UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Alércio Charles Silva

Correção Automática de questões discursivas de resposta curta: Uma abordagem baseada em Extração de Informação

São Luís - MA

2023

Alércio Charles Silva

Correção Automática de questões discursivas de resposta curta: Uma abordagem baseada em Extração de Informação

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Curso de Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Antônio de Abreu Batista Jr

São Luís - MA

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Silva, Alércio Charles.
Correção Automática de questões discursivas de resposta curta: Uma abordagem baseada em Extração de Informação / Alércio Charles Silva. - 2023.
46 f.

Orientador(a): Antônio de Abreu Batista Jr.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís - MA, 2023.

1. Avaliação de Estudantes. 2. Correção automática. 3. Correção de questões discursivas. 4. Extração de Informação. 5. Processamento de linguagem natural. I. Batista Jr, Antônio de Abreu. II. Título.

Alércio Charles Silva

Correção Automática de questões discursivas de resposta curta: Uma abordagem baseada em Extração de Informação

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho Aprovado, em 12 de Janeiro de 2023:

Prof. Dr. Antônio de Abreu Batista Jr
Orientador
Universidade Federal do Maranhão

Prof. Dr. Luciano Reis Coutinho
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Robson da Cunha Santos
Examinador Externo
Instituto Federal Fluminense (IFF/Macaé)

São Luís - MA
2023

Agradecimentos

Agradeço primeiramente à Deus, por tudo em minha vida. Pela saúde e proteção, por me capacitar e por guiar meu passos nos caminhos da vida.

Agradeço em especial aos meus pais, Laércio e Ednalva, à quem devo tudo em minha vida. Agradeço pelo dom da vida, pela educação, pelos exemplos, pelos conselhos, por toda a dedicação e por todos os esforços feitos para garantir a minha educação e por todo suporte dado em toda minha vida.

Agradeço também aos meus avós, por todo amor e por toda sabedoria dada a mim. Ao meu avô Anastácio (in memoriam) a quem serei eternamente grato, pelos conselhos e ensinamentos, por ser meu maior exemplo de força e sabedoria e agradeço por toda sua atuação na formação do meu caráter como homem e cidadão. Agradeço também a minha avó Maria Carmem(in memoriam) por ser meu maior exemplo de amor e de dedicação à família e por ser minha principal inspiração em momentos de fraqueza.

Agradeço também a toda minha família, por todo apoio dado em momentos que precisei. Em especial aos meus tios Marylanda e Benonilson, por toda ajuda e auxílio dados em minha formação acadêmica. Agradeço também aos meus irmãos pelo companheirismo, aos meus tios, aos meus primos e todos os meus amigos.

Agradeço aos professores da Universidade Federal do Maranhão, pelo conhecimento e por todo suporte dado durante o período de graduação. Em especial, ao professor Antônio de Abreu Batista Jr, pelos conselhos, pela paciência, pela disponibilidade e por todo suporte dado durante o desenvolvimento do presente trabalho.

Agradeço por fim aos meus amigos, pela ajuda dada e pelas experiências compartilhadas durante o período de graduação, em especial a Alan marques, Pedro Moraes e Samuel Silva.

"Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer."

(Alan Turing)

Resumo

A avaliação de estudantes por meio de questões discursivas é um recurso pedagógico importantíssimo utilizado para estimar o nível de conhecimento retidos pelos alunos, para identificar as dificuldades do educando e posteriormente para aprimorar as ferramentas educacionais utilizadas pelo educador. Contudo, apesar de sua importância, sua aplicação resulta em algumas problemáticas, pois sua correção não é uma tarefa trivial. A medida em que cresce o número de alunos dentro de uma turma, tal tarefa torna-se inviável, consumindo um tempo excessivo do professor que poderia ser gasto em outras tarefas pedagógicas. Com o objetivo de auxiliar neste problema, este trabalho propõe um novo método de correção automática de questões discursivas utilizando uma abordagem baseada em extração de informação. Com o auxílio das bibliotecas mais recentes de Processamento de Linguagem natural e utilizando dados reais de alunos do ensino fundamental, foi computada a acurácia do método proposto e comparada contra o método do estado da arte. A abordagem proposta mostra uma clara vantagem (desempenho 12 por cento superior) sobre as técnicas do estado da arte que usam aprendizado de máquina, apresentando resultados animadores e encorajadores, como um índice de concordância substancial/forte (coeficiente Kappa 0,674) entre o método e avaliadores humanos.

Palavras-chave: Avaliação de estudantes, Extração de informação, Correção de questões discursivas, Correção automática, Processamento de Linguagem Natural.

Abstract

The students assesment throught essay questions is a very important pedagogical resource used to estimate the level of knowledge retained by students, to identify the student's difficulties and later to improve the educational tools used by the educator. However, despite its importance, its application results in some problems, because its correction is not a trivial task. As the number of students in a class grows, this task becomes unfeasible, consuming excessive teacher time that could be spent on other pedagogical tasks. In order to help with this problem, this work proposes a new method of automatic correction of essay questions using an approach based on information extraction. With the aid of the most recent Natural Language Processing libraries and using real data from elementary school students, the accuracy of the proposed method was computed and compared against the state-of-the-art method. The proposed approach shows a clear advantage (12 percent higher performance) over state-of-the-art techniques that use machine learning, showing interesting and encouraging results, such as a substantial/strong concordance index (Kappa coefficient 0.674) between the method and human grading.

Keywords: Student assessment, Information extraction, Essay questions grading, Automatic grading, Natural language processing.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

EI	*Extração de Informação*
PLN	*Processamento de Linguagem Natural*
POS	*Part-of-Speech (Parte do discurso)*

Sumário

1 Introdução

A avaliação da aprendizagem é um recurso pedagógico de suma importância que permite ao educador auxiliar o educando na busca de sua autoconstrução e de seu modo de estar na vida mediante a aprendizagens bem sucedidas, além disso dá ao educador a oportunidade de reconhecer a eficácia ou ineficácia de seus recursos pedagógicos utilizados, pois a mesma permite estimar a porcentagem de conhecimento retido pelos alunos, permitindo portanto a intervenção de correção dos rumos da atividade e dos seus resultados (LUCKESI, 2013).

Mesmo diante das concepções pedagógicas mais modernas, a forma mais comum de avaliação de aprendizagem continua sendo através da aplicação de provas ou testes escritos, pois as mesmas mostram-se instrumentos importantíssimos para identificar a realidade de conhecimento de cada aluno e verificar suas habilidades ou dificuldades de aprendizagem. As provas ou testes escritos podem ser compostas por questões objetivas que solicitam aos alunos que selecionem uma alternativa correta para a questão a partir de um conjunto apresentado e as questões dissertativas que exigem que os alunos construam uma resposta para a questão, a partir de seus conhecimentos (KAIPA, 2020).

A aplicação de provas com questões objetivas não é um problema, porém seu uso excessivo ou exclusivo no processo de avaliação é desaconselhado, pois as questões objetivas podem incentivar uma tendência superficial de aprendizagem, além disso os alunos tendem a adivinhar as respostas em alguns casos e são menos propensos a aplicar seus conhecimentos e raciocínio ao responder às perguntas (WOODFORD; BANCROFT, 2004). A utilização de questões discursivas no processo de ensino permite ao professor, avaliar os processos cognitivos mais elevados quando comparado a aplicação de questões objetivas, pois permitem avaliar capacidade de leitura, interpretação e escrita de um aluno.

Contudo, a correção manual das questões discursivas não é uma tarefa trivial, ao contrário das questões objetivas a correção desse tipo de questão demanda um consumo demasiado de tempo de trabalho do professor, onde em turmas com número elevado de alunos tal tarefa pode ser tornar inviável. Tal situação sobrecarrega o professor e reduz seu tempo, que poderia ser gasto em outras atividades como elaboração de aulas (SAKAGUCHI; HEILMAN; MADNANI, 2015).

Além disso, a correção manual de questões discursivas está a mercê de subjetividade humana, desde cansaço, preconceitos baseados na relação com alunos ou a simples ordenação dos testes resulta na presença de um viés tendencioso no processo de correção humana (HALEY et al., 2007).

Portanto, o uso de métodos computacionais na criação de ferramentas capazes

de automatizar a correção dessas avaliações são de grande valia, visto que tratam-se de ferramentas incansáveis e são isentas de subjetividade humana. A correção de questões objetivas é uma tarefa simples de se resolver com os métodos computacionais, visto que há apenas uma única simples resposta correta para cada questão, que facilita a comparação e atribuição de nota (BURROWS; GUREVYCH; STEIN, 2014).

Já a correção de questões discursivas é uma tarefa cuja solução completa é algo ainda não resolvido pelos métodos computacionais, sobretudo devido às inúmeras possibilidades de uso da língua natural para expressar um mesmo conceito, tal questão dificulta a execução das técnicas de PLN (Processamento de Linguagem Natural), pois a mesma encontra dificuldade em fazer a interpretação de conteúdo, representação de conhecimento, comparação e consequentemente a atribuição de notas.

Com o objetivo de encontrar uma solução para auxiliar o processo de correção automática de questões discursivas, o presente trabalho utiliza as ferramentas avançadas de processamento de linguagem natural disponibilizadas pelas bibliotecas mais recentes, para estudar e propor um método de correção automática que utiliza as técnicas de EI (Extração de Informação) para extrair e avaliar o conteúdo semântico presente nas respostas analisadas.

Objetivos

O objetivo geral deste trabalho consiste em propor um método computacional para a correção automática de questões discursivas baseado na aplicação de técnicas de EI e no uso de padrões sintáticos. E, como objetivos específicos:

- Identificar e construir regras de extração baseadas em padrões de características sintáticas rasas e profundas;
- Estudar e investigar as bibliotecas disponíveis para a execução de tarefas comuns ao PLN;
- Investigar a existência de bases de dados em português disponíveis para aplicação de testes em trabalhos que abordam correção automática;
- Comparar os resultados do método proposto com outros métodos do estado da arte;e
- Validar o método proposto.

2 Fundamentação Teórica

Neste capítulo, é apresentado o arcabouço de extração de informação e as principais etapas do pipeline de processamento de linguagem natural. Além disso, discute-se as duas principais formas de provas escritas: discursivas e objetivas, e por fim, é apresentado um coeficiente capaz de medir a correlação entre dois avaliadores.

2.1 Avaliações de estudantes

Existem muitas formas de avaliar os estudantes. Prova escrita é a mais comum. Existem dois tipos: (i) prova objetiva e (ii) prova subjetiva. Provas de múltipla escolha são um exemplo da primeira (AALAEI; AHMADI; AALAEI, 2016). Elas já trazem enunciadas as possibilidades de resposta, e o aluno deve escolher uma única opção. O esforço do avaliador para corrigir uma prova objetiva é quase nenhum. Além disso, neste tipo de avaliação, elimina-se qualquer viés humano no processo de correção.

Por outro lado, as questões subjetivas são mais duras de corrigir porque seus enunciados e respostas apresentam uma alta subjetividade, sendo menos diretos. O estudante apresenta sua resposta em um pequeno texto escrito em linguagem natural sobre um determinado assunto.

As questões subjetivas podem ser divididas em duas formas: discursivas de resposta longa e as discursivas de resposta curta. As discursivas de resposta longa exigem que o aluno dê a resposta em dois ou mais parágrafos onde a correção geralmente é focada em características baseadas no estilo de escrita, enquanto as discursivas de resposta curta contém respostas expressas entre uma ou três frases, e possuem correção baseada no conteúdo semântico. Mais especificamente, de acordo com Burrows, Gurevych e Stein (2014) uma questão de resposta curta é aquela que pode ser considerada como atendendo aos seguintes critérios específicos:

- A pergunta deve exigir uma resposta que relembre o conhecimento externo, em vez de exigir que a resposta seja reconhecida de dentro da pergunta.
- A pergunta deve exigir uma resposta dada em linguagem natural.
- O comprimento da resposta deve ser aproximadamente entre uma frase e um parágrafo.
- O foco principal da avaliação é o conteúdo da resposta, em vez do estilo de escrita.

- Um desenho de pergunta objetiva é necessário para restringir o nível de abertura em respostas abertas versus respostas fechadas.

Por isso, neste trabalho, o foco está em avaliar provas discursivas com respostas curtas. Provas deste tipo são muito vistas nos anos iniciais e finais do ensino fundamental. Essa decisão foi tomada para controlar a complexidade do processo de correção de provas, em especial, em circunstâncias em que isso for feito por um algoritmo.

2.2 Extração de Informação

A extração de informação (EI) é a subárea da inteligência artificial que trata da extração automática de informações estruturadas, como entidades, relacionamentos entre entidades e atributos que descrevem entidades de fontes não estruturadas (SARAWAGI, 2008). O objetivo geral dos sistemas de extração de informação não é interpretar o texto inteiro, mas apenas extrair suas partes e informações mais relevantes, essas informações são registradas em representações estruturadas denominadas *template*, que são definidas durante a construção do sistema e devem ser capazes de representar todas as informações de interesse do sistema. A figura 1 mostra o exemplo de um *template* preenchido com informações extraídas de um texto livre.

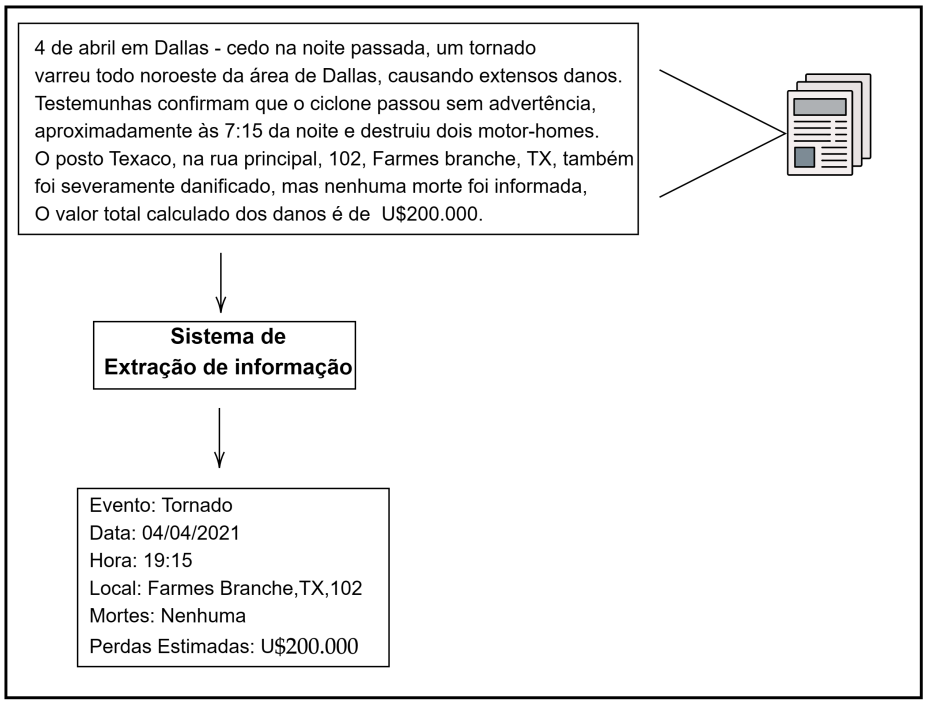

Figura 1 – Template de um sistema de extração de informação

De acordo com Appelt e Israel (1999) existem duas abordagens básicas para a construção de sistemas de extração de informação, a abordagem baseada em engenharia

de conhecimento e a abordagem baseada em treinamento automático. A abordagem de engenharia do conhecimento, também referida como abordagem baseada em regras (SYNTACTIC..., 2021), é caracterizada pelo desenvolvimento manual de regras gramaticais por um "engenheiro do conhecimento", que sozinho ou em consulta com um especialista no domínio de aplicação, escreve regras para o componente do sistema que irá marcar ou extrair as informações procuradas.

Já na abordagem de treinamento automático, não é necessária a existência de um especialista por perto com conhecimento detalhado de como o sistema de extração funciona ou como escrever regras para construção do sistema de EI, é necessário apenas ter alguém que saiba o suficiente sobre o domínio e a tarefa para pegar um corpus e anotar os textos adequadamente para as informações que estão sendo extraídas. Uma vez anotado um corpus de treinamento adequado, o algoritmo de treinamento pode ser executado, resultando em um conjunto de informações que um sistema pode empregar na análise de novos corpus de textos (APPELT; ISRAEL, 1999).

Outra abordagem para obter dados de treinamento é de forma supervisionada, onde o sistema deve interagir com o usuário durante o processamento de um texto, onde o mesmo tem permissão para indicar se o as hipóteses do sistema sobre o texto estão corretas, caso contrário, o sistema modifica suas próprias regras para acomodar as novas informações, e portanto aumenta sua eficiência na tarefa de extração.

Abordagem baseada em Engenharia do Conhecimento

A abordagem baseada em engenharia do conhecimento é comumente a abordagem mais utilizada na construção de sistemas de extração de informação, sobretudo, devido ao fato da abordagem baseada em engenharia do conhecimento tender a produzir maior desempenho, porque a experiência humana geralmente resulta em padrões e regras de extração mais precisas (ZHOU; EL-GOHARY, 2017), além disso a abordagem baseada em aprendizado de máquina requer um uma quantidade razoavelmente grande de dados de treinamento para obter uma performance aceitável, de acordo com (APPELT; ISRAEL, 1999) em alguns casos pode-se desejar desenvolver um sistema de extração para um tópico para o qual existem poucos exemplos relevantes em um corpus de treinamento, tais situações valorizam a intuição humana de um bom criador de regras para construção manual das mesmas.

O nível de desempenho alcançado por um sistema de extração que utiliza essa abordagem, está diretamente ligada a habilidade do engenheiro de conhecimento em identificar e definir as regras de extração. As regras de extração são construídas a partir da análise humana dos recursos de texto em um conjunto relativamente pequeno de corpus de texto (que é análogo aos dados de treinamento no caso da abordagem de aprendizado de máquina), após a análise o engenheiro deve identificar a partir da sua intuição os padrões

de texto em termos dos recursos de texto e em seguida, desenvolver as regras de extração com base nos padrões definidos (ZHOU; EL-GOHARY, 2017).

Para atingir um alto desempenho, a construção dessas regras geralmente requer um processo iterativo onde após a definição do conjunto de regras inicial, o sistema é executado em um corpus de textos de treinamento e a saída é examinada para ver onde as regras são geradas de forma insuficiente ou excessiva (APPELT; ISRAEL, 1999).

Arquitetura de sistemas de Extração de Informação

Operacionalmente, os sistemas de extração necessitam de dois principais processos para identificar e interpretar o texto de destino, o pré-processamento de documentos e a aplicação de regras de extração (normalmente expressões regulares ou padrões) (NéDELLEC; NAZARENKO; BOSSY, 2009). De acordo com Appelt e Israel (1999) a arquitetura e os módulos que compõem um sistema de extração de informação, variam de acordo com a tarefa que o mesmo se propõe a resolver. Porém existem módulos e elementos compartilhados por quase todos os sistemas de extração, independentemente da tarefa que o mesmo se propõe a resolver ou do paradigma de construção de regras que o mesmo utiliza. De maneira genérica a arquitetura dos sistemas de EI apresenta os módulos de processamento apresentados na figura 2.

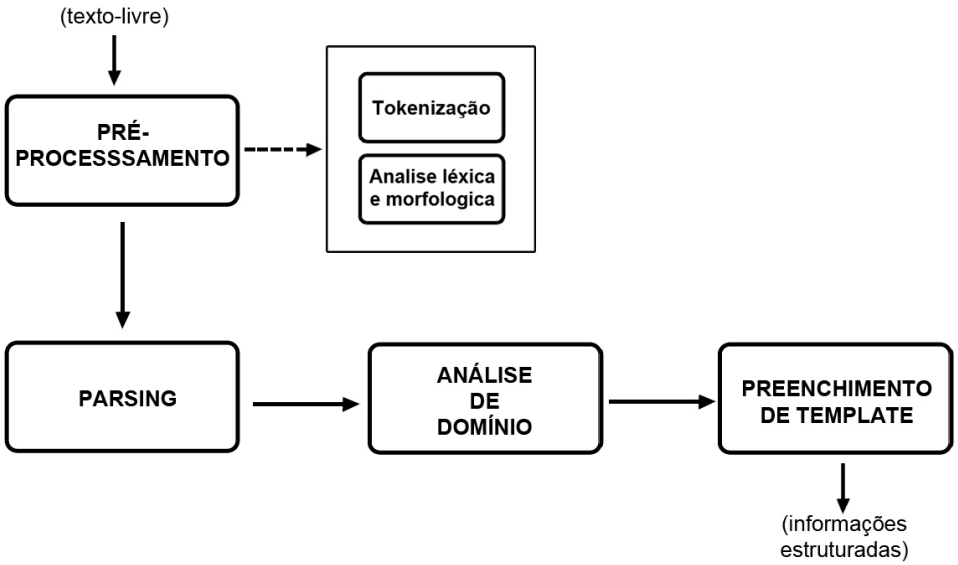

Figura 2 – Fases do processamento de um sistema de EI

O pré-processamento do texto é o primeiro processo comumente executado em sistemas de extração de informação, o pré-processamento dos dados ajuda a garantir que o processo de extração seja o mais preciso e eficiente possível. Durante essa etapa

são executadas diversas tarefas comuns ao processamento de linguagem natural como tokenização, normalização, marcação de parte da fala e lematização. As tarefas e processos referentes ao processamento de linguagem natural serão melhor descritas na seção 2.3.

Parsing

Durante a etapa de parsing os sistemas de EI realizam a análise sintática das sentenças nos textos. A análise sintática decompõe os termos constituintes de uma oração e verifica a função de cada termo dentro de uma sentença, onde esses termos são classificados de forma diferente de acordo com a função que a mesma desempenha, portanto a análise sintática é o processo de identificação e investigação da sintaxe das palavras nas orações (SANTANA, 2020).

A identificação dos aspectos da estrutura sintática dos textos simplifica as fases subsequentes de extração de fatos, pois os argumentos a serem extraídos geralmente correspondem a sintagmas nominais no texto, e as relações a serem extraídas geralmente correspondem a relações funcionais gramaticais (GRISHMAN, 1997).

No processamento de linguagem natural, uma das formas de análise sintática é o parsing. No processo de parsing, é construída uma árvore sintática para cada sentença, de modo a identificar as dependências sintáticas entre as palavras constituintes da sentença. A figura 3 apresenta uma árvore de dependência gerada pela biblioteca *Spacy*.

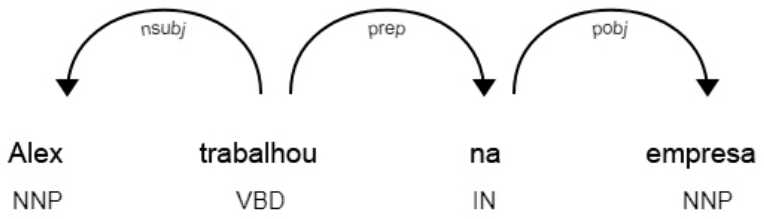

Figura 3 – Arvore de dependência sintática

Análise de domínio

Para que ocorra a extração de fatos e eventos de interesse, o sistema necessita de uma análise sobre o domínio de aplicação. Durante a etapa de analise de domínio, é feita uma análise sobre as informações textuais obtidas durante a execução dos módulos anteriores e são definidas e aplicadas as regras de extração necessárias para a extração das informações de interesse.

A primeira tarefa comum a sistemas de EI executada durante o módulo de análise domínio é a de resolução de correferências. As entidades relevantes de uma aplicação serão referidas de muitas maneiras diferentes ao longo de um determinado texto (APPELT; ISRAEL, 1999). Portanto, a análise de correferência de um texto visa determinar quando

uma frase nominal se refere a mesma entidade que outra frase nominal, ou seja identificar todas as maneiras pelas quais a entidade é nomeada ao longo do texto.

A principal tarefa realizada durante o módulo de análise de domínio é combinar todas as informações coletadas dos componentes anteriores e criar quadros completos que descrevem eventos e relacionamentos entre entidades (FELDMAN; SANGER, 2006). A principal abordagem utilizada em sistemas de EI para a extração dessas informações é baseada na aplicação de regras de domínio. As regras de domínios são regras obtidas a partir da análise de um engenheiro do conhecimento (engenharia do conhecimento) ou automaticamente (treinamento automático).

As regras de extração especificam as condições que o texto pré-processado deve atender e como os fragmentos textuais relevantes podem ser interpretados para preencher os *templates*. A construção de padrões para um sistema baseado em regras pode utilizar dos mais diversos recursos resultantes do processamento de linguagem natural, as características lexicais, gramaticais, sintáticas, semânticas ou de discursos dos termos que compõem as sentenças podem ser utilizadas na criação de padrões que podem ser aplicadas no processo de extração.

A tabela 1 exemplifica a detecção de padrões léxicos-semânticos em um conjunto de sentenças e a respectivas tuplas extraídas a partir do reconhecimento dessas regras, os padrões estão escritos na forma de expressões regulares que detectam tanto a presença e disposição de classes gramaticais como NP(nome próprio) ou V(verbo) na sentença, como a presença de termos comumente presente em sentenças argumentativas como 'Na' ou 'Dos'. A presença e ordenação desses termos ou dessas classe gramaticais presentes na sentença validam a regra e extraem informações estruturadas das mesmas.

Tabela 1 – Padrões léxico-semânticos usados para extrair relações de sentenças

Sentença	**Padrão**	**Tupla**
Um ramo central da metafísica é a ontologia	UM [NP] DA [NP] É [NP]	(ontologia, é um ramo central, metafísica)
Biologia é o estudo dos seres vivos	[NP] É O [EXP] DOS [NP]	(Biologia, estudo, seres vivos)
Aristóteles nasceu em Stageira	[NP] [V] (EM/NA) [NP]	(Aristóteles, nasceu, Stageira)

Preenchimento de template

Os sistemas de extração de informação não tem como objetivo interpretar todo o documento que está sendo processado, mas apenas identificar os trechos desse documento que preenchem corretamente um formulário pré-especificado que contém informações de interesses, esses formulários de saída que representam um conjunto de informações estruturadas são chamados de *templates*. Os *templates* são estruturas que definem tipos específicos de evento com um conjunto de funções semânticas (ou slots) que representam as entidades típicas envolvidas em tal evento (CHAMBERS; JURAFSKY, 2011).

Durante o módulo de preenchimento de *template*, os sistemas de EI reúnem todas as informações referentes a entidades, eventos ou relações identificadas durante o processamento do módulos anteriores e instanciam essas informações extraídas dentro dos determinados *slots* que representam as informações de interesse definidas pelo *template*. A figura 1 mostra o exemplo de um *template*, preenchido após o processamento do sistema de EI, contendo um conjunto de informações acerca de um desastre natural.

2.3 Processamento de linguagem natural

A linguagem natural é qualquer linguagem desenvolvida naturalmente pelo ser humano, de forma não premeditada e sem planejamento consciente, com objetivo de permitir a comunicação entre indivíduos. O Processamento de Linguagem Natural (PLN) é um campo da Inteligência Artificial cujo o objetivo é compreender, analisar e gerar a língua natural para os Humanos (PINTO, 2015), o processamento de linguagem natural permite que os computadores entendam e extraiam informações expressas em linguagem natural como humanos, utilizando técnicas para transformar a linguagem humana não estruturada em representações mais formais, mais facilmente manipuláveis por programas de computador.

O grande desafio do processamento de linguagem natural é lidar com problemas decorrentes das regras gramaticais e ambiguidades presentes na língua natural tanto na forma falada, quanto na forma escrita, o papel do PLN é desenvolver técnicas capazes de lidar com a forma que falamos, superando nossos erros de ortografia, ambiguidades, abreviações, gírias e expressões coloquiais. Para tratar de textos não estruturados expressos em linguagem natural, a IE aplica técnicas de PLN como (POS) tagging, marcação, análise morfológica, parsing, tokenização, e outras técnicas para reconhecer informações de dados não estruturados e formalizá-los em dados estruturados.

Níveis de entendimento da linguagem natural

A definição das etapas ou das fases de processamento da linguagem natural se baseia nos conhecimentos linguísticos necessários à compreensão da linguagem. Segundo Kostareva, Chuprina e Nam (2016), qualquer sistema de processamento de linguagem natural deve possuir os componentes para as seguintes etapas de análise: tokenização, análise morfológica, análise sintática e análise semântica.

Tokenização

A tokenização é o primeiro passo executado em qualquer pipeline de PLN, a etapa de tokenização, visa identificar e decompor o texto em pequenas partes para facilitar e permitir a execução de etapas futuras do processamento de linguagem natural como

por exemplo a análise morfológica, a principal tarefa realizada durante essa etapa é a segmentação do texto de consulta em pequenas unidades chamadas tokens.

Análise morfológica

A morfologia é o estudo da estrutura e da formação das palavras. Assim, a análise morfológica é a etapa responsável pela análise de palavras isoladas, onde cada palavra é classificada em uma categoria de acordo com as regras que regem a língua portuguesa (substantivo, adjetivo, numeral, verbo, advérbio, pronome, artigo, preposição, conjunção e interjeição) (FARINON, 2015).

Análise sintática

A sintaxe é a parte da gramática que estuda as palavras enquanto elementos de uma frase, portanto a análise sintática decompõe os termos constituintes de uma sentença e verifica a função de cada termo dentro da mesma, onde esses termos são classificados de forma diferente de acordo com a posição que ocupam e a função que o mesmo desempenha.

Análise semântica

A semântica é a área da linguística focada no significado das palavras bem como na relação entre o sentido e a estrutura desses elementos dentro de uma sentença. Portanto, de maneira bem geral, análise semântica é dar significado às estruturas criadas na análise sintática.

2.4 Métrica kappa de Cohen

Proposto por Jacob Cohen em 1960, trata-se de um método estatístico para avaliar o nível de concordância ou reprodutibilidade entre dois conjuntos de dados. O coeficiente Kappa de Cohen é geralmente utilizado para medir a concordância entre avaliadores que estão julgando uma mesma coisa. Geralmente é considerada uma medida mais robusta do que o simples cálculo percentual de concordância, pois o coeficiente de Kappa leva em consideração a probabilidade de que os avaliadores podem concordar por acaso.

Semelhante aos coeficientes de correlação, o coeficiente Kappa de Cohen pode variar de -1 a +1, onde 0 representa a quantidade de concordância que pode ser esperada da chance aleatória e 1 representa concordância perfeita entre os avaliadores (MCHUGH, 2012).

Coeficiente kappa ponderado

O Kappa ponderado é uma variante do Kappa de Cohen, que permite o uso de esquemas de ponderação para levar em conta a proximidade de concordância entre as categorias, bastante utilizado em situações que possuem variáveis ordinais ou classificadas. O objetivo desse coeficiente é distinguir uma discordância grave(por exemplo, um avaliador classifica uma questão como 1 e outro avaliador como 3) de uma discordância leve(por exemplo, um avaliador classifica uma questão como 1 e outro avaliador como 2).

3 Trabalhos Relacionados

A pesquisa referente ao uso de métodos computacionais para correção automática de questões discursivas teve início na década de 60 com o PEG (Project Essay Grader) por Page (1966). O sistema contudo, não obteve resultados satisfatórios, fazendo com que o mesmo fosse alvo de críticas pela comunidade acadêmica. Com a evolução dos métodos de PLN e do campo de extração de informação, a pesquisa no campo de correção automática aumentou significativamente e novas abordagens focadas em conteúdo semânticos puderam surgir. Como em Mitchell et al. (2002) que abordou o problema usando técnicas de extração de informações. As respostas são representadas através de modelos sintáticos-semânticos. Cada modelo sintático-semântico define uma forma particular de resposta: aceitável ou não aceitável. Os autores adotaram duas abordagens de avaliação: a cega e a moderada. A primeira apresenta um processo de correção totalmente automático, enquanto na segunda existe a possibilidade de intervenção humana, onde um avaliador pode revisar o modelo gerado. A taxa de correlação de atribuição de notas entre o sistema e avaliadores humanos alcançou 92,5%.

Mais recentemente, Thomas (2003) explorou o problema de correção automática utilizando uma abordagem baseada em padrões booleanos. A abordagem consistia em um algoritmo simples de correspondência de frases-chave entre a resposta de um aluno a uma pergunta e a solução definida pelo avaliador. O método apresentou-se adequado para contextos simples. O sistema eletrônico de correção desenvolvido obteve uma correlação média de 0.86 com três avaliadores humanos.

Por outro lado, Hahn e Meurers (2012) introduziram uma abordagem semântica para correção de questões de interpretação de leitura, que utilizava o método LRS(Lexical Resource Semantics) para criar representações abstratas de textos, capazes de representar distinções semânticas com precisão, robustez e modularidade. As representações LRS das respostas do professor e do aluno são modeladas como gráficos e um alinhamento baseado em limite é realizado para detectar significados equivalentes. A abordagem obteve uma acurácia de 86,3% para um conjunto de 1032 respostas de alunos.

Apesar de ser uma área de pesquisa ainda pouco explorada, a correção automática de questões discursivas que explora conjuntos de dados compostos por respostas escritas em língua portuguesa, tem recebido muita atenção nos últimos seis anos. Dentre os trabalhos que exploram dados em português, podemos citar o trabalho de Galhardi, Souza e Brancher (2020) que apresenta uma abordagem baseada em aprendizado de máquina, ao todo seis abordagem de modelagem foram testadas e um modelo final foi criado combinando as seis abordagens, o modelo criado apresentou coeficiente de correlação satisfatório entre o

sistema e a correção humana. Além disso, o trabalho de Galhardi, Souza e Brancher (2020) criou e disponibilizou publicamente o primeiro conjunto de dados contendo respostas escritas em língua portuguesa, contendo ao todo mais de 7000 respostas, facilitando o surgimento de pesquisas futuras na área de correção automática. O conjunto de dados disponibilizado foi utilizado no desenvolvimento do presente trabalho.

Dentre outros trabalhos de correção que exploram dados em português, podemos citar Santos (2016) que apresentou uma abordagem baseada em bigramas, que utiliza Latent Semantic Analysis(LSA) na avaliação automática de respostas curtas. Podemos também citar o trabalho de Oliveira, Pozzebon e Santos (2020) que propôs uma solução para a criação de um sistema de correção automática de questões discursivas baseada no uso de medidas similaridade como Cosine similarity, para medir os valores semânticos de cada palavra presente na resposta do aluno e do professor, o sistema de correção obteve um erro relativo de 15% e uma acurácia de 88,7%, resultados interessantes e animadores, porém obtidos através de experimentos aplicados sobre um conjunto relativamente pequeno de dados de teste.

Assim como em Mitchell et al. (2002) e em Hahn e Meurers (2012), o presente trabalho utiliza uma abordagem baseada em extração de informação, contudo diferentemente dos trabalhos citados, utiliza além de características sintáticas rasas, características sintáticas profundas para criação de regras de extração, bem como possui uma abordagem semelhante a sistemas de extração abertos, que extraem relações independentes de domínio.

4 Método Proposto

Neste capítulo, será descrita a metodologia utilizada para a realização das tarefas referentes à abordagem proposta, suas etapas e os motivos que levaram à sua escolha. Além disso, discorrerá sobre os materiais e ferramentas utilizadas durante o desenvolvimento do trabalho.

As questões discursivas de resposta curta geralmente incluem em suas soluções ideias específicas e conceitos factuais (BURROWS; GUREVYCH; STEIN, 2014), portanto, o uso das técnicas de EI no contexto da abordagem proposta, possibilita que essas ideias possam ser pesquisadas e modeladas por templates, ou seja de forma mais geral, permite extrair dados estruturados de fontes não estruturadas. Essas informações de interesses que serão buscadas e posteriormente comparadas dentro das respostas, tratam-se de argumentos binários que serão modelados na forma de tuplas do tipo: *(Argumento,Relação,Argumento)*.

O método proposto para correção de questões discursivas baseia-se em extração de informação, e segue as etapas descritas no fluxograma apresentado na Figura 4. No restante desta seção serão descritos os passos chaves de cada uma das etapas, bem como todas as sub-tarefas presentes em cada uma delas.

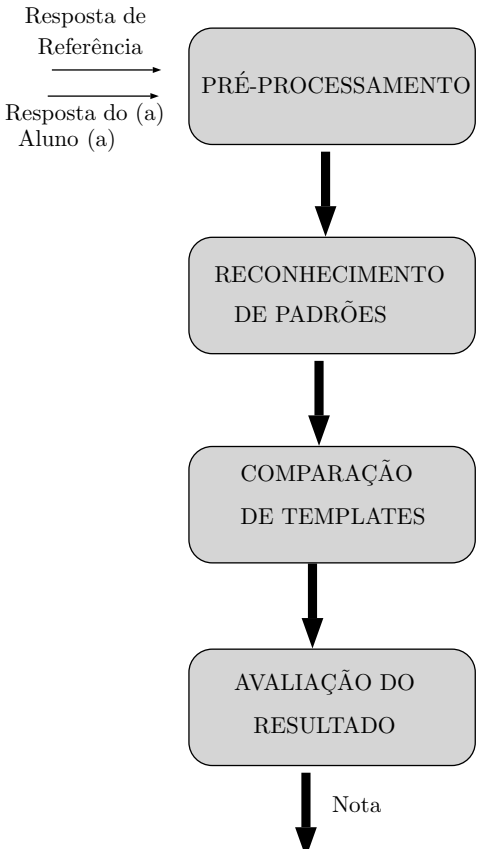

Figura 4 – Pipeline (Fluxograma) de processamento de respostas de alunos.

4.1 Pré-processamento

Em posse da resposta dada pelo aluno e da resposta de referência definida pelo professor, o sistema de correção automática inicia o processo de correção com o pré-processamento do texto.

Assim como a resposta de referência definida pelo professor, a resposta cedida pelo aluno estará expressa em linguagem natural, portanto nessa etapa ocorrem os processos referentes ao tratamento desse texto, que em sua maioria estarão expressos de forma não estruturada, para uma forma mais organizada, padronizada, com sentenças bem formadas e delimitadas com objetivo de facilitar a execução das etapas futuras e portanto facilitar sua interpretação.

Além disso, durante esta etapa serão executadas as tarefas referentes à identificação da classe morfológica de cada palavra que compõe os textos e das dependências sintáticas entre as palavras constituintes das sentenças, bem como do reconhecimento de entidades presentes nas sentenças e da resolução de co-referências das mesmas.

4.1.1 Tokenização e normalização

O primeiro processo executado após o recebimento do texto de entrada, é o de segmentação de texto ou tokenização, durante esta etapa o texto é dividido em unidades menores chamadas de tokens. Para o método proposto, a etapa de tokenização do texto é feita em duas partes, a tokenização a nível de sentenças que compõem os textos e a tokenização a nível de palavras, o processo de separação do texto a nível de sentença facilita a execução das etapas futuras, pois a análise de fragmentos menores de texto tende a gerar melhores resultados. Além disso, também é feita a separação do texto a nível de palavras, visando estruturar o texto para a execução de tarefas futuras e remover caracteres e símbolos não necessários para a interpretação do texto. Para a execução das duas tarefas mencionadas acima foram utilizados módulos de segmentação de textos disponíveis dentro da biblioteca *SpaCy* [1].

Para simplificar a interpretação, extração e a comparação de conceitos presentes dentro dos textos processados, durante o pré-processamento, são aplicadas técnicas de normalização. A normalização busca padronizar o texto para facilitar a interpretação e comparação entre conceitos extraídos, sem que sejam perdidas as informações semânticas das palavras.

Portanto, durante esta etapa são aplicadas técnicas como a Lematização, que reduz palavras a sua raiz ou forma, onde o objetivo é se esquivar das várias formas de representação de uma palavra associada a um mesmo conceito. A Lematização aplicada

[1] SpaCy é uma biblioteca Python para processamento avançado de linguagem natural

durante essa etapa facilita a correspondência entre conceitos extraídos da resposta dos alunos e da resposta de referência durante as etapas futuras.

A Lematização e normalização do texto são executados com componentes disponíveis dentro da biblioteca SpaCy. A figura 5 apresenta um exemplo de lematização executada pela biblioteca *Spacy* sobre um conjunto de termos, como é visível, os termos possuem uma mesma forma base, apesar de estarem escritas em diferentes flexões do verbo correr.

$$LEMATIZAÇÃO$$
$$correu \rightarrow correr$$
$$corrreriam \rightarrow correr$$
$$correm \rightarrow correr$$
$$corre \rightarrow correr$$
$$correr \rightarrow correr$$
$$corria \rightarrow correr$$
$$correndo \rightarrow correr$$

Figura 5 – Exemplo de Lematização.

4.1.2 Análise léxica

Para a execução de tarefas comuns a sistemas de extração de informação baseados em regra é necessário compreender cada palavra que compõem a estrutura. Conhecer a classe gramatical que cada uma possui, baseada no contexto em que ela está inserida é vital para extrair determinados fatos e informações presentes dentro dos textos em questão.

O reconhecimento de tais características gramaticais é feito com a utilização de um POS Tagger. Um POS(Part-of-Speech) tagger é um sistema que usa o contexto para atribuir partes do discurso ou classe gramaticais às palavras (CUTTING et al., 1992). Para a execução das tarefas referentes ao POS Tagging foi utilizada uma pipeline pré-treinada em português disponível na biblioteca SpaCy, a Figura 6 apresenta um exemplo de marcação de classes gramaticais das palavras que compõem uma sentença.

A	**célula**	**possui**	**formato**	**irregular.**
artigo	substantivo	verbo	substantivo	adjetivo

Figura 6 – Exemplo de POS TAGGING.

4.1.3 Análise sintática

Após a fase de análise léxica é realizada a análise sintática dos componentes da sentença. A abordagem de extração de informação baseada em regras adotada neste trabalho, utiliza tanto de informações a nível léxico quanto de informações a nível sintático dos termos presentes nas sentenças, fazendo-se necessário conhecer a função sintática de cada termo presente nas sentenças e sua ligação com outras palavras, para que possam

ser executadas as tarefas referentes ao reconhecimento de padrões durante as etapas subsequentes.

Para identificar e extrair essas características textuais, durante esta etapa é executada a tarefa de *parsing*, que é o procedimento que avalia os vários modos de como combinar as regras gramaticais, com a finalidade de extrair informações relacionadas a estrutura sintática da sentença analisada.

O *parsing* é executado com o auxílio de um componente analisador de dependência baseado em transição disponível dentro da biblioteca SpaCy, o módulo de parsing dentro da biblioteca analisa as sentenças de entrada e disponibiliza para cada palavra o tipo de relação que a mesma exerce bem como informa os termos aos quais a mesma possui relação de dependência. A Figura 7 mostra o exemplo de uma árvore sintática gerada pela biblioteca Spacy.

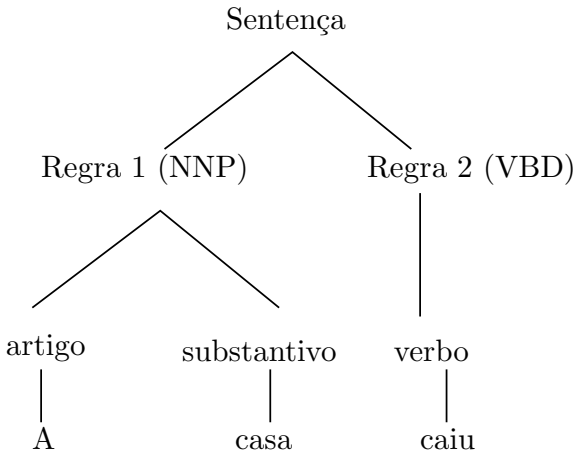

Figura 7 – Exemplo de árvore sintática gerada pelo SpaCy .

A figura 8 exemplifica o conjunto de informações de interesse obtidas após a execução da etapa de pré-processamento sobre uma sentença, utilizando as etapas anteriormente descritas na abordagem proposta. O sistema recebe o texto livre de entrada e executa as etapas de análise descritas anteriormente na seção, a figura apresenta o conjunto de informações obtidas após a execução, onde o texto de entrada é primeiramente dividido baseado nas sentença que o compõe, e depois separado á nível de palavras que compõe cada sentença. A parte inferior da figura apresenta o conjunto de informações obtidas para cada sentença, onde o primeiro conjunto de rótulos apresenta os termos obtidos após a tokenização, lematização e normalização, o segundo conjunto de rótulos apresenta a classe gramatical atribuída a cada palavra, e por fim, o último conjunto de rótulos apresenta as característica de dependência sintática de cada termo.

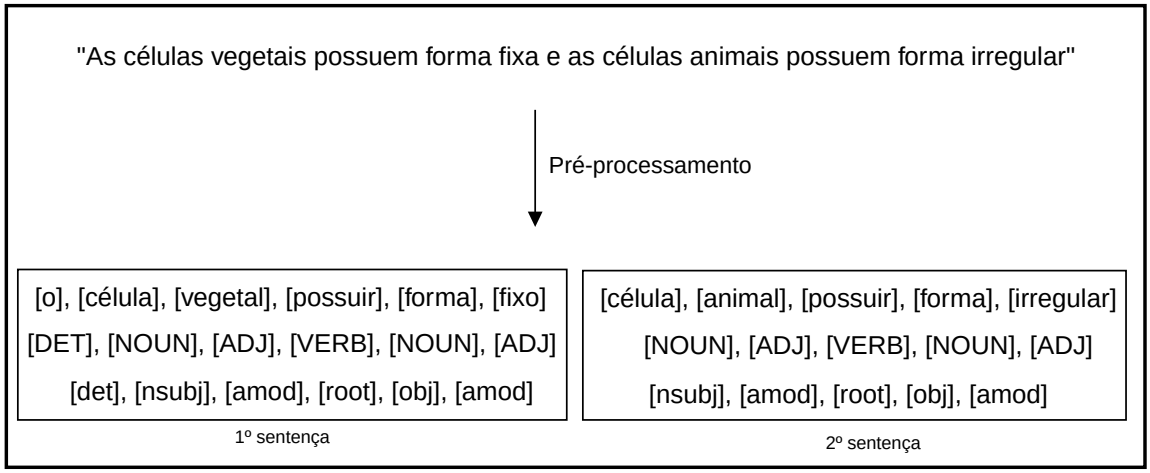

Figura 8 – Execução do pré-processamento.

4.2 Reconhecimento de padrões

Em posse das informações sintáticas, léxicas e morfológicas dos termos que compõem as sentenças presentes nos textos processados, o sistema agora precisa extrair as informações factuais expressas tanto na resposta do aluno quanto na resposta de referência definida pelo professor, e posteriormente representar essas informações de maneira estruturada, para que possa então ser feita a comparação e posteriormente a atribuição de nota.

Na abordagem proposta, a extração de informação é feita utilizando uma abordagem baseada em regras, onde tal se caracteriza pela criação manual de regras de extração, que são baseada em características sintáticas rasas obtidas durante a execução do POS tagging e em características sintáticas profundas obtidas durante a execução do *parsing*, observadas no corpus trabalhado, onde tais regras geram saídas capazes de representar o conhecimento existente no trecho analisado e validado pela regra.

Cada regra aplicada sobre as sentenças, durante essa etapa, foi identificada e definida manualmente, através da observação de padrões de elementos sintáticos e morfológicos que se repetiam em diferentes sentenças com uma mesma estrutura de afirmação lógica expressa. Esses padrões referem-se a ordem na qual as palavras ou elementos são organizados para formar elementos maiores, como frases, cláusulas ou declarações. As regras construídas durante esta etapa foram definidas e posteriormente refinadas iterativamente para melhorar a precisão do processamento de texto.

Tabela 2 – Exemplos de regras de extração criada para o sistema de EI

Padrão	**Template de Saída**
[NOUN|nsubj:pass], [VERB|root], [ADP|case], [NPROP|obl:agent]	[NPROP, VERB, NOUN]
[NOUN], [VERB:root], [ART:OP], [NOUN|obj]	[1ºNOUN, VERB, NOUN]
[NOUN or NPROP], [AUX:OP], [VERB|root], [NOUN], [CC|conj], [NOUN]	1ºNOUN, VERB , 2ºNOUN] [1ºNOUN, VERB, 3ºNOUN]
[NOUN|nsubj], [VERB|root], [VERB:xcomp], [ADV|advmov], [cc], [NOUN],[VERB|root], [NOUN]	1ºNOUN, VERB, ADV] [2ºNOUN, VERB, 3ºVERB]

A tabela 2 apresenta um conjunto de regras que foram criadas para a tarefa de extração, cada elemento delimitado por colchetes representa um termo dentro de uma sentença, os símbolos escritos em maiúsculos representam as classes gramaticais de cada termo, e os termos escritos em minúsculos representam as características de dependência sintática de cada termo, a presença desses termos contendo essas seguintes características definidas pelo regra, validam a mesma, e o template de saída mapeia e modela os termos de forma que consiga representar o conteúdo semântico expresso naquele trecho.

As informações de dependência sintática, são utilizadas para identificar e definir a estrutura argumentativa das sentenças analisadas, a localização do sujeito, da raiz e do predicado, e das cláusulas ajudam a identificar a maneira na qual a estruturação argumentativa é organizada para construir a sentença, pois os argumentos a serem extraídos geralmente correspondem a sintagmas nominais no texto, e as relações a serem extraídas geralmente correspondem a relações funcionais gramaticais.

Já as características gramaticais presentes nos padrões definidos, além de auxiliar a execução da tarefa anterior, ajudam a identificar os termos principais da sentença, auxiliando na tarefa de identificar os principais termos envolvidos, essas características são utilizadas para mapear os termos na tarefa de preenchimento de templates.

Para a tarefa de identificação e criação das regras, Inicialmente foi selecionado um conjunto de respostas aleatórias contendo pelo menos três respostas de cada uma das 15 questões presentes na base de dados e então foi definido um conjunto de regras e padrões que fossem capazes de extrair todos os predicados lógicos dentro das sentenças presentes nas respostas. Posteriormente, essas regras criadas foram testadas em outro conjunto de respostas para verificar se as mesmas eram capazes de contemplar todas as respostas. O passo anterior foi repetido, adicionando um segundo conjunto de regras para que todas as respostas que não tivessem tido suas afirmações lógicas detectadas pelo interpretador pudessem ser contempladas.

A figura 9 apresenta um exemplo de detecção de uma das regra de extração criadas, a regra de extração é destacada na parte superior da figura, e as informações acerca do padrão sintático e da estrutura de saída é demonstrada, a sentença testada possui suas informações gramaticais depois suas informações de dependência sintática apresentadas, e como é possível verificar, tais características correspondem às características definida pela regra, a saída gerada é mostrada na parte inferior da figura, apresenta as argumentações extraídas, contendo os termos identificados.

A estratégia para extração adotada na abordagem proposta é baseada em sistemas de extração de informação abertos. Os sistemas de extração de informações abertos, são sistemas de EI que visam extrair informações estruturadas de texto não estruturado sem limitações no tipo de relação ou no domínio do texto, essas informações geralmente são expressas em um grande número de triplas do tipo *(Arg1, Rel, Arg2)* que representam

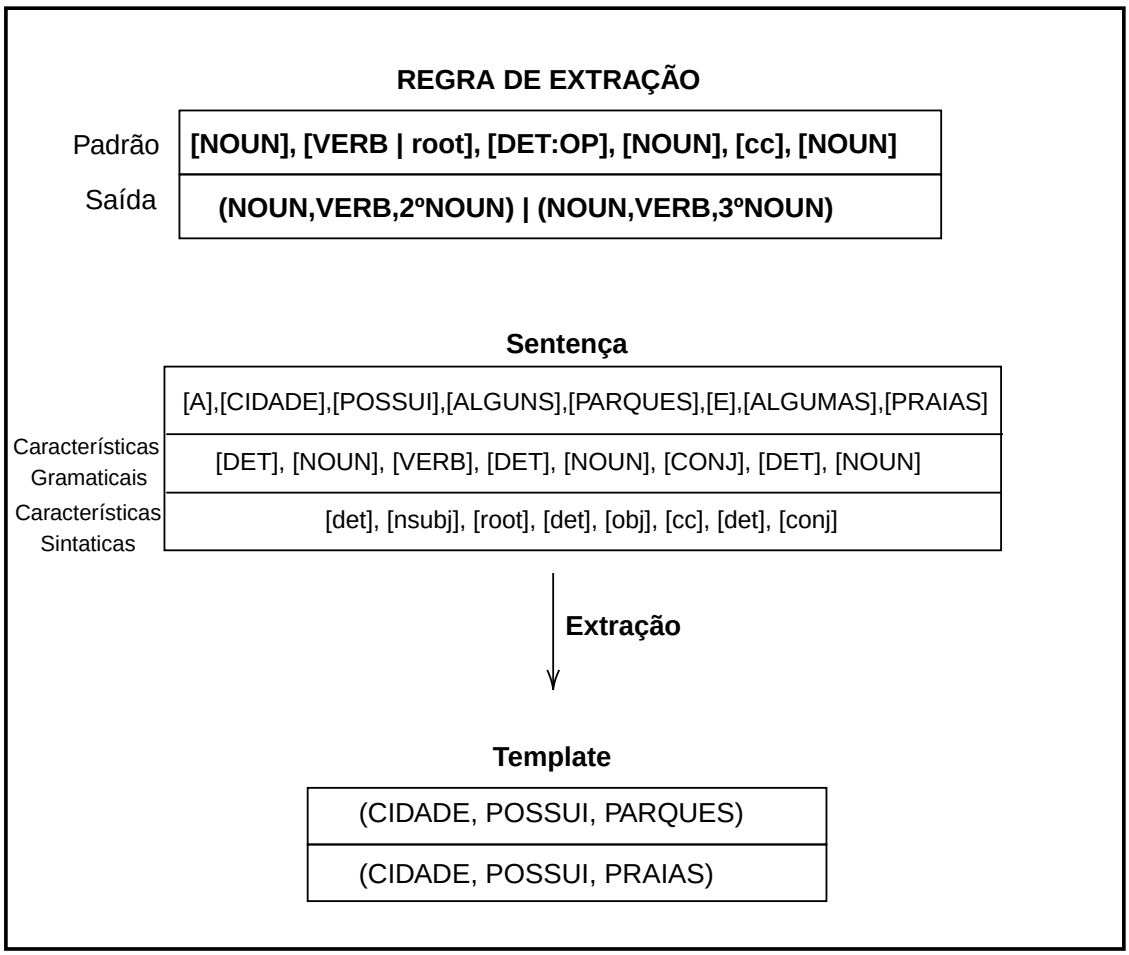

Figura 9 – Exemplo de detecção de regra e extração em uma sentença.

qualquer relação binária encontrada no texto (GAMALLO; GARCIA; FERNÁNDEZ-LANZA, 2012).

4.3 Comparação e preenchimento de templates

O objetivo geral dos sistemas de extração de informação é extrair informações sobre um conjunto pré-especificado de entidades, relações ou eventos de textos em linguagem natural e registrar essas informações em representações estruturadas denominadas *template* (GAIZAUSKAS; WILKS, 1998). O objetivo do sistema desenvolvido neste trabalho é extrair argumentos e conceitos específicos da resposta do aluno, expressos em afirmações declarativas contendo relações binárias.

Portanto, o formato de saída escolhido para representar essas afirmações extraídas das sentenças presentes nos textos processados, foi através de tuplas, que são representações simples capazes de expressar o conhecimento extraído daquele trecho. As tuplas definidas para a extração durante esta etapa, são compostas por verbos que representam o argumento

central de uma cláusula extraída, juntamente com seus argumentos(sujeito,objeto ou predicado) que compõem a estrutura de argumentação, a figura 10 apresenta o exemplo de uma sequência de tuplas extraídas de uma sentença.

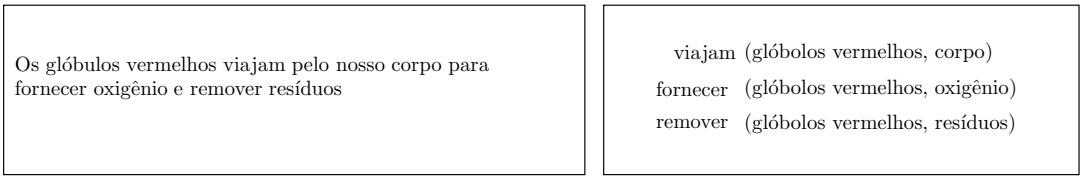

Figura 10 – Exemplo de tuplas extraídas de uma sentença.

Comparação de Conceitos

Em posse das afirmações lógicas extraídas e representadas de forma estruturada obtida após a execução da etapa anterior, o sistema agora precisa comparar e correlacionar as afirmações e conceitos apresentados pelo aluno com os conceitos e afirmações esperados pelo professor.

O sistema realiza buscas dentre os conceitos extraídos da resposta apresentada pelo aluno, e verifica a presença ou ausência dos mesmos no conjunto de informações extraídas da resposta definida pelo professor. A presença de tuplas argumentativas, contendo o termo central da relação, e os dois argumentos que compõem a estrutura extraída, é verificada, e a atribuição de notas é realizada baseada na contagem da mesma. A lematização executada durante a etapa de pré-processamento, facilita o casamento de termos, que apesar de estarem escritos em flexões diferentes, possuem uma mesma raiz em comum.

4.4 Avaliação dos resultados

A última etapa do método proposto para correção automática é a de atribuição de notas. A atribuição de notas é feita baseada na presença ou ausência dentro das respostas dos alunos, dos conceitos ou ideias esperadas e determinadas pelo professor. O sistema proposto utiliza o mesmo critério e mesma escala para atribuição de notas utilizado por Galhardi, Souza e Brancher (2020), onde a atribuição de notas é feita da seguinte maneira:

- **Zero**: quando a resposta está pelo menos na maior parte errada, fora do escopo ou sem sentido.
- **Um**: se a resposta tiver algo correto, mas ainda estiver errado ou incompleto.
- **Dois**: se a resposta estiver correta, mas tiver algum detalhe errado ou faltando conteúdo importante.

- **Três**: se a resposta estiver correta, com os pontos importantes apresentados.

Utilizando a escala anterior, o sistema analisa e atribui as notas levando em conta o número de conceitos esperados pelo professor, que foram apresentados pelo aluno em sua resposta, penalizando o mesmo tanto pela ausência de conceitos esperados, como pela presença de conceitos incorretos ou irrelevantes, portanto para o bom funcionamento do sistema é necessário que o professor defina cuidadosamente e apresente todos os conceitos possíveis que podem ser aceitos como respostas aceitáveis para aquela questão.

É importante ressaltar que o critério para correção adotada nesta abordagem tem foco no conteúdo presente nas respostas apresentadas e não no estilo de escrita, portanto alguns erros de ortografia são tolerados pelo sistema de correção. Com o uso do módulo de correção presente na biblioteca *HunsSpell* [2], alguns erros de ortografia cometidos pelo aluno são tratados, portanto o sistema foca a correção apenas na presença ou ausência de conceitos e fatos esperados.

A figura 11 exemplifica o processo de correção realizada pelo método proposto. A figura apresenta um processo de extração, onde templates são gerados a partir do processamento da resposta de referência definida pelo professor, e da resposta dada pelo aluno. Os templates são comparados e por fim é atribuída a nota, no exemplo apresentado a resposta do aluno apresentou alguns conceitos, esperados pelo professor, porém, além de deixar de abordar certas ideias definidas na resposta do professor, apresentou conceitos irrelevantes em sua resposta, por fim, seguindo os critérios apresentados no início da seção, o sistema atribui nota 1 para a resposta apresentada, mesma nota atribuída pelos avaliadores humanos.

[2] O Hunspell é um corretor ortográfico e analisador morfológico desenvolvido para idiomas com morfologia rica, palavras compostas e codificação de caracteres complexos

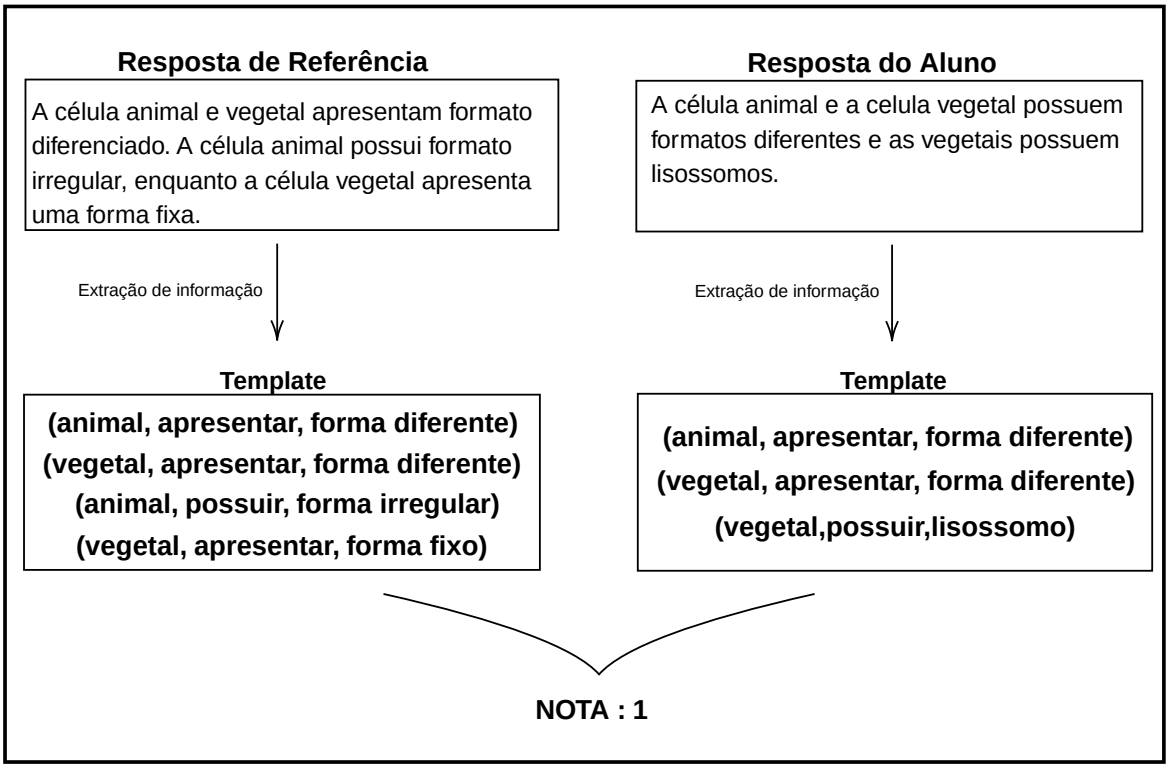

Figura 11 – Processo de correção realizada pelo sistema.

5 Avaliação do Método Proposto

Nesta seção, descreve-se o processo para avaliar o método proposto. A avaliação é executada sobre um conjunto de dados reais descrito na próxima seção, e a métrica usada para avaliar a sua acurácia é descrita na seção 5.2. Os resultados são apresentados na seção 5.3.

5.1 Conjunto de Dados

Para a realização dos testes e obtenção dos resultados apresentados nesta seção, foi utilizada uma base de dados pública disponibilizada por por Galhardi, Souza e Brancher (2020) em seus estudos, e que se encontra disponível no site Kaggle [1].

O conjunto de dados é composto por dados reais extraídos de alunos e professores brasileiros. A base de questões do Kaggle contém 7473 respostas coletadas de um teste aplicado a 659 estudantes brasileiros, onde para cada uma dessas respostas estão atribuídas notas definidas pela correção de 14 avaliadores humanos. A aplicação do teste para a coleta de respostas foi feita com a supervisão dos professores, onde os alunos foram instruídos a tentar o seu melhor, mesmo que envolvesse adivinhação, a fim de coletar todos os tipos de respostas e notas (GALHARDI; SOUZA; BRANCHER, 2020).

O conjunto de respostas presentes na base de dados é oriundo de um teste elaborado por 5 professores do ensino fundamental que contém 15 questões de biologia, onde o assunto abordado pelas questões consiste principalmente em tópicos sobre o corpo humano, vistos principalmente na 8ª série do ensino fundamental, alguns exemplos de questões presentes na base de dados são: “Explique o mecanismo de inspiração e de expiração do ar no corpo humano” e “Quais são as diferenças entre as células vegetais e as células animais?”.

A base de dados contém, além de arquivos compostos pelas respostas dadas pelos alunos com as respectivas notas atribuídas pelos avaliadores, arquivos contendo respostas de referências definidas por professores como solução para cada questão presente na base de dados. A Tabela 3 apresenta um conjunto de dados de entrada presente nos arquivos disponíveis que exemplificam o conjunto de informações presentes na base de dados, a tabela possui um conjunto de respostas para a questão: "Qual a diferença entre veias e artérias?".

[1] Kagle é uma plataforma de aprendizagem e competição para cientistas de dados.

Tabela 3 – Exemplo de dados de entrada da base de dados

Id	Texto de Resposta	Nota
46	A diferença é que a veia passa sangue e as arterias proteínas.	0
46	As artérias levam o sangue e as veias trazem	1
46	E que a artéria leva para o resto do corpo e a veia traz para o coração.	1
46	Veia leva sangue para o coração e é mais fina, artéria leva sangue para o corpo e é mais resistente	3

5.2 Métricas de Avaliação

A métrica escolhida para avaliar a correlação entre os dois avaliadores foi o coeficiente Kappa de Cohen. O Kappa de Cohen é uma estatística que varia entre -1 e 1, muito utilizada para analisar concordância entre classificadores, onde quanto mais próximo de 1, maior a concordância entre os avaliadores (MCHUGH, 2012). Além de ser uma das métricas mais comuns e utilizadas para realizar esse tipo de avaliação, a escolha do coeficiente de Kappa como métrica se dá pelo fato da mesma ser uma medida mais robusta do que o cálculo de concordância percentual simples, pois o Kappa de Cohen leva em consideração a possibilidade da concordância ocorrer por acaso.

Calculou-se o percentual de concordância entre as notas atribuídas pelos avaliadores humanos e o algoritmo proposto. Quanto maior a proximidade entre as notas determinadas pelo sistema e as notas definidas pelos avaliadores humanos, maior o coeficiente de correlação entre esses dois avaliadores e portanto maior seria a confiabilidade do método proposto. Contudo, é importante ressaltar a existência de casos em que a presença de subjetividade na correção humana, resulta na discordância de notas entre humano e sistema, mesmo que o algoritmo tenha corrigido a questão de forma correta.

5.3 Resultados e Discussão

Esta seção apresenta os resultados obtidos e os compara com os achados de outros trabalhos anteriores usados como referência.

A Tabela 4 apresenta o grau de concordância entre as notas atribuídas pelos avaliadores humanos e as notas atribuídas pelo algoritmo desenvolvido, para as 15 questões, utilizando o coeficiente Kappa de Cohen ponderado com pesos quadráticos.

A abordagem proposta obteve um bom desempenho (Coeficiente Kappa médio de 0,674) para as 15 questões testadas durante os experimentos, de acordo com a classificação de nível de concordância definida por Landis e Koch (1977) que atribui um grau de concordância, substancial ou forte, para um coeficiente de Kappa entre 0,61 e 0,8. Analisando as questões separadamente, pode-se observar que a abordagem proposta atingiu um nível de concordância considerado substancial para 10 das 15 questões testadas, em apenas 1 das questões apresentou um percentual de concordância considerado quase-

Tabela 4 – Nível de concordância de notas entre o algoritmo desenvolvido e os avaliadores humanos.

Questão	**Coeficiente Kappa**
1	0,812
2	0,627
3	0,698
4	0,684
5	0,785
6	0,655
7	0,578
8	0,597
9	0,542
10	0,677
11	0,803
12	0,673
13	0,701
14	0,587
15	0,691
Média	**0,674**

perfeito e em outras 4 questões atingiu um nível de concordância considerado moderado.

De acordo com McHugh (2012) valores de kappa abaixo de 0,60 podem indicar concordância inadequada entre os avaliadores e pouca confiança deve ser depositada nos resultados do estudo, e como observado na Tabela 4 pode-se constatar que em 4 das 15 questões testadas a abordagem proposta não conseguiu atingir números confiáveis de concordância com avaliadores humanos. Essas questões nas quais o método não conseguiu obter resultados satisfatórios tratam-se de questões que requerem do aluno, em média, respostas expressas em no mínimo três ou mais sentenças, além disso essas questões possuem um domínio muito aberto de respostas possíveis e aceitáveis, como por exemplo: *"Por que é necessário comer vários tipos diferentes de alimentos?"*, ao contrário das outras questões que requerem do aluno respostas concisas com conceitos específicos que podem ser expressos em poucas sentenças.

É importante também ressaltar a presença de inconsistências no conjunto de dados em relação a correção dos avaliadores humanos, que provavelmente afetaram o desempenho da abordagem desenvolvida. Como discutido anteriormente, a subjetividade do ser humano pode ter um grande impacto na avaliação de questões não-objetivas, e tal subjetividade se mostra presente no conjunto de dados utilizados.

Para quatro questões dentro do conjunto de dados, foram coletadas correções de mais de 1 avaliador humano, e como foi mostrado por Galhardi, Souza e Brancher (2020) em seu estudo, a correlação encontrada entre os avaliadores humanos para as quatro questões variam em um intervalo de 0,52 a 0,57, o que de acordo com Landis e Koch (1977) é um nível apenas moderado de concordância. Tal discordância encontrada entre os avaliadores humanos pode explicar em grande parte os resultados não tão satisfatórios obtidos pelo algoritmo para essas questões em que os avaliadores humanos discordam em

maior grau.

A Tabela 5 apresenta um comparativo entre o novo método baseado em regras e o método proposto por Galhardi, Souza e Brancher (2020) baseado em Ngrams. A métrica de desempenho usada para comparar as abordagens foi o *bk value*, uma medida do grau de concordância calculada a partir da média entre o coeficiente Kappa de Cohen utilizando pesos quadráticos e utilizando pesos lineares.

Tabela 5 – Comparação entre o método proposto e a abordagem baseada em ngrams

	bk value	
Questão	**Método proposto**	**Método de Referência**
1	0,773	0,612
2	0,616	0,645
3	0,686	0,575
4	0,677	0,775
5	0,773	0,685
6	0,638	0,487
7	0,561	0,364
8	0,592	0,495
9	0,536	0,484
10	0,668	0,820
11	0,798	0,517
12	0,667	0,592
13	0,698	0,362
14	0,579	0,639
15	0,684	0,808
Média	**0,663**	**0,591**

Observa-se na Tabela 5 que o novo método tem uma clara vantagem sobre o método baseado em Ngrams (GALHARDI; SOUZA; BRANCHER, 2020) considerado o estado da arte (média 0,663 contra 0,591), nota-se também que em 10 das 15 questões testadas o novo método obteve um coeficiente de concordância superior em relação ao método de referência. Além disso é perceptível que o número de questões em que o método proposto obteve concordância inadequada com avaliadores humanos foi substancialmente menor que o método de referência. No geral, as questões nas quais o método proposto obteve pior desempenho são as mesmas nas quais o método de referência obteve coeficientes de concordâncias menos confiáveis.

6 Conclusão

Neste trabalho, foi apresentado um novo método computacional para correção automática de questões discursivas textuais simples. O método usa técnicas de extração de informação baseadas em regras que utilizam padrões sintáticos e gramaticais para extrair e estruturar o conteúdo semântico presente na resposta dada pelo aluno, facilitando a posterior tarefa de comparação com a resposta definida pelo professor.

Os resultados apresentados reforçam a importância da análise do conteúdo semântico no processo de correção, em relação a análise de outras características textuais . E, embora o desempenho do método, diante de questões que possuem domínio muito aberto de respostas, não foi o ideal, acredita-se que o método pode ser útil.

Por último, é importante frisar que os resultados apresentados foram obtidos a partir de testes realizados sob um conjunto de respostas pertencentes a questões da área de estudos da biologia, no entanto, espera-se que o desempenho continue constante para questões pertencentes a outros campos de estudo. Para isso será necessário a criação de um conjunto mais robustos de regras de extração que possam contemplar textos mais subjetivos.

6.1 Trabalhos futuros

Progresso considerável tem sido feito com relação a correção automática de questões discursivas da língua portuguesa. Entretanto, mais trabalho focado e aprofundado precisam ser realizados, utilizando tanto as descobertas mais recentes desenvolvidas na área de inteligência artificial, quanto utilizando outros métodos para extração, modelagem e comparação de informações, como por exemplo a aplicação de métodos de aprendizado de máquina para a detecção de padrões e criação regras que possam adicionar eficiência na extração de conteúdo semântico.

Propõe-se também, o aprimoramento do método apresentado no presente trabalho em novos estudos, utilizando uma abordagem semi-automática que utiliza supervisão humana para aumentar a eficiência da correção. O método apresentado neste trabalho, realiza o processo de correção de forma totalmente automática, porém estima-se que o auxílio humano em algumas etapas possam aumentar a eficiência do processo de correção, Portanto a intervenção humana, para por exemplo adicionar termos semanticamente parecidos aos termos presentes no modelo de resposta gerado, facilitando a correlação com argumentos semanticamente parecidos expressos na respostas dos alunos, que porém não utilizam termos iguais ou que contenham raízes parecidas, além disso recomenda-

se a aplicação da intervenção humana adicionando pesos para os argumentos extraídos da resposta do professor, definindo quais conceitos são de mais valia na construção da argumentação expressos pelos alunos.

Por último, propõe-se a criação e disponibilização pública de novas base de dados que contenham um conjunto de perguntas e respostas escritas em português, a existência de tal conjunto de informações fomentaria o surgimento de novas pesquisas relacionadas ao campo de correção automática. A língua portuguesa contém uma relevante variedade de similaridades, tanto no que tange a composição léxica, variação morfológica ou estruturação sintática, portanto o surgimento de pesquisa nessa área, podem revelar descobertas quanto a métodos mais eficientes para a extração de conteúdo semântico baseado nos recursos linguísticos do idioma, além de possibilitar a aplicação de testes que possam validar a eficácia dos novos métodos propostos.

Referências

AALAEI, S.; AHMADI, M. A. T.; AALAEI, A. A comparison of multiple-choice and essay questions in the evaluation of dental students. *International Journal of Advanced Biotechnology and Research*, v. 7, n. 5, p. 1674–1680, 2016. Citado na página 15.

APPELT, D. E.; ISRAEL, D. J. Introduction to information extraction technology. In: *International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 1999. Citado 4 vezes nas páginas 16, 17, 18 e 19.

BURROWS, S.; GUREVYCH, I.; STEIN, B. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, Springer Science and Business Media LLC, v. 25, n. 1, p. 60–117, out. 2014. Citado 3 vezes nas páginas 14, 15 e 26.

CHAMBERS, N.; JURAFSKY, D. Template-based information extraction without the templates. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, 2011. p. 976–986. Citado na página 20.

CUTTING, D.; KUPIEC, J.; PEDERSEN, J.; SIBUN, P. A practical part-of-speech tagger. In: *Third Conference on Applied Natural Language Processing*. Trento, Italy: Association for Computational Linguistics, 1992. p. 133–140. Citado na página 28.

FARINON, J. L. *Análise E Classificação de conteúdo textual*. 2015. Disponível em: <https://repositorio.unisc.br/jspui/handle/11624/1038>. Citado na página 22.

FELDMAN, R.; SANGER, J. Information extraction. In: _____. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. [S.l.]: Cambridge University Press, 2006. p. 94–130. Citado na página 20.

GAIZAUSKAS, R.; WILKS, Y. Information extraction: beyond document retrieval. *Journal of Documentation*, Emerald, v. 54, n. 1, p. 70–105, mar. 1998. Citado na página 32.

GALHARDI, L.; SOUZA, R. de; BRANCHER, J. Automatic grading of portuguese short answers using a machine learning approach. In: *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*. Porto Alegre, RS, Brasil: SBC, 2020. p. 109–124. ISSN 0000-0000. Citado 6 vezes nas páginas 24, 25, 33, 36, 38 e 39.

GAMALLO, P.; GARCIA, M.; FERNÁNDEZ-LANZA, S. Dependency-based open information extraction. In: *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*. Avignon, France: Association for Computational Linguistics, 2012. p. 10–18. Citado na página 32.

GRISHMAN, R. Information extraction: Techniques and challenges. In: PAZIENZA, M. T. (Ed.). *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p. 10–27. ISBN 978-3-540-69548-6. Citado na página 19.

HAHN, M.; MEURERS, D. Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In: . [S.l.: s.n.], 2012. p. 326–336. Citado 2 vezes nas páginas 24 e 25.

HALEY, D.; THOMAS, P.; ROECK, A. D.; PETRE, M. Measuring improvement in latent semantic analysis-based marking systems: Using a computer to mark questions about html. *Conferences in Research and Practice in Information Technology Series*, v. 66, 01 2007. Citado na página 13.

KAIPA, R. M. Multiple choice questions and essay questions in curriculum. *Journal of Applied Research in Higher Education*, Emerald, v. 13, n. 1, p. 16–32, abr. 2020. Citado na página 13.

KOSTAREVA, T.; CHUPRINA, S.; NAM, A. Using ontology-driven methods to develop frameworks for tackling nlp problems. In: *AIST*. [S.l.: s.n.], 2016. Citado na página 21.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, [Wiley, International Biometric Society], v. 33, n. 1, p. 159–174, 1977. Citado 2 vezes nas páginas 37 e 38.

LUCKESI, C. C. *Avaliação da aprendizagem: componente do ato pedagógico*. [S.l.]: Editora Cortez, 2013. Citado na página 13.

MCHUGH, M. L. Interrater reliability: the kappa statistic. *Biochemia Medica*, v. 22, p. 276 – 282, 2012. Citado 3 vezes nas páginas 22, 37 e 38.

MITCHELL, T.; RUSSELL, T.; BROOMHEAD, P.; ALDRIDGE, N. Towards robust computerised marking of free-text responses. In: . [S.l.: s.n.], 2002. Citado 2 vezes nas páginas 24 e 25.

NéDELLEC, C.; NAZARENKO, A.; BOSSY, R. Information extraction. In: _____. [S.l.: s.n.], 2009. p. 663–685. Citado na página 18.

OLIVEIRA, D.; POZZEBON, E.; SANTOS, T. Aplicação das técnicas de processamento de linguagem natural cosine similarity e word movers distance para auxiliar na correção de questões discursivas em um tutor inteligente. In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2020. p. 1243–1252. Citado na página 25.

PAGE, E. B. The imminence of... grading essays by computer. *The Phi Delta Kappan*, Phi Delta Kappa International, v. 47, n. 5, p. 238–243, 1966. Citado na página 24.

PINTO, S. C. S. *Estudo Geral*. 2015. Citado na página 21.

SAKAGUCHI, K.; HEILMAN, M.; MADNANI, N. Effective feature integration for automated short answer scoring. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 2015. p. 1049–1054. Citado na página 13.

SANTANA, R. S. *Análise Sintática - Português*. 2020. Citado na página 19.

SANTOS, J. C. A. d. *Avaliação automática de Questões discursivas USANDO LSA*. [S.l.]: Universidade Federal do Pará, 2016. Citado na página 25.

SARAWAGI, S. Information extraction. *Foundations and Trends in Databases*, v. 1, p. 261–377, 01 2008. Citado na página 16.

SYNTACTIC and semantic information extraction from NPP procedures utilizing natural language processing integrated with rules. *Nuclear Engineering and Technology*, v. 53, n. 3, p. 866–878, 2021. ISSN 1738-5733. Citado na página 17.

THOMAS, P. The evaluation of electronic marking of examinations. 09 2003. Citado na página 24.

WOODFORD, K.; BANCROFT, P. Using multiple choice questions effectively in information technology education. In: . [S.l.: s.n.], 2004. Citado na página 13.

ZHOU, P.; EL-GOHARY, N. Ontology-based automated information extraction from building energy conservation codes. *Automation in Construction*, v. 74, p. 103–117, 2017. ISSN 0926-5805. Citado 2 vezes nas páginas 17 e 18.

Apêndice A - Questões da base de dados

ID	Questão
1	Qual a diferença entre a célula animal e a célula vegetal?
2	Quais são as partes que compõe a célula e suas funções?
3	O corpo humano possui vários tipos de células que se organizam, de acordo com suas especializações e funções, formando os tecidos. Quais são as características do tecido epitelial?
4	Correr, estudar e dançar são atividades que necessitam de muita energia e que podemos realizar graças às nossas células que trabalham sem parar. Como chamamos o conjunto de reações químicas que ocorrem ao nível das células e como ele acontece?
5	Qual a diferença entre fenótipo e genótipo?
6	O que significa transmissão de caracteres hereditários?
7	"O que define o sexo na espécie humana são as características sexuais primárias, ou seja, os órgãos sexuais. Tanto os homens quanto as mulheres têm órgãos sexuais externos e internos. Quais são e quais as funções dos órgãos sexuais internos do homem?
8	Explique o mecanismo de inspiração e de expiração do ar no corpo humano:
9	Em condições normais e estando acordada uma pessoa pode suspender a respiração temporariamente ou acelerar o ritmo respiratório na hora em que desejar fazê-lo. Mas uma pessoa não consegue mesmo que queira, provocar a falta total de gás oxigênio no organismo simplesmente parando de respirar. Por quê?
10	Os cromossomos humanos podem ser estudados em células extraídas do sangue. Em qual das células sanguíneas deve ser feito este estudo. Por quê?
11	Quais são as diferenças entre veias e artérias?
12	Qual é a função do fígado no processo digestivo?
13	Apesar do avanço que a medicina vem apresentando no início do século XXI, ainda nos deparamos com grandes desafios na área da medicina preventiva. Devemos sempre ficar atentos ao calendário de vacinas e às campanhas de vacinação. Quais são as funções das vacinas no corpo humano?
14	Por que é necessário comer vários tipos diferentes de alimentos?
15	O que a hemodiálise faz no corpo humano?

Apêndice B - Repositório contendo algoritmo desenvolvido utilizando a abordagem proposta

<github.com/Alerciosilva9/correcao-automatica>