U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DEI DEPARTAMENTO DE
ENGENHARIA INFORMÁTICA

LIACC

Neural Question Generation for the Portuguese Language: A Preliminary Study

Bernardo Leite, Henrique Lopes Cardoso
Faculty of Engineering - U. Porto, Portugal (FEUP)
Artificial Intelligence and Computer Science Laboratory (LIACC)
{bernardo.leite, hlc}@fe.up.pt

21st EPIA Conference on Artificial Intelligence
August 31 - September 2, 2022, Lisbon, Portugal

Illustrative examples of Neural Question Generation

Passage: *Afonso Henriques, também chamado de Afonsinho, e cognominado de "o Conquistador", foi o primeiro Rei de Portugal. Passa a intitular-se "Rei dos Portugueses" a partir de 1140 e reinou de jure a partir de 5 de outubro de 1143, com a celebração do Tratado de Zamora, até à sua morte. Era filho de Henrique, Conde de Portucale e sua esposa Teresa de Leão.*

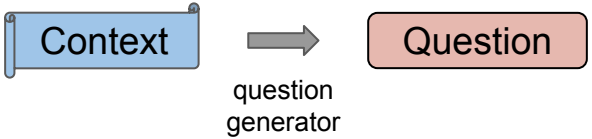
Question 1: *Quem foi o primeiro rei de Portugal? Afonso Henriques*
Question 2: *Qual era o outro nome para Afonso Henriques? Afonsinho*
Question 3: *Qual era o apelido de Afonso Henriques? "o Conquistador"*
Question 4: *Em que ano foi celebrado o Tratado de Zamora? 1143*
Question 5: *Quem era o pai de Afonso Henriques? Conde de Portucale*

Agenda

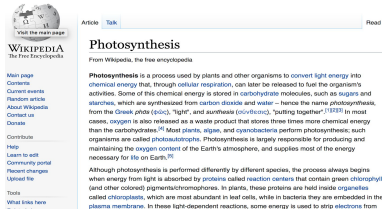
- **Introduction**
- Related Work
- Proposal
- Experimental Setup
- Results
- Final Remarks

Introduction → Background of Question Generation

What is Question Generation (QG)?

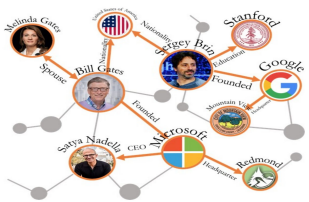


❑ Texts



Why do humans take O2 to produce CO2 while plants do the opposite?

❑ Knowledge Graphs



Were Bill Gates and Satya Nadella once colleagues?

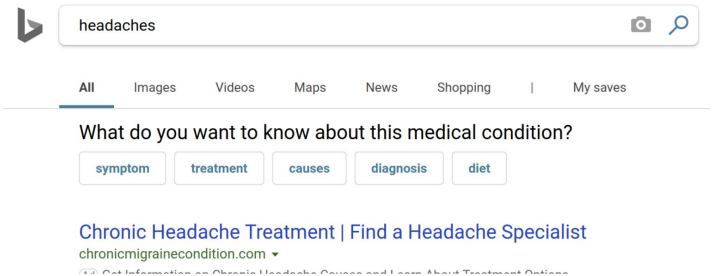
❑ Images

Who is having a birthday party?

Introduction → Background of Question Generation

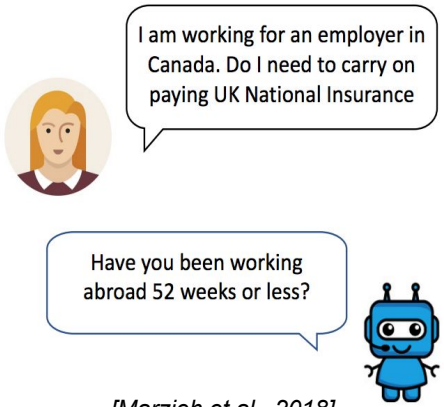
What are the main applications of Question Generation?

❑ Information Retrieval



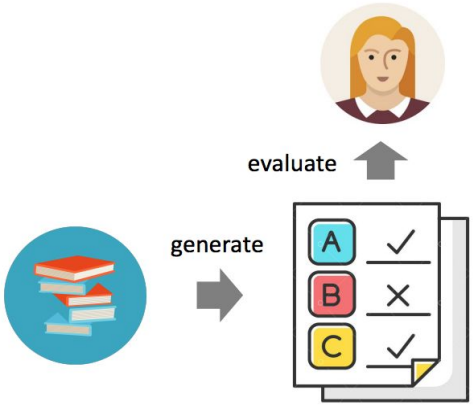
[Zamani et al., 2020]

❑ Dialogue Systems



[Marzieh et al., 2018]

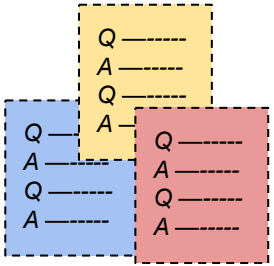
❑ Education



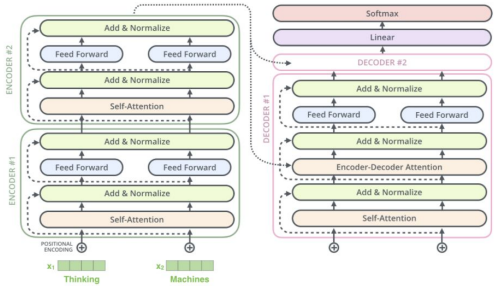
[Rocha et al., 2017]

Introduction → Motivation

Basic "ingredients" for Neural Question Generation



QA corpora:
multiple question-answer pairs



Text generation neural models

after training

Question Generation system

Lower-resourced languages (e.g. Portuguese) lack large-scale quality QA data

Problem!

Introduction → Research Goals

Preliminary analysis of a neural approach to the QG task for Portuguese

- **What type of data?**
  - Portuguese machine-translated version of the SQuAD v1.1 dataset [Rajpurkar *et al.*, 2016]
- **What type of questions?**
  - Open-domain factoid (or *wh*)-questions, e.g, “Who was the first King of Portugal?”
- **What type of model?**
  - State-of-the-art text generation model – T5 [Raffel *et al.*, 2020]
- **How do we evaluate?**
  - Automatic evaluation metrics
  - Error analysis

Agenda

- Introduction
- **Related Work**
- Proposal
- Experimental Setup
- Results
- Final Remarks

Related Work → Question Generation

Traditional and **Neural** Question Generation

- Rule-based [Liu *et al.*, 2012]; [Gates, 2009]; [Mazidi *et al.*, 2014];
- RNN-based [Du *et al.*, 2017]; [Zhao *et al.*, 2018];
- Transformer-based [Dong *et al.*, 2019]; [Xiao *et al.*, 2020];

Question Generation for **Portuguese**

- Rule-based [Curto, 2010]; [Diéguez *et al.*, 2011]; [Correia *et al.*, 2012]; [Pirovani *et al.*, 2017]; [Ferreira *et al.*, 2020]; [Leite *et al.*, 2020];
- RNN-based (there is no work published)
- Transformer-based (there is no work published)

Agenda

- Introduction
- Related Work
- **Proposal**
- Experimental Setup
- Results
- Final Remarks

Proposal → T5 for Question Generation

- **T**ext-**t**o-**T**ext-**T**ransfer-**T**ransformer (T5) was released in 2020 [Raffel *et al.*, 2020]
- Underlying idea for T5: treat every text processing problem as a “text-to-text”
- T5 achieved SOTA results for multiple NLP tasks (e.g., summarization, QA)
- T5 is pre-trained on the C4 corpus: around 750GB of clean English text
- T5 has a multilingual version: mT5 which includes Portuguese [Xue *et al.*, 2020]
- T5 has a portuguese (pt-br) version: PTT5 [Carmo *et al.*, 2020]

We fine-tune T5 for Question Generation

By fine-tuning, we mean that the model has been pre-trained in an initial task (e.g., language modelling) using large-scale corpora (C4) and then retrained on a specific dataset (i.e., SQuAD v1.1) and downstream task (QG)

Proposal → T5 for Question Generation

Code & models: https://github.com/bernardoleite/question-generation-t5-pytorch-lightning

Proposal → T5 for Question Generation

Formally, given the context ***C*** and answer ***A***, the goal is to generate a question ***Q*** that satisfies:

$$\hat{Q} = \underset{Q}{\operatorname{argmax}}\, p(Q|C, A)$$

During training, the model parameters ***θ*** are optimized considering the sequence cross-entropy loss:

$$l(\theta) = -\sum_{t=1}^{T} log\, p(y_t|X, y_{<t}; \theta)$$

Proposal → Experimental Setup

- SQuAD v1.1 Dataset
  - ≈100K pairs of $\langle question, answer \rangle$
  - Original $\langle train, validation, test \rangle$ splits → train (≈76K), validation (≈11K) and test (≈12K)
  - Different dataset versions → original[1] (English) + machine-translated[2] (Br-Portuguese)

- Neural Question Generation (pre-trained models)
  - T5 (English) [Raffel *et al.*, 2020]
  - PTT5 (Portuguese) [Carmo *et al.*, 2020]
  - mT5 (English & Portuguese) [Xue *et al.*, 2020]

- Automatic Evaluation
  - BLEU 1-4
  - $ROUGE_L$
  - METEOR

1. github.com/rajpurkar/SQuAD-explorer/tree/master/dataset
2. github.com/nunorc/squad-v1.1-pt

Agenda

- Introduction
- Related Work
- Proposal
- Experiments
- **Results**
- Final Remarks

Results → Question Generation

Model			Lang.	BL 1	BL 2	BL 3	BL 4	RL	MTR
Seq2Seq	Prior Work	(Du et al., 2017)	EN	43.09	25.96	17.50	12.28	39.75	16.62
		(Zhao et al., 2018)		45.07	29.58	21.60	16.38	44.48	20.25
		(Li et al., 2019)		45.66	30.21	21.82	16.27	44.35	20.36
Pre-trained Seq2Seq		(Chan and Fan, 2019)		**49.73**	**34.60**	26.13	20.33	48.23	23.88
		(Dong et al., 2019)		-	-	-	22.12	51.07	25.06
		(Xiao et al., 2020)		-	-	-	**25.40**	**52.84**	**26.92**
		(Bao et al., 2020)		-	-	-	24.70	52.13	26.33
		(Qi et al., 2020)		-	-	-	23.91	52.26	26.60
		(Wang et al., 2020b)		-	-	-	21.07	49.14	24.09
	Our Proposal								

- Pre-trained seq2seq models generally outperform the other models for all metrics
- BLEU 4 values are in the range **12-25** → low values
- Both English & Portuguese fine-tuned models are within the range of **12-25** for BLEU 4
- **What does a low BLEU value mean in this context?**

What does a low BLEU value mean in this context?

- Generated questions **<u>are valid</u>** paraphrases of the human-written questions
  - e.g **human-written**: *Que tipo de desastre natural o centro da cidade de Lisboa foi projetado para resistir?*
  - e.g **machine-generated**: *O novo centro da cidade de Lisboa foi projetado para resistir a quê?*

BLEU 4	**ROUGE$_L$**
0.54	0.73

- Generated questions **<u>are not valid</u>** paraphrases of human-written questions
  - e.g **human-written**: *Lisboa sofreu alguma epidemia da calamidade?*
  - e.g **machine-generated**: *Por que a cidade de Lisboa não sofreu nenhuma epidemia?*

BLEU 4	**ROUGE$_L$**
0.0	0.57

Error Analysis → Translation problem of the SQuAD v1.1

Context:

"…The band entire will play a concert at the steps of Bond Hall…"

answer

Context fully translated:

"...A banda inteira fará um show nos **degraus** de Bond Hall…"

Individually translated answer:

"as **etapas** de Bond Hall…"

different meanings!

Results **with** translation problems (≈4K)	Results **without** translation problems (≈7K)
BLEU 4: 16.0 ROUGE$_L$: 41.9	BLEU 4: 18.4 ROUGE$_L$: 44.6

Error Analysis → Translation problem of the SQuAD v1.1

Machine translations can cause the text to change its meaning
Example of bad translation:

	English	**Machine-Translated**
Context	The band entire will play a concert at the **steps** of Bond Hall.	A banda inteira fará um show nos **degraus** de Bond Hall.
Answer	the **steps** of Bond Hall	as **etapas** de Bond Hall

different meanings!

Impact on results:

Results **with** translation problems (≈4K)	Results **without** translation problems (≈7K)
BLEU 4: 16.0 ROUGE$_L$: 41.9	BLEU 4: 18.4 ROUGE$_L$: 44.6

Error Analysis → Manual Sample Analysis

We manually analyze 100 generated Portuguese questions randomly extracted

Problems found

- Semantic errors (8 occurrences)
  - e.g question: "Quem Liszt escreveu para Liszt?”
- Answer misalignments (5 occurrences)
  - e.g question: “O terramoto de Sichuan é um dos desastres naturais mais caros da história?” and answer: “história chinesa”
- Syntactic errors (3 occurrences)
  - e.g question: “Quem o Carrefour acusou de doar fundos para?”
- Absence of information (1 occurrence)
  - e.g question: “Quantos alunos estavam lá quando o terramoto aconteceu?”

Agenda

- Introduction
- Related Work
- Proposal
- Experimental Setup
- Results
- **Final Remarks**

Final Remarks

Conclusions

- Neural models for generating questions in Portuguese present (at least) **quantitative results** that align with prior work in English
- Automatic translation of the dataset can lead to unexpected errors
- Automatic evaluation metrics may **not correlate** well with the *actual* quality of the questions

Future Work

- Perform human-evaluation
- Compare rule-based vs neural approaches for question generation in Portuguese
- Explore alternative seq2seq models

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DEI DEPARTAMENTO DE
ENGENHARIA INFORMÁTICA

LIACC

Neural Question Generation for the Portuguese Language: A Preliminary Study

Bernardo Leite, Henrique Lopes Cardoso
Faculty of Engineering - U. Porto, Portugal (FEUP)
Artificial Intelligence and Computer Science Laboratory (LIACC)
{bernardo.leite, hlc}@fe.up.pt

21st EPIA Conference on Artificial Intelligence
August 31- September 2, 2022, Lisbon, Portugal