RMP
Revista dos Mestrados Profissionais
UFPE / CCSA – MGP
https://periodicos.ufpe.br/revistas/rmp
ISSN – 2317 - 0115
V. 14 - n. 1 (2025)
Recife – PE

# Revisão sobre a geração automática de questões na Educação: técnicas, conjuntos de dados e métricas de avaliação

Alana Viana Borges da Silva Neo – Instituto Federal do Mato Grosso do Sul
https://orcid.org/0009-0000-1910-1598
alana.neo@ifms.edu.br
Giseldo da Silva Neo – Instituto Federal de Alagoas
https://orcid.org/0000-0001-5574-9260
giseldo.neo@ifal.edu.br
Mario Diego Ferreira dos Santos – Universidade Federal de Pernambuco
https://orcid.org/0000-0003-2014-3191
mdfs@cin.ufpe.br
Kleber Jose Araújo Galvão Filho – Universidade Federal de Alagoas
https://orcid.org/0009-0001-1591-0272
kleber.filho@ifal.edu.br
Olival de Gusmão Freitas Junior – Universidade Federal de Alagoas
https://orcid.org/0000-0003-4418-8386
olival@ic.ufal.br

**Resumo** – A geração automática de questões (AQG) é uma área interdisciplinar que combina processamento de linguagem natural, aprendizado de máquina para criar sistemas que geram perguntas. Este artigo é uma revisão sistemática relacionada a AQG aplicada à educação. A abordagem mais utilizada neste domínio, segundo resultados da revisão, é a técnicas "baseada em regras", seguida por aprendizado de máquina. Os conjuntos de dados mais comuns incluem o SQuAD, enquanto as métricas de avaliação variam entre BLEU, acurácia e precisão. A revisão também destaca as limitações das abordagens e aponta oportunidades de pesquisa, como a necessidade de conjuntos de dados abertos e melhorias nas técnicas de avaliação automática. Esperamos que isto auxilie pesquisadores na exploração de novas oportunidades e avanços na área de AQG.

**Palavras-chave:** geração automática de questões, processamento de linguagem natural, aprendizado de máquina.

# Review of automatic question generation in Education: assessment techniques, datasets and metrics

**Abstract** – AQG is an interdisciplinary area that combines natural language processing, machine learning and educational techniques to create systems that generate questions automatically. This article presents a systematic review on automatic question generation (AQG) applied to education. We analyzed the main approaches, identifying that the rule-based approach is the most used, followed by machine learning techniques. The most common datasets include SQuAD, while evaluation metrics range from BLEU, accuracy, and precision. The review also highlights the limitations of current methodologies and highlights research opportunities, such as the need for open datasets and improvements in automatic assessment techniques. We hope that this study will assist researchers in exploring new opportunities and advances in AQG.

**Keywords:** automatic question generation, natural language processing, machine learning.

**Data da Submissão:** 17/11/2024 - **Data de aceitação:** 26/07/2025

DOI: https://doi.org/10.51359/2317-0115.2025.265432



## 1. Introdução

A geração automática de questões (AQG), em inglês *automatic question generation*, é a tarefa de criar questões de forma automática a partir de algum tipo de entrada, podendo, opcionalmente, gerar ou avaliar as respostas dos usuários (ZHANG et al., 2021). Ela é utilizada em aplicativos de linguagem natural, na indústria de tecnologia e na área de pesquisa para melhorar a capacidade de compreensão e resposta a perguntas e similares. Além disso pode ser aplicada em diferentes áreas, como, linguística, psicologia, educacional e outras, onde a compreensão e o raciocínio são fundamentais.

Particularmente, este campo de pesquisa tem despertado crescente interesse acadêmico devido ao seu potencial para melhorar o processo de ensino e aprendizagem em ambientes virtuais e tutores inteligentes. A AQG tem o potencial de aliviar o esforço manual dos educadores ao criar atividades avaliativas e de autoaprendizagem. No entanto, apesar dos progressos significativos, há lacunas na variabilidade dos dados, nas métricas de avaliação e na padronização dos métodos, o que justifica a necessidade de uma revisão sistemática. Este estudo se propõe a mapear o estado da arte, identificar desafios e apontar direções futuras que possam aprimorar a aplicabilidade e eficácia das soluções de AQG em contextos educacionais diversificados

O presente estudo tem como objetivo realizar esta revisão sobre AQG aplicada ao contexto educacional. A investigação busca identificar as principais abordagens utilizadas para a geração de questões, bem como os conjuntos de dados mais comumente empregados e as métricas adotadas para avaliar os sistemas de AQG.

Os objetivos secundários são:

- Identificar as abordagens para AQG aplicadas na educação.
- Em mais detalhes: identificar as principais técnicas utilizadas na construção dos AQGs, quais as abordagens e quais os principais frameworks e tecnologias empregadas. Outra questão é
- Identificar os conjuntos de dados utilizados na AQG.
- Em mais detalhes: identificar os conjuntos de dados utilizadas na construção dos AQGs; verificar se eles estão disponíveis para serem utilizados em outras

pesquisas e identificar também se foram desenvolvidos internamente ou foram coletados externamente (de terceiros).

- Indicar as métricas utilizadas para avaliar a AQG.
- Em mais detalhes: identificar como estes sistemas de AQG foram avaliadas e comparadas entre si e identificar quais as métricas utilizadas.
- Verificar quais as limitações dos artigos relacionados a AQG.
- Verificar quais as oportunidades de pesquisa no tema do AQG.

Este artigo está organizado da seguinte forma. A seção 2 aborda os conceitos teóricos que sustentam a pesquisa, com foco na definição da AQG, exemplos de aplicação em diferentes áreas, descrição de um conjunto de dados e explicação de algumas métrica. A seção 3 apresenta os detalhes sobre o processo de revisão sistemática da literatura, incluindo os critérios de inclusão e exclusão de artigos, os termos de busca utilizados, as bases de dados consultadas e o software empregado. A seção 4 apresenta os resultados da revisão, as principais abordagens de AQG, as técnicas de PLN mais utilizadas, as bibliotecas mais empregadas, os conjuntos de dados utilizados, os métodos de avaliação e as métricas de avaliação. Além disso esta seção apresenta uma limitação dos artigos incluídos e as oportunidades de pesquisa com sugestões de áreas promissoras para futuras pesquisas em AQG. Por fim, a seção 5 conclui com os principais achados da revisão.

## 2. Revisão Conceitual

Listam-se em sequência a maior parte dos conceitos necessários à execução da pesquisa.

### 2.1 AQG na educação

Um AQG é um sistema de inteligência artificial projetado para criar perguntas de forma autônoma a partir de textos, dados ou contextos fornecidos. Ele usa algoritmos e técnicas de processamento de linguagem natural para entender o conteúdo e gerar perguntas relevantes, que podem ser utilizadas para testes, educação ou mesmo para melhorar a interação em agentes conversacionais.

Existem vários exemplos da AQG na construção de agentes conversacionais (PIWEK et al., 2007), em ambientes virtuais de aprendizagem, nos sistemas de tutores inteligentes e Cursos Online Abertos e Massivos, conhecidos como MOOCs (*Massive Open Online Courses*), nos sistemas de busca de informação e em vários outros contextos (RUS et al., 2010). Um exemplo de aplicação da AQG em agentes conversacionais pode ser observado em assistentes virtuais como *Siri*, *Cortana*, *Alexa* e *Google Assistant*. Esses sistemas utilizam a AQG para iniciar interações com perguntas ou para manter a fluidez de diálogos já em andamento, aprimorando assim a interatividade e prolongando engajamentos com os usuários (ZHANG et al., 2021).

Diversas áreas de aplicação são beneficiadas pelo uso da AQG. Na área educacional, a AQG pode ser incorporada em sistemas de tutoria inteligente para cumprir objetivos de avaliação da aprendizagem e promoção da autorregulação (ZHANG et al., 2021). Além disso, a AQG permite a geração automática de questões, aliviando o intenso

trabalho manual do professor na elaboração de perguntas e respostas, além disso, pode ser utilizada pelo aluno em atividades voltadas à autorregulação da aprendizagem.

Como um exemplo de incentivo ao uso, a comunidade de geração de texto em linguagem natural organiza competições com o objetivo de incrementar pesquisas em geração AQG e aumentar a visibilidade dessa área (RUS et al., 2010). Uma dessas iniciativas que foram propostas para fomentar avanços na pesquisa em AQG é o "*The First Shared Task Evaluation Challenge on Question Generation* (QG-STEC)" (RUS et al., 2010). O QG-STEC disponibiliza um conjunto de dados padronizado e métricas de avaliação específicas, permitindo a comparação do desempenho entre diferentes participantes e ressaltando inovações tecnológicas em desenvolvimento nesta área emergente.

Por outro lado, a utilização de sistemas de AQG, especialmente por parte de docentes, demanda várias decisões técnicas, dada a ampla gama de artefatos que podem ser empregados em sua construção. Entre os artefatos comumente utilizados incluem-se: gráficos de conhecimento (ELSAHAR et al., 2018; INDURTHI et al., 2017; KUMAR et al., 2021; SERBAN et al., 2016; SONG; ZHAO, 2016), bancos de dados, sites, imagens (FAN et al., 2018a; FAN et al., 2018b; MOSTAFAZADEH et al., 2016), vídeos e textos em linguagem natural.

Além disso, as questões geradas por tais sistemas podem ser classificadas como objetivas ou subjetivas. Relativamente às questões objetivas, estas podem assumir formatos como múltipla escolha, verdadeiro ou falso, e sim ou não. Em contraste, as questões subjetivas podem requerer respostas tanto curtas quanto extensas. Assim, a utilização de um sistema de AQG proporciona uma variedade de opções para sua configuração e aplicação.

Por fim, ao gerar perguntas, a AQG pode ajudar a avaliar a capacidade de um modelo de linguagem em entender o contexto, os conceitos e relacionamentos entre os elementos do texto. Isso permite identificar lacunas e limitações do modelo, bem como melhorar a precisão e a confiabilidade da resposta. Além disso, a AQG também pode ser usada para preparar materiais de estudo, testes e atividades de aprendizado, bem como para desenvolver habilidades de raciocínio crítico e de resolução de problemas.

## 2.2 Conjunto de dados

Em AQG um conjunto de dados refere-se a uma coleção estruturada de exemplos que alimentam o sistema. Este conjunto é composto por pares de perguntas e respostas, bem como por textos ou contextos que servem como base para a formulação de novas perguntas.

Essencialmente, o conjunto de dados funciona como um corpus de treinamento, permitindo que o modelo aprenda a identificar padrões na estrutura das perguntas, a pertinência contextualmente relevante e a adequação das respostas geradas.

A qualidade e a diversidade deste conjunto de dados são cruciais para o desempenho do sistema, pois influenciam diretamente a capacidade do modelo de gerar perguntas claras, concisas e relevantes. Um destes conjuntos de dados é o *SQuAD*.

O *SQuAD*, é um conjunto de dados com mais de 100.000 perguntas, extraídas por pessoas, de um conjunto de artigos da Wikipedia, onde cada resposta é extraída de um segmento do texto selecionado (RAJPURKAR et al., 2016).

A seguir é apresentado uma pergunta extraída do *SQuAD*, com tradução nossa. Onde "Trecho" é o artefato original de entrada, do tipo texto em linguagem natural. "Pergunta" é a pergunta que foi gerada por um humano a partir do "Trecho". Por fim, "resposta" é a resposta criada pelo humano.

- Trecho: "Em meteorologia, precipitação é qualquer produto da condensação do vapor da água da atmosfera que cai devido à gravidade."
- Pergunta: Qual é a causa de a precipitação cair?
- Resposta: Gravidade.

## 2.3 Métricas utilizadas

Em um Ambiente AQG, várias métricas são utilizadas para avaliar a qualidade e eficácia dos modelos de geração de perguntas. Uma das principais métricas é a Acurácia, que mede a proporção de perguntas geradas que são precisas e relevantes para o conteúdo de treinamento. Outra métrica importante é a "Relevância", que avalia a relação entre as perguntas geradas e o conteúdo de treinamento, verificando se as perguntas estão ligadas ao contexto e ao tema. Outra métrica é a "*Baseline BLEU Metric*" (BLEU).

O BLEU, é um método automático de avaliação utilizado em tradução de máquina (PAPINENI et al., 2002). Esse método de avaliação é barato, rápido, independente de linguagem, e altamente correlacionado com as avaliações realizadas pelos seres humanos. A ideia principal é que quanto mais próximo à tradução da máquina de uma tradução profissional realizada por um ser humano, melhor é a tradução. Portanto, para essa avaliação é necessário um corpus com boas traduções de textos na língua escolhida.

Nos parágrafos seguintes apresentaremos como o BLEU funciona sucintamente.

Vamos supor duas traduções do inglês para português, a tradução a) e a tradução b), de um mesmo texto T (que foi omitido) qualquer. Ambas realizadas por uma AQG (ou seja, de forma automática).

- Tradução máquina a) Suas palavras são tão vazias quanto o seu futuro.
- Tradução máquina b) O escrito é sem nada e acontecerá.

Por uma análise semântica (de sentido) um humano consegue inferir que a opção a) é uma tradução melhor que a opção b) para o texto T (mesmo sem ver o texto original, que foi omitido). Dado este cenário o BLEU avaliará entre a) e b) qual a melhor tradução.

O BLEU, como premissa, possui acesso a um corpus com a tradução do mesmo texto T realizada por 3 especialistas diferentes, ou seja, ele possui 3 traduções para o texto T.

- Tradução humana A) As tuas palavras são vazias assim como o seu futuro.
- Tradução humana B) No futuro suas palavras não têm sentido.

- Tradução humana C) O sentido das suas palavras não diz nada nem aqui no presente, nem no futuro.

Note que o termo "palavras" da opção a) é compartilhado pelas 3 traduções feitas por especialistas A) B) e C). Bem como, o termo "futuro" da opção a) é compartilhado por A) e C). A intuição por trás do algoritmo é que quanto mais próximo os termos das opções a) ou b) dos termos das traduções A), B) e C) melhor é a opção de tradução.

O método BLEU compara os *n-grams* das opções a) e b) com os *n-grams* das traduções e atribui mais pontos quanto mais igual e/ou próximos estiverem. A precisão então é o somatório dos pontos calculados para cada tradução que combinaram com cada uma delas divididos pelo total de palavras das traduções.

A métrica do BLEU varia entre 0 e 1 para cada sentença e representa um número do quanto a tradução está mais próxima de uma boa tradução, dificilmente uma opção de tradução tem uma medida de 1, pois isso significa que está idêntico ao corpus e possivelmente até traduções de outros seres humanos do mesmo texto não será igual à do corpus (PAPINENI et al., 2002). Detalhes relacionados ao BLEU, tais como as fórmulas, podem ser encontrados em Papineni *et al.,* (2002)

## 2.4 Técnicas Utilizadas

Existem alguns conceitos fundamentais para diversas aplicações em processamento de linguagem natural (PLN) e aprendizado de máquina. Á seguir algumas delas relacionadas aos resultados são apresentadas sucintamente á seguir (JURAFSKY; MARTIM, 2024):

- A *tokenização* refere-se ao processo de segmentação de um texto em unidades menores, chamadas *tokens*;
- Gramática livre de contexto - *Context-free grammar*: é um sistema formal usado para descrever a estrutura de sentenças em uma linguagem natural, que é composta por regras de produção que definem como símbolos podem ser substituídos por outros símbolos ou sequências de símbolos, permitindo a geração de todas as frases gramaticalmente corretas de uma linguagem;
- *Semantic role labeling* é o processo de atribuir papéis semânticos às palavras ou frases em uma sentença. Ele identifica o relacionamento entre os elementos de uma sentença e os eventos descritos, categorizando os papéis de sujeito, objeto, instrumento etc.;
- *N-grams* são sequências contínuas de n itens (geralmente palavras) de um dado texto ou discurso. Eles são usados em modelos de linguagem para prever a próxima palavra em uma sequência, baseando-se na probabilidade das ocorrências dos *n-grams* em um *corpus* de texto.
- *Part of Speech Tagging* é a tarefa de atribuir rótulos gramaticais a cada palavra em um texto, tais como substantivo, verbo, adjetivo. Isso ajuda na compreensão da estrutura gramatical e do significado das palavras em contexto;

- *Named entity recognition* é a técnica de identificar e classificar entidades mencionadas em um texto em categorias predefinidas, como nomes de pessoas, organizações, locais, datas e outros.

SEQ2SEQ é um modelo de aprendizado profundo que mapeia uma sequência de entrada para uma sequência de saída, sendo amplamente utilizado em tarefas como tradução automática, onde uma frase em um idioma é convertida em uma frase em outro idioma.TF-IDF são técnicas de mineração de texto que quantificam a importância de palavras em um documento. Enquanto TF mede a frequência de um termo em um documento o TF-IDF ajusta a frequência do termo com base na frequência do termo em uma coleção de documentos, destacando termos relevantes em documentos individuais (BIRD et al., 2019).

O *Word2Vec* é uma técnica de aprendizado de vetores de palavras, onde palavras são mapeadas para vetores de alta dimensão. Esses vetores capturam relações semânticas entre palavras, facilitando tarefas como similaridade semântica e análise de texto. O LSTM é um tipo de rede neural recorrente projetada para aprender dependências de longo prazo em sequências de dados. *GloVe* é um método de aprendizado de vetores de palavras baseado na matriz de ocorrência de palavras em um grande corpus de texto. Diferentemente de Word2Vec, que aprende vetores de palavras baseados em contextos locais, *GloVe* captura a estatística global de ocorrência, resultando em representações vetoriais de palavras semântica e contextualizadas (BIRD et al., 2019).

O *NLTK* é uma das bibliotecas em Python® mais antigas e completas para o PLN, usada para análise de texto, *tokenização*, *stemming*, *tagging*, *parsing*. Já o *spaCy é u*ma biblioteca, também para Python®, de PLN moderna e eficiente, focada em velocidade e usabilidade, oferecendo ferramentas para *tokenização,* etiquetagem, *parsing* de dependências, reconhecimento de entidades nomeadas, e vetores de palavras. O *PyTorch* é uma biblioteca de aprendizado profundo utilizada para construção e treinamento de modelos de aprendizado de máquina, com suporte para computação em GPU. O *AllenNLP* é uma plataforma de aprendizado profundo construída sobre *PyTorch*, projetada especificamente para tarefas de PLN, oferecendo modelos prontos para uso. Por fim, o *Stanford parser* é um *parser* robusto desenvolvido pela Universidade de Stanford, utilizado para análise sintática de textos, incluindo *parsing* de dependências e análise de constituintes, com suporte para múltiplas línguas (BIRD et al., 2019).

## 3. Metodologia

Foi realizada uma revisão sistemática, seguindo os critérios estabelecidos por Kitchenham (2007). A revisão sistemática configura-se como um estudo secundário cujo objetivo é identificar, analisar e interpretar informações relevantes aos objetivos de pesquisa formulados pelos investigadores, com ênfase na sistematização para garantir a reprodutibilidade. As atividades envolvidas nesse processo incluem planejamento, execução e relato dos resultados (figura 1).

Inicialmente, elaborou-se um protocolo para a revisão, abrangendo a definição das questões de pesquisa, a especificação dos termos de busca, a filtragem dos resultados baseando-se nos títulos e resumos que atendem aos critérios de inclusão e exclusão

previamente estabelecidos, assim como a remoção de artigos duplicados. Para assegurar a reprodutibilidade da pesquisa, utilizou-se o *software* https://parsif.al/, que permitiu o registro detalhado de todas as etapas da revisão.

**Figura 1** – Atividades do processo de revisão


Os termos de busca utilizados na pesquisa são:

("*automatic question generation*" or "*AQG*") and ("*education technology*" or "*education*") and ("*natural language processing*" or "*NLP*")

Os critérios de inclusão utilizados foram:

1. O artigo é um experimento sobre geração de perguntas automáticas em contextos educacionais;
2. O Artigo está escrito na língua inglesa.
3. Os critérios de exclusão aplicados foram:
4. Artigos duplicados;
5. Artigos com baixa qualidade;
6. Artigos não disponíveis ou incompletos;
7. Livros;
8. Capítulos;
9. Tutoriais;
10. Revisões sistemáticas.

O PICOC (*population, intervention, comparison, outcome, context*) desta revisão sistemática foi:

- *Population* (População): Pesquisadores e desenvolvedores que trabalham com geração automática de perguntas, especialmente em contextos educacionais e de aprendizado de máquina.
- *Intervention* (Intervenção): Aplicação de técnicas de PLN e aprendizado de máquina para gerar questões automaticamente em ambientes educacionais.
- *Comparison* (Comparação): Comparação entre diferentes abordagens para AQG, como métodos baseados em regras (rule-based) e aprendizado de máquina (machine learning).

- *Outcome* (Resultado): A identificação de quais abordagens são mais eficazes, os conjuntos de dados utilizados, e as métricas de avaliação que medem a qualidade das perguntas geradas.
- *Context* (Contexto): O uso de AQG em ambientes educacionais, com foco em melhorar o processo de ensino e aprendizagem em tutores inteligentes e plataformas de aprendizado virtual, além de destacar lacunas e oportunidades de pesquisa nessa área.

O procedimento de extração de dados da revisão sistemática seguiu as etapas definidas por Kitchenham (2007) com o objetivo de garantir a coleta precisa e padronizada das informações relevantes dos estudos incluídos. A revisão foi realizada por dois revisores, seguindo os critérios estabelecidos e garantindo a reprodutibilidade do processo de análise. A primeira etapa envolveu a definição dos dados, que serve como guia para a coleta de informações essenciais. Estes dados foram registrados no parsif.al que foi estruturado para contemplar todos os elementos importantes para a revisão, permitindo que os revisores capturem de forma consistente as informações necessárias.

Em seguida, os revisores foram treinados para garantir a compreensão dos critérios e utilizar corretamente o software parsif.al para reduzir a ocorrência de erros ou interpretações divergentes na coleta dos dados. Após o treinamento, os dados foram extraídos em duplicidade pelos dois revisores. Ambos realizam a extração dos dados de cada estudo de forma autônoma, e, em caso de discrepâncias, foram resolvidas por consenso.

O próximo passo foi a análise dos dados extraídos de forma quantitativa para os três primeiros objetivos secundários e qualitativa para os dois últimos dois objetivos secundários. Por fim, o resultado da validação foi revisado para assegurar que estavam completos e corretos antes de avançar para a fase de síntese e discussão dos resultados.

## 4. Resultados

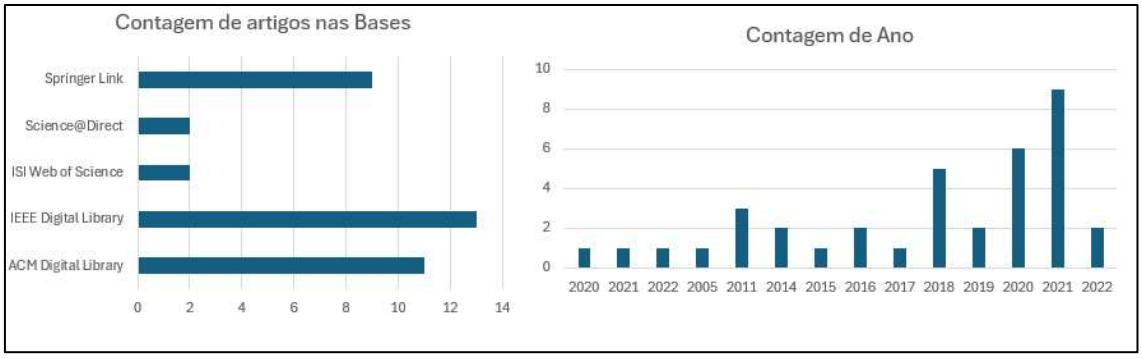
Nas bases secionadas, 178 artigos foram retornados inicialmente para os termos de busca. Após o processo de leitura do título, resumo e leitura completa restaram 31 artigos. Portanto foram aceitos na revisão 31 artigos, de onde foram extraídos os dados. A extração das limitações e oportunidades ocorreu nos 9 mais recentes, dado limitações de tempo. O resultado detalhado da execução do protocolo de revisão categorizado pelo repositório, além dos repositórios utilizados, é apresentado no quadro 1. O quadro 2, apresenta os 31 artigos remanescentes após leitura do título e resumo.

**Quadro 1 –** Resultado protocolo revisão.

| Repositório | Artigos Identificados | Remanescentes após leitura do título e resumo | Remanescentes após leitura completa |
|---|---|---|---|
| *ACM Digital Library* | 28 | 11 | 9 |
| *IEEE Xplore DL* | 13 | 13 | 13 |
| *ISI Webn of Science* | 14 | 2 | 2 |
| *Science@Direct* | 10 | 2 | 2 |
| *Springer Link* | 113 | 9 | 5 |
| Total | 178 | 37 | 31 |

A figura 2 da esquerda apresenta a contagem de artigo nas bases de busca utilizadas. A base IEEE e ACM foi a origem da maior quantidade dos artigos encontrados. Ainda na figura 2, no gráfico da esquerda, é possível perceber um aumento no número de publicações sobre o tema nos últimos anos.

**Figura 2** – Estatísticas da contagem dos artigos.




Fonte: os Autores (2024).

**Quadro 2** – Os 31 artigos da revisão, antes da leitura completa (parte 1).

| # | Título | Período | Ano |
|---|---|---|---|
| 1 | [33] Generating QA from Rule-based Algorithms | 2022 International Conference on Electronics and Renewable Systems (ICEARS). IEEE | 2022 |
| 2 | [9] Automatic question generation and answer assessment for subjective examination | Cognitive Systems Research | 2022 |
| 3 | [23] A Tool for Automatic Question Generation for Teaching English to Beginner Students | 2021 40th International Conference of the Chilean Computer Science Society (SCCC). IEEE | 2021 |
| 4 | [37] Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment | Multimedia Tools and Applications | 2021 |
| 5 | [45] A Review on Question Generation from Natural Language Text | ACM Trans. Inf. Syst. | 2021 |
| 6 | [40] Automatic Question-Answer Generation from Video Lecture using Neural Machine Translation | 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE | 2021 |
| 7 | [21] Expert-Authored and Machine-Generated Short-Answer Questions forAssessing Students' Learning Performance | Educational technology & society | 2021 |
| 8 | [44] Developing Question Generation System for Bahasa Indonesia Using Indonesian Standard Language Regulation | 2021 10th International Conference on Software and Computer Applications | 2021 |
| 9 | [6] Rule Based Question Generation for Arabic Text: Question Answering System | Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence | 2021 |
| 10 | [19] Factual Question Generation for the Portuguese Language | 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE | 2020 |

**Quadro 2** – Os 31 artigos da revisão, antes da leitura completa (parte 2).

| # | Título | Periódico | Ano |
|---|---|---|---|
| 11 | [7] Controllable Question Generation via Sequence-to-Sequence Neural Model with Auxiliary Information | 2020 International Joint Conference on Neural Networks (IJCNN). IEEE | 2020 |
| 12 | [4] Applicable strategy to choose and deploy a MOOC platform with multilingual AQG feature | 2020 21st International Arab Conference on Information Technology (ACIT). IEEE | 2020 |
| 13 | [41] Questionator-Automated Question Generation using Deep Learning | 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE | 2020 |
| 14 | [43] Key Phrase Extraction for Generating Educational Question-Answer Pairs | Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale | 2019 |
| 15 | [10] Bilingual Ontology-Based Automatic Question Generation | 2019 IEEE Global Engineering Education Conference (EDUCON). IEEE | 2019 |
| 16 | [27] Automatic distractor generation for multiple-choice English vocabulary questions | Research and Practice in Technology Enhanced Learning | 2018 |
| 17 | [26] Video-Based Question Generation for Mobile Learning | Proceedings of the 2nd International Conference on Education and Multimedia Technology | 2018 |
| 18 | [31] Automatic Chinese Multiple-Choice Question Generation for Human Resource Performance Appraisal | Procedia Computer Science | 2018 |
| 19 | [46] Domain specific automatic Chinese multiple-type question generation | 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE | 2018 |
| 20 | [31] Automatic Chinese Multiple Choice Question | Procedia Computer Science | 2018 |
| 21 | [20] Generation Using Mixed Similarity Strategy | IEEE Transactions on Learning Technologies | 2018 |
| 22 | [8] Factual open cloze question generation for assessment of learner's knowledge | International Journal of Educational Technology in Higher Education | 2017 |
| 23 | [16] Content-Dependent Question Generation Using LOD for History Learning in Open Learning Space | New Generation Computing | 2016 |
| 24 | [11] Ill-formed to well-formed question generator | 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE). IEEE | 2016 |
| 25 | [29] Evaluation of a question generation approach using semantic web for supporting argumentation | Research and Practice in Technology Enhanced Learning | 2015 |
| 26 | [22] Automatic generation of multiple-choice questions using dependency-based semantic relations | Soft Computing | 2014 |
| 27 | [1] Automatic Gap-Fill Question Generation from Text Books | Proceedings of the sixth workshop on innovative use of NLP for building educational applications. | 2011 |

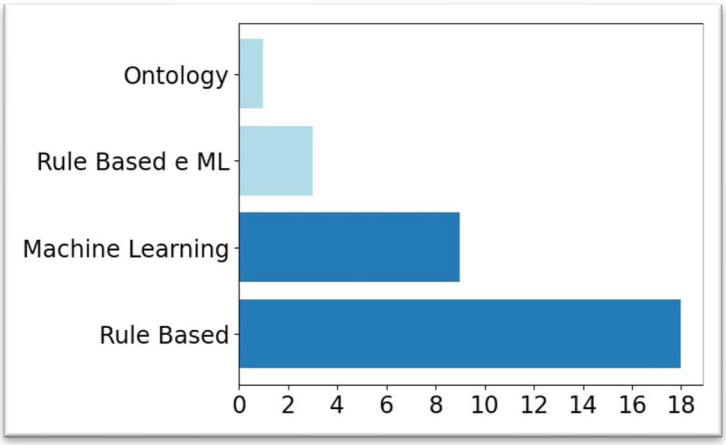
**Quadro 2** – Os 31 artigos da revisão, antes da leitura completa (parte 3).

| # | Título | Periódico | Ano |
|---|---|---|---|
| 28 | [3] Automatic Question Generation Using Discourse Cues | Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications | 2011 |
| 29 | [25] Math Word Question Generation for Training the Students with Learning Difficulties | Proceedings of the International Conference & Workshop on Emerging Trends in Technology. | 2011 |
| 30 | [42] Automatic Question Generation for Repeated Testing to Improve Student Learning Outcome | 2021 International Conference on Advanced Learning Technologies (ICALT). IEEE | 2021 |
| 31 | [15] Focused Questions and Answer Generation by Key Content Selection | 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM) | 2020 |

## 4.1 Quais as abordagens mais utilizadas?

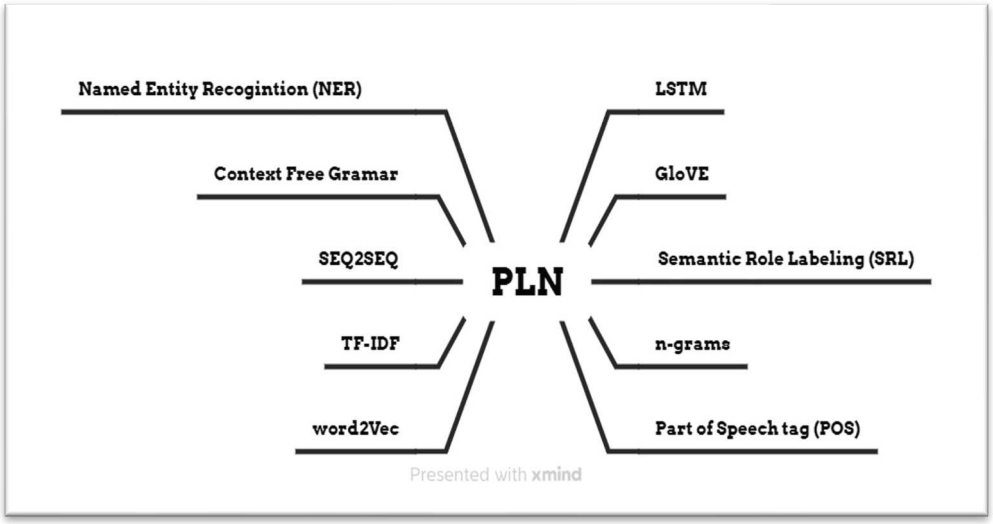
As abordagens mais utilizadas nos artigos selecionados foram agrupadas em dois grupos: *Rule-Based* e *Machine Learning*, conforme a figura 3, sendo a *Rule-Based* a abordagem dominante. A abordagem *Rule-Based* é caracterizada pelo uso de métodos heurísticos e é a abordagem mais tradicional, enquanto a abordagem *Machine Learning* é mais recente, sendo caracterizada pelo uso de técnicas de aprendizagem de máquina, redes neurais, ou mesmo estatísticas.

**Figura 3.** Contagem dos artigos agrupados pelas abordagens.




As técnicas de processamento de linguagem natural mais utilizadas, apresentadas na figura 4, foram: a tokenização (de sentenças, palavras ou termos); o uso de *Context Free Gramar*; SEQ2SEQ; TF/TF-IDF; *word2Vec*; LSTM; GloVE; *Semantic Role Labeling* (SRL); *n-grams*; *Part of Speech tag* (POS) e o uso de *Named Entity Recogintion* (NER); Modelos do tipo: “O quê?”, “Como?”, e “Onde?” também foram utilizados.

**Figura 4.** Técnicas de PLN utilizadas.

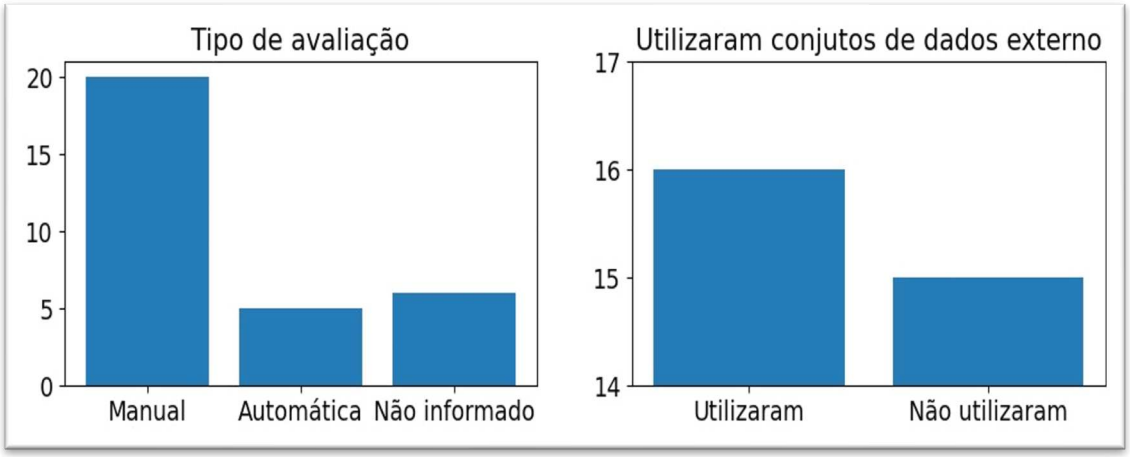



Já as bibliotecas mais utilizadas foram: *NLTK*, spaCy, *PyTorch*, *AllenNLP* e o *Stanford Parser*.

## 4.2 Quais os conjuntos de dados?

Dos artigos selecionados nessa revisão, 51,6% dos estudos não utilizaram conjuntos de dados externos e 48,4% utilizaram conjuntos de dados de outras pesquisas (**Figura 5**). Em relação às pesquisas que utilizaram conjunto de dados externos, o conjunto de dados *SQuAD* foi o mais utilizado, em segundo lugar o *WordNET*, os outros só foram mencionados em um único artigo. O quadro 3 apresenta uma lista com um *link* para ter acesso ao conjunto de dados.

**Figura 5.** Contagem das pesquisas que utilizaram conjunto de dados externos e tipos de avaliação.



**Quadro 3**– Conjuntos de dados.

| Nome | Links para download |
|---|---|
| *SQuAD* | https://rajpurkar.github.io/SQuAD-explorer/ |
| *WordNET* | https://wordnet.princeton.edu/ |
| *GENIA* | http://www.nactem.ac.uk/meta-knowledge/download.php |
| *ILSVRC* | https://image-net.org/challenges/LSVRC/2012/ |
| *QGSTEC* | https://github.com/bjwyse/QGSTEC2010 |
| *Wikipedia Dataset* | https://www.tensorflow.org/datasets/catalog/wikipedia |
| *wiki2vec* | https://github.com/idio/wiki2vec/ |
| *WikiNER* | https://spacy.io/models/pt |
| *Kagle Dataset StackOverflow* | https://www.kaggle.com/stackoverflow/pythonquestions |

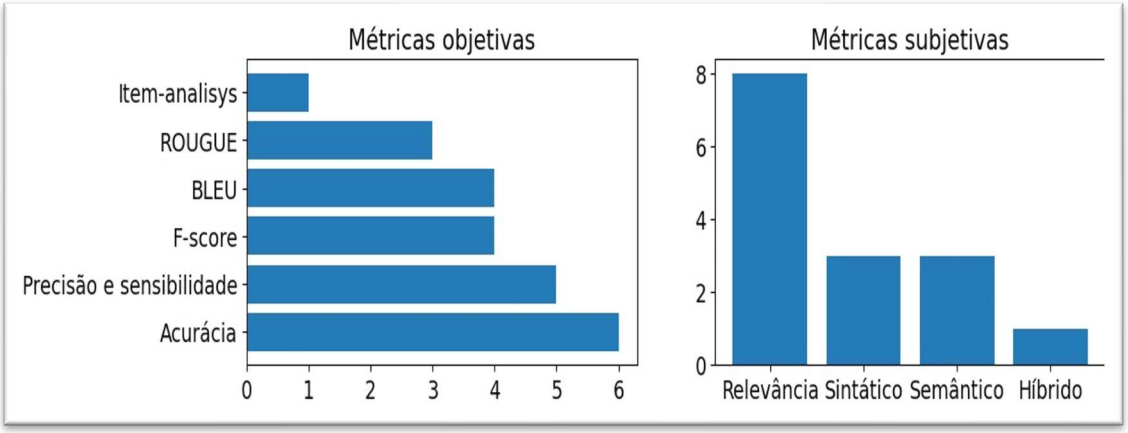
## 4.3 Quais os métodos de avaliação?

Dos artigos selecionados para a revisão, 64,5% utilizaram humanos (Método manual) para realizar a validação das questões geradas. Já 16,1% utilizaram métodos de validação automáticos. Por fim, para 19,4% dos artigos não foi possível identificar os métodos de avaliação. Avaliar se uma questão é adequada ou não que pode ser fácil para humanos, mas envolve ainda muita subjetividade para a máquina.

As métricas mais utilizadas para avaliar as questões foram: Acurácia, Precisão, Sensibilidade e F-Score, seguido por BLEU e Rogue. Conforme figura 6, na imagem da esquerda, foi utilizada a métrica acurácia em 6 pesquisas; a precisão e sensibilidade foi utilizada por 5; a métrica *FScore* foi utilizada por 4; a métrica BLEU foi utilizada por 4 pesquisas; a métrica ROUGE por 3; e por fim, uma pesquisa utilizou *item-analisys*.

Também foram utilizados critérios mais subjetivos na avaliação das questões, são eles: análise sintática, análise semântica, qualidade com relevância e legibilidade com usabilidade. Conforme a **figura 6**, na imagem da direita, foi utilizada análise sintática em 3 pesquisas; a análise semântica também foi utilizada por 3; critérios como qualidade ou relevância (no gráfico chamado de relevância) foram utilizados por 8; por fim, uma pesquisa utilizou quatro métricas: legibilidade, semântica, aceitabilidade e usabilidade (no gráfico chamado de híbrido).

**Figura 6.** Contagem das métricas objetivas e subjetivas.



### 4.4 Limitações

Esta seção apresenta as limitações dos 9 artigos mais recentes incluídos na revisão.

#1 (DAS et al., 2022): Este estudo apresenta um sistema automatizado de geração de perguntas do tipo subjetiva e propõe um sistema de avaliação destas respostas. Para a geração das perguntas, palavras-chave mais frequentes são selecionadas do texto, e a partir dessas palavras questões do tipo subjetiva são criadas. Para avaliar as respostas é utilizado similaridade de *String*, similaridade semântica e similaridade de palavra-chave.

A principal limitação do método proposto é a seleção das palavras-chave utilizadas na geração das perguntas. A seleção do número exato de palavras-chave configurável é sensível ao conjunto de dados. Em alguns casos, aumentar o número de palavras-chave melhora o desempenho.

#2 (RAO et al., 2022): Este estudo utilizou um modelo que gera 3 tipos de perguntas: Perguntas Complexas, Perguntas de Múltipla Escolha e Preencher os espaços em branco. Esse modelo facilita o processo de correção de respostas para os professores, pois qualquer pessoa com a folha de respostas fornecida pelo modelo pode verificá-lo. Como o modelo recebe a entrada como texto e cria perguntas, qualquer pessoa sem o domínio do texto de entrada pode usar o sistema e criar perguntas sobre o tópico em segundos.

O modelo faz pares de perguntas e respostas eficazes a partir de um texto sem perder tempo e requer menos esforço do que a forma tradicional de fazer perguntas. Ele também cria diversas perguntas que são fáceis de entender e gramaticalmente corretas. O modelo pode ser usado para fins educacionais, por exemplo, para praticar um capítulo específico por meio de questionários. Os alunos tendem a se interessar mais quando há atividades externas e ensino baseado em jogos envolvidos. O modelo foi bem-sucedido na geração de uma variedade de questões do tipo Questões Complexas, Questões de Múltipla Escolha e Questões de Preenchimento de Espaços. Ao reconhecer as entidades nomeadas no texto de entrada fornecido e usar os algoritmos mencionados no artigo, o modelo foi bastante eficaz na produção de questões de qualidade.

Uma limitação encontrada é que a correção sintática ainda pode ser melhorada, trabalhando com os algoritmos no pós-processamento. A eficiência em obter distratores (as outras respostas, fora as corretas) de qualidade também pode ser aprimorada, aumentando a variedade de tuplas no conjunto de dados.

#3 (TSAI et al., 2021): Este estudo concentrou-se no processamento de linguagem natural, e implementou um AQG embutido em um aplicativo para fins educacionais. Após um semestre de experimento, os autores concluíram que: (1) o método de geração de perguntas baseados em sintaxe e semântica teve bom desempenho em termos de qualidade das perguntas geradas; (2) os testes repetidos (gerados com AQG) foram benéficos à memória de longo prazo do aluno.

Foram encontradas duas limitações: uma foi a qualidade dos materiais didáticos e a outra foi a configuração do formato que pode afetar a qualidade das perguntas geradas pela máquina, na geração de tipos de perguntas, pois só foram usadas apenas um tipo de pergunta.

#4 (SONIA et al., 2021): Este estudo apresentou um protótipo para geração de pergunta com resposta curta, automaticamente, a partir de videoaulas utilizando Redes

Neurais. O texto gerado a partir da transcrição do vídeo é recuperado e aplicado para gerar perguntas e respostas. Basicamente, o AQG captura uma videoaula do YouTube e gera o texto. A principal vantagem dessa abordagem é que ela pode ser usada para autoavaliação de alunos ou qualquer tipo de aplicativo educacional e pode ser usada em várias palestras e aprendizados online.

Uma limitação é que este AQG não trabalha a qualidade da geração de questões. O algoritmo do sistema necessita de pré-processamento após a geração do texto a partir da videoaula. Pois nem todo o texto gerado é útil na geração da Pergunta e Resposta.

#5 (DAS, 2021): Neste estudo foi proposta uma abordagem para gerar questões de múltipla escolha com distratores (respostas alternativas que não são a resposta correta) gerados automaticamente. Primeiramente, as frases simples foram identificadas e classificadas com base nas palavras-chave. A frase da pergunta (raiz) é gerada substituindo a chave de resposta por uma palavra apropriada ou um espaço em branco (lacuna). Para gerar o conjunto de opções, os distratores são selecionados de forma que estejam intimamente relacionados à chave de resposta. A técnica de agrupamento baseada em recursos é empregada para selecionar o conjunto candidato de distratores. A categoria de distratores é identificada automaticamente usando o método cotovelo no agrupamento *K-means*. A distância de *Levenshtein* e a análise semântica latente são exploradas para combinar similaridade de *String* e similaridade semântica para selecionar o conjunto final de distratores. Os resultados sugerem que a abordagem proposta pode produzir questões de boa qualidade e ser útil em vários níveis de avaliação. Uma limitação encontrada é que ele não lida com questões subjetivas.

#6 (MORÓN et al., 2021): Neste estudo foi desenvolvido um protótipo para Geração Automática de Perguntas implementando um conjunto de regras, utilizando textos que podem ser usados para o ensino de inglês como segunda língua para crianças de nível iniciante. A ferramenta permite que os professores insiram um texto e gere um conjunto de pares de perguntas/respostas classificados por uma heurística que tenta adicionar diversidade ao conjunto. Os alunos podem abrir um exercício, responder às perguntas e receber feedback imediato sobre suas respostas. Foi realizada uma avaliação inicial da geração de perguntas e respostas com resultados animadores, principalmente para as questões “o quê” e “quem”, embora sejam necessárias mais pesquisas a esse respeito.

As limitações encontradas são a falta de inclusão de novos tipos de perguntas e exploração de técnicas de Machine Learning, porém o artigo cita que irá fazê-lo em trabalhos futuros através da utilização do protótipo e sua funcionalidade de edição para construir um corpus com textos e questões associadas, selecionando textos adequados para o nível iniciante. A estratégia de trabalho priorizou ter um protótipo rápido que implementasse um conjunto inicial de regras, para que pudesse ser utilizado para a geração de um corpus de textos com questões associadas.

#7 (GANGOPADHYAY; RAVIKIRAN, 2020): Neste estudo foi apresentado um método para geração de diversas questões propondo um módulo chamado “*Focus Generator*”. O objetivo deste AQG é gerar perguntas de diferentes níveis de dificuldade correspondentes a uma determinada resposta. Para gerar essas diversas perguntas, o módulo gera três conjuntos diferentes de foco e as perguntas são criadas com base nesses conteúdos de foco.

O módulo desenvolvido também adiciona a capacidade de gerar automaticamente as *tags* de resposta ao contrário de outros modelos em que as respostas devem ser fornecidas manualmente. O modelo obteve uma melhor pontuação BLEU e menos tempo de treinamento do que outros trabalhos anteriores. Ele também simplifica a arquitetura usando apenas um módulo. O modelo funciona como um *plug and play* onde qualquer corpus pode ser usado para derivar inferência do modelo.

Uma limitação encontrada é a dificuldade em justificar a diversidade das questões geradas. Isso se deve à falta de intervenção humana neste processo. Além disso, frases extremamente complexas com menos palavras-chave não geram perguntas diversas devido à falta de conteúdo de foco significativo. Como trabalho futuro, propõe-se usar arquiteturas de última geração como BERT e GPT para esta tarefa.

#8 (LEITE et al., 2020): Estes autores apontaram um aspecto central que é entender qual abordagem nos permite gerar perguntas de maior qualidade. A resposta certa para isso não é trivial. Com relação ao método baseado em sintaxe, podemos verificar que ele considera sua estrutura sintática, ou seja, posições de palavras e *tags*. Outra vantagem é que esta abordagem, quando combinada com entidades, permite a criação de questões muito específicas. Por outro lado, o modelo é claramente dependente do desempenho do componente da entidade.

Uma limitação é que este modelo com entidades, identifica apenas três entidades (pessoas, locais e empresas). O método baseado em semântica parece muito promissor, porque captura informações semânticas das frases extraídas. Usando seus papéis semânticos podemos verificar que esta abordagem fornece um nível mais profundo de análise quando comparado ao método baseado em sintaxe. Essa abordagem tem a vantagem de levantar questões sobre informações relevantes, como horários, causas, costumes, locais e propósitos. Esse tipo de informação permite a geração de uma variedade de perguntas, de diferentes contextos.

Outra abordagem utilizada realiza a análise de dependência e gera o menor número de perguntas, mas levanta uma possibilidade: a capacidade de lidar com novas funções sintáticas para mais segmentos dentro de nossas frases (além do complemento direto e indireto). A identificação do verbo também é crucial, ao permitir identificar caracterizações de coisas e pessoas em sentenças declarativas. Uma limitação desta abordagem é que ela depende da correta identificação dos rótulos.

#9 (BACHIRI; MOUNCIF, 2020): Neste estudo foram estabelecidos critérios para a escolha de um sistema de gestão de aprendizagem adequado e integrar novos processamentos automáticos de linguagem natural, utilizando técnicas de inteligência artificial para tornar os MOOCs mais atrativos. Para isso, foi criado um plugin de AQG que utiliza gamificação, transformando o curso em um jogo com pontos e medalhas. Este sistema de AQG gera o seguinte componente: o radical (a pergunta), uma palavra alvo, o qual é a resposta chave, e um trecho de leitura como resumo do texto e quatro distratores. Na pesquisa, foi proposto um modelo que permite integração com diversos sistemas virtuais de aprendizagem com a integração do plugin desenvolvido. Uma limitação identificada é o modelo ser incapaz de gerenciar textos complexos e encontrar os distratores adequados automaticamente.

### 4.5 Oportunidades de pesquisa

A geração automática de questões é uma área de pesquisa com bastante potencial, especialmente quando aplicada à educação. Embora as abordagens baseadas em regras (*Rule-Based*) sejam as mais utilizadas, observamos uma crescente adoção de técnicas de aprendizado de máquina (*Machine Learning*). Novas pesquisas podem explorar o uso de redes neurais e outras metodologias avançadas para aprimorar a qualidade e a eficácia das questões geradas automaticamente.

Além disso, a disponibilidade e o desenvolvimento de conjuntos de dados abertos representam uma necessidade crucial para o progresso da área. Conjuntos de dados padronizados e acessíveis permitem uma comparação justa entre diferentes abordagens e promovem a replicabilidade dos estudos. Pesquisas futuras podem focar na criação e utilização de novos conjuntos de dados, bem como na aplicação de métricas consistentes para garantir a comparabilidade entre as diversas técnicas empregadas.

A avaliação automática das questões geradas é outra área dos AQGs que pode ser pesquisada. A dificuldade das máquinas em identificar e avaliar a qualidade das questões limita a automatização completa do processo, exigindo avaliações manuais adicionais. Melhorias nas técnicas de avaliação automática são essenciais para superar essas limitações, proporcionando uma análise mais precisa e eficiente das questões geradas.

Explorar a diversidade de artefatos que podem alimentar os sistemas de AQG também se apresenta como uma promissora linha de pesquisa. Gráficos de conhecimento, bancos de dados, sites, imagens, vídeos e textos em linguagem natural são apenas alguns exemplos de entradas que podem ser utilizadas. A capacidade de gerar diferentes tipos de questões, tanto objetivas quanto subjetivas, a partir de diversas fontes de dados, pode ampliar significativamente as aplicações e a utilidade dos sistemas de AQG.

Finalmente, as métricas atuais, como Acurácia, precisão, sensibilidade, F-Score, BLEU e ROUGE, apresentam limitações que podem ser superadas com o desenvolvimento de novos critérios de avaliação. Pesquisas futuras devem buscar métricas que avaliem de maneira mais abrangente e precisa a qualidade e a relevância das questões geradas, contribuindo para a evolução contínua desta área de estudo.

## 5. Conclusão

Existe uma crescente demanda de AQGs para exames competitivos e de avaliação educacional, especialmente em ambientes virtuais de aprendizagem. A geração automática de perguntas e respostas a partir de diversos artefatos de entrada é uma área de pesquisa popular entre educadores e pesquisadores de processamento de linguagem natural. Neste artigo realizamos uma revisão sistemática da literatura para extrair dados que auxiliem os pesquisadores em busca de novas oportunidades de pesquisa. Extraímos as principais abordagens, os conjuntos de dados utilizados e as principais métricas de avaliação.

Nos artigos selecionados nessa revisão o uso da abordagem *Rule-Based* vem perdendo espaço para outras abordagens que usam *Machine Learning*. O uso de conjuntos de dados abertos possibilita uma comparação entre as abordagens, sendo desejado seu uso quando possível. Porém, mesmo com o uso de dados abertos, as mesmas métricas devem ser utilizadas para comparabilidade entre as abordagens.

Além disso, a dificuldade da máquina em identificar uma boa questão impede comparações automáticas da abordagem, trazendo um esforço adicional para avaliação manual das questões geradas. Os critérios subjetivos da análise da qualidade da questão estão relacionados a própria dificuldade de inferência de texto em linguagem natural pelo computador. Esperamos que esta revisão auxilie os futuros pesquisadores interessados na construção de questões automáticas e que possa apontar novos caminhos de pesquisa.

## 6. Disponibilidade dos artefatos

Artefatos complementares, tais como o código-fonte para a geração dos gráficos deste artigo, a lista de todos os artigos e tabelas adicionais, podem ser acessados em: https://github.com/giseldo/artigo-sbie-rev-autorregulacao-artefatos

**Referências**

BACHIRI, Younes-aziz; MOUNCIF, Hicham. Applicable strategy to choose and deploy a MOOC platform with multilingual AQG feature. In: **2020 21st International Arab conference on information technology (ACIT)**. IEEE, 2020. p. 1-6.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward.Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.

CAO, Zhen; TATINATI, Sivanagaraja; KHONG, Andy WH. Controllable question generation via sequence-to-sequence neural model with auxiliary information. In: **2020 International Joint Conference on Neural Networks (IJCNN)**. IEEE, 2020. p. 1-7.

DAS, Bidyut et al. Automatic question generation and answer assessment for subjective examination. **Cognitive systems research**, v. 72, p. 14-22, 2022.

ELSAHAR, Hady; GRAVIER, Christophe; LAFOREST, Frederique. Zero-Shot Question Generation from Knowledge Graphs for Unseen Predicates and Entity Types. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. 2018. p. 218-228.

FAN, Zhihao et al. A Question Type Driven Framework to Diversify Visual Question Generation. In: **IJCAI**. 2018. p. 4048-4054.

FAN, Zhihao et al. A reinforcement learning framework for natural question generation using bi-discriminators. In: **Proceedings of the 27th International Conference on Computational Linguistics**. 2018. p. 1763-1774.

GANGOPADHYAY, Susmita; RAVIKIRAN, S. M. Focused questions and answer generation by key content selection. In: **2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)**. IEEE, 2020. p. 45-53.

GHIDINI, Chiara et al. **The semantic web-ISWC 2019**. Springer International Publishing, 2019.

INDURTHI, Sathish Reddy et al. Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**. 2017. p. 376-385.

JURAFSKY, D.; MARTIM J. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. **Online manuscript** released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3.

KITCHENHAM B Guidelines for performing Systematic Literature Reviews in Software Engineering, Version 2.3, **EBSE Technical Report** EBSE-2007-01 (2007), Keele University and University of Durham

KUMAR, Vishwajeet et al. Difficulty-controllable multi-hop question generation from knowledge graphs. In: **The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18**. Springer International Publishing, 2019. p. 382-398.

LEITE, Bernardo et al. Factual question generation for the Portuguese language. In: **2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)**. IEEE, 2020. p. 1-7.

MORÓN, Martín et al. A tool for automatic question generation for teaching english to beginner students. In: **2021 40th International Conference of the Chilean Computer Science Society (SCCC)**. IEEE, 2021. p. 1-5.

MOSTAFAZADEH, Nasrin et al. Generating natural questions about an image. **arXiv preprint arXiv:1603.06059**, 2016.

PAPINENI, Kishore et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**. 2002. p. 311-318.

PIWEK, Paul et al. T2D: Generating dialogues between virtual agents automatically from text. In: **Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7**. Springer Berlin Heidelberg, 2007. p. 161-174.

RAJPURKAR, P. Squad: 100,000+ questions for machine comprehension of text. **arXiv preprint arXiv:1606.05250**, 2016.

RAO, Pratiksha Rajesh et al. Generating QA from Rule-based Algorithms. In: **2022 International Conference on Electronics and Renewable Systems (ICEARS)**. IEEE, 2022. p. 1697-1703.

RUS, Vasile et al. The first question generation shared task evaluation challenge. In: **Proceedings of the 6th International Natural Language Generation Conference**. 2010. p. 251-257.

SERBAN, Iulian Vlad et al. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. **arXiv preprint arXiv:1603.06807**, 2016.

SONG, Linfeng; ZHAO, Lin. Domain-specific question generation from a knowledge base. **arX-iv.**, 2016.

SONIA, Sonam; KUMAR, Praveen; SAHA, Amal. Automatic question-answer generation from video lecture using neural machine translation. In: **2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)**. IEEE, 2021. p. 661-665.

TSAI, Danny CL et al. Automatic question generation for repeated testing to improve student learning outcome. In: **2021 International Conference on Advanced Learning Technologies (ICALT)**. IEEE, 2021. p. 339-341.

ZHANG, Ruqing et al. A review on question generation from natural language text. **ACM Transactions on Information Systems (TOIS)**, v. 40, n. 1, p. 1-43, 2021.