U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
DEI DEPARTAMENTO DE ENGENHARIA INFORMÁTICA
LIACC
fct Fundação para a Ciência e a Tecnologia
REPÚBLICA PORTUGUESA

FairytaleQA Translated: Enabling Educational Question and Answer Generation in Less-Resourced Languages

Bernardo Leite, Tomás Freitas Osório, Henrique Lopes Cardoso
LIACC/FEUP

19th European Conference on Technology Enhanced Learning
16-20 September, 2024, Krems, Austria

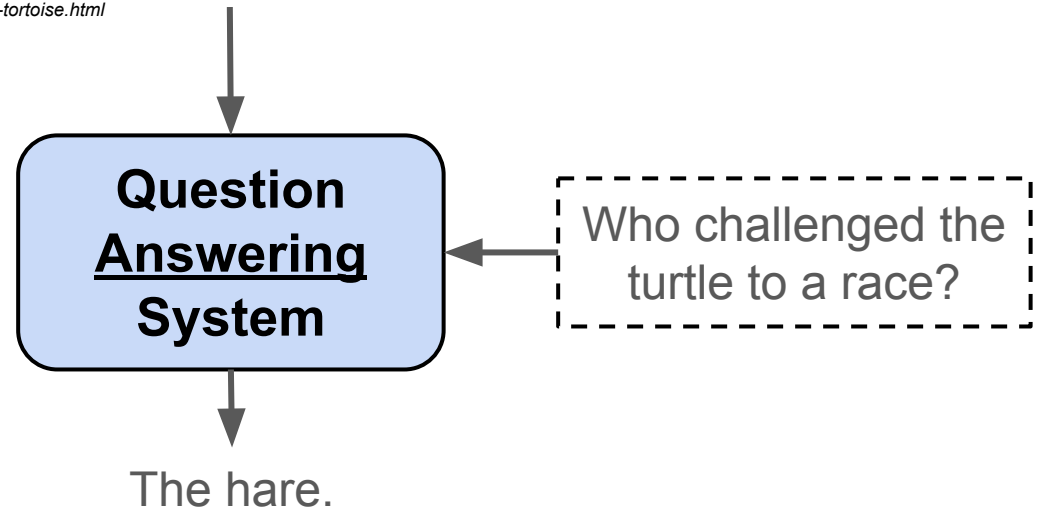
Agenda

- **Background: QA & QG Tasks**
- Introduction
- Related Work
- Translating FairytaleQA
- Baseline Benchmarks
- Case Study
- Conclusions and Future Work

Background → Question Answering

"Once there were a hare and a turtle. The hare was proud of his speed. He asked the turtle to race..."

source: http://simpleessaywriting.blogspot.com/2017/09/a-short-story-on-hare-and-tortoise.html



Background → Question Generation

"Once there were a hare and a turtle. The hare was proud of his speed. He asked the turtle to race..."

source: http://simpleessaywriting.blogspot.com/2017/09/a-short-story-on-hare-and-tortoise.html

Question Generation System

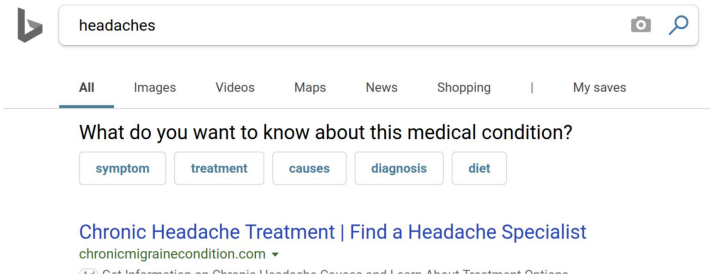
The hare.

Who challenged the turtle to a race?

Introduction → Motivation

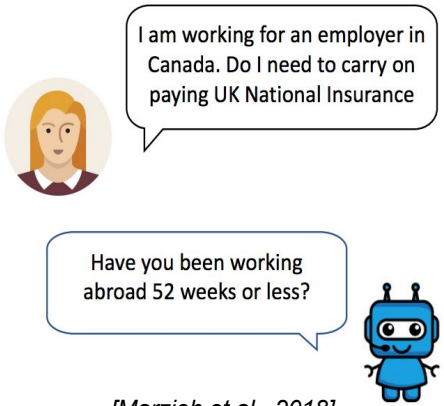
What are the main applications of Question Answer & Generation?

❑ Information Retrieval



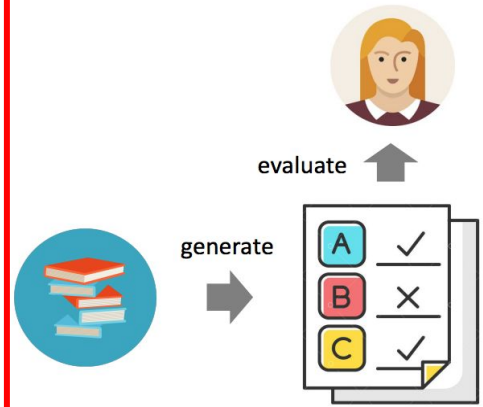
[Zamani et al., 2020]

❑ Dialogue Systems



[Marzieh et al., 2018]

❑ Education

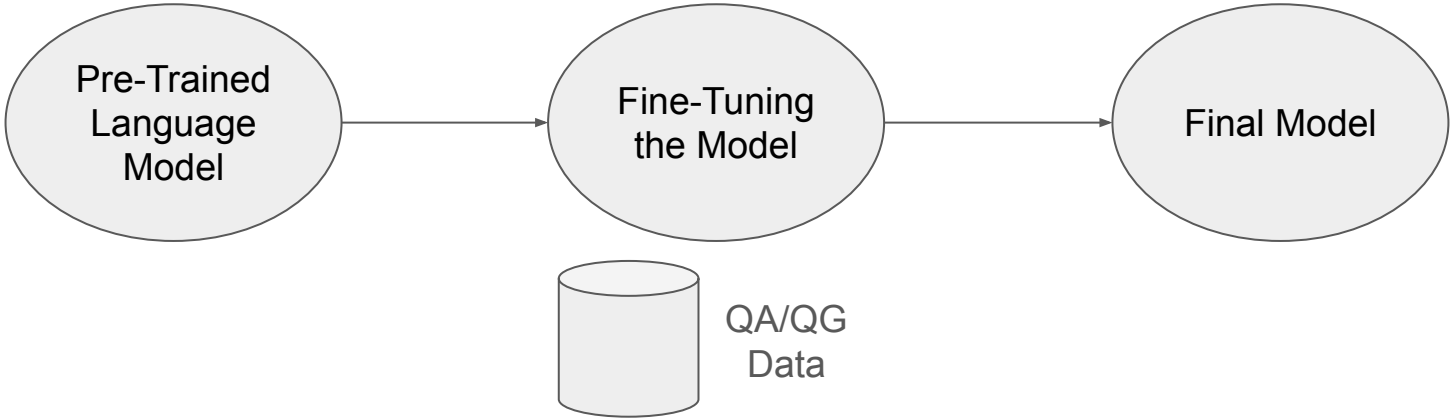


[Rocha et al., 2017]

based on slides: Towards Generating Human-like Deep Questions by Liangming Pan

Introduction → Motivation

Traditional Approach for Building QA/QG Systems



Most of the data is in English.

What about less-resourced languages?

Introduction → Research Goals

Main Goals:

1. Produce **Machine-Translated versions of FairytaleQA**
   Spanish, Portuguese (pt-PT and pt-BR), and French

2. Establish **Baseline Benchmarks** for both QA and QG tasks

3. Present a Case Study where a **Question-Answer Pair Generation Model (QAPG) is Evaluated** by Humans

Agenda

- Background: QA & QG Tasks
- Introduction
- **Related Work**
- Translating FairytaleQA
- Baseline Benchmarks
- Case Study
- Conclusions and Future Work

Related Work → Question & Answer Generation Corpora

Numerous QA/QG datasets have been proposed…

- SQuAD [Rajpurkar et al., 2016]
  - SQuAD-spanish [Carrino et al., 2020]
  - SQuAD-slovak [Staš et al., 2023]
- RACE [Lai et al., 2017]
- CLOTH [Xie et al., 2018]
- …
- **FairytaleQA** [Xu et al., 2022]

Background → FairytaleQA dataset

> *"Once there were a hare and a turtle. The hare was proud of his speed. He asked the turtle to race (...) The hare ran very fast, and the turtle was left behind. The hare thought he should take some rest…"*

source: http://simpleessaywriting.blogspot.com/2017/09/a-short-story-on-hare-and-tortoise.html

Simplistic Example of dataset:

FairytaleQA
(Xu et al., 2022)

Question	Answer	Narrative Label
Who challenged the turtle to a race?	The hare.	Character
(...)	(...)	(...)
Why did the hare decide to take some rest?	The turtle was far behind.	Causal Relation

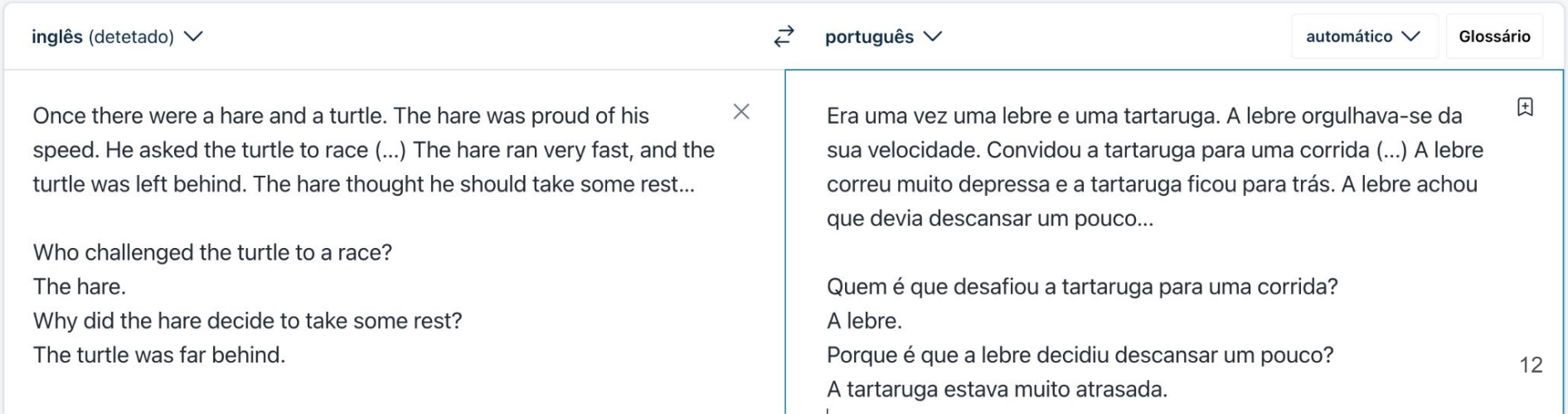
Background → FairytaleQA dataset

Narrative Labels

- Character
- Setting
- Action
- Feeling
- Causal Relationship
- Outcome Resolution
- Prediction

Translating FairytaleQA → About Data

- 278 stories
- 10,580 QA pairs
- We use DeepL for machine-translation (contextualized translation):
  - Spanish, Portuguese (pt-PT and pt-BR), and French



Translating FairytaleQA → Sample Evaluation

- Manual Error Analysis: **150 QA pairs**
- Focus on the **European Portuguese (pt-PT)** translated version

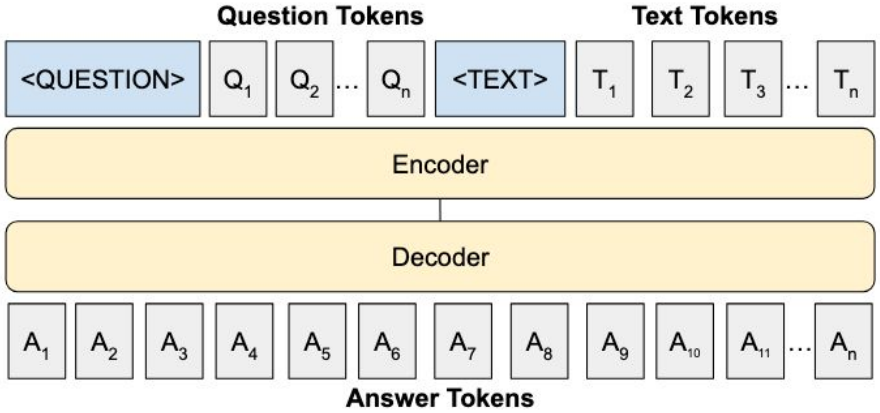
Incidence of issues resulting from machine translation

Issue	Nr. Texts	Nr. Questions	Nr. Answers
Translating Names	7/10	22/150	10/150
Change of Gender	0/10	1/150	2/150
Lost in Translation	0/10	1/150	0/150
Outdated Spelling Agreement	1/10	1/150	0/150

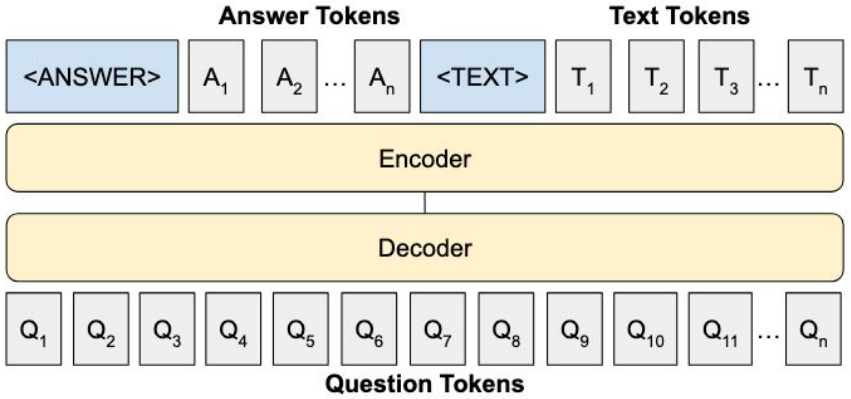
Agenda

- Background: QA & QG Tasks
- Introduction
- Related Work
- Translating FairytaleQA
- **Baseline Benchmarks**
- Case Study
- Conclusions and Future Work

Baseline Benchmarks → Implementation Details



(a) QA setup.



(b) QG setup.

- T5 model [Raffel et al., 2020]
- 8,548/1,025/1,007 as train/val/test QA pairs

Baseline Benchmarks → Results for QA & QG

$ROUGE_L$-F1 values (0-1)

Language	Model	QG	QA
English	T5 (baseline)	0.530	0.551

Baseline Benchmarks → Results for QA & QG

$ROUGE_L$-F1 values (0-1)

Language	Model	QG	QA
English	T5 (baseline)	0.530	0.551
Portuguese (pt-PT)	PTT5	0.496	0.436
Portuguese (pt-BR)	PTT5	0.470	0.448
French	T5-fr	0.404	0.431
Spanish	T5S	0.445	0.382

*What happens if we **back-translate** the Spanish dataset into English?*

Baseline Benchmarks → Results for QA & QG

$ROUGE_L$-F1 values (0-1)

Language	Model	QG	QA
English	T5 (baseline)	0.530	0.551
Portuguese (pt-PT)	PTT5	0.496	0.436
Portuguese (pt-BR)	PTT5	0.470	0.448
French	T5-fr	0.404	0.431
Spanish	T5S	0.445	0.382
Spanish (back-translated)	T5	0.497	0.478

*What happens if we **back-translate** the Spanish dataset into English?*

Agenda

- Background: QA & QG Tasks
- Introduction
- Related Work
- Translating FairytaleQA
- Baseline Benchmarks
- **Case Study: QAPG**
- Conclusions and Future Work

Case Study→ Evaluating a QA Pair Gen. (QAPG) Model

Our Research Question:

<1B par.

Can a <u>modest-scale</u> QAPG model, trained on translated data, generate QA pairs that are qualitatively similar to those used in real exams in a less-resourced language?

Case Study→ Evaluation Protocol for QAPG Model

- 60 QA pairs evaluated (pt-PT)
  - 20 from QAPG
  - 20 from Real Exams
  - 20 from GPT-4 Turbo
- 15 participants
- Metrics:
  - Well-formedness (question only)
  - Relevance with Text (question only)
  - Answerability (question only)
  - Answer Alignment (question and answer)

Case Study→ Results (Well-Formedness)

	QA Pairs Provenance		
Metric	**Real-Exam**	**GPT-4**	**QAPG**
Well-formedness	20/0	19/1	20/0
Relevance with Text	20/0	20/0	19/1
Answerability	20/0	20/0	14/6
Answer Alignment	18/2	20/0	8/12

- **No errors** (grammar or orthography)
- 1 question voted as ill-formed for GPT-4
  Question was reported to be written in a **different language variant**

Case Study→ Results (Relevance with Text)

	QA Pairs Provenance		
Metric	**Real-Exam**	**GPT-4**	**QAPG**
Well-formedness	20/0	19/1	20/0
Relevance with Text	20/0	20/0	19/1
Answerability	20/0	20/0	14/6
Answer Alignment	18/2	20/0	8/12

- 1 question voted-wrong for QAPG
  *"Question inquired about the **feelings mistakenly attributed** to a wrong character"*

Case Study→ Results (Answerability)

	QA Pairs Provenance		
Metric	**Real-Exam**	**GPT-4**	**QAPG**
Well-formedness	20/0	19/1	20/0
Relevance with Text	20/0	20/0	19/1
Answerability	20/0	20/0	14/6
Answer Alignment	18/2	20/0	8/12

6 questions voted-wrong for QAPG

- 1 question demand an answer that pertains to an **unclear story event**
- 2 questions demand answers that are **not explicitly** provided in the story
- 3 questions were reported to be **fully unanswerable**

  e.g., "**Who was the bear?**" (bear description is not provided)

Case Study→ Results (Answer Alignment)

	QA Pairs Provenance		
Metric	**Real-Exam**	**GPT-4**	**QAPG**
Well-formedness	20/0	19/1	20/0
Relevance with Text	20/0	20/0	19/1
Answerability	20/0	20/0	14/6
Answer Alignment	18/2	20/0	8/12

6 (previous) + 6 (new) QA pairs voted-wrong for QAPG

- inaccurate (2 cases)
- incomplete (1 case)
- wrong/nonsensical (3 cases)

Case Study→ Discussion

Revisiting our Research Question:

Can a modest-scale QAPG model, trained on translated data, generate QA pairs that are qualitatively similar to those used in real exams in a less-resourced language?

- Well-formedness
  On par with real exams and GPT-4 😊
- Relevance with Text / Answerability
  Slightly lower performance than real exams and GPT-4 😐
- Answer Alignment
  Significantly lower performance than real exams and GPT-4 😕

Future Directions 🔭

- Exploring **alternative modest-scale** models
- **Double-checking answer** existence
- Employing **QA pair alignment verification** models
- Comparing between **synthetic vs translated** data

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
DEI DEPARTAMENTO DE ENGENHARIA INFORMÁTICA
LIACC
fct Fundação para a Ciência e a Tecnologia
REPÚBLICA PORTUGUESA

FairytaleQA Translated: Enabling Educational Question and Answer Generation in Less-Resourced Languages

Bernardo Leite, Tomás Freitas Osório, Henrique Lopes Cardoso
LIACC/FEUP

SCAN ME

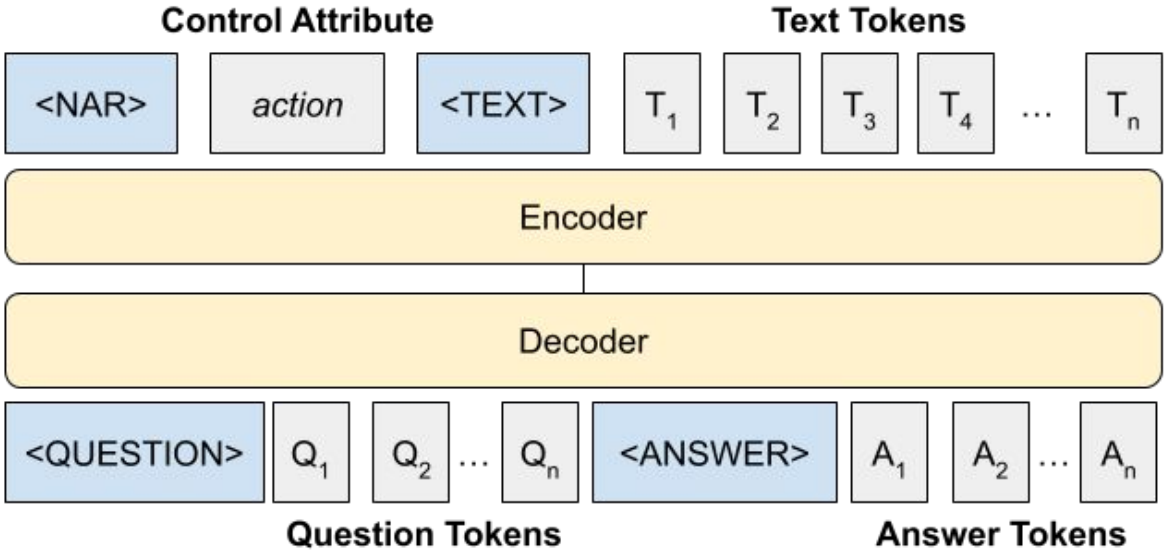
All datasets: Spanish, Portuguese (pt-PT and pt-BR), French + Italian and Romanian

All QA & QG models

> 800 downloads

Appendix

Case Study→ Introducing QAPG Model



Setup for Question-Answer Pair Generation (QAPG)