Um Estudo Sobre Métricas de Avaliação para Sumarização de Acórdãos

Gustavo Rufino Feltrin[1], **Daniela Vianna**[1], **Altigran da Silva**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brazil

{feltrin,dvianna,alti}@icomp.ufam.edu.br

Abstract. *Various evaluation metrics for text generation have been proposed in recent years. However, many questions have emerged about how well they can evaluate the accuracy and quality of the text generated. In this work, we study how some of the most popular text generation metrics behave when dealing with the text summarization task in the Portuguese legal domain. More specifically, we evaluated five metrics – ROUGE, BERTScore, BARTScore, BLEURT, and MoverScore –, using a dataset containing 892 rulings from the Brazilian Superior Court of Justice. Each item in the dataset is composed of a ruling, which is the original legal document, and a syllabus, which corresponds to a manually generated summary of the original legal document. Our study revealed that, for the Brazilian legal domain, none of the metrics evaluated were capable of fully measuring the quality of manually generated summaries when compared with their original documents, and that, among the evaluated metrics, ROUGE and BERTScore presented the most promising results.*

Resumo. *Várias métricas de avaliação para geração de texto foram propostas nos últimos anos. No entanto, muitas questões surgiram sobre o quão bem elas podem avaliar a acurácia e a qualidade do texto gerado. Neste trabalho, estudamos como algumas das métricas de geração de texto mais populares se comportam ao lidar com a tarefa de sumarização de texto no domínio jurídico em Português. Mais especificamente, avaliamos cinco métricas – ROUGE, BERTScore, BARTScore, BLEURT e MoverScore –, usando um dataset contendo 892 acórdãos do Superior Tribunal de Justiça. Cada item do dataset é composto por um acórdão, que é o documento jurídico original, e uma ementa, que corresponde a um resumo manualmente gerado do documento jurídico original. Nosso estudo revelou que, para o domínio jurídico brasileiro, nenhuma das métricas avaliadas foi capaz de mensurar totalmente a qualidade dos resumos gerados manualmente quando comparados com seus documentos originais, e que, dentre as métricas avaliadas, ROUGE e BERTScore apresentaram os resultados mais promissores.*

1. Introdução

A sumarização de texto é a tarefa de criar uma representação textual concisa e reduzida do texto, mantendo as informações relevantes do documento original. Existem duas abordagens conhecidas para esta tarefa, a abordagem *extrativa*, que extrai frases diretamente do texto, e a abordagem *abstrativa*, que visa extrair conceitos e ideias do texto original

criando uma nova e distinta versão do texto, de forma semelhante a um resumo feito por um humano. A sumarização de textos tornou-se uma tarefa essencial no domínio jurídico, com casos envolvendo uma vasta quantidade de documentos e arquivos. Em particular, temos o sistema judiciário brasileiro, um sistema muito grande e complexo com 27,7 milhões de novos processos judiciais somente no ano de 2021[1]. Portanto, a capacidade de resumir e condensar documentos legais tornou-se muito mais importante e urgente ao longo dos anos. Um caso muito especial de sumarização no sistema judiciário brasileiro é a obrigatoriedade de criação de uma ementa para cada novo acórdão.

Um acórdão é uma sentença coletiva de um tribunal que exibe uma posição argumentada sobre a aplicabilidade de um determinado direito legal a uma situação fática específica. Uma ementa, por sua vez, é a síntese do acórdão em que seus pontos fundamentais são sintetizados manualmente por um operador do direito. Dado sua brevidade, a ementa introduz e esclarece o contexto do acórdão, permitindo assim aos operadores do direito tomar conhecimento do mesmo, com menor esforço dado o tamanho da ementa comparado ao tamanho do acórdão. Além disso, possibilita ao operador do direito embasar/endossar um posicionamento semelhante, acolhido anteriormente por outras Turmas e Tribunais [Guimarães 2004].

Atualmente, a elaboração de ementas é feita de forma assistida; portanto, é necessário que um profissional do direito a elabore manualmente. Esse processo, por ser manual, implica em acórdãos sem ementa e resulta em falta de padronização do texto, uma vez que são elaborados por diferentes operadores do direito. Uma alternativa para lidar com esse problema é o uso de técnicas de sumarização automática de texto [Liu 2019, Jain et al. 2021, Fabbri et al. 2021, Pandya 2019, Farzindar e Lapalme 2004, Polsley et al. 2016, Feijó e Moreira 2019]. Argumentamos que a sumarização automática pode ser muito útil no domínio jurídico, auxiliando profissionais do direito e pessoas comuns a terem um rápido entendimento de ações judiciais, uma rápida contextualização das ações judiciais prévias por outras partes interessadas e um alto nível de transparência [Jain et al. 2021].

A ementa apresenta-se como uma síntese, um sumário do acórdão, e para comparar a eficácia das técnicas de sumarização automática, especialmente das abordagens abstrativas [Feijó e Moreira 2023, Zhang et al. 2020, Lewis et al. 2020], se faz necessário identificar qual métrica de avaliação de sumarização melhor capturará a equivalência semântica entre os acórdãos e suas respectivas ementas. É importante destacar que os documentos jurídicos, além de muito extensos, são compostos por termos e expressões próprias do domínio jurídico, o que representa um desafio para qualquer tarefa de Processamento de Linguagem Natural (PLN), incluindo a representação de texto. Além disso, em relação às métricas de avaliação mais recentes, introduzidas com modelos de linguagem baseados em *transformers* [Vaswani et al. 2017], há uma grande lacuna em soluções voltadas para o Português.

Tendo em vista os desafios impostos pela sumarização de textos jurídicos em Português, neste trabalho, investigamos uma variedade de métricas de avaliação para geração de texto, com o objetivo de investigar como cada métrica avalia a ementa de um acórdão. Isso é importante para identificar métricas adequadas a serem aplicadas na avaliação de

[1] `https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/`

novas técnicas de sumarização desenvolvidas para o domínio jurídico, principalmente em situações onde uma coleção de referência composta por sumários não está disponível.

Um total de cinco métricas de avaliação foram consideradas: ROUGE [Lin 2004], BERTScore [Zhang et al. 2019], BARTScore [Yuan et al. 2021], BLEURT [Sellam et al. 2020] e MoverScore [Zhao et al. 2019]. O estudo foi realizado a partir de um *dataset* composto por 892 acórdãos do STJ (Superior Tribunal de Justiça). Nossos experimentos demonstram que, das cinco métricas avaliadas, apenas *precision* do ROUGE e *precision* do BERTScore podem, até certo ponto, capturar a alta qualidade das ementas. As métricas restantes, BARTScore, BLEURT e MoverScore, falharam completamente ao avaliar resumos no domínio jurídico Português.

Este artigo está organizado da seguinte forma. Na Seção 2 revisamos brevemente as métricas de avaliação de geração de texto consideradas neste trabalho e também discutimos alguns trabalhos relacionados. A metodologia experimental, que inclui uma descrição do *dataset* jurídico usado durante a avaliação, é apresentada na Seção 3. Uma avaliação experimental das métricas de geração de texto também é apresentada na Seção 3. Finalmente, as conclusões e direções de pesquisas futuras são apresentadas na Seção 5.

2. Fundamentação e Trabalhos relacionados

Uma ampla variedade de métricas de avaliação para geração de texto foram propostas com técnicas que variam desde as abordagens mais tradicionais de correspondência de n-gramas até as recentes abordagens baseadas em aprendizado.

A família de métricas ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin 2004], considerada a métrica de avaliação padrão para sumarização de texto, conta o número de sobreposição de unidades, como n-gramas, sequências de palavras e pares de palavras entre a referência e o candidato a ser avaliado. O BERTScore [Zhang et al. 2019], por outro lado, calcula um score de similaridade entre vetores de representação contextual ou *embeddings* contextuais para os tokens presentes no texto de referência e os *embeddings* contextuais para os tokens no texto candidato.

O BARTScore [Yuan et al. 2021] conceitua a avaliação do texto gerado como um problema de geração de texto, portanto, modelos treinados para gerar texto de/para uma saída de referência ou texto de origem alcançarão pontuações mais altas quando o texto gerado for melhor. O BLEURT [Sellam et al. 2020], diferente de outras abordagens, apresenta uma métrica de avaliação aprendida baseada no BERT capaz de modelar julgamentos humanos. Com uma abordagem de pré-treinamento extra que utiliza um conjunto de perturbações aleatórias, aumentadas com diferentes sinais a nível lexical e semântico. Outra métrica que também tem sido utilizada para avaliar os textos gerados é o MoverScore [Zhao et al. 2019], essa métrica combina representações contextualizadas com a medida de distância para realizar a avaliação.

Existem alguns estudos na literatura que tentam avaliar a adequação das métricas de avaliação atuais para métodos de sumarização. Por exemplo, Kryściński et al. [Kryściński et al. 2019] apresentam uma análise crítica de vários *datasets*, métricas e modelos de linguagem previamente propostos na literatura. Eles mostram que a sumarização de texto requer restrições adicionais para fornecer resumos bem formados, que

os métodos atuais aprendem a confiar demais em vieses associados a textos de domínios específicos que estão sendo resumidos e que o protocolo de avaliação atual é fracamente relacionado ao julgamento feito por um ser humano ao mesmo tempo não avalia certas características de sumarização de texto.

Mais recentemente, Fabbri et al. [Fabbri et al. 2021] argumentaram que a escassez de estudos que abranjam e atualizem a avaliação de sumarização de textos e a falta de consenso sobre protocolos de avaliação continuam inibindo o progresso na área. Eles estudam as deficiências atuais dos métodos de sumarização em um relatório que considera cinco dimensões: 1) reavaliando 14 métricas de avaliação automática junto com anotação humana de especialistas e *crowdsourcing*; 2) comparando 23 modelos de sumarização com métricas de avaliação; 3) estruturando e compartilhando uma coleção de resumos gerados por modelos treinados com o *dataset* de notícias CNN/DailyMail; 4) desenvolvendo e compartilhando o SummEval (*toolkit* extensível unificado para avaliar modelos de sumarização); e 5) reunindo e compartilhando uma coleção de julgamentos humanos de resumos gerados por um modelo sob o *dataset* CNN/Daily.

Em contraste com esses trabalhos anteriores, iniciamos este estudo compreendendo a diferença entre cada métrica de avaliação, permitindo-nos analisar como elas lidam com este domínio específico. Para isso, usamos um *dataset* com curadoria humana que já possui ementas feitas por humanos, tomando-os como nosso *golden set*. Com isso, objetivamos identificar qual dessas métricas de avaliação pode melhor avaliar/expressar a qualidade de ementas escritas em Português no domínio jurídico quando comparadas aos documentos originais (acórdãos).

3. Metodologia

Este estudo foi realizado em parceria com uma grande empresa de tecnologia jurídica no Brasil, Jusbrasil[2], que nos forneceu o *dataset* de acórdãos.

Este *dataset* contém 892 acórdãos do Superior Tribunal de Justiça (STJ) e suas respectivas ementas. Cada item do *dataset* é composto por um inteiro teor e uma ementa. O inteiro teor é o documento original, a fonte da verdade e contém todos os detalhes de uma decisão, incluindo todos os votos individuais de cada juiz e a decisão final. A ementa é a síntese/resumo do acórdão. Cada ementa também pode ser dividida em duas partes conforme ilustrado na Tabela 1, um cabeçalho que consiste em um conjunto de palavras-chave representativas da temática geral do acórdão; e um texto dispositivo, que pode ser definido de forma lógica e clara, como o resumo do acórdão.

A Tabela 2 apresenta a distribuição de tokens dos inteiros teores e ementas no *dataset*, apresenta também a distribuição de tokens entre as duas partes da ementa: cabeçalho e dispositivo. Para cada parte do acórdão, informamos o mínimo (Min), o máximo (Max), a média (Mean) e o desvio padrão (Std Dev) em relação ao número de tokens. Neste trabalho, o documento é tokenizado usando o caractere espaço como delimitador, ou seja, o número de tokens é o número de palavras no documento. Podemos observar que, em média, o inteiro teor é bastante grande quando comparado com a ementa, sendo que o dispositivo da ementa contém a maior parte do resumo. Ao mesmo tempo, olhando para o desvio padrão (Std Dev), vemos que o tamanho de um inteiro teor varia consideravelmente.

[2]`https://sobre.jusbrasil.com.br/`

Tabela 1. Exemplos de ementas

Cabeçalho	Dispositivo
PROCESSO CIVIL. EMBARGOS DO DEVEDOR.	A pessoa jurídica não tem legitimidade para interpor recurso no interesse do sócio. Recurso especial desprovido. Acórdão submetido ao regime do art. 543-C do CPC e da Resolução STJ n. 8/08.
RECURSO ESPECIAL. REPRESENTATIVO DE CONTROVÉRSIA. JUROS DE MORA LEGAIS. NATUREZA INDENIZATÓRIA. NÃO INCIDÊNCIA DE IMPOSTO DE RENDA.	- Não incide imposto de renda sobre os juros moratórios legais em decorrência de sua natureza e função indenizatória ampla. Recurso especial, julgado sob o rito do art. 543-C do CPC, improvido.
TRIBUTÁRIO. PARCELAMENTO DE DÉBITO. DENÚNCIA ESPONTÂNEA. INAPLICABILIDADE. RECURSO REPETITIVO. ART. 543-C DO CPC.	1. O instituto da denúncia espontânea (art. 138 do CTN) não se aplica nos casos de parcelamento de débito tributário. 2. Recurso Especial provido. Acórdão sujeito ao regime do art. 543-C do CPC e da Resolução 8/2008 do STJ.
PROCESSUAL CIVIL. RECURSO ESPECIAL REPRESENTATIVO DE CONTROVÉRSIA. ART. 543-C DO CPC. AÇÃO MONITÓRIA APARELHADA EM CHEQUE PRESCRITO. DISPENSA DA MENÇÃO À ORIGEM DA DÍVIDA.	1. Para fins do art. 543-C do CPC: Em ação monitória fundada em cheque prescrito, ajuizada em face do emitente, é dispensável menção ao negócio jurídico subjacente à emissão da cártula. 2. No caso concreto, recurso especial parcialmente provido.

Tabela 2. Distribuição de tokens no *Dataset*

	Mean	Std Dev	Min	Max
Inteiro Teor	5891,39	6155,84	37	45657
Ementa	542,19	481,27	36	4205
Cabeçalho	55,52	49,53	5	713
Dispositivo	486,67	466,53	31	4131

Para analisar e entender o que as métricas expressam em relação aos acórdãos, foram realizadas avaliações das ementas juntamente com seus inteiros teores (fonte da verdade) utilizando diversas métricas, desde a tradicional ROUGE até as mais recentes BERTScore, MoverScore, BLEURT e BARTScore. Essas métricas foram revisadas na Seção 2.

4. Resultados

Nesta seção, apresentamos uma análise aprofundada das cinco métricas de avaliação, ROUGE, BERTScore, BARTScore, BLEURT e MoverScore. As métricas serão estudadas usando *dataset* de acórdãos introduzido na Seção 3. É importante observar que todas as métricas de avaliação foram aplicadas usando os parâmetros padrões propostos em suas implementações originais.

Apensar dessas métricas terem sido desenvolvidas para avaliar a qualidade de novos sumários com relação a um sumário referência, nesse estudo nós intencionalmente comparamos sumários (ementas) com o documento original (acórdãos). Nosso objetivo é estudar o comportamento dessas métricas em um cenário onde uma coleção de referência com sumários não está disponível.

A Tabela 3 apresenta a métrica ROUGE-1 que inclui as medidas de precisão (R1-P) (do Inglês *Precision*), e revocação (R1-R) (do Inglês *Recall*) e *F-measure* (R1-F) para n-gramas de tamanho 1. A métrica de avaliação foi aplicada não apenas entre o inteiro teor (texto original) e a ementa (resumo), mas também entre o inteiro teor e ambas as partes da ementa, cabeçalho e dispositivo. Podemos observar que, para a métrica ROUGE (baseada em sobreposição), a revocação atinge valores baixos em relação a precisão, isso se deve à significativa diferença de tamanho entre a ementa candidata e o inteiro teor (como pode ser visto na Tabela 2), levando à conclusão de que, no ROUGE-1, a revocação não é uma métrica confiável em nosso cenário. A métrica *F-measure* alcança valores baixos porque é uma média ponderada entre a revocação e a precisão. Além disso, o alto valor obtido com a precisão (alta sobreposição) pode indicar que, embora a ementa seja criada manualmente por humanos, ela possui características mais extrativas do que abstrativas, uma vez que muitos termos usados na ementa são frequentemente encontrados no inteiro teor.

Tabela 3. ROUGE-1 - F-measure (R1-F), Precisão (R1-P), and Revocação (R1-R)

Candidato	R1-F	R1-P	R1-R
Ementa	0,1862	0,8570	0,1459
Cabeçalho	0,0402	0,8741	0,0278
Dispositivo	0,1692	0,8622	0,1323

Ao separar a ementa em cabeçalho e dispositivo, pretendemos analisar como as métricas de avaliação se comportam quando o resumo é composto não apenas por um texto humano comum, mas também por um conjunto de palavras-chave que se assemelham a termos mais extrativos e frases-chave curtas. Como pode ser visto na Tabela 3, não houve diferença considerável entre a Precisão (R1-P) obtida para o cabeçalho e o dispositivo.

Conforme apresentado na Seção 2, as demais métricas de avaliação analisadas nesta seção utilizam modelos de linguagem baseados em *transformers*. Na ausência de modelos de linguagem pré-treinados, treinados especificamente com textos em Português, que sejam compatíveis com o BERTScore, neste estudo exploramos dois modelos de linguagem multilíngue: bert-base-multilingual-cased e distilbert-base-multilingual-cased. Os resultados são apresentados na Tabela 4 com colunas relatando *F1-score* (F1), precisão e revocação.

Tabela 4. BERTScore - F-measure, Precisão, e Revocação

Modelo	Candidato	F1	Precisão	Revocação
bert-base-multilingual-cased	Ementa	0,6854	0,6889	0,6832
bert-base-multilingual-cased	Cabeçalho	0,6576	0,7029	0,6188
bert-base-multilingual-cased	Dispositivo	0,6112	0,6429	0,5845
distilbert-base-multilingual-cased	Ementa	0,7775	0,7811	0,7749
distilbert-base-multilingual-cased	Cabeçalho	0,7661	0,8101	0,7275
distilbert-base-multilingual-cased	Dispositivo	0,7079	0,7497	0,6724

Ao comparar o BERTScore com ambos os modelos de linguagem, distilbert-base-multilingual-cased supera o bert-base-multilingual-cased para todos os resumos candidatos, ementa, cabeçalho e dispositivo, e também para todas as métricas, F1, precisão e revocação. Novamente, a precisão é a métrica que melhor representa a similaridade entre o inteiro teor (documento original) e sua ementa (resumo). O BERTScore, uma métrica de avaliação baseada em *embeddings* contextuais e similaridade de cosseno, é mais adequada para sumarização abstrativa do que abordagens de sobreposição, como ROUGE. Isso se deve à capacidade dos *embeddings* contextuais de entender o contexto e a semântica de uma mesma palavra em frases diferentes, dependendo das palavras ao redor. No entanto, quando comparamos a precisão obtida com o BERTScore junto do distilbert-base-multilingual-cased, Tabela 4, com a precisão do ROUGE (R1-P), Tabela 3, podemos ver que o ROUGE parece capturar melhor a qualidade do resumo gerado manualmente do que a métrica BERTScore. Uma suposição aberta a investigação é se o desempenho do BERTScore pode ser melhorado com a adoção de um modelo de linguagem pré-treinado ou ajustado com textos jurídicos do Português. No entanto, haveria alterações no código do BERTScore, já que atualmente só funciona com modelos específicos. É importante ressaltar que métricas como o BERTScore, baseadas em modelos de linguagem, sofrem com a limitação do tamanho da entrada. Ou seja, no cenário jurídico, onde documentos são costumeiramente longos, apenas uma parte do documento será considerada na análise. Apesar dessa limitação, o BERTScore apresentou um desempenho competitivo em relação ao ROUGE.

A Tabela 5 apresenta os resultados obtidos para ementa, cabeçalho e dispositivo ao usar o BARTScore com o modelo de linguagem facebook/bart-large-cnn. Diferente do ROUGE e do BERTScore, o BARTScore relata a verossimilhança média de log para tokens alvos, com pontuações calculadas chegando a valores menores que zero. Como pode ser visto na Tabela 5, a parte dispositiva da ementa é a que mais se aproxima do inteiro teor; no entanto, a diferença é tão pequena quando comparada com o cabeçalho e toda ementa, que pode ser desconsiderada. Para fins de comparação, calculamos o BARTScore em dois outros cenários: (1) Próprio, que mede a qualidade de uma ementa

quando comparada com si mesma, que deveria ser a pontuação mais alta possível; (2) Aleatório, que mede a qualidade de uma ementa quando comparada com qualquer outra ementa escolhida aleatoriamente. Com os cenários Próprio and Aleatório, nossa intenção era definir possíveis limites inferiores e superiores para o BARTScore. Como pode ser observado na Tabela 5, o BARTScore obtido para o acórdão está muito próximo da abordagem aleatória, mostrando que o BARTScore não é a métrica ideal para comparar a qualidade de um resumo com o documento de origem.

Tabela 5. BARTScore - probabilidade logarítmica ponderada

Modelo	Candidato	Próprio	Aleatório	Inteiro Teor
facebook/bart-large-cnn	Ementa	-0,4365	-3,6823	-3,3969
facebook/bart-large-cnn	Cabeçalho	-0,2627	-3,9532	-3,6849
facebook/bart-large-cnn	Dispositivo	-0,3902	-3,5967	-3,2966

Os resultados obtidos com o BLEURT, uma métrica treinada como um modelo de regressão, podem ser encontrados na Tabela 6. Similarmente ao que foi feito nas avaliações anteriores, um modelo multilíngue (BLEURT-20 [Pu et al. 2021]) foi adotado para validar o desempenho do BLEURT na geração de textos jurídicos em Português. Também calculamos as pontuações dos limites inferior (abordagem Aleatória) e superior (abordagem Próprio) para nosso *dataset*. Assim como foi observado para o BARTScore, a métrica BLEURT não foi capaz de captar a alta qualidade dos resumos (coluna Inteiro Teor, Tabela 6) quando comparados com seus respectivos acórdãos. Este resultado pode ser confirmado comparando as colunas Aleatório e Inteiro Teor na Tabela 6, com Inteiro Teor tendo um desempenho tão ruim quanto Aleatório. Lembrando que na abordagem Aleatória, comparamos a ementa alvo com outra ementa escolhida aleatoriamente.

Tabela 6. BLEURT - métrica de avaliação aprendida

Modelo	Candidato	Próprio	Aleatório	Inteiro Teor
BLEURT-20	Ementa	0,7928	-0,1755	-0,1251
BLEURT-20	Cabeçalho	0,9083	-0,1057	-0,1050
BLEURT-20	Dispositivo	0,8307	-0,1577	-0,1137

A métrica MoverScore foi avaliada usando o modelo de linguagem padrão distilbert-base-uncased e dois outros modelos multilíngues compatíveis, bert-base-multilingual-cased e distilbert-base-multilingual-cased . Observe que ambos os modelos multilíngues também foram usados na avaliação do BERTScore. Como o MoverScore é uma métrica baseada em distância, os resultados apresentados na Tabela 7 correspondem à distância média entre os *embeddings* dos inteiros teores e das ementas, incluindo suas partes, cabeçalho e dispositivo. Podemos observar que, independente do modelo linguístico adotado, os resultados foram muito semelhantes, com os três modelos tendo um desempenho ruim no cenário jurídico em Português. O fato do MoverScore com o modelo distilbert-base-uncased se sair um pouco melhor do que os outros dois modelos multilíngues foi muito inesperado, pois o modelo distilbert-base-uncased foi treinado apenas com textos em Inglês e, por isso, deveria estar menos equipado para lidar com documentos legais em Português.

Tabela 7. MoverScore - medida de distância

Modelo	Candidato	Distancia
distilbert-base-uncased	Ementa	0,5378
distilbert-base-uncased	Cabeçalho	0,5173
distilbert-base-uncased	Dispositivo	0,5350
bert-base-multilingual-cased	Ementa	0,5032
bert-base-multilingual-cased	Cabeçalho	0,4809
bert-base-multilingual-cased	Dispositivo	0,4999
distilbert-base-multilingual-cased	Ementa	0,5309
distilbert-base-multilingual-cased	Cabeçalho	0,5071
distilbert-base-multilingual-cased	Dispositivo	0,5268

De todas as métricas avaliadas, a precisão do ROUGE (R1-P) e a precisão do BERTScore foram as únicas capazes, até certo ponto, de capturar a qualidade dos resumos gerados manualmente. Indicando que, para o cenário considerado neste trabalho, textos jurídicos do Português, precisamos modificar as métricas existentes ou desenvolver novas voltadas para o domínio jurídico em Português.

5. Conclusões

Neste estudo, analisamos uma variedade de métricas de avaliação para geração de texto, com a intenção de avaliar sua capacidade de capturar a qualidade dos resumos no domínio jurídico em Português. Um total de cinco métricas de avaliação foram consideradas, incluindo ROUGE, BERTScore, BARTScore, BLEURT e MoverScore. Para avaliar a capacidade das métricas, utilizamos um *dataset* composto por 892 acórdãos do STJ. As ementas dos acórdãos são geradas manualmente por operadores do direito e são consideradas como o melhor resumo possível para seus respectivos inteiros teores. Com este estudo concluímos que o ROUGE e o BERTScore podem, até certo ponto, serem utilizados para avaliar a qualidade de um resumo; no entanto, eles ainda apresentam baixo desempenho. As demais métricas, BARTScore, BLEURT e MoverScore não foram capazes de capturar a qualidade das ementas quando comparadas com as mesmas, apresentando desempenho tão baixo quanto a abordagem aleatória, que compara uma ementa com outra escolhida aleatoriamente. Concluindo, existe atualmente a necessidade de métricas capazes de aferir a qualidade dos resumos no domínio jurídico Português, principalmente na ausência de coleções de referência.

Agradecimentos

Este trabalho foi parcialmente apoiado pela Jusbrasil, pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001 financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD, e pelo CNPq através de uma bolsa PQ para Altigran da Silva (Proc. 307248/2019-4).

Referências

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., e Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Farzindar, A. e Lapalme, G. (2004). LetSum, an automatic legal text summarizing system. In *Jurix*, pages 11–18.

Feijó, D. d. V. e Moreira, V. P. (2019). Summarizing legal rulings: Comparative experiments. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019*, pages 313–322.

Feijó, D. d. V. e Moreira, V. P. (2023). Improving abstractive summarization of legal rulings through textual entailment. *Artificial Intelligence and Law*, 31(1):91–113.

Guimarães, J. A. C. (2004). *Elaboração de ementas jurisprudenciais: elementos teórico-metodológicos*, volume 9. Subsecretaria de Divulgação e Editoração da Secretaria de Pesquisa e Informação Jurídicas do Centro de Estudos Judiciários.

Jain, D., Borah, M. D., e Biswas, A. (2021). Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., e Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., e Zettlemoyer, L. (2020). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, Y. (2019). Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

Pandya, V. (2019). Automatic text summarization of legal cases: A hybrid approach. In *5th International Conference on Advances in Computer Science and Information Technology (ACSTY-2019)*.

Polsley, S., Jhunjhunwala, P., e Huang, R. (2016). Casesummarizer: A system for automated summarization of legal texts. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations*, pages 258–262.

Pu, A., Chung, H. W., Parikh, A. P., Gehrmann, S., e Sellam, T. (2021). Learning compact metrics for MT. In *Conference on Empirical Methods in Natural Language Processing*.

Sellam, T., Das, D., e Parikh, A. P. (2020). BLEURT: learning robust metrics for text generation. *CoRR*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., e Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Yuan, W., Neubig, G., e Liu, P. (2021). Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, pages 27263–27277.

Zhang, J., Zhao, Y., Saleh, M., e Liu, P. J. (2020). PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., e Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. *CoRR*.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., e Eger, S. (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *CoRR*.