U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

DEI DEPARTAMENTO DE ENGENHARIA INFORMÁTICA

fct Fundação para a Ciência e a Tecnologia

LIACC

Do Rules Still Rule? Comprehensive Evaluation of a Rule-Based Question Generation System

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

15th International Conference on Computer Supported Education
21-23 April, 2023, Prague, Czech Republic

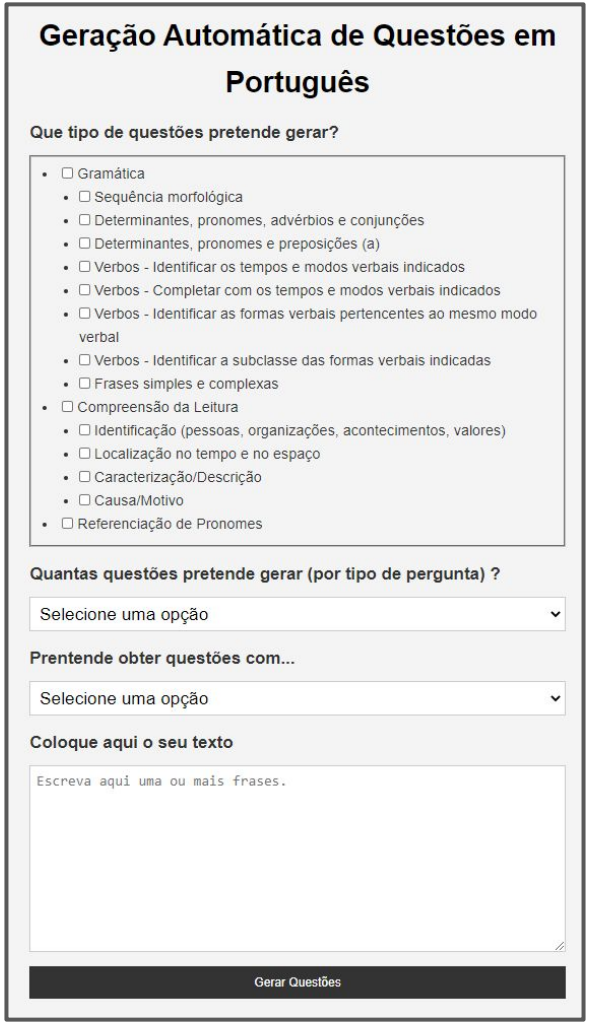
Select Question Type: Grammar

Select Question Type: *wh*-questions

Nr. of Questions

Desired Difficulty

Text Input



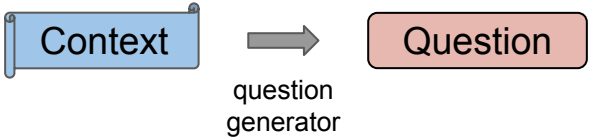
code: github.com/bernardoleite/question-generation-portuguese

Agenda

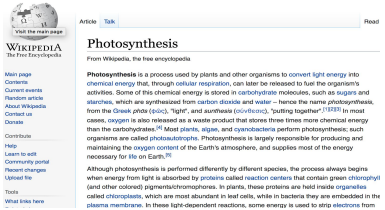
- Introduction
- Related Work
- Question Generation Framework
- Comprehensive Evaluation
  - Study 1: Similarity between Machine-Generated and Human-Authored Questions
  - Study 2: Quality of Machine-Generated and Human-Authored Questions
- Final Remarks

Introduction → Background of Question Generation

What is Question Generation (QG)?

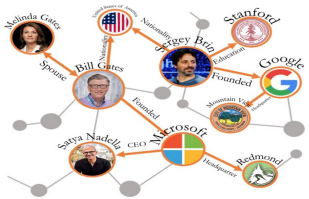


❑ Texts



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

❑ Knowledge Graphs



Were Bill Gates and Satya Nadella once colleagues?

❑ Images

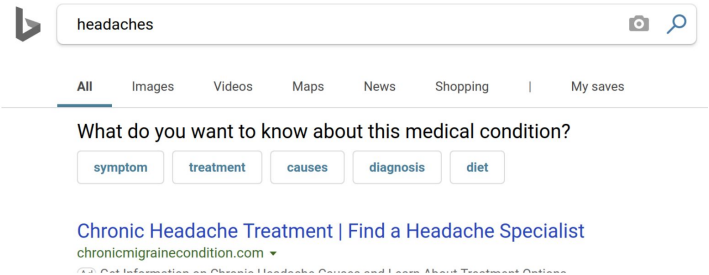
Who is having a birthday party?

based on slides: [Towards Generating Human-like Deep Questions](#) by Liangming Pan

Introduction → Background of Question Generation

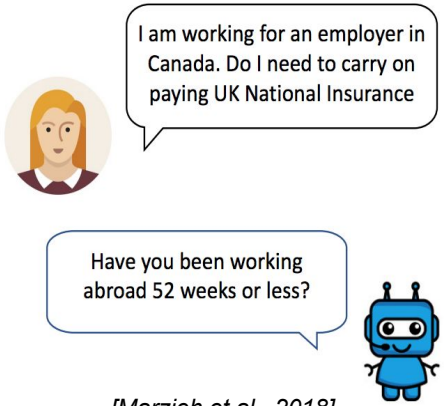
What are the main applications of Question Generation?

- **Information Retrieval**



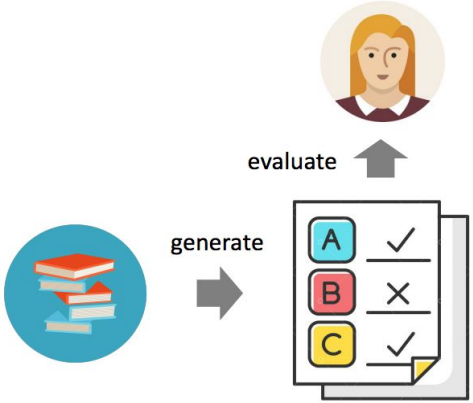
[Zamani et al., 2020]

- **Dialogue Systems**



[Marzieh et al., 2018]

- **Education**

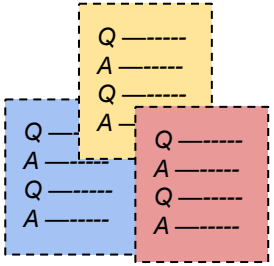


[Rocha et al., 2017]

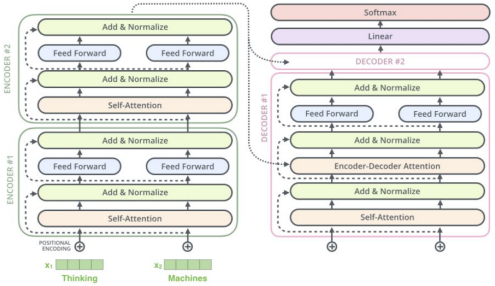
based on slides: [Towards Generating Human-like Deep Questions](#) by Liangming Pan

Introduction → Motivation

Basic “ingredients” for Neural Question Generation



QA corpora:
multiple question-answer pairs



Text generation neural models

after training

Question Generation system

Large-scale quality QA datasets **are not** available for lower-resourced languages (e.g. Portuguese)

Problem!

Possible solutions?

Introduction → Research Goals

To perform a feasibility analysis of a traditional rule-based method for QG in a context of a lower-resourced language.

- **What is the target language?**
  - (European) Portuguese 🇵🇹
- **What type of questions?**
  - Factoid (w*h*)-questions, e.g, “Who was the first King of Portugal?”
- **What is the purpose of generated questions?**
  - Pedagogical goal is to *identify themes, main ideas, facts, causes and effects*
- **What are the main contributions?**
  - QG Framework
  - List of rules used to find patterns
  - Evaluation and comparison between 5 main linguistic aspects for QG

Agenda

- Introduction
- **Related Work**
- Question Generation Framework
- Comprehensive Evaluation
  - Study 1: Similarity between Machine-Generated and Human-Authored Questions
  - Study 2: Quality of Machine-Generated and Human-Authored Questions
- Final Remarks

Related Work → Question Generation Methods

- **Neural** Question Generation
  - RNN-based [Du *et al.*, 2017]; [Zhao *et al.*, 2018]
  - Transformer-based [Dong *et al.*, 2019]; [Xiao *et al.*, 2020]

- **Rule-based** Question Generation
  - **3** linguistics aspects have been extensively explored:
    - Syntactic information [Liu *et al.*, 2010]; [*Heilman and Smith*, 2010]
    - Semantic information [*Lindberg et al.*, 2013]; [*Mazidi and Nielsen*, 2014]
    - Dependency information [*Mazidi and Tarau,* 2016a,b]
  - This research considers **+2** linguistic aspects that have **not** been much explored:
    - Discourse connectors [*Agarwal et al., 2011*]
    - Relative pronouns & adverbs [*Khullar et al., 2018*]

Related Work → Evaluation of QG systems

- **Automatic** Evaluation
  - *N*-gram similarity
    - BLEU 1-4 [*Papineni et al., 2002*]
    - $ROUGE_{L,}$ [*Lin, 2004*]
  - Embedding similarity
    - BERTScore [*Zhang et al., 2019*]
    - BLEURT [*Sellam et al., 2020*]

- **Human** Evaluation
  - Well-formedness
  - Answerability
  - Difficulty
  - …

Agenda

- Introduction
- Related Work
- **Question Generation Framework**
- Comprehensive Evaluation
  - Study 1: Similarity between Machine-Generated and Human-Authored Questions
  - Study 2: Quality of Machine-Generated and Human-Authored Questions
- Final Remarks

Question Generation Framework → The 4 steps

1 — **Pre-Processing**

2 — **Sequence Generation**

3 — **Search Pattern**

4 — **Question Formulation**

Question Generation Framework → Full Example

Passage: *The year of 1917 was difficulty for all belligerents.*

Sequence Generation

with **dependency** information

det-nsubj-case-nmod-cop-root-case-det-det-obl-punct

Search Pattern

det-nsubj-case-nmod-cop-root.?punct*

Question Formulation

'How do you characterize' + match(det-nsubj-case-nmod-cop-root.?punct) + '?'*

Generated Question: *How do you characterize the year of 1917?*

Note: Example is translated from Portuguese.

Question Generation Framework → More Examples

Using **syntactic** information:

Francisco Pizarro discovered the Inca Empire in South America.

Gen. Question: *Who has discovered the Inca Empire in South America?*

Using **semantic** information:

With a kiss, the Morning rubs out each star as it continues its walk towards the horizon.

Gen. Question: *How does Morning rub out each star?*

Using **relative pronouns and adverbs** information:

The Cat took the direction of the narrow paths which lead to the crossroads at the end of the world.

Gen. Question: *What leads to the crossroads at the end of the world?*

Using **discourse connectors** information:

The Portuguese would be in numerical advantage, because of the occupying forces dispersed…

Gen. Question: *Why would the Portuguese be in numerical advantage?*

Note: Examples are translated from Portuguese.

Agenda

- Introduction
- Related Work
- Question Generation Framework
- **Comprehensive Evaluation**
  - Study 1: Similarity between Machine-Generated and Human-Authored Questions
  - Study 2: Quality of Machine-Generated and Human-Authored Questions
- Final Remarks

Comprehensive Evaluation → Study 1 and 2

- **Study 1**

  RQ1. *Are rule-based generated questions similar to those written by humans?*

- **Study 2**

  RQ2. *Are rule-based generated questions comparable to those written by humans in terms of well-formedness and answerability?*

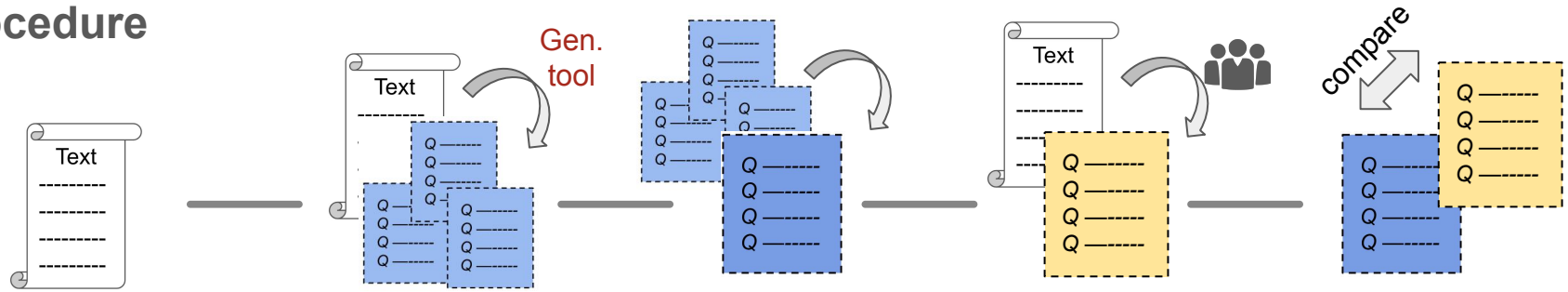
  RQ3. *How well can humans distinguish questions generated by a machine from those written by a person?*

Comprehensive Evaluation → Study 1

Research Question

RQ. 1 *Are rule-based generated questions similar to those written by humans?*

Procedure



Text Collection

5 educational stories

Automatic QG

Gen. all possible questions

Sentence-Question sampling

Sample 150 machine-generated questions

Human QG

150 human questions

Automatic Evaluation

Machine-generated vs Human questions

Comprehensive Evaluation → Study 1

Results

		BLEU 4		ROUGE$_L$		BERTScore	
Aspect	Nr.	Worst	Best	Worst	Best	Worst	Best
Syn.	30	8.61	27.39	17.80	47.16	75.10	83.39
Sem.	30	8.28	30.33	19.76	51.16	77.40	85.54
Dep.	30	**14.09**	**43.64**	25.99	**61.23**	**80.98**	**90.38**
Rel.	30	7.52	25.46	14.90	48.44	74.75	82.85
Disc.	30	13.10	35.06	**26.91**	57.69	75.67	85.30
Overall	150	10.79	32.33	21.07	53.14	76.78	85.49

Neural QG	≈12 to ≈25	≈32 to ≈53	≈75 to ≈90

☺ Dependency information yields consistently better results

☹ Relative Pronouns & Adverbs yields consistently worst results

☺ Scores are quantitatively aligned with those obtained in the literature

Comprehensive Evaluation → Study 2

Research Questions

RQ. 2 *Are rule-based generated questions comparable to those written by humans in terms of well-formedness and answerability?*

RQ. 3 *How well can humans distinguish questions generated by a machine from those written by a person?*

Procedure

- 98 machine-generated + 97 human-written questions
- Each question has been rated by >= 3
- Inquiry metrics

Well-formedness - How well-formed is this question item?	[1-5]
Answerability - How many answers does the question have?	[One, Two+ (ambiguous), None]
Distinguishability - What is the question provenance?	[Human, Machine, Doubt]

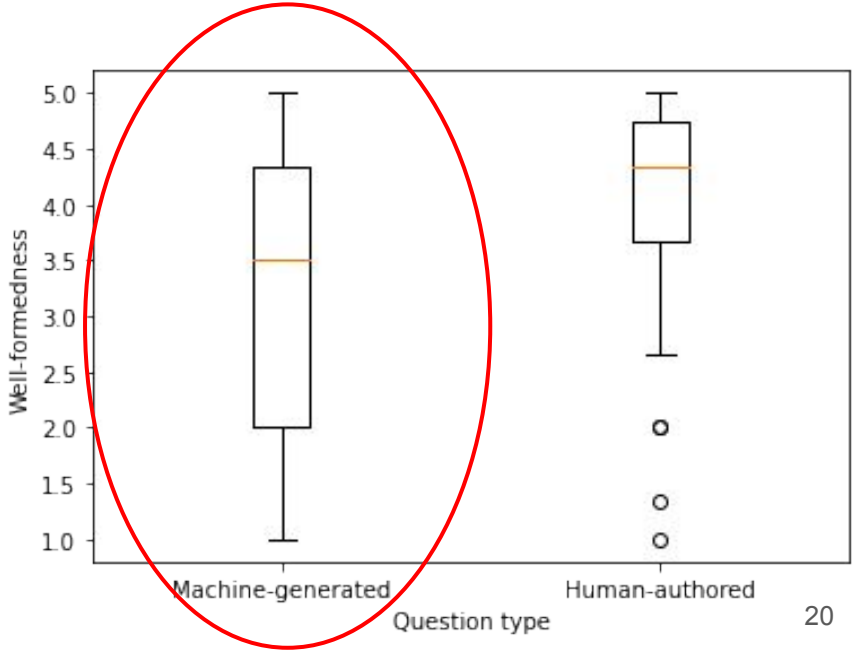
Comprehensive Evaluation → Study 2

Research Question

RQ. 2 *Are rule-based generated questions comparable to those written by humans in terms of well-formedness and answerability?*

Results → *Well-formedness*

- Human: **4.12** ±.86 || Machine: **3.24** ±1.32
- Human-authored questions **fall short**?
- Machine-generated questions is still above the scale's avg.



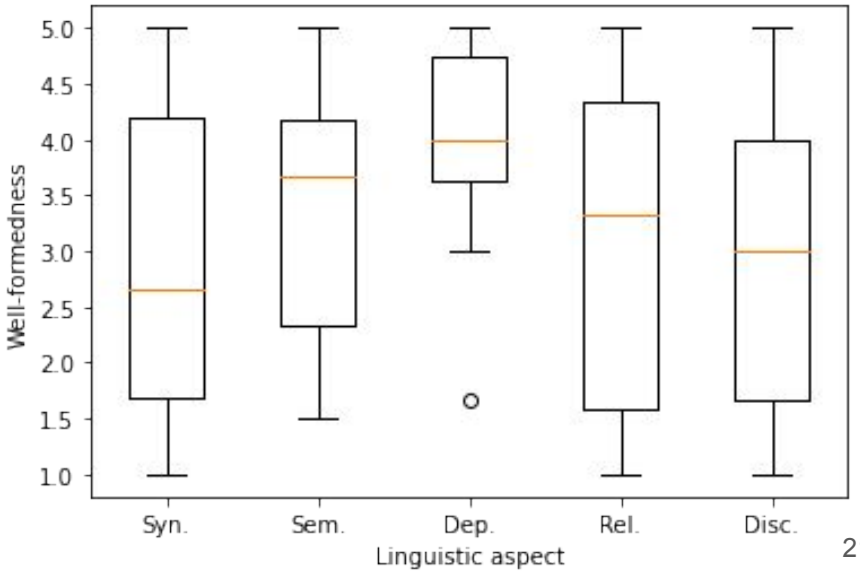
Comprehensive Evaluation → Study 2

Research Question

RQ. 2 *Are rule-based generated questions comparable to those written by humans in terms of well-formedness and answerability?*

Results → *Well-formedness*

Detailed values per linguistic aspect



Comprehensive Evaluation → Study 2

Research Question

RQ. 2 *Are rule-based generated questions comparable to those written by humans in terms of well-formedness and answerability?*

Results → *Answerability*

	Question provenance	
Responses	Human-authored	Machine-generated
One answer	84	65
Two or more answers (ambiguous)	1	1
None (badly formulated question)	3	20
None (answer not in the excerpt)	5	4

Comprehensive Evaluation → Study 2

Research Question

RQ. 3 *How well can humans distinguish questions generated by a machine from those written by a person?*

Results → *Distinguishability*

Over half of the cases (52%) where participants **cannot distinguish** or present **doubts** between question provenance.

	Question provenance	
Responses	HA	MG
Human-authored	118	76
Doubt	91	54
Machine-generated	114	192

Final Remarks

RQ1. Are rule-based generated questions similar to those written by humans?

For *n*-gram and embedding similarity metrics, they are aligned in SOTA.

RQ2. Are rule-based generated questions comparable to those written by humans in terms of well-formedness and answerability?

Yes, when exploring dependency information as a linguistic aspect.

RQ3. How well can humans distinguish questions generated by a machine from those written by a person?

In most cases humans did not distinguish or have doubts.

Do Rules Still Rule? 🤔

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

DEI DEPARTAMENTO DE ENGENHARIA INFORMÁTICA

fct Fundação para a Ciência e a Tecnologia

LIACC

Do Rules Still Rule? Comprehensive Evaluation of a Rule-Based Question Generation System

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

15th International Conference on Computer Supported Education
21-23 April, 2023, Prague, Czech Republic