Capítulo 19

Correção automática de redação

Amanda Pontes Rassi
Priscilla de Abreu Lopes

19.1 Introdução

A Correção Automática de Redação (CAR) é uma das várias aplicações do PLN e pode ser definida como "o processo de avaliação e atribuição de nota em textos escritos em prosa, via programas computacionais" (Shermis; Burstein, 2013) [1].

A correção manual de redações é uma prática bastante antiga, mas esse processo feito de forma automática data da década de 60, em inglês, e é ainda mais recente para o português.

Em inglês, as áreas de *Automated Essay Scoring* (AES) e *Automated Essay Evaluation* (AEE) surgem como distintas, porém complementares e, às vezes, com alguma intersecção. A primeira tem como desafio a automatização de atribuição de nota para redação, enquanto a segunda está preocupada, também, em automatizar o retorno ou feedback para o aluno, colaborando para o processo de aprendizagem da escrita.

A AES costuma ser traduzida para o português como **Avaliação Automática de Redação** (AAR) (Bittencourt Jr., 2020; Da Silva Jr., 2021; Lima et al., 2023), enquanto a AEE está associada ao termo **Correção Automática de Redação** (CAR), apesar do falso cognato. Neste capítulo, adotamos o segundo, por entendermos que ele abarca as duas áreas AEE e AES, ou seja, trata-se de uma solução completa. Para que seja considerada como solução completa de CAR, a aplicação deve contemplar pelo menos três etapas básicas:

i) a detecção de desvios no texto;
ii) a atribuição da nota, seja ela global ou por critério; e
iii) um feedback para o aluno.

Cada uma dessas etapas pode ser vista como uma aplicação independente no PLN. Por exemplo, existem várias ferramentas de auxílio à escrita, bem como corretores ortográficos e gramaticais, que executam exclusivamente a tarefa de identificação de desvios no texto; e isso constitui uma aplicação em si. Da mesma forma, a tarefa de dar um feedback com sugestões para o aluno é semelhante a outras aplicações de PLN que envolvem geração de linguagem natural (ou *Natural Language Generation*).

Apesar de poderem figurar como ferramentas e/ou aplicações independentes, consideramos que a correção de redação, para ser entendida como uma solução completa

[1] Tradução nossa. Do original: "*the process of evaluating and scoring written prose via computer programs*".

do ponto de vista pedagógico, exige o cumprimento dessas três etapas, que serão bem detalhadas ao longo deste capítulo.

Antes de abordar cada uma das etapas, porém, faremos uma breve explicação sobre o objeto de estudo da CAR, que é a redação escolar, definindo e exemplificando os principais gêneros e tipos textuais, os critérios avaliados e alguns modelos brasileiros de correção de redação.

19.1.1 O que é uma redação escolar?

A redação escolar é considerada um gênero textual, mas também pode ser distribuída em vários tipos e gêneros textuais. As redações, ou textos[2] de redação escolar, são geralmente utilizadas para avaliar as habilidades de escrita, interpretação, argumentação e criatividade dos alunos, bem como para desenvolver o pensamento crítico e a capacidade de expressão escrita. As redações podem abordar temas diversos, desde assuntos cotidianos até questões mais complexas e abstratas, e são uma forma importante de avaliar o progresso dos alunos ao longo do tempo.

Para fins didáticos e de correção de redação, é importante salientar a diferença entre tipo textual e gênero textual, já que as redações devem atender a um tipo específico e a algum gênero específico, a depender da proposta de redação. Por exemplo, a redação do Enem[3] é sempre do tipo argumentativo e do gênero dissertação-argumentativa.

Os **tipos textuais** (ou "modos textuais", para Marcuschi (2008, p. 154)) se referem à forma como o texto é organizado, ou seja, a sequência linguística e os aspectos lexicais, sintáticos, tempos verbais, relações lógicas que são mobilizados para constituir o texto. Existe um conjunto bastante limitado de tipos textuais, o qual abrange: narração, argumentação, descrição, exposição e injunção. Ressaltamos que um texto raramente apresenta apenas características de um mesmo tipo. Deste modo, classificamos um texto como sendo de um determinado tipo quando há predominância de elementos que o caracterizam.

Já os **gêneros textuais** são formas de comunicação que se desenvolvem em diferentes contextos sociais e culturais e se caracterizam pelo seu propósito ou objetivo comunicativo. Em outras palavras, cada gênero tem uma finalidade específica e é utilizado em determinadas situações comunicativas, dependendo de fatores sociais, culturais, dos falantes, da relação entre eles, do contexto, da finalidade da comunicação, dentre vários outros. Dependendo da situação comununicativa, cada gênero pode exigir um registro ou vocabulário específico, a norma culta ou coloquial, na modalidade escrita ou oral da língua.

Por esse motivo, os gêneros são mais fluidos, podendo surgir, modificar-se, mesclar com outros, desaparecer e reaparecer com outra roupagem em outro contexto ou época. São exemplos de gêneros textuais: bula de remédio, carta pessoal, diálogo informal, e-mail,

[2] Existe uma longa discussão conceitual e técnica sobre a definição do termo "texto" em Linguística Textual. Para o propósito deste capítulo, adotaremos como conceito de "texto" um conjunto de palavras e frases organizadas de forma coerente e coesa, com o objetivo de transmitir uma mensagem ou ideia. Em outras palavras, o texto é uma unidade de linguagem que tem um sentido completo e pode ser compreendido em um contexto específico.

[3] O Exame Nacional do Ensino Médio (Enem) é uma prova do Governo Federal que avalia o desempenho escolar dos estudantes ao término do Ensino Médio. Essa prova avalia várias áreas do conhecimento e também a produção de uma redação.

edital de concurso, inquérito policial, piada, receita culinária, reportagem, resenha, sermão etc.

As redações podem apresentar diversos formatos e objetivos, dependendo do nível de ensino e do tema proposto pelo professor ou pela instituição de ensino. Dentre os tipos e gêneros textuais mais comuns associados à redação escolar, convém mencionar:

- **Dissertação**, em que o autor apresenta e disserta sobre um determinado tema, apresentando informações e argumentos relacionados ao assunto;
- **Narração**, em que o autor conta uma história, relatando fatos e acontecimentos em alguma ordem que pode ser cronológica ou não;
- **Carta**, em que o autor se dirige a um destinatário específico, fazendo requisições, solicitações ou expressando suas opiniões e sentimentos;
- **Artigo de opinião**, em que o autor defende um ponto de vista sobre um tema específico, utilizando argumentos e evidências para sustentá-lo;
- **Resenha**, em que o autor faz uma análise crítica de um texto, obra ou produto.

Esses são apenas alguns dos tipos e gêneros de redação escolar mais comuns. Os demais incluem a descrição, a exposição, a crônica, o relatório, o conto, a fábula, entre outros.

19.1.2 O que é avaliado?

Vários aspectos do texto são avaliados em uma correção de redação, tais como o uso da norma padrão da língua portuguesa, a adequação ao tema e ao gênero, questões relacionadas à coesão, à coerência, à progressão textual etc. Cada modelo de correção organiza e nomeia seus critérios de avaliação de formas distintas, mas, basicamente, todos eles analisam:

Língua portuguesa Avalia a linguagem usada para expressar o conteúdo, verificando se há desvios ortográficos e/ou gramaticais, se a norma (culta ou coloquial) está de acordo com o tipo de texto exigido, se há problemas de estrutura sintática nas frases, orações e períodos, se o vocabulário foi usado adequadamente etc. Outros nomes para esse critério incluem "Escrita", "Modalidade escrita", "Norma culta", "Norma padrão", "Correção gramatical e adequação vocabular" ou "Expressão (modalidade)".

Tema Esse aspecto avalia a adequação da redação em relação à temática proposta, verificando se a abordagem do tema foi completa, se tangenciou ou fugiu do tema proposto, se o abordou de forma superficial ou profunda etc. Também é chamado de "Abordagem temática", "Desenvolvimento do tema", "Proposta temática" ou "Progressão temática".

Gênero Esse critério considera a adequação da redação em relação ao tipo textual e ao gênero textual exigidos na proposta. Também pode ser chamado de "Gênero textual", "Adequação ao tipo textual", "Organização do texto dissertativo-argumentativo" ou "Estrutura (gênero/tipo de texto)".

Coerência Neste quesito, avalia-se a coerência entre as ideias, a ordem dos argumentos, a profundidade da argumentação, a clareza e autoria das ideias desenvolvidas, assim como verifica-se se há contradições no texto, se as informações são vagas e/ou muito generalistas, se falta informação, dentre outros. Também pode ser chamado

de “Progressão textual”, “Defesa do ponto de vista”, “Coerência dos argumentos”, “Estrutura (coerência)”, “Indícios de autoria” e outros termos.

Coesão Avalia o uso correto ou incorreto, presença ou ausência, pertinência ou não de operadores coesivos, tais como conjunções, preposições, pronomes e expressões discursivas. O critério é também chamado de “Coesão e articulação”, “Articulação das partes do texto”, “Expressão (coesão)”, “Conexão entre os parágrafos”, “Uso de operadores argumentativos”, “Recursos coesivos”, dentre outros termos.

Além desses aspectos que são comuns a todos os modelos de correção, alguns professores, instituições de ensino e vestibulares também podem avaliar a “Leitura”, ou seja, o uso e interpretação dos textos motivadores ou da coletânea que embasa a proposta de redação, e também a presença e adequação da “Proposta de intervenção”, que é um critério exclusivo do Enem.

19.1.3 Alguns modelos brasileiros de correção

O principal modelo de correção de redação, no Brasil, é o Enem, responsável pela avaliação anual de cerca de 4 milhões de alunos[4]. Mas também existem outros modelos de correção relacionados a vestibulares e universidades específicas, tais como Fuvest, Unesp, Unicamp, FGV e outros igualmente relevantes. Apesar de haver critérios gerais que são avaliados por todos eles, cada um tem autonomia para definir sua grade específica, os pesos de cada critério e sua própria forma de avaliação.

O Quadro 19.1 apresenta quatro modelos brasileiros de correção relacionados a vestibulares, indicando o gênero textual exigido, seus critérios de avaliação e faixas de nota possíveis.

Quadro 19.1: Modelos de correção de vestibulares.

Modelo	Tipo/Gênero	Critérios avaliativos	Nota mín. critério	Nota máx. critério	Nota global
Enem	Dissertação-argumentativa	Língua Portuguesa (Competência 1)	0	200	1000
		Abordagem temática e adequação ao tipo textual (Competência 2)	0	200	
		Progressão textual e defesa do ponto de vista (Competência 3)	0	200	
		Coesão e articulação (Competência 4)	0	200	
		Proposta de intervenção (Competência 5)	0	200	
Fuvest	Dissertação-argumentativa	Desenvolvimento do tema e organização do texto dissertativo-argumentativo	4	20	50

[4]Média aproximada de inscritos por ano no Enem entre 2015 e 2023.

		Coerência dos argumentos e articulação das partes do texto	3	15	
		Correção gramatical e adequação vocabular	3	15	
Unesp	Dissertação-argumentativa	Tema	–	–	28
		Estrutura (gênero/tipo de texto e coerência)	–	–	
		Expressão (modalidade e coesão)	–	–	
Unicamp	variados	Proposta temática	0	2	12
		Gênero	0	3	
		Leitura	0	3	
		Articulação escrita	1	4	

No modelo de correção do **Enem**[5], o aluno é avaliado quanto à produção de um texto do tipo dissertativo-argumentativo para um tema específico, que muda todo ano. A avaliação é dividida em 5 competências (critérios avaliativos), cada uma no intervalo de notas de 0 a 200. A soma direta das notas das competências leva à nota total, que fica no intervalo de 0 a 1000. Considerando os critérios básicos descritos na Seção 19.1.2, vale dizer que o Enem os divide da seguinte forma: (i) Língua Portuguesa, (ii) Tema e Gênero, (iii) Coerência e (iv) Coesão. Além desses, o modelo ainda avalia um quinto critério, que é a presença e adequação da "Proposta de Intervenção", que consiste na sugestão de ação ou medida interventiva para solucionar ou minimizar o problema associado ao tema proposto.

O vestibular da **Fuvest**[6] também exige um texto do gênero dissertativo-argumentativo para um tema específico que muda todo ano. O modelo de avaliação agrupa os critérios básicos em 3, sendo: (i) Tema/Gênero, (ii) Coerência/Coesão e (iii) Língua Portuguesa. Para cada um dos três aspectos, cada avaliador atribui pontuação de 1 a 5. Os pontos atribuídos a cada aspecto são multiplicados por 4, 3 e 3, respectivamente, obtendo-se, assim, uma nota ponderada para a redação, que varia entre 10 e 50 pontos.

Já no modelo de correção da **Unesp**[7], os textos, que devem seguir o gênero dissertação-argumentativa, também são avaliados em três eixos, agrupados da seguinte forma: (i) Tema, (ii) Gênero/Coerência e (iii) Língua Portuguesa/Coesão. A pontuação individual ou peso por critério não é divulgado no material do candidato, mas é definido que a pontuação final fica entre 0 e 28 pontos.

A redação da **Unicamp**[8], a cada ano, varia a exigência dos tipos e gêneros textuais[9],

[5]Cartilha do participante do Enem 2022: https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf.

[6]Manual do candidato do vestibular Fuvest 2023: https://www.fuvest.br/wp-content/uploads/fuvest2023_manual_candidato_retificado_29112022.pdf.

[7]Manual do candidato do vestibular Unesp 2023: https://documento.vunesp.com.br/documento/stream/MzQxOTk5NA%3d%3d.

[8]Manual do ingresso (https://www.comvest.unicamp.br/wp-content/uploads/2023/02/Manual_do_Ingresso_2023_Atualizado.pdf) e grade de redação (https://www.comvest.unicamp.br/vestibular-2023/grade-da-redacao/) do vestibular 2023 da Unicamp.

[9]Além da dissertação, outros gêneros textuais já exigidos pela Unicamp são: síntese e carta-convite (2015), resenha e texto de divulgação científica (2016), carta argumentativa e texto de apresentação (2017), palestra e artigo de opinião (2018), abaixo-assinado e postagem em fórum (2019), roteiro de podcast

geralmente oferecendo duas alternativas das quais o candidato deve escolher uma para execução. A Unicamp agrupa os critérios básicos da seguinte forma: (i) Tema, (ii) Gênero e (iii) Língua Portuguesa/Coerência/Coesão. Além desses três eixos, também avalia a "Leitura", que corresponde à leitura e interpretação crítica dos textos fornecidos na proposta, sem contudo copiá-los ou parafraseá-los. Na avaliação, cada critério possui pesos diferentes: Tema varia entre 0 e 2 pontos, Gênero entre 0 e 3, Leitura entre 0 e 3 e Língua Portuguesa/Coerência/Coesão varia entre 1 e 4 pontos. A soma dos pontos de cada critério leva à nota final, cujo valor máximo é de 12 pontos.

Fora todas essas diferenças já apontadas, convém ressaltar que todos os modelos penalizam o aluno (zerando a redação) no caso de falhas graves. No entanto, cada modelo define um conjunto específico de falhas graves, que podem ser: fuga ao tema, fuga ao gênero, assinatura na prova, desenho ou sinal gráfico, redação em língua estrangeira, caligrafia ilegível, recado para o corretor, parte desconectada, texto insuficiente, dentre outras situações[10].

Quanto ao título da redação, o Enem e a Unesp não exigem, mas também não proíbem; simplesmente desconsideram para a avaliação da redação. Já a Fuvest não menciona a exigência de título no Manual do candidato do vestibular Fuvest 2023[11], mas coloca como instrução no Caderno de prova[12]. Já a Unicamp pode ou não exigir título, a depender do gênero textual proposto.

Essa grande variedade de modelos de correção é considerada um dos grandes desafios para a CAR, não sendo recomendado treinar um modelo computacional que abarque todos os tipos de correção ou que misture redações dos vários tipos como dados de treinamento para algum modelo. A exigência por modelagens de nota específicas por modelo de correção não impede, no entanto, o reaproveitamento de parte das ferramentas de correção, como a detecção automática de desvios no texto, desde que modelos de correção distintos tenham diretrizes similares para esse tipo de tarefa.

Todas essas questões serão mais detalhadas ao longo deste capítulo, que está organizado da seguinte forma: a Seção 19.2 descreve como fazer a detecção de desvios em textos em português, demonstrando alguns tipos de desvios e os formalismos usados. A Seção 19.3 apresenta os principais trabalhos da literatura que realizam a atribuição da nota para redações em português. A Seção 19.4 demonstra as possibilidades de geração de um feedback para o aluno. Na Seção 19.5, discutimos as vantagens e desvantagens da correção manual e da correção automática, a fim de esclarecer ao leitor que ambas possuem potencialidades, mas também limitações. Por fim, nas Considerações finais (Seção 19.6), retomamos os pontos principais do capítulo, indicando também o que está previsto para a revisão deste capítulo na próxima versão do livro.

e crônica (2020), discurso político e diário (2021) e postagem para redes sociais e manifesto coletivo (2022).

[10]Para uma descrição completa e exemplos de todos os casos que zeram a redação em cada modelo de correção, sugere-se consultar os respectivos manuais do candidato ou cartilhas do participante.

[11]https://www.fuvest.br/wp-content/uploads/fuvest2023_manual_candidato_retificado_29112022.pdf

[12]https://acervo.fuvest.br/fuvest/2020/fuv2020_2fase_dia_1.pdf

19.2 Detecção de desvios no texto

Conforme apresentado na Introdução (Seção 19.1), consideramos que uma das etapas da Correção Automática de Redação (CAR) é a detecção ou identificação de desvios[13]. Essa etapa nem sempre é realizada nos trabalhos de Avaliação Automática de Redação, ou, por vezes, os desvios são contabilizados para o cálculo da nota, porém não são apresentados ao aluno.

A detecção desses desvios pode ser feita por meio de duas abordagens distintas: baseada em regras (**abordagem simbólica**) e baseada em modelos estatísticos (**abordagem estatística**). Os sistemas baseados em regras são mais adequados para identificar desvios gramaticais, o que é mais comum de ser cometido por falantes nativos da própria língua, enquanto os estatísticos capturam melhor os desvios de uso, que são erros mais comuns por não-nativos[14].

Embora a abordagem simbólica (baseada em regras) seja considerada obsoleta para tarefas mais complexas, ainda é a mais utilizada ainda hoje para detectar desvios na área de CAR. Para outros tipos de tarefas, modelos estatísticos e neurais performam melhor e são mais escaláveis do que modelos simbólicos. No entanto, para a tarefa de identificação de desvios em textos, ainda se usa a abordagem simbólica baseada em regras porque ela permite mostrar o erro ao aluno, explicar por que está errado e ainda fazer sugestões de correção.

Para o português, existem recursos disponíveis, tais como o CoGroo[15] e o LanguageTool[16], que são repositórios contendo regras gramaticais para a língua portuguesa. Esses recursos têm versões livres, gratuitas e de código-aberto, com extensão para navegadores web e também acopláveis a editores de texto.

Também há plataformas de correção de redação que desenvolveram seu próprio conjunto de recursos linguísticos e regras gramaticais, o que é uma boa opção quando há um padrão muito claro e estruturado que se possa expressar com regras simbólicas ou expressões regulares, que é o caso dos desvios mais comuns em redações.

Na Seção 19.2.1 caracterizamos alguns dos tipos de desvios mais comuns em redações. Posteriormente, na 19.2.2, apresentamos duas alternativas de formalismo para a definição de regras de detecção de desvios.

19.2.1 Tipos de desvios

Existem diversos tipos de desvios que podem ser marcados em uma redação, como os ortográficos, os gramaticais (ou sintáticos), os de uso de vocabulário ou registro, os desvios no uso de recursos coesivos, dentre outros. Para cada modelo de correção de redação, é possível criar uma taxonomia própria de tipos de desvios que se pretende identificar em um texto.

Ressaltamos que a criação de recursos para esse tipo de tarefa é um processo difícil, moroso e custoso, que depende de especialistas. Deste modo, é importante estabelecer um

[13] Adotaremos o conceito de "desvio" como sinônimo de "erro", mas evitaremos esse segundo termo para evitar preconceitos e julgamentos contidos na palavra "erro".

[14] Para uma explicação detalhada dos vários sistemas que usam cada uma das abordagens simbólica e estatística para detecção de desvios, ver (Leacock et al., 2010) e (Gamon et al., 2013).

[15] https://cogroo.sourceforge.net/

[16] https://languagetool.org/pt-BR/

planejamento criterioso caso seja necessário criar recursos próprios.

Nesta seção exploramos alguns tipos de desvios, por serem os mais comuns, mas é importante esclarecer que os tipos de desvios não se limitam aos indicados neste capítulo. Na Seção 19.2.1.1 descrevemos os desvios ortográficos, na Seção 19.2.1.2 os gramaticais, na Seção 19.2.1.3 os lexicais, relacionados ao vocabulário utilizado e na Seção 19.2.1.4 os desvios no uso de conectivos.

19.2.1.1 Desvios ortográficos

A grande maioria dos desvios ortográficos é facilmente detectável e tratável. O simples uso de um bom dicionário de língua portuguesa já indica quais palavras existem e quais não existem na língua. Portanto, identificar palavras com grafia desviante do léxico da língua é uma tarefa relativamente simples.

O Unitex[17], por exemplo, dispõe de três dicionários muito completos para o português: o **Delas** (com cerca de 75.000 canônicas), o **Delaf** (com cerca de 9.000.000 entradas) e o **Delacf** (com cerca de 4.000 entradas). Esse recurso pode ser usado como uma primeira etapa de identificação de desvios ortográficos, a fim de identificar palavras que existem no léxico do português e palavras desviantes.

Outros desvios ortográficos se dividem em:

- problemas de falta de acentuação ou uso indevido de acentuação (ex: “prática” x “pratica”);
- problemas de capitalização (uso de maiúscula onde deveria ser minúscula ou uso de minúscula onde deveria ser maiúscula);
- grafia incorreta de palavras homônimas ou parônimas (ex: “mas” x “mais” ou “há” x “a”);
- problemas de segmentação (uso de hífen, palavras juntas que deveriam ser separadas ou palavras separadas que deveriam ser escritas juntas); e
- desvios com relação à nova ortografia.

Para todos esses casos, a abordagem baseada em regras precisa identificar corretamente o contexto em que a palavra-alvo está inserida. O que torna essa tarefa complexa e nem sempre bem sucedida é que a identificação do contexto linguístico muitas vezes depende de um bom *parser* e um bom *tagger*. Conforme foi apresentado em capítulos anteriores (Capítulo 4 e Capítulo 7), essas ferramentas nem sempre têm uma ótima performance em português.

19.2.1.2 Desvios gramaticais

Os desvios gramaticais, também chamados de desvios sintáticos, correspondem aos problemas de estrutura sintática, ou seja, nas relações entre as palavras, que podem estar no escopo de uma sentença, um sintagma, um grupo ou uma *string*. Por exemplo, na sentença “As menina dançam”, existe um desvio de concordância nominal entre “As” (plural) e “menina” (singular) e/ou um desvio de concordância verbal entre “menina” (singular) e “dançam” (plural).

[17] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html

Além dos desvios de concordância (que correspondem a cerca de 18,9%), também são comuns em redações escolares: os de vírgula e pontuação (44%), de formas verbais (6,8%), pronomes (5,8%), preposições (5,7%), crase (4,2%), segmentação (4,1%), regência (3,4%), outros (2,3%), conjunções (2,3%) e determinantes (2%)[18].

Apesar de os desvios de pontuação e vírgula serem os mais frequentes, são também os mais difíceis de serem tratados, pois em geral a vírgula separa constituintes sintagmáticos e são raros os *parsers* de constituência para o Português[19].

As regras gramaticais que exigem um contexto linguístico local, por exemplo, para avaliação da crase ou concordância nominal, geralmente funcionam melhor, ao passo que as regras que dependem de um contexto linguístico maior, com macrorrelações de dependência (Capítulo 6), ou quando um *token* está muito distante do outro, tendem a performar mal.

19.2.1.3 Desvios de vocabulário, registro e gênero

Conforme explicado na Seção 19.1.1, cada gênero textual pode exigir um léxico ou vocabulário próprio, alguns podem exigir apenas o registro formal (norma culta) da língua portuguesa, enquanto outros admitem registro informal (linguagem coloquial).

Quando o texto da redação apresenta vocabulário, léxico ou estruturas não condizentes com o gênero textual exigido, é possível identificar desvios de vocabulário, registro ou gênero, por meio de regras formais que associam determinadas palavras e expressões a determinados gêneros.

Por exemplo, quando a proposta de redação pede uma dissertação argumentativa e o aluno usa muitos verbos e pronomes em primeira pessoa do singular, e.g. "eu", "acho", "penso", "creio" ou expressões opinativas (e.g. "na minha opinião"), todos eles podem ser considerados desvios e a redação ser penalizada em relação à adequação ao Gênero.

Uma solução possível para esses casos seria usar uma lista simples de pronomes e verbos em primeira pessoa e mais algumas expressões. Mas isso poderia trazer outros problemas, como a marcação incorreta dos casos a seguir:

- quando o aluno faz uma citação dentro ou fora de aspas (ex: **Penso**, logo **existo**);
- quando indica um livro ou filme (ex: **Eu**, **eu** mesmo e Irene);
- quando faz referência a algum perfil de rede social ou hashtag (ex: #temmeuvoto escrito como "Tem **meu** voto") ou nomes de campanhas (ex: "**Eu quero minha** biblioteca");
- quando o aluno comete um desvio gramatical e acaba gerando uma forma de primeira pessoa (ex: "Os usuários podem **vim** a ter problema" em vez de "Os usuários podem **vir** a ter problema"), onde não há desvio de gênero, e sim de conjugação verbal.

Nesses casos, a abordagem mais adequada seria por meio de regras que identifiquem o contexto em que essas palavras e expressões estão inseridas e, assim, restrinjam o contexto linguístico a fim de marcar corretamente os desvios.

[18] Esses percentuais foram calculados a partir dos números absolutos da Tabela 8 de Ramisch (2020, p. 76), que anotou os desvios sintáticos em uma amostra de 1.045 redações.

[19] O *parser* PALAVRAS (Bick, 2000) dispõe de um módulo de análise sintática por constituintes.

19.2.1.4 Desvios no uso de recursos coesivos

Quando o aluno usa algum conectivo inadequadamente, também pode ser considerado um desvio no uso de recursos coesivos e a redação ser penalizada em relação à Coesão.

Um dos desvios mais comuns em redações escolares é o uso do "contudo" com o sentido de conclusão, em vez do seu sentido original de adversidade. Também é bastante recorrente os alunos usarem o pronome "onde" em contextos não locativos, por exemplo, para se referirem a épocas, histórias, pessoas ou instituições. Nesses casos, as regras formais devem identificar o contexto em que o pronome foi usado, verificar se ele faz referência a uma palavra locativa e, se não estiver usado corretamente, sugerir que o aluno use "em que" em vez de "onde".

Também é possível criar regras para identificação de uso correto de alguns conectivos e elogiar ou contabilizar positivamente para a nota de Coesão. Nesse sentido, as regras servem não apenas para detectar desvios, mas também para detectar usos corretos e elogiáveis dos recursos coesivos.

Estes são apenas alguns dos exemplos de desvios que podem ser identificados automaticamente por meio de regras e outros recursos linguísticos, mas vários outros são igualmente possíveis.

Na seção a seguir, exemplificaremos brevemente a abordagem baseada em regras a partir da indicação de dois formalismos de regras, com exemplos em português.

19.2.2 Formalismos de regras

Há inúmeras maneiras de escrever regras de forma que o computador consiga lê-las e interpretá-las. Cada ferramenta pode criar seu próprio formalismo e mecanismo de inferência, mas também há alguns disponíveis gratuitamente e que podem ser usados para um projeto inicial.

O LanguageTool[20] implementa um mecanismo de inferência para regras formalizadas em XML (*Extensible Markup Language*). Para o português, o software disponibiliza cerca de 2.880 regras abrangendo várias categorias linguísticas, tais como: gramática geral, ortografia, pontuação, capitalização, tipografia, estilo, redundância, palavra composta, semântica, repetição, linguagem informal, uso de pronomes, dentre outras. Essas regras podem ser consultadas via repositório Language Tool Community[21].

Como exemplo, reproduzimos o formalismo de uma regra para identificar redundância quando se escreve "gelo gelado", na Figura 19.1[22].

Neste exemplo, consta o id e o nome da regra (linha 1), seguidos do padrão a ser buscado (linhas de 2 a 7), seguido da mensagem a ser mostrada (linhas 8 a 10), e de um exemplo de uso (linhas 11 a 13).

A complexidade das regras pode variar dependendo da complexidade do problema linguístico ou do padrão a ser buscado. No caso do código na Figura 19.1, o problema linguístico em questão – a redundância – é muito simples, e isso se reflete na simplicidade da regra, a qual procura basicamente dois *tokens*: o primeiro é "gelo", imediatamente seguido

[20] https://languagetool.org/pt-BR/

[21] https://community.languagetool.org/rule/list?lang=pt

[22] Fonte: regras para português no repositório languagetool (Github) (https://github.com/languagetool-org/languagetool/blob/50c9a5eb145f6289762fc64a2b8773629ca085e1/languagetool-language-modules/pt/src/main/resources/org/languagetool/rules/pt/pt-BR/style.xml#L142-L150).

Figura 19.1: Exemplo de formalismo de regra do LanguageTool

```
<rule id="GELO_GELADO" name="Gelo ">
    <pattern>
        <marker>
            <token>gelo</token>
            <token>gelado</token>
        </marker>
    </pattern>
    <message>
        Substitua «gelo gelado» por <suggestion>gelo</suggestion>.
    </message>
    <example correction="gelo">
        Vendemos uma pedra de <marker>gelo gelado</marker>.
    </example>
</rule>
```

do segundo, que é "gelado". Por outro lado, problemas linguísticos mais complexos também exigem regras mais complexas que podem usar lemas, *tokens*, etiquetas morfológicas, morfossintáticas, expressões regulares, relações de dependência, entidades nomeadas, dentre outros.

Ressaltamos que a performance das regras do LanguageTool não é ótima, mas é um recurso útil para quem não quer começar essa tarefa do zero. Considerando que o software possui versão aberta[23], é possível corrigir e definir novas regras usando o mesmo formalismo e avaliá-las por meio da própria ferramenta.

Outra ferramenta que podemos indicar para esse tipo de tarefa é o módulo Python spaCy[24], que implementa três mecanismos para identificação de padrões em textos que podem ser bastante úteis na tarefa de detecção de desvios. Esses mecanismos fazem parte do sub-módulo chamado *Rule-based matching*, que permite a busca por um *token* em determinado contexto (chamado *Matcher*), por uma frase ou sintagma (chamado *Phrase matcher*), ou ainda por relações de dependências entre elementos da sentença (chamado *Dependency matcher*). Eles podem ser usados separadamente ou combinados entre si para garantir melhor acurácia na busca por padrões linguísticos.

Na Figura 19.2, apresentamos código Python que utiliza a classe `Matcher` do spaCy para a definição da regra que identifica a redundância "gelo gelado", que foi reescrita, e executa a busca por padrões em um texto.

O exemplo inicia importando o módulo `spacy` e, especificamente, a classe `Matcher` (linhas 1 e 2). Em seguida um *pipeline* pré-treinado do spaCy para português é carregado (linha 4) e seu vocabulário é utilizado para inicializar uma instância da classe `Matcher` (linha 5). Em seguida, são definidos um identificador para a regra (linha 7), o padrão buscado de dois *tokens* (linhas 9 a 12), cada um representado por um dicionário, e a mensagem que deve ser impressa na tela caso o padrão seja identificado no texto (linha 14). Nas linhas de 16 a 20, a regra é adicionada à instância `matcher`, incluindo a definição de uma função para a impressão da mensagem na tela quando o padrão é encontrado (`on_match`). As linhas 22 a 23 definem um texto para teste de busca do padrão e a execução dessa busca.

[23]https://github.com/languagetool-org/languagetool
[24]https://spacy.io/

Figura 19.2: Exemplo de formalismo da regra “gelo gelado” usando a biblioteca spaCy

```
import spacy
from spacy.matcher import Matcher

nlp = spacy.load('pt_core_news_sm')
matcher = Matcher(nlp.vocab)

rule_id = 'GELO_GELADO'

pattern = [
    {'LOWER': 'gelo'},
    {'LOWER': 'gelado'}
]

message = 'Substitua «gelo gelado» por "gelo"'

matcher.add(
    rule_id,
    [pattern],
    on_match=lambda matcher, doc, i, matches: print(message)
)

doc = nlp('Gelo gelado gela a garganta.')
matches = matcher(doc)
```

O spaCy não conta com um repositório de regras pré-definidas para detecção de desvios. Contudo, por ser uma ferramenta de PLN, disponibiliza uma série de funcionalidades que podem contribuir para essa tarefa de maneira mais simples, i.e. sem a necessidade de alterar a implementação dos mecanismos de busca já disponíveis.

A detecção de desvios no texto é uma etapa importante em CAR, especialmente por indicar e colaborar para aprendizagem da escrita. Os desvios encontrados podem, inclusive, ser utilizados na etapa de atribuição de nota à redação. Na Seção 19.3 apresentamos as principais abordagens e tendências nessa área, além de citar os principais trabalhos dedicados a redações em português.

19.3 Atribuição de nota

A atribuição de nota a uma redação pode ser feita de forma global, ou seja, uma nota única para a redação inteira, ou por meio de notas individuais para cada critério de avaliação. No geral, as abordagens fazem uso de *corpus* rotulado, i.e., conjuntos de redações que já foram avaliadas manualmente e possuem indicação de nota e/ou adequação da redação em relação ao critério avaliado. Desse modo, as técnicas utilizadas para atribuição de nota se encaixam na área de aprendizado supervisionado por classificação ou regressão.

O *Project Essay Grade* (PEG) (Ajay; Tillet; Page, 1973) foi uma das primeiras ferramentas estáveis para a atribuição de notas em redações com boa performance dentro do contexto aplicado: redações universitárias curtas em inglês. No entanto, a falta de acesso a computadores foi, por algum tempo, impedimento para o desenvolvimento de

outras soluções. Na metade da década de 90, dados os avanços tecnológicos de hardware e software, a área de AES viu um reaquecimento e, desde então, surgiram novos trabalhos consistentemente, inclusive apoiados por abordagens que tiveram ascensão a partir da década de 2010, como *deep learning* e Transformers.

Como mencionado na Seção 19.1, é importante conhecer o contexto e modelo de correção para realizar a atribuição de nota de forma efetiva. A despeito disso, diferentes estratégias podem ser reaproveitadas e combinadas para a avaliação de redações de modelos de correção distintos. Na Seção 19.3.1 apresentamos uma visão geral de técnicas e estratégias para a atribuição de notas em redações. Dada a relevância do Enem para o contexto de redações em português, a Seção 19.3.2 traz trabalhos especificamente voltados para a automatização da avaliação em redações desse modelo de correção.

19.3.1 Como atribuir nota a redações?

A abordagem clássica para atribuição de notas envolve a extração de atributos (*features*) a partir do texto, que são utilizados para descrever redações de um conjunto de treinamento, além da transformação e seleção desses atributos, em um processo nomeado engenharia de atributos. Os dados extraídos servem de entrada para um algoritmo de aprendizado para a geração de modelos capazes de atribuir nota a novas redações a partir dos valores de seus atributos.

A primeira versão do PEG (Ajay; Tillet; Page, 1973) utilizava atributos baseados em contagens de diferentes elementos do texto, categorizadas em: (i) **simples** (e.g. número de adjetivos na redação): redações com mais adjetivos são avaliadas com notas maiores por humanos (relação linear); (ii) **enganosamente simples** (e.g. número de palavras na redação): redações muito curtas são penalizadas, porém, conforme o tamanho da redação aumenta, esse atributo perde importância para atribuição de nota (relação logarítmica); e (iii) **sofisticadas** (e.g. número de palavras que podem representar contextos maiores): o número de conectivos, por exemplo, pode indicar a complexidade de uma sentença.

Page; Petersen (1995) introduzem a terminologia de *proxes*, o que é de fato mensurável ou variáveis observáveis, e *trins*, o que se está tentando medir ou variáveis latentes. Nesse contexto, o nível de coerência de uma redação, por exemplo, pode ser considerado uma variável latente, enquanto os atributos potencialmente relacionados à coerência do texto são as variáveis observáveis. Trabalhos que exploram a atribuição de nota global podem incluir *proxes* especificamente relacionados a critérios de avaliação a fim de considerar diferentes *trins* em sua modelagem.

Considerando a tarefa de atribuição de nota, é possível utilizar atributos que sejam independentes. Ferramentas como Coh-Metrix[25] e Linguistic Inquiry Word Count (LIWC)[26], são utilizadas em trabalhos como Ferreira et al. (2021) e Ferreira Mello et al. (2022) para a extração de informações linguísticas, como legibilidade e coesão.

Trabalhos que utilizam métricas independentes de conteúdo são capazes de representar critérios de avaliação como Coerência e Coesão. No entanto, critérios como Tema são melhor

[25]Coh-Metrix é uma ferramenta computacional que calcula métricas e índices para aspectos linguísticos e discursivos em um texto e que será melhor explorada neste capítulo na Seção 19.4.1. Disponível em: http://cohmetrix.memphis.edu/cohmetrixhome.

[26]LIWC é uma ferramenta computacional, que realiza análise de textos baseada em métricas. Disponível em: https://www.liwc.app/.

avaliados por atributos dependentes de conteúdo, como exemplo as matrizes de termos, descritas no Capítulo 14, e métricas calculadas a partir dessas matrizes, como a similaridade de cosseno utilizada entre tema e redação em Amorim; Veloso (2017). Em Louis; Higgins (2010) e Persing; Ng (2014), são propostos cálculos de atributos dependentes de conteúdo com base em recursos linguísticos pré-definidos e associados aos temas relacionados às redações utilizadas nos experimentos.

Ainda sobre a extração de atributos, vale mencionar o trabalho de Sousa et al. (2021) que, além de aspectos linguísticos, explora aspectos relacionados à construção da argumentação, por meio de mineração de argumentos. A combinação de diferentes estratégias para extração de atributos é bastante comum, conforme realizado por Amorim; Veloso (2017) que, além de aspectos linguísticos e associados ao tema, incluem métricas associadas ao correto uso da língua, calculadas com base em desvios identificados por ferramentas externas, como as mencionadas na Seção 19.2.

É importante ressaltar que a inclusão de atributos relacionados a critérios de avaliação específicos não é imprescindível para atribuição de nota global. No entanto, a partir do momento em que se propõe atribuir notas por critérios avaliativos, é interessante incluir atributos que representem cada critério, ou poderá haver discrepância significativa no resultado obtido entre critérios, como observado em alguns trabalhos (Amorim; Veloso, 2017; Fonseca et al., 2018).

Selecionado um conjunto de atributos e realizada a análise estatística dos dados, podemos seguir à etapa de treinamento de modelos. Não convém aqui sugerirmos esta ou aquela técnica ou algoritmo, uma vez que conjuntos de dados distintos podem apresentar resultados também diferentes para os mesmos algoritmos (Ferreira Mello et al., 2022; Ferreira et al., 2021; Fonseca et al., 2018; Marinho et al., 2022). Ao treinar modelos para atribuição de nota, assim como modelos com outros objetivos, é fundamental definir mais de um algoritmo e configurações para, então, realizar uma comparação estatística entre os resultados obtidos.

Entre os trabalhos que utilizam a abordagem de extração de atributos, há modelos de classificação e regressão treinados com diversos algoritmos, como: regressão linear (Fonseca et al., 2018), *Suppport Vector Machines* (SVM) (Haendchen Filho et al., 2018, 2019), *Gradient Boosting* (Fonseca et al., 2018; Marinho et al., 2022). A comparação entre os modelos se dá, principalmente, pela avaliação dos valores obtidos para métricas como precisão, revocação, medida-F, RMSE e Kappa de Cohen.

Embora seja possível obter resultados satisfatórios pela engenharia de atributos e treinamento de modelos por algoritmos clássicos de aprendizado de máquina, é notável o esforço humano necessário para o processo de extração e seleção de atributos, considerando que muitos dos conjuntos de atributos são compostos por algumas centenas de métricas. Com isso em vista, surgem trabalhos que utilizam outras técnicas para a representação de textos e algoritmos de redes neurais profundas para a tarefa de atribuição de nota.

Alikaniotis; Yannakoudakis; Rei (2016) propõem uma técnica de *word embeddings* treinada com base em notas de redações, que é utilizada com redes neurais LSTM. O trabalho relata melhores resultados obtidos em comparação com outras abordagens.

Em Fonseca et al. (2018), as *word embeddings* GloVe são combinadas com redes LSTM bidirecionais e os resultados são comparados, também, com uma abordagem que utiliza engenharia de atributos. Os autores relatam que, embora a técnica de redes neurais tenha gerado bons resultados, o modelo gerado a partir de atributos se mostrou superior em diferentes aspectos.

Mayfield; Black (2020) realizam *fine-tuning* de modelos pré-treinados (BERT e variações) para a atribuição automática de notas. Apesar de relatar resultados até 5% melhores do que modelos baseados em n-gramas, os autores discutem sobre o tempo de treinamento deste tipo de modelo, que é cerca de 100 vezes mais demorado do que outras abordagens, e sobre o impacto que isso pode ter em fluxos mais dinâmicos de trabalho.

Bittencourt Jr. (2020) define 14 técnicas baseadas em combinações de diferentes representações de palavras e arquiteturas de redes neurais profundas para a execução da tarefa de atribuição de nota a redações. Os experimentos são realizados com um conjunto composto por redações de 18 temas, sendo que cada técnica é utilizada para o treinamento de um modelo por tema (18 modelos por técnica). Também é proposta uma abordagem para treinamento de modelo multi-tema, ou seja, um modelo único para a atribuição de notas para redações de mais de um tema.

O trabalho de Marinho et al. (2022) compara 3 tipos de abordagens: (i) engenharia de atributos com algoritmo de regressão, (ii) *doc embeddings* com algoritmo de regressão e (iii) *word embeddings* com LSTM. As abordagens (i) e (iii) apresentaram melhores resultados para critérios de avaliação distintos, sendo a abordagem (iii) eleita pelos autores como a melhor. Os resultados da abordagem (iii) ainda foram comparados com resultados de Amorim; Cançado; Veloso (2018) e Fonseca et al. (2018), sendo relatado melhor desempenho desta abordagem na atribuição de nota por critério de avaliação.

Quanto a abordagens para atribuição de notas, entendemos que muito já foi desenvolvido, especialmente para o inglês, que conta com ferramentas comerciais bem estabelecidas para atribuição de nota e correção de redações. No entanto, para o português, a limitação de recursos, modelos de anotação morfossintática e, inclusive, de conjuntos de dados, podem ser obstáculos para os trabalhos nessa área.

19.3.2 Atribuição de nota para redações do Enem

Especificamente para o português, a maior parte dos trabalhos relacionados à atribuição de notas (e CAR) utiliza *corpora* compostos por redações do modelo de correção do Enem como base de treinamento. Dada a importância e dimensão do exame no Brasil, há interesse particular em encontrar soluções para a atribuição de nota exclusivamente para esse modelo de correção.

Como descrito na Seção 19.1.3, o Enem exige a produção de um texto do gênero dissertativo-argumentativo sobre um tema específico que é avaliado em 5 critérios, também chamados de competências: (1) Língua portuguesa, (2) Abordagem temática e adequação ao tipo textual, (3) Progressão textual e defesa do ponto de vista, (4) Coesão e articulação e (5) Proposta de intervenção. Os trabalhos que treinam modelos de atribuição de nota para o Enem predizem uma nota global, porém alguns tentam também aperfeiçoar por competência.

Em Amorim; Veloso (2017), Fonseca et al. (2018), Marinho et al. (2022) e Bittencourt Jr. (2020), o foco está na atribuição de notas para cada uma das competências, o que pode ser feito com base em modelos treinados para cada competência ou um modelo único que prediz as notas para cada uma delas. Já em Haendchen Filho et al. (2018), é explorada a atribuição de notas para a competência 2, especificamente.

Ao realizar estudo sobre a predição de notas para cada uma das competências do Enem, Haendchen Filho et al. (2019) notaram o significativo desbalanceamento do conjunto de

redações e tornaram esse o foco de seu trabalho, a fim de analisar o impacto e tratamento de conjuntos de dados desbalanceados na tarefa de atribuição de nota.

Alguns trabalhos que utilizam redações em português não têm como foco direto a atribuição de nota, mas a proposta de técnicas mensuráveis relacionadas a critérios cuja avaliação pode ser mais complexa. Como exemplo, citam-se as contribuições de Ferreira et al. (2021), Sousa et al. (2021) e Ferreira Mello et al. (2022) para a avaliação das competências 3 e 4.

É notável que, no momento da escrita deste capítulo, não pudemos encontrar nenhum trabalho em que se dê atenção em particular para a melhoria de atribuição de nota na competência 5 do Enem.

Vale ressaltar que os conjuntos de dados utilizados pelos referidos trabalhos não são muito representativos, possuindo até alguns milhares de redações de uma baixa diversidade de temas. O maior conjunto relatado possui 56.644 redações, sem indicação de número de temas (Fonseca et al., 2018). O conjunto de redações com maior número de temas relatado, que também é o segundo em número de redações, conta com 27.184 redações distribuídas entre 18 temas, sendo que o número de redações por tema varia entre 3.070 e 710 (Bittencourt Jr., 2020). Além disso, ambos os maiores conjuntos foram fornecidos por empresas privadas e, portanto, não são públicos.

O tamanho e distribuição do conjunto de dados são considerados obstáculos para o treinamento de modelos de atribuição de notas, especialmente quando utilizadas técnicas de *deep learning*. Mesmo com a aplicação de outras técnicas, nesse contexto, a comprovação e generalização de resultados é um desafio. No entanto, há uma iniciativa para a criação de um conjunto público de redações do modelo Enem para utilização em trabalhos de CAR: até agosto de 2022 era composto por 6.579 redações pré-processadas e divididas em 151 temas (Marinho; Anchiêta; Moura, 2022) [27].

Também não encontramos nenhum trabalho de PLN que tenha relatado a atribuição de notas em redações de outros modelos de correção, como Fuvest, Unicamp, FGV ou outros.

Enfim, acreditamos que ainda há espaço para trabalhos quanto à tarefa de atribuição de notas em redações em português. Contudo, para atingir a meta de soluções completas de correção de redação, apenas a nota é insuficiente do ponto de vista do processo de ensino e aprendizagem. Para suprir essa lacuna, a Seção 19.4 discute a terceira tarefa de CAR, referente ao provimento de feedback relacionado ao texto.

19.4 Feedback para o aluno

Conforme apresentado na Introdução (Seção 19.1), a última etapa da Correção Automática de Redação (CAR) é o fornecimento de um feedback para o aluno. Até pouco tempo atrás, a correção automática produzia basicamente uma nota como resultado da avaliação da redação. Mas isso já não era mais suficiente e foi surgindo a necessidade de explicar ou justificar essa nota. De acordo com Shermis; Burstein (2013), os primeiros trabalhos se limitavam a dar feedbacks sobre as características e propriedades linguísticas do texto. Pesquisas mais recentes vêm focando em aspectos mais complexos e profundos da língua, que vão além da superficialidade do texto[28].

[27] https://github.com/lplnufpi/essay-br

[28] Vale ressaltar que os feedbacks baseados em características e propriedades linguísticas do texto ainda são os mais praticados hoje pelas plataformas brasileiras ou que processam o português, então focaremos

Em uma **correção manual**, esse feedback é feito pelo próprio corretor da redação, na forma de comentário livre, em linguagem natural, sem seguir nenhum tipo de padronização, podendo tecer críticas, fazer sugestões, elencar pontos fortes e pontos a melhorar, abordar questões gerais ou específicas da redação, enfim, de formas bastante variadas.

Já em uma **correção automática**, as plataformas que dão algum tipo de feedback sobre a correção o fazem de forma sistematizada. Porém, são raras as empresas que fornecem esse tipo de devolutiva ao aluno. Lima et al. (2023) fizeram uma revisão sistemática da literatura sobre CAR e uma das lacunas que identificaram nos trabalhos para o português é o baixo detalhamento nos feedbacks retornados pelos modelos de avaliação.

Na prática, os corretores automáticos costumam apontar apenas estatísticas básicas do texto, tais como quantidade de conectivos (conjunções), variação lexical (taxa de *types* por *tokens*), quantidade de palavras de conteúdo (substantivos, adjetivos, verbos e alguns advérbios), tamanho médio das palavras, frases e parágrafos, dentre outros, o que geralmente não tem utilidade pedagógica para o aluno. A Seção 19.4.1 apresenta como essas informações são calculadas e exibidas.

Algumas plataformas de CAR também disponibilizam para o aluno sistemas ou *bots* baseados em assistentes de escrita ou ferramentas computacionais de auxílio à escrita. Na Seção 19.4.2 apresentamos como esses recursos e ferramentas são utilizadas em sistemas de CAR.

Mais recentemente, com o surgimento e popularização do ChatGPT, algumas empresas também já começaram a fornecer feedbacks gerados automaticamente por esses modelos gerativos. Também é possível gerar automaticamente as devolutivas a partir de elementos encontrados ou não encontrados no texto, instanciando palavras ou trechos do texto da redação. Mas isso só é possível se for usada uma abordagem simbólica. Nesse sentido, o feedback pode conter críticas referenciando os desvios apresentados na Seção 19.2 e/ou elogios aos pontos fortes, como será apresentado na Seção 19.4.3.

19.4.1 Estatísticas básicas do texto

Algumas plataformas e empresas privadas que oferecem serviço de CAR apresentam para o aluno contagens básicas do texto, tais como a quantidade de palavras, caracteres, sentenças, parágrafos e até a quantidade de palavras por classe gramatical (verbos, substantivos, adjetivos, preposições, conjunções etc.). Outras oferecem um pouco mais de informação baseada em estatísticas simples, como a proporção de palavras únicas (*types*) em relação à quantidade total de palavras no texto (*tokens*), alguma medida de similaridade entre as sentenças, desvio padrão dos parágrafos, dentre outras.

Um dos recursos disponíveis para recuperar essas informações é o NILC-Metrix (Leal et al., 2021), uma versão brasileira do Coh-Metrix. O NILC-Metrix[29] é a atualização mais recente do Coh-Metrix-Port (Scarton; Aluísio, 2010), contendo 200 métricas[30] distribuídas nas 14 categorias apresentadas no Quadro 19.2, as quais avaliam a coerência, a coesão, a inteligibilidade, a complexidade e outros aspectos:

nessa abordagem ao longo desta seção.

[29] http://fw.nilc.icmc.usp.br:23380/metrixdoc

[30] Definição, explicação e exemplos das métricas podem ser conferidos na Documentação do NILC-Metrix (http://fw.nilc.icmc.usp.br:23380/metrixdoc).

Quadro 19.2: Categorias de métricas disponíveis no NILC-Metrix.

Categoria	Qtde.	Descrição das métricas
Medidas Descritivas	10	Quantificam sílabas por palavra, palavras por sentença, sentenças por parágrafo, assim como quantidades absolutas de palavras, sentenças e parágrafos.
Simplicidade Textual	8	Analisam proporção de palavras fáceis e difíceis em relação ao total de palavras, bem como sentenças longas e curtas.
Coesão Referencial	9	Avaliam quantidades de palavras de conteúdo, radicais de conteúdo, médias de referentes, de candidatos a referentes e proporção de pronomes anafóricos.
Coesão Semântica	11	Calculam similaridade, entropia, desvio padrão e taxa de *giveness*, ou seja, quantidade de informação dada e nova baseada em LSA (Capítulo 10).
Medidas Psicolinguísticas	24	Verificam 4 critérios: concretude, imageabilidade, familiaridade e idade de aquisição das palavras.
Diversidade Lexical	15	Calculam proporções e desvio padrão de *types* e *tokens*, tanto gerais quanto por PoS e por categorias de palavras lexicais e funcionais.
Conectivos	12	Verificam a proporção de vários tipos de conectivos (aditivos, causais, lógicos) e operadores lógicos (positivos e negativos) em relação ao total de palavras do texto.
Léxico Temporal	12	Consideram verbos flexionados nos diferentes tempos e modos verbais, bem como as formas regular e irregular de particípio.
Complexidade Sintática	27	São bastante diversas, considerando desde a quantidade de orações por sentença, coordenação e subordinação, até aspectos como voz ativa e passiva, aposto, distância na árvore de dependência, fórmulas de Frazier e Yngve, dentre vários outros.
Densidade de Padrões Sintáticos	4	Calculam proporção de verbos no gerúndio e também tamanhos de sintagmas nominais.
Informações Morfossintáticas de Palavras	42	Verificam quantidade média, proporção ou desvio padrão das palavras por PoS.
Informações Semânticas de Palavras	11	Incluem dados sobre substantivos abstratos, palavras polissêmicas, hiperônimos, nomes próprios e polaridade (positiva e negativa) das palavras.
Frequência de Palavras	10	Calculam frequências de vários tipos de palavra (de conteúdo, e.g.) com base na curva de Zipf e considerando diferentes *corpora*.
Índices de Leiturabilidade	5	Incluem índices e fórmulas já consolidados em PLN, tais como Fórmula Dale Chall, Índice de Brunet, Flesch, Gunning Fog e Estatística de Honoré.

Os cálculos dessas métricas geralmente resultam em um valor numérico, o qual não se faz útil para o aluno. Porém, há diferentes maneiras de devolver ao aluno um feedback textual com a interpretação de algumas dessas métricas. Por exemplo, se considerarmos os valores de 4 métricas de Simplicidade Textual, referentes a tamanho de sentença (a saber: *long_sentence_ratio*[31], *medium_long_sentence_ratio*[32], *medium_short_sentence_ratio*[33], *short_sentence_ratio*[34]), é possível criar um resultado interpretável para dizer ao aluno que ele constrói sentenças muito longas e isso pode prejudicar a compreensão das ideias do texto.

Tanto as estatísticas básicas quanto as métricas do NILC-Metrix podem ser utilizadas não apenas para devolver feedbacks aos alunos, mas também como atributos para calcular a nota da redação ou de alguns aspectos da redação, conforme apresentado na Seção 19.3.

[31]Proporção de sentenças muito longas em relação a todas as sentenças do texto.
[32]Proporção de sentenças longas em relação a todas as sentenças do texto.
[33]Proporção de sentenças médias em relação a todas as sentenças do texto.
[34]Proporção de sentenças curtas em relação a todas as sentenças do texto.

19.4.2 Assistentes de escrita e ferramentas de auxílio à escrita

Para prover uma devolutiva ao estudante, também é possível recorrer a sistemas prontos de PLN, como os assistentes virtuais, assistentes de escrita ou ferramentas de auxílio à escrita. Essas soluções podem ser entendidas como aplicações finais, mas, na área de CAR, elas são usadas como recursos ou ferramentas intermediárias para subsidiar a solução completa de CAR.

Essas ferramentas são capazes de gerar, melhorar, reformular e personalizar qualquer tipo de conteúdo textual, incluindo redações. Algumas delas funcionam de forma síncrona *real-time*, fazendo correções e dando sugestões à medida que o texto está sendo escrito, enquanto outras funcionam a posteriori, ou seja, depois que o aluno submete sua redação à plataforma de correção, ele recebe uma devolutiva com críticas e/ou elogios.

Para a língua inglesa, há inúmeros assistentes de escrita e muitos deles conhecidos no Brasil porque as pessoas usam o inglês para escrever, por exemplo, artigos científicos. Um dos mais populares é o Grammarly[35], mas também há outros bastante usados, como Linguix[36], Ginger[37], Reverso[38], Writer[39], Hemingway App[40] e outros.

Para o português, também existem vários softwares comerciais, sendo a maioria paga. As ferramentas de auxílio à escrita, ao lado dos simplificadores textuais e dos sumarizadores automáticos, podem contribuir com a área de CAR, pois fornecem:

- **Correção ortográfica e gramatical:** Os sistemas podem usar regras ou modelos de linguagem treinados em um grande volume de textos em português para identificar erros ortográficos e gramaticais comuns.
- **Análise de contexto:** As ferramentas não apenas verificam palavras isoladas, mas também consideram o contexto da frase em que uma palavra está inserida. Isso ajuda a evitar falsos positivos e permite que o sistema forneça sugestões de correção mais precisas.
- **Sugestões de melhoria:** Quando uma palavra é identificada como incorreta ou quando uma construção gramatical suspeita é detectada, o assistente de escrita oferece sugestões para corrigir o problema. Essas sugestões podem incluir substituições de palavras, ajustes na estrutura da frase ou correções de pontuação.
- **Detecção de estilo:** Além de corrigir erros básicos, um assistente de escrita também pode oferecer sugestões para melhorar o estilo de escrita. Isso inclui alertas sobre uso excessivo de palavras, repetições, uso inadequado de voz passiva, entre outros aspectos.
- **Feedback de clareza:** As ferramentas também podem avaliar a clareza do texto, identificando frases longas e complexas que podem ser difíceis de entender, podendo sugerir dividir essas frases ou reformulá-las para tornar o conteúdo mais acessível.
- **Verificação de plágio:** Algumas soluções comerciais oferecem uma funcionalidade adicional para verificar a originalidade do texto, identificando trechos que possam ser semelhantes a outras fontes online. Isso é especialmente útil para evitar acidentalmente

[35]https://www.grammarly.com/
[36]https://linguix.com/
[37]https://www.gingersoftware.com/
[38]https://www.reverso.net/tradu%C3%A7%C3%A3o-texto
[39]https://writer.com/grammar-checker/
[40]https://hemingwayapp.com/

usar conteúdo plagiado.

- **Aprendizado contínuo:** Assim como outras ferramentas de PLN, os assistentes de escrita também continuam aprendendo e melhorando com o tempo. Eles são atualizados com novos dados e feedbacks dos usuários, o que ajuda a aprimorar seus modelos e a abordagem dos problemas linguísticos.
- **Extensões e integrações:** Muitos deles oferecem extensões para navegadores, complementos para processadores de texto e aplicativos móveis, o que permite aos usuários verificar seu conteúdo em tempo real enquanto escrevem em várias plataformas.
- **Personalização:** Em alguns desses sistemas, o usuário pode personalizar as configurações com base em suas preferências de estilo e escrita. Isso permite adaptar as sugestões e correções de acordo com o contexto e o público-alvo.

Conforme dito anteriormente, as melhores ferramentas de auxílio à escrita que existem hoje para o português são soluções comerciais de empresas privadas. Existem também alguns sistemas desenvolvidos a partir de pesquisas acadêmicas e científicas, mas nenhuma focada em redação. Por exemplo, o SciPo[41] (Feltrim et al., 2003), que é um sistema de auxílio à escrita de resumos acadêmicos em português, especialmente para teses e dissertações da área da Ciência da Computação. Outro exemplo é o WRITEME[42] (Leite et al., 2020), que é ferramenta de auxílio à escrita de READMEs que usa dados abertos dos repositórios mais populares do GitHub para gerar recomendações de seções, mas também não é focada e redação.

19.4.3 Identificação de pontos fortes e elogiáveis

Na Seção 19.2 falamos da detecção de pontos fracos e desvios no texto. Por outro lado, também é importante detectar pontos fortes e elogiáveis e demonstrá-los ao aluno para que ele continue usando a mesma estratégia nos próximos textos.

Esses pontos fortes podem ser identificados por meio de regras formais, mas também é possível usar diferentes estratégias para cada aspecto da avaliação.

Tendo identificado todos ou alguns aspectos (positivos ou negativos) do texto, é possível retornar essas informações ao aluno na forma de feedbacks construtivos para auxiliá-lo a se tornar um escritor mais habilidoso e confiante.

Avaliação de coesão Na Seção 19.2.1.4, falamos brevemente de como identificar usos corretos de recursos coesivos usando regras em contextos linguísticos menores, como dentro de uma sentença.

Também é possível criar regras formais que percorrem todo o texto procurando as ocorrências de conectivos, avaliar a sua distribuição ao longo do texto, calcular a variabilidade e diversificação deles e até procurar conectivos em pontos específicos da redação, como no início da conclusão, por exemplo.

Com o objetivo de fornecer um feedback baseado na avaliação da coesão do texto, uma solução simples é usar um *tagger* que identifique palavras etiquetadas como conjunções, preposições e advérbios, ou usando listas e léxicos específicos. A outra

[41] https://escritacientifica.sc.usp.br/scipo/
[42] https://repositorio.ufpe.br/handle/123456789/50043

solução, que é um pouco mais rebuscada, é recorrer às métricas do NILC-Metrix que incidem sobre a coesão textual.

Identificação de repertórios Para avaliar a abordagem temática, referente à competência 2 do Enem, podemos elogiar a presença (ou criticar a ausência) de repertórios socioculturais, que são informações, fatos, citações, definições ou termos de alguma área do conhecimento, ou ainda experiências pessoais que, de alguma forma, contribuem como argumento para defender um ponto de vista.

Pelo Manual de leitura do Enem[43], os repertórios socioculturais podem ser legitimados (com citação da fonte) ou não legitimados (sem citação da fonte), ter uso produtivo (pertinente à discussão em mais de um momento do texto) ou não, pertencente ao tema ou não e ainda devem ser penalizados se forem exclusivamente baseados nos textos motivadores. Identificar automaticamente todos esses tipos e usos (corretos ou não) dos repertórios não é uma tarefa simples. Porém isso pode ser feito usando modelos de extração de entidades nomeadas (Capítulo 17), buscando, por exemplo, as citações de filósofos, sociólogos e outros estudiosos, ou buscando as menções a livros, filmes, séries, dentre outras entidades que funcionem como repertórios legitimados.

Avaliação da Progressão textual Para a avaliação da progressão textual, é possível treinar e usar modelos de tópico, a exemplo do *Hidden Topic Markov Models* (HTMM) (Gruber; Weiss; Rosen-Zvi, 2007), que classificam as sentenças de um texto por tópicos ou assuntos, o que nos permite avaliar a progressão, a continuidade, a retomada e até a circularidade entre os assuntos, a partir da distribuição dos tópicos em um texto.

Blei; Moreno (2001) apresentam resultados dessa abordagem de segmentar um texto não estruturado em tópicos, testando em notícias do New York Times. Os autores propuseram uma combinação do tradicional modelo oculto de Markov (*Hidden Markov Model* – HMM) com o modelo de semântica latente de Hofmann (Hofmann, 1999), resultando em um novo método probabilístico que segmenta um texto em tópicos. Essa abordagem pode ser muito útil para avaliar o encadeamento das ideias de um texto, principalmente no caso de redação do Enem, pois o gênero dissertativo-argumentativo costuma seguir um padrão bem definido de: (i) apresentação do tema, introdução ao ponto de vista a ser defendido e breve menção aos argumentos a serem utilizados (no primeiro parágrafo); (ii) exposição do primeiro argumento (no segundo parágrafo); (iii) exposição do segundo argumento (no terceiro parágrafo); (iv) proposta de intervenção para solucionar o problema discorrido e retomada da tese na forma de conclusão (no quarto parágrafo).

Identificação de proposta de intervenção No Enem, para que uma redação receba nota máxima na Competência 5, o estudante precisa criar uma proposta de intervenção que contenha pelo menos 5 elementos: o agente (quem?), a ação (o quê?), o modo ou meio (como?), a finalidade (para quê?) e o detalhamento de algum dos elementos anteriores.

Para a correta identificação desses elementos, pode-se usar modelos de extração de informação (Capítulo 17) ou recorrer a extração de entidades nomeadas ou recursos linguísticos como listas e léxicos específicos.

[43] https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_2.pdf

O Inep disponibiliza a Cartilha do participante[44] com instruções sobre agentes que devem ser considerados nulos, ações interventivas que devem ser consideradas nulas, propostas de intervenção negativas ou condicionais, dentre outras orientações que podem se transformar em atributos para modelos.

Não encontramos nenhum trabalho para o português que reporte bons resultados quanto à identificação da proposta de intervenção e que valha ser replicado. É um dos campos de CAR que merece ser mais explorado.

Ao longo desta Seção 19.4, discutimos algumas formas possíveis de devolver um feedback ao aluno, que podem ser: indicando números, percentuais e estatísticas básicas do texto, ou acoplando um assistente de escrita ao corretor automático para fazer isso em tempo real, ou ainda instanciando elementos da redação (recursos coesivos, repertórios, sequência de tópicos, elementos da proposta de intervenção) em uma mensagem gerada automaticamente. Mas as possibilidades não se limitam a essas indicadas neste capítulo. Para outras formas de geração de feedbacks em redações escolares, ver Gamon et al. (2013).

Tendo em vista todo o conteúdo apresentado na Seção 19.2, Seção 19.3 e Seção 19.4, pode surgir o questionamento sobre o papel (ou até extinção) das correções manuais, dados os avanços em CAR. Para tanto, na Seção 19.5, propomos uma discussão sobre prós e contras de cada uma das abordagens de correção: a manual e a automática, apresentando alguns casos de sucesso e defendendo uma correção híbrida que se beneficie do potencial de cada abordagem.

19.5 Correção manual vs(?) correção automática

A correção automática de redações (CAR) divide opiniões entre estudantes, escritores, professores de redação, bancas de avaliação em série, especialistas em Linguística Computacional, cientistas de dados e desenvolvedores de sistemas. Ainda existe muito preconceito quando se trata de correção automática de redação, mas já é consensual aceitar as vantagens dos corretores ortográficos e gramaticais quando embutidos em outras soluções, como no pacote Office, no Gdrive, em redes sociais ou nos teclados dos smartphones.

A discussão principal gira em torno de seus prós e contras, se a correção automática deve substituir ou complementar a correção humana, sobre questões éticas relacionadas à correção automática, sobre a subversão dos valores pedagógicos e educacionais da avaliação manual para uma avaliação automática de textos; enfim, para uma discussão mais filosófica e profunda sobre todos esses aspectos, ver Elliot; Klobucar (2013) e Hakuta (2013).

Nesta seção, abordaremos apenas questões práticas relacionadas à correção manual e à correção automática de redações para, ao final, defender uma correção híbrida, que utilize as principais potencialidades de cada tipo, reconhecendo-se também suas limitações.

19.5.1 Avanços dos últimos anos

Até as décadas de 80 e 90, as avaliações de redação no Brasil eram holísticas, ou seja, o avaliador do texto atribuía uma nota global (de 0 a 100, por exemplo) para a redação,

[44] https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf

sem seguir rigorosamente nenhum critério previamente estabelecido. Por volta dos anos 2000, essas avaliações passaram a ser analíticas, tendo que explicitar todos os critérios e todos os conceitos que deveriam ser avaliados. Ao mesmo tempo, as avaliações passaram a ser em duplas às cegas, ou seja, cada redação deveria ser avaliada por dois corretores independentes, o que exigia maior sistematicidade e coerência entre eles.

Nessa transição de avaliação holística para analítica, as grades de correção de redações se tornaram mais padronizadas. E sabe-se que tarefas mais padronizadas são melhor executadas por máquinas do que por humanos.

Mesmo com a tentativa (por vezes, falha) de padronização das grades, ainda se percebe a falta de objetividade na definição de critérios por parte de alguns modelos de correção. Quando a grade de correção é muito aberta ou não apresenta os critérios bem definidos para cada faixa de nota, aumentam as chances de haver divergência entre duas avaliações cegas. Por outro lado, quando os corretores humanos passam por treinamentos rigorosos, tal como é feito no Enem, isso pode reduzir o número de inconsistências nas avaliações, mas ainda assim não elimina as divergências, já que pessoas diferentes podem ter interpretações diferentes sobre a mesma instrução. Prova disso são os índices de redações do Enem que vão para uma terceira correção[45], nos casos de discrepância de 80 ou mais pontos em uma competência ou de 100 ou mais pontos na nota final.

Posto isso, a correção automática no Brasil passou a ser considerada como uma alternativa à manual, já que esta última sempre foi passível de subjetividade e viés.

19.5.2 Vantagens da correção automática

Correções manuais estão sujeitas a subjetividade e viés, além do cansaço humano, a pressão por produtividade, a cobrança por eficiência, o desinteresse pela tarefa, dentre outros fatores que podem prejudicar a qualidade da avaliação ou comprometer sua validade.

Para além dessas questões de limitação humana, é necessário mencionar também o tempo e o custo da correção manual. De acordo com uma matéria veiculada no Portal G1[46] em 2016, os corretores humanos conseguem corrigir, em média, 74 redações por dia. Já Bittencourt Jr. (2020, p. 19) apresenta uma média de 12 minutos por correção, o que daria 40 redações por dia, considerando-se 8 horas de trabalho. E o custo de cada correção de redação do Enem para o Governo Federal era de R$15,88 em 2015. No mesmo ano foram corrigidas 6.54 milhões de redações, perfazendo um custo aproximado de R$104 milhões para o governo. Esse valor provavelmente está defasado, mas foi o último registro oficial encontrado.

Automatizar a correção de redações traz como vantagem a redução do custo de correção e elimina os fatores problemáticos relacionados ao trabalho humano.

[45] Os índices de terceira correção variam a cada ano, pois dependem de vários aspectos, inclusive a mudança dos critérios do Inep para a terceira correção. A título de exemplificação, podemos citar o índice de 20,10% em 2012 disponível no Portal do MEC (http://portal.mec.gov.br/component/tags/tag/correcao). Também se pode inferir o índice de 43,52% em 2014, a partir DE "Ao todo, foram corrigidos 6.193.565 textos. [...] foram encaminhadas 2.695.949 redações para um terceiro corretor." disponível no Portal do MEC (http://portal.mec.gov.br/component/tags/tag/espelho-da-redacao). Ou uma estimativa de 29% em 2017 "O Inep estima que das 4,1 milhões de redações corrigidas, cerca de 1,2 milhão receberão a terceira correção." disponível no Portal do MEC (http://portal.mec.gov.br/component/tags/tag/correcao).

[46] https://g1.globo.com/educacao/enem/2016/noticia/corretores-de-redacao-do-enem-avaliam-em-media-74-redacoes-por-dia.ghtml

Outro aspecto da correção que merece ser comparado é a confiança (ou *reliability*, em inglês). Os sistemas automáticos têm confiança de 100%, o que não pode ser afirmado para a correção manual. Isso significa que toda vez que a mesma redação passar pelo mesmo sistema de correção automática, receberá a mesma correção e a mesma nota. Isso parece óbvio, mas não é o que acontece na correção humana. Diferentes pessoas que corrigirem a mesma redação poderão naturalmente atribuir diferentes notas e/ou apontar diferentes aspectos a serem melhorados. O que também ocorre é que a mesma redação, quando corrigida pelo mesmo corretor humano em diferentes momentos, também pode receber avaliações muito diferentes, o que abre brecha para reclamações.

19.5.3 Vantagens da correção manual

Apesar de todos esses aspectos negativos em relação à correção manual, deve-se ressaltar o ponto forte desse tipo de correção, que é a possibilidade que o humano tem de observar todo e qualquer aspecto relacionado ao processo de construção de sentidos em um texto, o que a máquina não é capaz de fazer.

A produção textual é um processo sócio-cognitivo muito complexo que vai além da capacidade dos sistemas computacionais. A máquina não entende a redação, não interpreta o conteúdo veiculado pelo texto, mas apenas se comporta da forma como ela foi treinada para fazê-lo. Por mais que alguns modelos computacionais possam ser "interpretáveis", é impossível identificar e definir todos os fatores sociais, psicológicos, cognitivos, emocionais etc. que podem interferir tanto no processo de escrita por parte do aluno quanto no processo de correção por parte do avaliador.

Nesse sentido, considerando que a correção automática é limitada, é passível de erros e está mais voltada para a avaliação da forma do que do conteúdo, levanta-se o seguinte questionamento: O uso da correção automática não levaria o aluno a focar sua atenção apenas nos aspectos formais da escrita, excluindo os aspectos mais ricos da construção de sentidos no texto? Por trás desse questionamento, existe uma preocupação legítima de que o aluno não construa sua própria autonomia enquanto escritor, mas apenas seja "adestrado" a escrever de uma forma que o algoritmo lhe dê uma nota boa.

19.5.4 O exemplo da língua inglesa

Para a língua inglesa, algumas instituições educacionais (e.g. **ETS** – *Educational Testing Service*) utilizam modelos de AEE para auxiliar (e não substituir) a correção manual. Esses modelos computacionais são usados como uma segunda avaliação, complementar à avaliação humana. Por exemplo, a avaliação do **TOEFL** (*Test of English as a Foreign Language Internet-based Test*) é dupla, sendo uma feita por humanos e outra feita por sistemas automáticos. A nota final do aluno é dada pela média das duas avaliações. No caso de divergência entre as notas, a redação é enviada a terceiro corretor, semelhante ao que ocorre na avaliação do Enem. Mas, no caso do Brasil, tanto o primeiro quanto o segundo avaliador são humanos. Vale ressaltar que, segundo Bridgeman (2013, p. 227), "a experiência com o programa TOEFL iBT sugere que, quando há discrepância e a redação é enviada a um avaliador humano adicional, esse avaliador tende a concordar com a máquina

mais do que com o outro humano"[47].

Um processo semelhante ocorre na avaliação do **GRE** (*Graduate Record Examination*), mas neste último caso a nota atribuída automaticamente é usada como se fosse uma validação para a avaliação humana. Em outras palavras, a correção automática é usada para monitorar a performance dos corretores humanos, a fim de identificar avaliadores desalinhados ou que precisam passar por novo treinamento.

A correção automática também pode auxiliar a correção manual no sentido de "nivelar" diferentes níveis de rigor. Sabe-se que diferentes avaliadores humanos podem ser sistematicamente mais rígidos ou mais permissivos em suas correções. Segundo Braun (1988, p. 1), "Quando o grau de leniência/severidade do avaliador pode ser atestado adequadamente, é possível calibrar estatisticamente os avaliadores e ajustar as pontuações corretamente [...] Essa calibração estatística parece ser uma abordagem econômica para aumentar a confiabilidade da nota quando comparada ao simples aumento do número de avaliadores por artigo."[48].

Nesse sentido, os modelos de AEE podem auxiliar a calibrar essas diferenças de rigidez, atribuindo um peso maior às correções dos avaliadores mais permissivos e um peso menor às correções dos avaliadores mais rigorosos.

19.5.5 O que defendemos

Levantamos todos esses questionamentos ao longo da Seção 19.5 a fim de tornar explícitas as potencialidades da área de CAR, mas, ao mesmo tempo, esclarecer ao leitor sobre suas limitações, da mesma forma que a correção humana também possui vantagens e desvantagens.

Tendo considerado os vários aspectos das duas abordagens, defendemos neste capítulo uma correção híbrida, semelhante ao que é praticado para a língua inglesa (Seção 19.5.4), que possa se beneficiar dos pontos positivos da correção automática, mas mantendo a correção manual para garantir a responsabilização do humano sobre a avaliação.

19.6 Considerações finais

Neste capítulo exploramos uma das várias aplicações do Processamento de Linguagem Natural (PLN), a chamada Correção Automática de Redação (CAR), a qual abarca duas áreas de PLN em inglês, representadas pelas siglas AES (*Automated Essay Scoring*) e AEE (*Automated Essay Evaluation*).

Ademais, defendemos uma abordagem holística para a CAR, abrangendo, no mínimo, três fases essenciais: (i) a detecção de desvios no texto, (ii) a atribuição de nota, e (iii) a geração de feedback construtivo para o estudante. Apesar de termos dividido essas etapas para fins didáticos, é crucial reconhecer sua interdependência no decorrer do processo. Por exemplo:

[47]Tradução nossa. Do inglês: "*In fact, experience with the TOEFL iBT program suggests that when flagged discrepant scores are sent to an additional human rater, that rater tends to agree with the machine more often than she or he agrees with the other human score.*"

[48]Tradução nossa. Do inglês: "*When rater leniency/severity can be adequately documented, it is possible to statistically calibrate raters and adjust scores accordingly [...] This statistical calibration appears to be a cost-effective approach to enhancing scoring reliability when compared to simply increasing the number of readings per paper.*"

1. os desvios gramaticais, ortográficos, de vocabulário etc. podem ser usados como atributos para o cálculo da nota;
2. as métricas e estatísticas básicas do texto, além de serem usadas para criar os feedbacks, também podem servir como atributos para o cálculo da nota;
3. a nota atribuída pelo modelo pode restringir, limitar ou ajudar a selecionar o feedback mais apropriado a ser exibido para o aluno;
4. A caracterização dos desvios (por tipos e quantidades) também pode ser usada para a geração do feedback.

Assim, ainda que tenhamos delimitado didaticamente essas três etapas, é importante ressaltar que, no contexto das tarefas de CAR, tais fases são intrinsecamente entrelaçadas e interdependentes, colaborando harmoniosamente para aprimorar a avaliação da redação.

Embora existam numerosos estudos nesses campos para o inglês e outras línguas, a documentação relevante para o português ainda é escassa e a maioria dos trabalhos acadêmicos confiáveis foi conduzida em pequenas amostras de dados. O progresso mais notável para textos em português provém de empresas e plataformas privadas que oferecem serviços de CAR. No entanto, os métodos e resultados dessas empresas nem sempre são divulgados, e, mesmo se o fossem, seria difícil compará-los devido à falta de uniformidade entre as soluções apresentadas.

Nesse sentido, a área de CAR ainda apresenta um vasto campo de trabalho a ser explorado por novos pesquisadores. Para o português, ainda faltam bons *datasets* de redações, que contenham, além dos textos, as notas por competência, anotação e apontamentos feitos por humanos; também faltam ferramentas robustas de detecção de desvios e de auxílio à escrita, bem como bons *parsers* e *taggers*; e faltam trabalhos que reportem bons resultados, com engenharia de atributos, comparação da performance dos algoritmos utilizados e uma análise aprofundada dos resultados.

Na próxima versão do livro, pretendemos incrementar este capítulo com algumas informações que consideramos relevantes, como: (i) atividades práticas para quem quer ingressar na área de CAR; (ii) limitações da correção automática, com relação a codificação de caracteres, tokenização, hifenização (em quebra de linha), paragrafação, presença de título e de outros elementos textuais externos à redação, como assinatura, turma e outros metadados; e (iii) a compilação de um *dataset* de redações que possam ser usadas para treinamento e testes de modelos.

Referências

AJAY, H. B.; TILLET, P.; PAGE, E. B. **Analysis of essays by computer (AEC-II)**. Storrs, CT: Univeristy of Connecticut, 1973.

ALIKANIOTIS, D.; YANNAKOUDAKIS, H.; REI, M. **Automatic Text Scoring Using Neural Networks**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. **Anais**...Association for Computational Linguistics, 2016.

AMORIM, E.; CANÇADO, M.; VELOSO, A. **Automated Essay Scoring in the Presence of Biased Ratings**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Association for Computational Linguistics, 2018.

AMORIM, E.; VELOSO, A. **A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese**. Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017.

BICK, E. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. tese de doutorado—[s.l.] Aarhus University Press, Denmark; University of Arhus, 2000.

BITTENCOURT JR., J. A. S. **Avaliação automática de redação em língua portuguesa empregando redes neurais profundas**. mathesis—[s.l.] Universidade Federal de Goiás, 2020.

BLEI, D. M.; MORENO, P. J. **Topic Segmentation with an Aspect Hidden Markov Model**. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. **Anais**...New York, NY, USA: Association for Computing Machinery, 2001.

BRAUN, H. I. Understanding Scoring Reliability: Experiments in Calibrating Essay Readers. **Journal of Educational Statistics**, v. 13, n. 1, p. 1–18, 1988.

BRIDGEMAN, B. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 221–232.

DA SILVA JR., J. A. **Um avaliador automático de redações**. mathesis—[s.l.] Universidade Federal do Espírito Santo, 2021.

ELLIOT, N.; KLOBUCAR, A. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 16–35.

FELTRIM, V. D. et al. **A Construção de uma Ferramenta de Auxílio à Escrita de Resumos Acadêmicos em Português**. Anais do Encontro Nacional de Inteligência Artificial (ENIA'2003). **Anais**...SBC, 2003.

FERREIRA MELLO, R. et al. **Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese**. (A. F. Wise, R. Martinez-Maldonado, I. Hilliger, Eds.)LAK22 Conference Proceedings. **Anais**...United States of America: Association for Computing Machinery (ACM), 2022.

FERREIRA, R. et al. Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays. Em: [s.l: s.n.]. p. 162–167.

FONSECA, E. R. et al. **Automatically Grading Brazilian Student Essays**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Springer International Publishing, 2018.

GAMON, M. et al. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 251–266.

GRUBER, A.; WEISS, Y.; ROSEN-ZVI, M. **Hidden Topic Markov Models**. Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. **Anais**...: Proceedings of Machine Learning Research.San Juan, Puerto Rico: PMLR, mar. 2007.

HAENDCHEN FILHO, A. et al. **An approach to evaluate adherence to the theme and the argumentative structure of essays**. International Conference on Knowledge-Based Intelligent Information & Engineering Systems. **Anais**...2018.

HAENDCHEN FILHO, A. et al. Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. **Procedia Computer Science**, v. 159, p. 764–773, jan. 2019.

HAKUTA, K. Handbook of Automated Essay Evaluation: Current Applications and New Directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 347–353.

HOFMANN, T. **Probabilistic Latent Semantic Indexing**. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99). **Anais**...New York, NY, USA: Association for Computing Machinery, 1999.

LEACOCK, C. et al. **Automated Grammatical Error Detection for Language Learners**. [s.l.] Morgan; Claypool Publishers, 2010.

LEAL, S. E. et al. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **CoRR**, v. abs/2201.03445, 2021.

LEITE, H. et al. **WRITEME: uma Ferramenta de Auxílio à Escrita de READMEs Baseada em Dados Abertos**. Anais do XVII Congresso Latino-Americano de Software Livre e Tecnologias Abertas. **Anais**...Porto Alegre, RS, Brasil: SBC, 2020.

LIMA, T. B. DE et al. Avaliação Automática de Redação: Uma revisáo sistemática. **Revista Brasileira de Informática na Educação**, v. 31, p. 205--221, maio 2023.

LOUIS, A.; HIGGINS, D. **Off-topic essay detection using short prompt texts**. Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. **Anais**...Los Angeles, California: Association for Computational Linguistics, jun. 2010.

MARCUSCHI, L. A. **Produção textual, análise de gêneros e compreensão**. [s.l.] Parábola Ed., 2008.

MARINHO, J. et al. **Automated Essay Scoring: An approach based on ENEM competencies**. Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...SBC, 2022.

MARINHO, J.; ANCHIÊTA, R.; MOURA, R. Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task. **Journal of Information and Data Management**, v. 13, n. 1, p. 65–76, 2022.

MAYFIELD, E.; BLACK, A. W. **Should You Fine-Tune BERT for Automated Essay Scoring?** Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. **Anais**...Association for Computational Linguistics, jul. 2020.

PAGE, E. B.; PETERSEN, N. S. The Computer Moves into Essay Grading: Updating the Ancient Test. **Phi Delta Kappan**, v. 76, p. 561–565, mar. 1995.

PERSING, I.; NG, V. **Modeling Prompt Adherence in Student Essays**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. **Anais**...Baltimore, Maryland: Association for Computational Linguistics, jun. 2014.

RAMISCH, R. **Caracterização de desvios sintáticos em redações de estudantes do ensino médio: subsídios para o processamento automático das línguas naturais**. mathesis—[s.l.] Universidade Federal de São Carlos, 2020.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. **Linguamática**, v. 2, n. 1, p. 45–61, abr. 2010.

SHERMIS, M. D.; BURSTEIN, J. **Handbook of Automated Essay Evaluation: Current Applications and New Directions**. [s.l.] Routledge/Taylor & Francis Group, 2013.

SOUSA, A. et al. **Cross-Lingual Annotation Projection for Argument Mining in Portuguese**. (G. Marreiros et al., Eds.)Progress in Artificial Intelligence. **Anais**...Springer International Publishing, 2021.