Investigando coerência em postagens de um fórum de dúvidas em ambiente virtual de aprendizagem com o BERT

Osmar de Oliveira Braz Junior[1,2], **Renato Fileto**[2]

[1]Universidade do Estado de Santa Catarina (UDESC)
[2]Universidade Federal de Santa Catarina (UFSC) – PPGCC/INE/CTC
Florianópolis, Santa Catarina, Brasil.

osmar.braz@{udesc.br,posgrad.ufsc.br}, r.fileto@ufsc.br

Abstract. *Incoherences can cause difficulties in interpreting discourses and impact the performance of conversational agents and intelligent tutoring systems, among others. Contextualized language models, such as BERT, have not yet been exploited in coherence analysis, despite their proven efficacy in several related tasks. This work employs Portuguese language variations of BERT to classify and measure text coherence. Experiments with news and an educational forum of student questions show that BERT supports sentence order discrimination with up to* 99.20% *accuracy and measures of (in)coherence consistent with such classification, being most of the best results for the forum texts.*

Resumo. *Incoerências podem causar dificuldades na interpretação de discursos e impactar o desempenho de agentes conversacionais e tutores inteligentes, entre outros. Modelos contextualizados de linguagem como o BERT não foram ainda explorados na análise de incoerência, a despeito de sua eficácia comprovada em diversas tarefas afins. Este trabalho usa variações do BERT em língua portuguesa para classificar e medir coerência textual. Experimentos com textos de notícias e de um fórum educacional de dúvidas de estudantes mostram que o BERT suporta discriminação da ordem de sentenças com até* $99,20\%$ *de acurácia e algumas medidas de (in)coerência consistentes com tal classificação, sendo a maioria dos melhores resultados para os textos do fórum.*

1. Introdução

Coerência é um importante traço da qualidade de um discurso e um dos critérios que distingue um texto bem escrito de outro confuso e difícil de compreender. Um discurso coerente consiste de agrupamentos semanticamente coesos e encadeados de componentes textuais, tais como sentenças e/ou tópicos nelas mencionados. Assim, modelos para medir coerência [Barzilay and Lapata 2008] são frequentemente baseados na similaridade de componentes textuais adjacentes. Por outro lado, modelos para classificar coerência de textos são usualmente produzidos via aprendizado supervisionado. O treinamento destes modelos costuma partir da premissa de que textos originalmente bem escritos são coerentes, enquanto permutações aleatórias de seus componentes os tornam incoerentes.

A análise de coerência pode auxiliar tarefas de Processamento de Linguagem Natural (PLN) [Xu et al. 2019] como sumarização de textos [Dias 2016], avaliação de atividades discursivas [Neto et al. 2020], conversação [Mohiuddin et al. 2018] e *Question Answering (QA)* [Xu et al. 2019]. O tratamento de incoerências pode melho-

DOI: 10.5753/sbie.2021.217397

rar o desempenho de um agente conversacional (*chatbot*) ou Sistema Tutor Inteligente (STI) [Oliveira et al. 2020], entre outros. Tais sistemas automatizados são cruciais, porque dispor de humanos para atendimento 24x7 é financeiramente inviável.

Alguns modelos recentes de coerência textual [Mohiuddin et al. 2018, Xu et al. 2019] e soluções para correção automática de respostas discursivas [Neto et al. 2020, Oliveira et al. 2020] medem similaridade/distância semântica de textos usando *embeddings* de palavras, tais como Word2Vec [Mikolov et al. 2013] e Glove [Pennington et al. 2014]. Um Modelo Contextualizado de Linguagem (MCL) como o BERT gera *embeddings* que variam de acordo com o contexto textual de cada ocorrência de um léxico, o que permite capturar nuances de significados. MCLs têm propiciado ganhos de desempenho em diversas tarefas de PLN [Devlin et al. 2019]. Porém, seu potencial ainda não foi explorado na análise de coerência textual.

Este trabalho usa o $BERTimbau$ [Souza et al. 2020] (BERT pre-treinado para a língua portuguesa) e o $BERT_{Multilingual}$ [Devlin et al. 2019] para classificar e medir a coerência de textos. O objetivo é comparar o desempenho dessas variações do BERT com o estado da arte e verificar se há diferenças de desempenho e de coerência entre textos de noticiário com redação profissional e postagens de um fórum de dúvidas de estudantes em disciplinas no Ambiente Virtual de Aprendizagem (AVA) Moodle de uma universidade brasileira. Primeiramente, essas variações do BERT com diferentes configurações de hiperparâmetros para ajuste fino são avaliadas na tarefa de discriminação da ordem de sentenças [Barzilay and Lapata 2008], i.e., diferenciar documentos originais (supostamente coerentes) de versões com suas sentenças aleatoriamente permutadas (supostamente incoerentes). Posteriormente, medidas de distância e similaridade entre *embeddings* de sentenças produzidos com o BERT são usadas para computar medidas de (in)coerência de documentos originais e de suas versões com sentenças permutadas.

Todas as variações do BERT usadas nos experimentos superaram o *baseline* [Dias 2016] na tarefa de discriminação da ordem de sentenças em língua portuguesa. O BERT também suportou o cálculo de medidas de (in)coerência que permitem discriminar a ordem de sentenças, apesar da proximidade dos valores dessas medidas para documentos originais e com sentenças permutadas. As melhores acurácias do nosso classificador foram obtidas com textos do AVA. Isso ilustra o potencial de MCLs como o BERT para lidar com incoerências em textos de estudantes, o que pode contribuir para superar desafios em sistemas conversacionais, tutores inteligentes e corretores de atividades discursivas.

Este artigo está organizado como se segue. A Seção 2 apresenta os fundamentos necessários ao seu entendimento. A Seção 3 discute os trabalhos relacionados. A Seção 4 descreve a proposta. A Seção 5 reporta os experimentos. A Seção 6 apresenta e discute os resultados. Finalmente, a Seção 7 enumera contribuições e os trabalhos futuros.

2. Fundamentos

2.1. *Embeddings* de texto

Diversas tarefas de PLN têm se beneficiado do uso de representações vetoriais de texto em *embeddings* que capturam propriedades semânticas [Zhang et al. 2020]. *Embeddings* de palavras podem ser contextualizados ou não [Qiu et al. 2020]. *Embeddings* clássicos (e.g., Word2vec, Glove) são representações estáticas não contextualizadas com um

problema fundamental: geram um mesmo *embedding* para cada léxico, independentemente de possíveis variações do seu significado em cada contexto textual em que aparece.

O BERT [Devlin et al. 2019] (acrônimo de *Bidirectional Encoder Representations from Transformers*) é um MCL constituído de uma rede neural profunda com processamento bidirecional. Ele é pré-treinado com grandes volumes de documentos não rotulados de cunho geral, mas permite ajustes finos com a adição de apenas uma camada de saída, visando otimizar o seu desempenho em tarefas como classificação de texto, reconhecimento de entidades nomeadas e QA [Devlin et al. 2019], em certos tipos de corpora (e.g., noticiário, diálogos entre tutores e estudantes).

MCLs como o BERT geram *embeddings* de tokens (palavras ou sub-palavras). Todavia, alguns modelos de coerência textual requerem *embeddings* de sentenças para medir distância ou similaridade entre elas. Duas estratégias de *pooling* [Reimers and Gurevych 2019] são frequentemente usadas na literatura para produzir *embeddings* de sentenças a partir dos *embeddings* dos tokens nelas contidos. A primeira calcula a média (MEAN) em cada posição dos *embeddings* a serem consolidados em um único, enquanto a segunda utiliza o maior (MAX) valor de cada posição. Neste trabalho avaliamos essas duas estratégias de *pooling*, considerando o número de dimensões (H) da camada oculta do BERT. Dadas duas sentenças distintas s_i e s_j ($i \neq j$) de um documento, o BERT mapeia seus respectivos tokens em *embeddings* que são consolidados usando essas estratégias de *pooling* nos respectivos *embeddings* de sentenças, $\tilde{s}_i$ e $\tilde{s}_j$. Funções de similaridade ou distância recebem $\tilde{s}_i$ e $\tilde{s}_j$ retornando uma medida a ser utilizada, por exemplo, no cálculo da (in)coerência. Medidas como distância Euclidiana (euc) e Manhattan (man) servem para avaliar incoerência entre sentenças, enquanto similaridade cosseno (cos) serve para avaliar coerência. Essas medidas estão entre as mais usadas no cálculo de (in)coerência usando uma variedade de modelos, como os discutidos a seguir.

2.2. Modelos de Coerência

Modelos de coerência visam distinguir discursos coerentes de incoerentes e/ou medir sua (in)coerência. A coerência pode ser analisada sob duas perspectivas: global e local. Coerência global analisa como os elementos de um documento em sua totalidade são vinculados, enquanto a coerência local analisa a coesão e consistência de sequências menores [Barzilay and Lapata 2008]. Neste trabalho, adotamos a perspectiva de coerência local devido ao contexto limitado dos diálogos na área de educação que pretendemos tratar e à limitação da quantidade de tokens (512) de entrada no BERT.

Entre os modelos de coerência local podemos destacar: a Teoria da Estrutura Retórica (*Rhetorical Structure Theory* - RST) [Mann and Thompson 1987] e a Grade de Entidades [Barzilay and Lapata 2008]. Esses modelos requerem análise sintática de discurso ao nível de sentenças e entidades, respectivamente, enquanto o nosso trabalho usa somente a proximidade semântica dos *embeddings* das palavras e sentenças.

As abordagens mais comuns para avaliar modelos de coerência são a discriminativa e a generativa [Li and Jurafsky 2017]. Na abordagem discriminativa o objetivo é distinguir documentos originais de suas versões com sentenças aleatoriamente permutadas. Um fator crucial nesta abordagem é a quantidade de permutações, recomenda-se em torno de 20 permutações distintas por documento [Barzilay and Lapata 2008]. Assim, cada qual precisa ter ao menos 4 sentenças. Neste trabalho optamos pela abordagem

discriminativa, por não necessitar rotular dados manualmente para ajuste fino do MCL. Calculamos a (in)coerência $C_m(D)$ de um documento D usando a Equação 1, adaptada de [Foltz et al. 1998], a qual define coerência e incoerência como a média de alguma medida m de distância ou similaridade, respectivamente, entre pares de sentenças adjacentes (s_i, s_{i+1}). Nos experimentos reportados neste trabalho utilizamos $m \in \{euc, man, cos\}$, com euc e man referindo-se respectivamente à distância Euclidiana (L2) e Manhathan (L1) e cos referindo-se à similaridade cosseno.

$$(in)Coherence_m(D) = C_m(D) = \frac{1}{n-1}\sum_{i=1}^{n-1} m(\tilde{s}_i, \tilde{s}_{i+1}) \quad (1)$$

3. Trabalhos relacionados

Várias propostas da literatura [Dias 2016, Mohiuddin et al. 2018, Xu et al. 2019], aplicam algoritmos de aprendizado de máquina ou redes neurais em modelos de coerência que exploram diferentes características (*features*) extraídas dos textos. Relações de coerência inspiradas na RST [Mann and Thompson 1987] são explorados com classificadores binários de coerência baseados em *Suport Vector Machines* (SVM) em [Dias 2016]. Tal trabalho não usa *embeddings* contextualizados, mas fornece o *baseline* e o conjunto de dados em língua portuguesa que utilizamos. Outro trabalho [Mohiuddin et al. 2018] aplica redes neurais convolucionais a representações em grades distribuídas das transições de entidades salientes (substantivos e objetos) em diálogos assíncronos para avaliar a sua coerência. Avalia o impacto de alternativas para determinar entidades salientes, mas não considera *embeddings* contextualizados. Além disso, nenhum desses trabalhos confronta classificação com mensuração de coerência ou investiga coerência em fóruns de AVA.

Coerência local e global são tratadas por [Xu et al. 2019] em um com modelo neural que emprega *embeddings* de sentenças do Glove, mas permite usar outros *embeddings*. Redes siamesas são exploradas por [Reimers and Gurevych 2019] para encontrar pares de sentenças semanticamente semelhantes, utilizando diferentes medidas de similaridade e distância entre *embeddings* do BERT. Tal trabalho fornece estratégias que usamos neste trabalho para gerar *embeddings* de sentenças com tamanho padronizado a partir dos *embeddings* de suas palavras, mas não considera coerência entre sentenças.

Outros trabalhos [Oliveira et al. 2020, Neto et al. 2020, Cavalcanti et al. 2020] manipulam textos em educação. Um sistema para correção automática de questões discursivas, que compara respostas de professores com as de alunos foi proposto por [Oliveira et al. 2020], usando diferentes medidas de similaridade e distância entre *embeddings*. Porém, sem avaliar coerência entre sentenças com *embeddings* contextualizados. A coerência de redações é avaliada automaticamente usando *Random Forests* em [Neto et al. 2020]. A qualidade de feedbacks de instrutores é analisada usando mineração de dados e características linguísticas como coerência em [Cavalcanti et al. 2020]. Porém, ambas as propostas dependem de dados rotulados e não usam *embeddings*. Nosso trabalho é o primeiro a explorar variações do BERT para classificar ordem de sentenças em língua portuguesa, além de avaliar várias medidas alternativas de (in)coerência.

4. Uma abordagem baseada no BERT para classificar e medir coerência

O nosso modelo de coerência baseado no BERT inclui um classificador binário de coerência e um medidor de (in)coerência. A Figura 1 ilustra o fluxo de informação proposto.

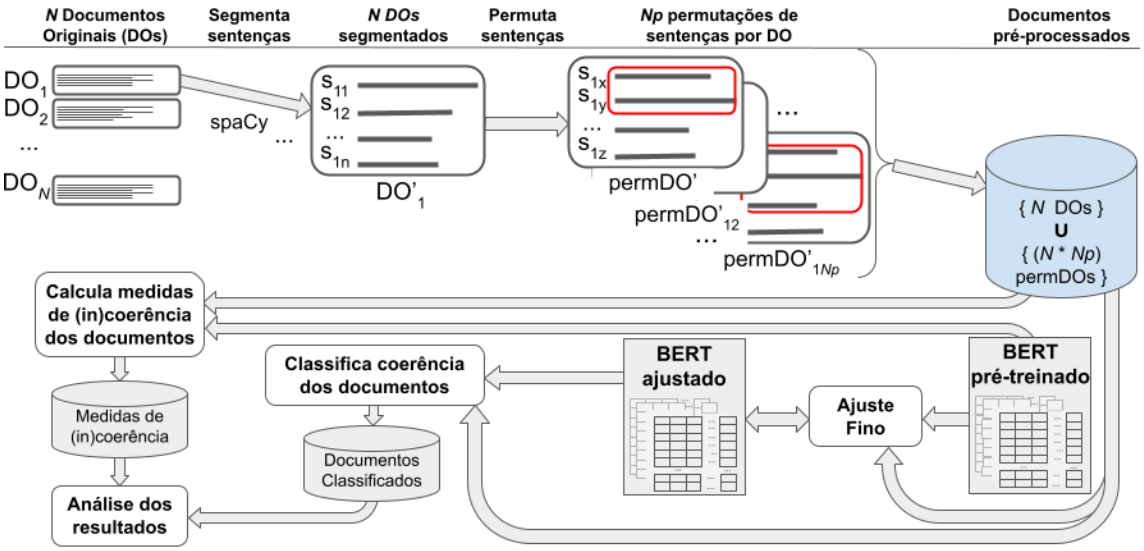


Figura 1. Classificação e medição de coerência usando o BERT.

Ele se inicia pela segmentação de N documentos originais (DO) em suas sentenças e seleção dos documentos com até 512 tokens (limitação de entrada do BERT). Para cada DO são então geradas Np permutações aleatórias distintas de suas sentenças, cada qual ($permDO$) supostamente menos coerente que o respectivo DO, resultando em $N * Np$ pares $\langle DO, permDO \rangle$ para treinar e avaliar classificadores e mensuradores de coerência.

O ajuste fino do BERT para classificação de coerência é feito com os documentos originais (DO) rotulados como coerentes e as suas versões com sentenças permutadas ($permDO$) rotuladas como incoerentes. O classificador de coerência usa o token de classificação ($[CLS]$) da última camada do BERT como discriminador do documento [Devlin et al. 2019]. O cálculo das medidas de (in)coerência dos documentos usa o BERT pré-treinado no modo avaliação. Seguindo recomendações de [Devlin et al. 2019], não utilizamos todas as camadas do BERT, mas somente a concatenação das 4 últimas. A função C_m recebe um documento D (DO ou $permDO$) e primeiramente o converte em uma representação por *embeddings* de sentenças. Todas as sentenças de cada documento são submetidas simultaneamente ao BERT por dois motivos: para obter *embeddings* de palavras que considerem todo o contexto e para capturar relações entre as sentenças. Os *embeddings* de sentenças podem ser produzidos aplicando as estratégias de *pooling* MAX e MEAN aos *embeddings* de todas as palavras ou somente as consideradas relevantes (e.g., exceto *stop words*, substantivos) nas respectivas sentenças. As medidas de (in)coerência são calculadas para cada documento de acordo com a Equação 1, usando diferentes medidas de distância e similaridade entre os *embeddings* de sentenças produzidos com variações do BERT e as estratégias de *pooling*. Ao final do processo, as distribuições das classificações e medidas de (in)coerência produzidas são analisadas.

5. Experimentos

A nossa proposta foi implementada na linguagem Python versão 3.6.9 e executada no ambiente Google Collaboratory, usando a arquitetura padrão do BERT implementada na biblioteca Hugginface versão 4.5.1. A segmentação dos documentos e a análise sintática das sentenças foi feita com a ferramenta spaCy versão 2.3.5. Utilizamos o BERT pré-

treinado na língua portuguesa ($BERTimbau$ [Souza et al. 2020][1]) nas dimensões $Base$ (768) e $Large$ (1.024) e o $BERT_{Multilingual}$[Devlin et al. 2019] somente $Base$, todos no formato "*cased*" (com caracteres maiúsculos e minúsculos). Os *embeddings* do BERT foram manipulados usando os métodos da biblioteca PyTorch versão 1.8.1. A implementação completa e resultados de alguns experimentos estão disponíveis no Github[2].

5.1. Conjuntos de dados

Os experimentos usam duas coleções de dados: OnlineEduc 1.0 (privado) e CSTNews[3] (público). O CSTNews foi criado para avaliar sumarização de documentos jornalísticos de diferentes fontes (Jornal do Brasil, Folha de São Paulo e O Estado de São Paulo). Originalmente possuía 1 resumo multidocumento para cada coleção de textos em certo assunto. Foi estendido por [Dias 2016], com mais 5 resumos para cada uma das 50 coleções, totalizando 300 resumos. Destes, removemos 2 resumos por terem mais tokens que o limite de entrada do BERT e 49 com menos de 4 sentenças, de modo a viabilizar 20 permutações aleatórias distintas das suas sentenças. Assim, restaram para os experimentos 249 resumos originais (DO) e suas 4.980 versões com sentenças permutadas ($permDO$). Usamos o CSTNews como conjunto de dados de controle, porque estamos interessados na classificação de textos curtos, visando futuras aplicações de QA no ensino a distância.

O OnlineEduc 1.0 contém postagens de fóruns de dúvidas do Moodle ao longo de 4 semestres de oferecimento de disciplinas de um curso superior a distância de uma universidade brasileira, entre os anos de 2017 e 2019. As postagens foram pré-processadas para remover tags html e xml, eliminar repetições de espaços e de pontuações (e.g., ???,!!!) e substituir urls pelo token "endereco_url". Do total de 8.492 postagens, foram selecionadas 1.080 $(12,72\%)$ por terem ao menos uma sentença terminando com '?' (pergunta). Dessas 1.080 postagens foram selecionadas 561 com 4 a 10 sentenças, O limite inferior (4) tem por objetivo gerar ao menos 20 permutações por postagem e o limite superior (20) se deve à limitação de tokens de entrada do BERT. Assim, restaram para os experimentos 561 postagens (DO), com as 11.220 permutações ($permDO$) aleatórias de suas sentenças (20 por DO).

5.2. Classificação e mensuração de coerência

O ajuste fino do BERT nos experimentos de classificação utilizam o agendador de taxa de aprendizado sem aquecimento, seguido por decaimento linear ($get\_linear\_schedule\_with\_warmup$) da taxa de aprendizado ao longo das etapas de treino. Usamos o otimizador AdamW da implementação Hugginface do BERT com $\beta_1 = 0,9$, $\beta_2 = 0,999$ (valor padrão). Os treinamentos foram interrompidos após certo número de épocas. Os hiperparâmetros são específicos para cada tarefa [Devlin et al. 2019] e sua busca exaustiva. Executamos 45 experimentos para cada conjunto de dados buscando os hiperparâmetros e o MCL ótimo para o classificador. As faixas de valores (sugeridas pelos autores) foram combinadas: taxa de aprendizagem de $1 * 10^{-5}$, $2 * 10^{-5}$, $3 * 10^{-5}$, $4 * 10^{-5}$, $5 * 10^{-5}$ (incluímos $1 * 10^{-5}$ e $4 * 10^{-5}$), número de épocas em {2, 3, 4} e as três variações do BERT. Devido a limitações de memória do ambiente de execução, utilizamos lotes de tamanho 4 para o treino e 8 para os testes, ao

[1] https://github.com/neuralmind-ai/portuguese-bert/
[2] https://github.com/osmarbraz/cohebert_v1/
[3] https://sites.icmc.usp.br/taspardo/Summary%20coherence%20models.zip

invés do sugerido pelos autores [Devlin et al. 2019]. O desempenho do classificador foi avaliado pela média das acurácias na validação cruzada (*10-fold cross validation*).

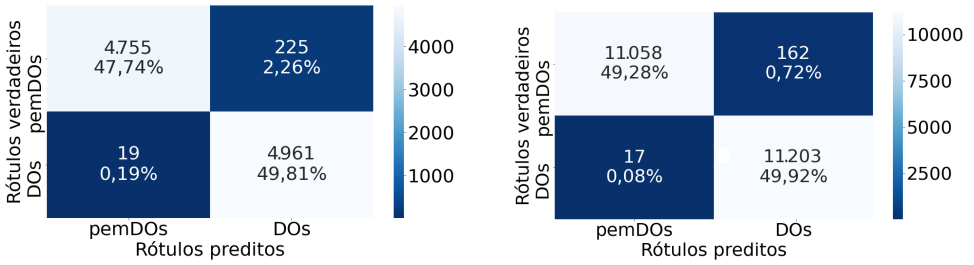
As medidas de (in)coerência foram calculadas com duas estratégias de *pooling* (MEAN e MAX) e três alternativas para selecionar palavras relevantes das sentenças: todas as palavras (ALL), remoção de *stop words* (CLEAN) e somente substantivos (NOUN). Combinando essas alternativas e as 3 variações do BERT foram realizados 18 experimentos para cada conjunto de dados. Cada C_m é calculada conforme especificado na Equação 1, com $m \in \{cos, euc, man\}$, i.e., similaridade cosseno (cos) para calcular coerência e distâncias Euclideana (euc) e Manhathan (man) para incoerência. Essas medidas foram calculadas para cada documento original (DO) do CSTNews e do OnlineEduc 1.0 e $Np = 20$ permutações aleatórias ($permDO$) distintas das sentenças de cada DO. Foram avaliadas a proporção de DOs com C_m melhor do que seus respectivos $permDOs$ e a distribuição das medidas C_m entre esses documentos.

6. Resultados e Discussão

A Tabela 1 apresenta a *Acurácia* média da avaliação *10-fold* de cada variação do classificador BERT na tarefa de distinguir documentos originais (DO) de permutados ($permDO$), com os respectivos números de *Épocas* e taxas de aprendizagem (*Taxa Apr.*) usados no treinamento. Os melhores resultados e respectivos hiperparâmetros são realçados em negrito. O desempenho foi sempre superior com o OnlineEduc 1.0. A Figura 2 mostra as matrizes de confusão dos classificadores para cada conjunto de dados.

Tabela 1. Hiperparâmetros e acurácias da classificação.

MCL	CSTNews			OnlineEduc 1.0		
	Épocas	Taxa Apr.	Acurácia	Épocas	Taxa Apr.	Acurácia
$BERTimbau_{Base}$	4	$3 * 10^{-5}$	97,44%	4	$3 * 10^{-5}$	99,14%
$BERTimbau_{Large}$	**4**	$\mathbf{10^{-5}}$	**97,55%**	2	10^{-5}	99,07%
$BERT_{Multilingual}$	4	10^{-5}	97,48%	**4**	$\mathbf{10^{-5}}$	**99,20%**



(a) CSTNews. (b) OnlineEduc 1.0.

Figura 2. Matrizes de confusão dos melhores resultados dos classificadores.

Todas as variações do classificador BERT superam os $92,69\%$ de acurácia do classificador baseado no modelo de relações discursivas de [Dias 2016]. Este foi escolhido como *baseline* por seu alto desempenho e por fornecer o CSTNews para os nossos experimentos. Infelizmente, não foi possível obter sua acurácia para o OnlineEduc 1.0, porque esse modelo requer a anotação manual das relações discursivas.

Outro experimento foi realizado para avaliar a generalização do classificador BERT para conjuntos de dados distintos, treinando o $BERTimbau_{Base}$ com um conjunto de dados e avaliando com outro. O ajuste fino utilizou 4 épocas e taxa de aprendizagem de 10^{-5}. Utilizando o CSTNews no treinamento e avaliando com o OlineEduc 1.0 a acurácia foi de $53,51\%$, enquanto para o inverso foi de $70,82\%$. Portanto, o treinamento com o OnlineEduc 1.0 permitiu melhores resultados, apesar dos textos serem do fórum do AVA, enquanto os do CSTNews são resumos de notícias da imprensa profissional.

A Tabela 2 apresenta os percentuais de pares $\langle DO, permDO \rangle$ (cada documento original pareado com cada uma de suas 20 permutações) onde a medida de (in)coerência C_m de DO é melhor que a de $permDO$. Os melhores percentuais para cada conjunto de dados e variação do BERT, destacados em negrito, foram alcançados com a medida de incoerência C_{euc} e o $BERTimbau_{Base}$, usando a estratégia de MEAN para consolidar os *embeddings* de todas as palavras de cada sentença. Assim, a seguir são analisadas distribuições de C_{euc} obtida com esta estratégia.

Tabela 2. Proporção de pares $\langle DO, permDO \rangle$ com melhor C_m para DO.

	CSTNews			OnlineEduc 1.0		
MCL	C_{cos}	C_{euc}	C_{man}	C_{cos}	C_{euc}	C_{man}
$BERTimbau_{Base}$	62,77	**64,18**	63,21	78,10	**80,01**	79,78
$BERTimbau_{Large}$	62,89	62,89	**62,91**	74,40	78,75	**79,47**
$BERT_{Multilingual}$	**58,39**	58,29	58,33	58,61	63,89	**68,12**

A Figura 3 apresenta a medida de incoerência C_{euc} de 10 documentos originais DO selecionados aleatoriamente do (a) CSTNews e do (b) OnlineEduc 1.0 (linhas horizontais contínuas) e de 20 permutações aleatórias $permDO$ das sentenças de cada DO (linhas tracejadas), calculadas com cada uma das três variações do BERT (cores distintas).

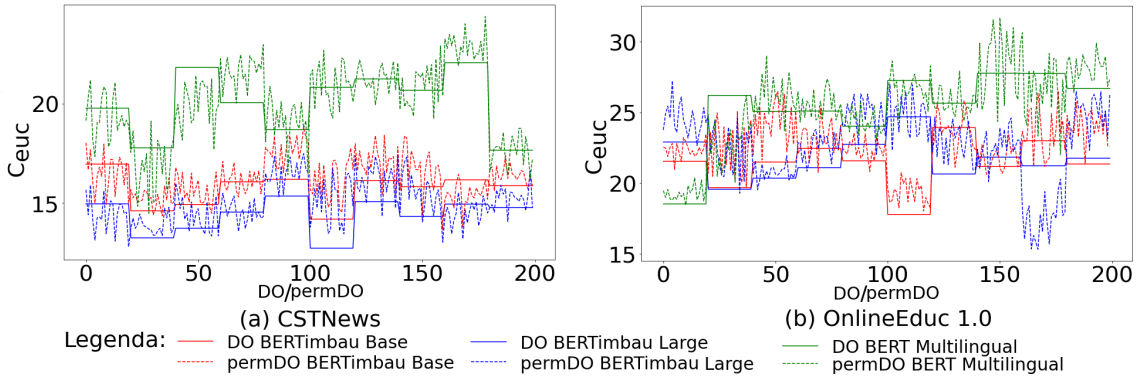


Figura 3. C_{euc} de 10 documentos originais e suas 20 permutações.

Observa-se na Figura 3 que a incoerência C_{euc} oscila para permutações distintas de cada documento, mas as versões permutadas tendem a ter C_{euc} maior do que os documentos originais correspondentes, para as três variações do BERT. Mais precisamente, das 200 versões permutadas dos documentos originais do CSTNews, 147 ($73,5\%$), 146 (73%) e 112 (56%) têm incoerência C_{euc} maior que os respectivos documentos originais, calculada com *embeddings* do $BERTimbau_{Base}$, $BERTimbau_{Large}$ e $BERT_{Multilingual}$,

respectivamente. Para os documentos do OnlineEduc 1.0 são 164 (82%), 161 $(80,5\%)$, 113 $(56,49\%)$ permutações com C_{euc} maior que os respectivos documentos originais, considerando as mesmas variações do BERT.

A Figura 4 detalha a variação da incoerência C_{euc} para diferentes permutações aleatórias de sentenças (linhas tracejadas) do documento original (DO) mais à esquerda nos respectivos gráficos para (a) CSTNews e (b) OnlineEduc 1.1 da Figura 3. Note que a maioria absoluta (19 em 20, i.e. 95%) das permutações das sentenças do DO do OnlineEduc 1.0 têm C_{euc} maior que a do DO (linha contínua), para todas as três variações do BERT. Porém, somente 8 (40%), 6 (30%) e 7 (35%) das permutações do DO do CSTNews têm C_{euc} superior à do DO, calculada com *embeddings* do $BERTimbau_{Base}$, $BERTimbau_{Large}$ e $BERT_{Multilingual}$, respectivamente.

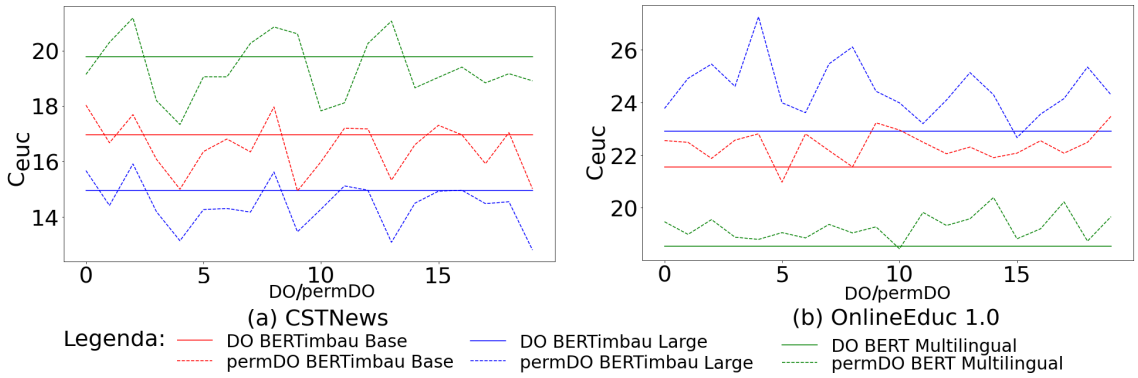


Figura 4. C_{euc} **de** 1 **documento original e suas** 20 **permutações.**

A Figura 5 apresenta em ordem crescente a incoerência C_{euc} dos documentos originais (linhas contínuas) do (a) CSTNews e do (b) OnlineEduc 1.0 e de suas permutações aleatórias (linhas tracejadas), calculada com as três variações do BERT (representadas por cores distintas). O maior somatório das diferenças de C_{euc} entre os documentos originais do CSTNews e suas permutações foi obtido com o $BERTimbau_{Base}$ $(1.407,68)$, seguido pelo $BERTimbau_{Large}$ $(1.387,75)$ e o $BERT_{Multilingual}$ $(1.053,99)$. No OnlineEduc 1.0 o $BERTimbau_{Large}$ apresentou maior somatório dessas diferenças $(13.273,59)$, seguido pelo $BERTimbau_{Base}$ $(13.065,58)$ e o $BERT_{Multilingual}$ $(7.427,81)$. Portanto, o $BERTimbau$ permitiu para ambos os conjuntos de documentos medida de incoerência mais discriminativa, além de maior acurácia na classificação de coerência.

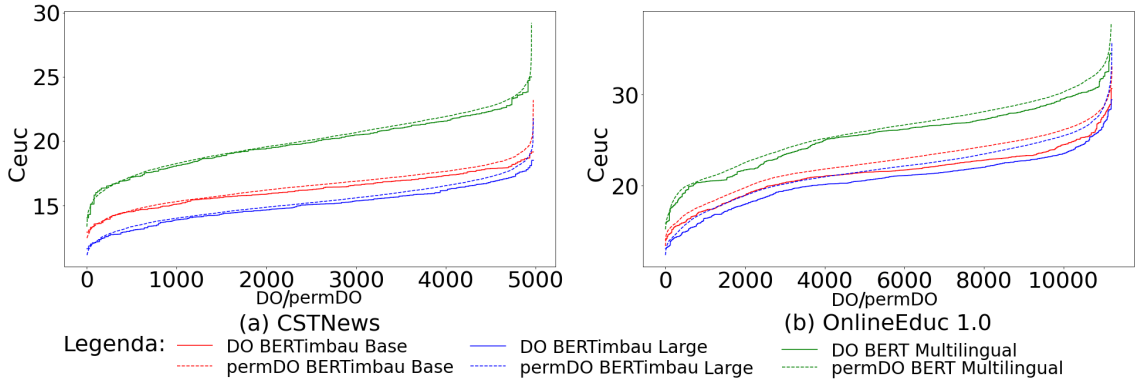


Figura 5. C_{euc} **em ordem crescente para todos os documentos.**

Finalmente, é importante ressaltar que os textos do CSTNews são jornalísticos e supostamente com boa redação, mas os melhores resultados tanto para classificação quanto mensuração discriminativa de (in)coerência foram obtidos com OnlineEduc 1.0, o que demonstra a viabilidade de aplicação da proposta a textos da área da educação, incluindo conversações informais entre alunos e professores.

7. Conclusões e trabalhos futuros

Este trabalho investigou o uso de variações do BERT para classificar e medir coerência em textos do OnlineEduc 1.0 (fórum de AVA) e do CSTNews (noticiário). As suas contribuições são: (i) novos classificadores de coerência textual em língua portuguesa baseados em três variações do BERT, com configurações de hiperparâmetros que conferem desempenho superior ao estado-da-arte; (ii) investigação do desempenho das variações do BERT, alternativas para consolidar *embeddings* contextualizados das palavras de cada sentença e diferentes medidas de distância e similaridade entre os *embeddings* resultantes para classificar coerência e calcular medidas de (in)coerência e (iii) análise comparativa dos resultados dos classificadores e de distribuições das medidas de (in)coerência dos documentos originais e de versões com suas sentenças permutadas.

Surpreendentemente, a acurácia dos classificadores de coerência baseados no BERT foi maior para textos do OnlineEduc 1.0 do que aqueles do CSTNews, talvez pelo fato dos últimos serem sumários, ainda que de textos com redação profissional. Finalmente, foram observadas medidas de incoerência (calculadas a partir das distâncias Euclidiana e Manhathan) superiores para os textos do OnlineEduc 1.0. Porém, a medida de coerência, calculada a partir da similaridade cosseno, foi um pouco mais discriminativa para os textos do CSTNews, embora com menor diferença absoluta.

Trabalhos futuros incluem: (*i*) analisar a influência de características sintáticas dos textos na classificação e mensuração de coerência; (*ii*) avaliar o desempenho da proposta frente a permutações de unidades textuais menores que sentenças, tais como suas orações e (*iii*) aplicar os recentes *embeddings* de caracteres na análise de coerência.

Agradecimentos

Este trabalho foi financiado pela Universidade do Estado de Santa Catarina (UDESC) e pelo Projeto PrInt CAPES-UFSC “Automação 4.0” da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Cavalcanti, A., Mello, R., Miranda, P., and Freitas, F. (2020). Análise automática de feedback em ambientes de aprendizagem online. In *Anais do XXXI Simp. Bras. de Informática na Educação*, pages 892–901, Porto Alegre, RS, Brasil. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. North American Chapter of the ACL: Human Language Technologies, Vol, 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics (ACL).

Dias, M. d. S. (2016). *Investigação de modelos de coerência local para sumários multi-documento*. PhD thesis, Universidade de São Paulo.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Li, J. and Jurafsky, D. (2017). Neural net models of open-domain discourse coherence. In *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles, Marina del Rey, California.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohiuddin, T., Joty, S., and Nguyen, D. T. (2018). Coherence modeling of asynchronous conversations: A neural entity grid approach. *arXiv preprint arXiv:1805.02275*, pages 558–568.

Neto, S. S. C., Favero, E., dos Santos, J. A., Freitas, S., and Júnior, M. N. (2020). Avaliação automática de redações na língua portuguesa baseada na coleta de atributos e aprendizagem de máquina. In *Anais do XXXI Simp. Bras. de Informática na Educação*, pages 1162–1171, Porto Alegre, RS, Brasil. SBC.

Oliveira, D., Pozzebon, E., and Santos, T. (2020). 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 Simp. Bras. de Informática na Educação*, pages 1243–1252, Porto Alegre, RS, Brasil. SBC.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proc. of the 2014 conf. on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing*, pages 3982–3992, Hong Kong, China. ACL.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417, Rio Grande, Brazil. Springer, Springer.

Xu, P., Saghir, H., Kang, J. S., Long, T., Bose, A. J., Cao, Y., and Cheung, J. C. K. (2019). A cross-domain transferable neural coherence model. In *Proc. 57th Annual Meeting of the ACL*, pages 678–687, Florence, Italy. Assoc. for Computational Linguistics (ACL).

Zhang, Z., Zhao, H., and Wang, R. (2020). Machine reading comprehension: The role of contextualized language models and beyond. *arXiv preprint arXiv:2005.06249*, abs/2005.06249.