarnadóttir,
- Ingibjörg Iðna Auðunardóttir, and
- Unnar Ingi Sæmundasson

Available on Clarin: <https://repository.clarin.is/repository/xmlui/handle/20.500.12537/143>

1 13,740 questions were created

2 Article found for 9,060 questions (65.9%)

3 18,378 labelled question-passage pairs from 1,400 unique Wikipedia articles with answer found in 5,405 cases

3 Agreement

Some question-passage pairs were annotated more than once

3,153 were 1-way annotated

2,721 were 2-way annotated - Agreement: 54.1%

2,817 were 3-way annotated - Agreement: 36.1%

333 were 4-way annotated - Agreement: 37.0%



1 Question words used

Question words	NQil	TyDi QA	SQuAD
What	27%	30%	51%
How	11%	19%	12%
When	10%	14%	8%
Where	7%	14%	5%
Yes/no	7%	10%	<1%
Who	20%	9%	11%
Which	12%	3%	5%
Why	3%	1%	2%
Other	3%	0%	0%

Most of the questions in the Other category start with a verb.

Reading Comprehension

We follow the methodology used to create TyDi QA (Clark et al., 2020).

Model	Dataset	F1	Exact match
IceBERT	NQil	76.0	58.4
XLMR-base	NQil	72.1	56.1
XLMR-base	TyDi English	67.7	56.6
XLMR-base	TyDi Finnish	70.3	44.4

Question type	Entries	F1	Exact match
What	182	65.4	37.4
How	25	78.8	60.0
When	60	64.0	41.7
Where	40	70.6	50.0
Is	18	79.6	72.2
Who	110	66.3	47.3
Which	43	62.0	46.5
Why	32	56.6	43.8
Other	16	97.9	87.5

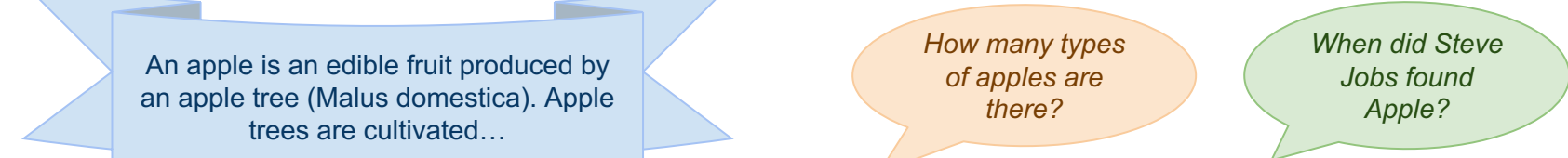
Method

We follow the methodology used to create TyDi QA (Clark et al., 2020).

1 Question elicitation

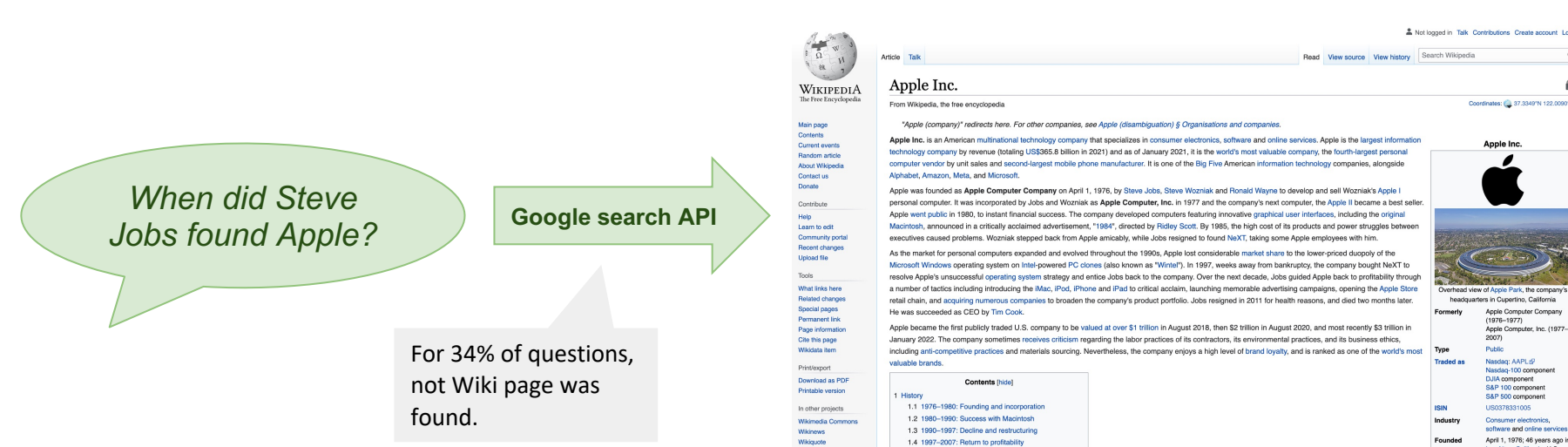
1. Annotators receive the first 100 characters of a Wikipedia article as a prompt.
2. Based on the prompt, the annotator should write a question that they want to know the answer to and that the prompt does not answer.

Note: The prompt serves as an inspiration, and the questions do not need to have a strong connection to the prompt.



2 Article retrieval

For a given question, we use the Google search API to select the top-ranked Icelandic Wikipedia page as a candidate that could contain the answer.

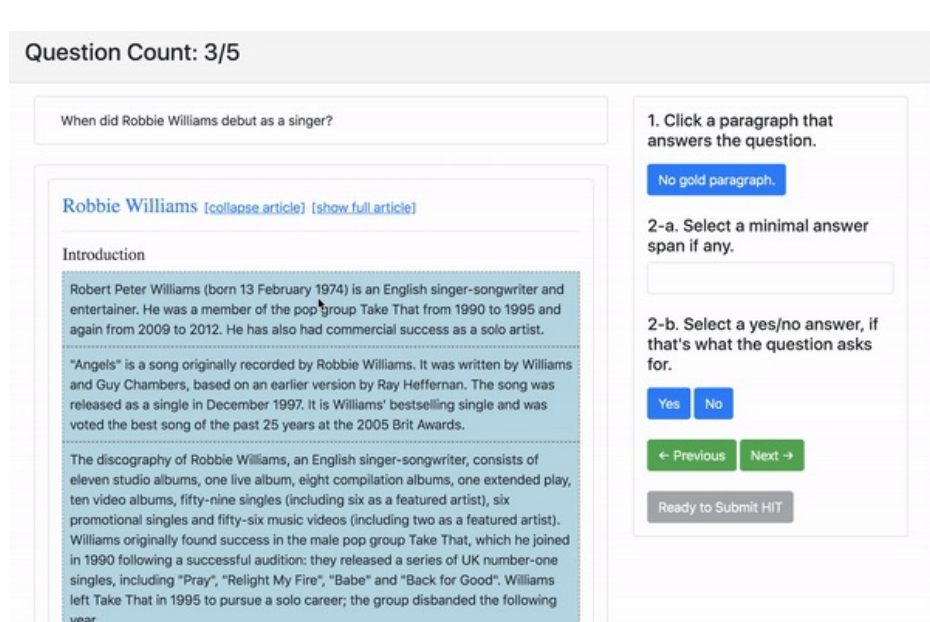


3 Answer labelling

Annotators received question-passage pairs

The annotators could:

- Label no answer.
- Select a paragraph containing the answer.
 - Mark as yes/no
 - Select short minimal answer span



We are grateful for receiving access to the interface from Asai et al. (2020).

Discussion

Wikipedia was almost too small for our setting

- Use alternative sources
- Build a larger Wikipedia

Crowdsourcing to improve diversity

Clearer instructions to reduce annotator disagreement

- Quality checks

Make the answer labelling task easier and more effective

- Rank passages by likelihood of containing answer
- Identify uncertain passages for labelling (active learning)

References

Asai, A., Kasai, J., Clark, J. H., Lee, K., Choi, E., and Hajishirzi, H. (2020). XOR QA: Cross-lingual Open-Retrieval Question Answering. arXiv:2010.11856 [cs], October. 00003 arXiv:2010.11856.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics.

De Bruyn, M., Loffi, E., Buhmann, J., and Daelemans, W. (2021). MFAQ: a Multilingual FAQ Dataset. arXiv:2109.12870 [cs], October. 00000 arXiv:2109.12870

Geirsson, P. (2013). Iceqa: Developing a question answering system for Icelandic.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453-466.

Rogers, A., Gardner, M., and Augenstein, I. (2021). QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. arXiv:2107.12708 [cs], July. 00008 arXiv:2107.12708.

Yang, Y., Yih, W.-t., and Meek, C. (2015). WikiQA: A Challenge Dataset for Open-Domain Question Answering. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2013-2018, Lisbon, Portugal, September. Association for Computational Linguistics. 00515.