Artificial Intelligence and Law (2023) 31:91–113
https://doi.org/10.1007/s10506-021-09305-4

ORIGINAL RESEARCH

Improving abstractive summarization of legal rulings through textual entailment

Diego de Vargas Feijo[1] · **Viviane P. Moreira**[1]

Accepted: 28 October 2021 / Published online: 27 November 2021

Abstract
The standard approach for abstractive text summarization is to use an encoder-decoder architecture. The encoder is responsible for capturing the general meaning from the source text, and the decoder is in charge of generating the final text summary. While this approach can compose summaries that resemble human writing, some may contain unrelated or unfaithful information. This problem is called "hallucination" and it represents a serious issue in legal texts as legal practitioners rely on these summaries when looking for precedents, used to support legal arguments. Another concern is that legal documents tend to be very long and may not be fed entirely to the encoder. We propose our method called LegalSumm for addressing these issues by creating different "views" over the source text, training summarization models to generate independent versions of summaries, and applying entailment module to judge how faithful these candidate summaries are with respect to the source text. We show that the proposed approach can select candidate summaries that improve ROUGE scores in all metrics evaluated.

Keywords Legal ruling summarization · Abstractive summarizer · Content digest · Legal case brief · Summary writing · Abstract generator · Automatic text summary · Textual entailment · Fact checking

1 Introduction

With the increasing availability of data, many areas face the need for text summarization. Such a requirement is felt especially in the legal field as texts are usually lengthy. Legal practitioners are expected to keep updated with relevant information ranging from news, jurisprudence changes, and rulings from many courts. Often,

✉ Diego de Vargas Feijo
dvfeijo@inf.ufrgs.br

Viviane P. Moreira
viviane@inf.ufrgs.br

[1] Institute of informatics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

when researching, a legal practitioner needs to seek within a large number of precedents looking for those that fit specific requirements. Each of these precedents may have dozens of pages with details that are specific to that case. For example, rulings from the Brazilian Supreme Court typically contain over 2,000 tokens on average (Feijo and Moreira 2018). Courts usually provide extracts of their most important decisions summarizing the main topics discussed and the outcomes. Currently, these legal summaries are generated by humans in a process that is time-consuming and labor-intensive.

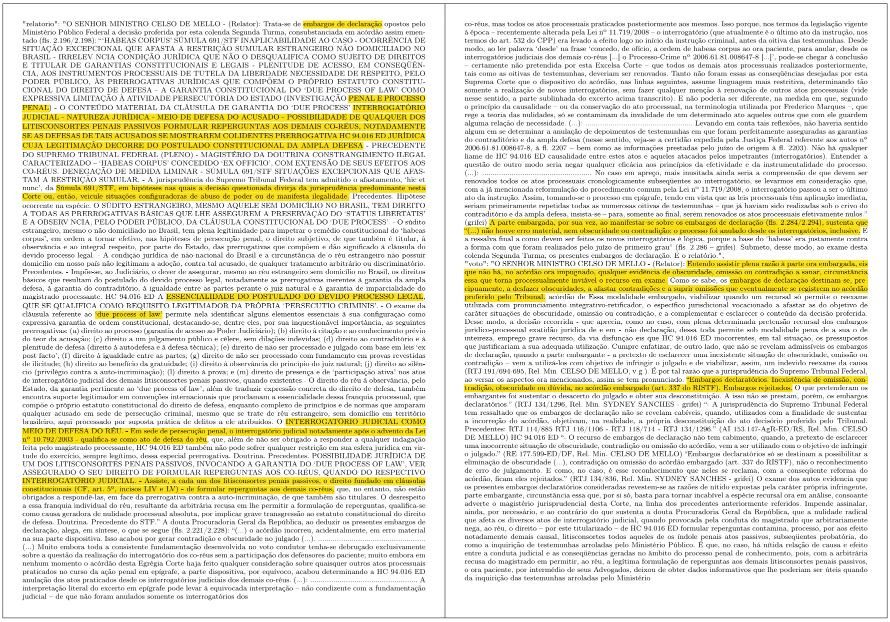
Beyond the difficulty of manually creating summaries, leaving this task to humans is prone to individual writing styles, which leads to a lack of standardization. A standardized way of writing is desirable as it yields more homogeneous summaries (Guimarães 2011). A summary should be concise, fluent, and contain paraphrased versions of the input text with a reduced length.

Most work on summarization has been done with news articles in mind. Dozens of methods have been proposed and tested on this type of document (See et al. 2017; Fan et al. 2018; Liu 2019; Zhang et al. 2020, 2019; Liu and Lapata 2019). Turtle (1995) points out that legal documents have some distinguishing characteristics compared to newspaper articles, namely (*i*) *size*, they tend to be longer; (*ii*) *structure*, they present an internal structure; (*iii*) *vocabulary*, many technical terms are often used; (*iv*) *ambiguity*, ambiguous terms may lead to different meanings for the same words; and (*v*) *citations*, play a prominent role in the legal domain more than in other areas.

News articles often start with some catch sentence called "lede" that summarizes the entire article, making the task of the automatic summarizer considerably easier. Legal texts do not follow this pattern; they are typically lengthier and contain sophisticated vocabulary and expressions.

Besides news articles, scientific papers have also been used for summarization (Chandrasekaran et al. 2019) but at a much smaller scale. Similar to legal documents, scientific papers are also long, have structure and citations. However, they typically have a conclusion section that summarizes the work. A summarizer could achieve a good result just by focusing on the conclusion. On the other hand, in legal documents, the relevant parts are scattered throughout the text. Figure 1 shows an example of a (shortened) legal ruling and how the critical information that should be included in the abstract is scattered through the text.

Another concern when generating the summary is to preserve the original meaning of the source text. Even a small difference in word order can make a big difference in the meaning of the sentence, *e.g., "denied an appeal that had accepted"* is very different from *"accepted an appeal that had denied"*. In an encoder-decoder architecture, the encoder builds a context vector that encompasses the relevant information from the input sequence, and the decoder generates the output summary word-per-word. The decoder is conditioned to generate texts following the characteristics that it has often seen during training. This bias might cause the decoder to generate frequent expressions or mention often seen facts that were not in the input sequence. State-of-the-art models in abstractive text summarization sometimes introduce false facts (changing numbers, locations, proper nouns) or even completely deviate from the original contents. The literature calls this phenomenon



"relatorio": "O SENHOR MINISTRO CELSO DE MELLO - (Relator): Trata-se de embargos de declaração opostos pelo Ministério Público Federal a decisão proferida por esta colenda Segunda Turma, consubstanciada em acórdão assim ementado (fls. 2.196/2.198): "'HABEAS CORPUS' SÚMULA 691/STF INAPLICABILIDADE AO CASO - OCORRÊNCIA DE SITUAÇÃO EXCEPCIONAL QUE AFASTA A RESTRIÇÃO SUMULAR ESTRANGEIRO NÃO DOMICILIADO NO BRASIL - IRRELEV NCIA CONDIÇÃO JURÍDICA QUE NÃO O DESQUALIFICA COMO SUJEITO DE DIREITOS E TITULAR DE GARANTIAS CONSTITUCIONAIS E LEGAIS - PLENITUDE DE ACESSO, EM CONSEQÜÊNCIA, AOS INSTRUMENTOS PROCESSUAIS DE TUTELA DA LIBERDADE NECESSIDADE DE RESPEITO, PELO PODER PÚBLICO, ÀS PRERROGATIVAS JURÍDICAS QUE COMPÕEM O PRÓPRIO ESTATUTO CONSTITUCIONAL DO DIREITO DE DEFESA - A GARANTIA CONSTITUCIONAL DO 'DUE PROCESS OF LAW' COMO EXPRESSIVA LIMITAÇÃO À ATIVIDADE PERSECUTÓRIA DO ESTADO (INVESTIGAÇÃO PENAL E PROCESSO PENAL) - O CONTEÚDO MATERIAL DA CLÁUSULA DE GARANTIA DO 'DUE PROCESS' INTERROGATÓRIO JUDICIAL - NATUREZA JURÍDICA - MEIO DE DEFESA DO ACUSADO - POSSIBILIDADE DE QUALQUER DOS LITISCONSORTES PENAIS PASSIVOS FORMULAR REPERGUNTAS AOS DEMAIS CO-RÉUS, NOTADAMENTE SE AS DEFESAS DE TAIS ACUSADOS SE MOSTRAREM COLIDENTES PRERROGATIVA HC 94.016 ED JURÍDICA CUJA LEGITIMAÇÃO DECORRE DO POSTULADO CONSTITUCIONAL DA AMPLA DEFESA - PRECEDENTE DO SUPREMO TRIBUNAL FEDERAL (PLENO) - MAGISTÉRIO DA DOUTRINA CONSTRANGIMENTO ILEGAL CARACTERIZADO – 'HABEAS CORPUS' CONCEDIDO 'EX OFFICIO', COM EXTENSÃO DE SEUS EFEITOS AOS CO-RÉUS. DENEGAÇÃO DE MEDIDA LIMINAR - SÚMULA 691/STF SITUAÇÕES EXCEPCIONAIS QUE AFASTAM A RESTRIÇÃO SUMULAR. - A jurisprudência do Supremo Tribunal Federal tem admitido o afastamento, 'hic et nunc', da Súmula 691/STF, em hipóteses nas quais a decisão questionada divirja da jurisprudência predominante nesta Corte ou, então, veicule situações configuradoras de abuso de poder ou de manifesta ilegalidade. Precedentes. Hipótese ocorrente na espécie. O SÚDITO ESTRANGEIRO, MESMO AQUELE SEM DOMICÍLIO NO BRASIL, TEM DIREITO A TODAS AS PRERROGATIVAS BÁSICAS QUE LHE ASSEGUREM A PRESERVAÇÃO DO 'STATUS LIBERTATIS' E A OBSERV NCIA, PELO PODER PÚBLICO, DA CLÁUSULA CONSTITUCIONAL DO 'DUE PROCESS'. - O súdito estrangeiro, mesmo o não domiciliado no Brasil, tem plena legitimidade para impetrar o remédio constitucional do 'habeas corpus', em ordem a tornar efetivo, nas hipóteses de persecução penal, o direito subjetivo, de que também é titular, à observância e ao integral respeito, por parte do Estado, das prerrogativas que compõem e dão significado à cláusula do devido processo legal. - A condição jurídica de não-nacional do Brasil e a circunstância de o réu estrangeiro não possuir domicílio em nosso país não legitimam a adoção, contra tal acusado, de qualquer tratamento arbitrário ou discriminatório. Precedentes. - Impõe-se, ao Judiciário, o dever de assegurar, mesmo ao réu estrangeiro sem domicílio no Brasil, os direitos básicos que resultam do postulado do devido processo legal, notadamente as prerrogativas inerentes à garantia da ampla defesa, à garantia do contraditório, à igualdade entre as partes perante o juiz natural e à garantia de imparcialidade do magistrado processante. HC 94.016 ED A ESSENCIALIDADE DO POSTULADO DO DEVIDO PROCESSO LEGAL, QUE SE QUALIFICA COMO REQUISITO LEGITIMADOR DA PRÓPRIA 'PERSECUTIO CRIMINIS'. - O exame da cláusula referente ao 'due process of law' permite nela identificar alguns elementos essenciais à sua configuração como expressiva garantia de ordem constitucional, destacando-se, dentre eles, por sua inquestionável importância, as seguintes prerrogativas: (a) direito ao processo (garantia de acesso ao Poder Judiciário); (b) direito à citação e ao conhecimento prévio do teor da acusação; (c) direito a um julgamento público e célere, sem dilações indevidas; (d) direito ao contraditório e à plenitude de defesa (direito à autodefesa e à defesa técnica); (e) direito de não ser processado e julgado com base em leis 'ex post facto'; (f) direito à igualdade entre as partes; (g) direito de não ser processado com fundamento em provas revestidas de ilicitude; (h) direito ao benefício da gratuidade; (i) direito à observância do princípio do juiz natural; (j) direito ao silêncio (privilégio contra a auto-incriminação); (l) direito à prova; e (m) direito de presença e de 'participação ativa' nos atos de interrogatório judicial dos demais litisconsortes penais passivos, quando existentes.- O direito do réu à observância, pelo Estado, da garantia pertinente ao 'due process of law', além de traduzir expressão concreta do direito de defesa, também encontra suporte legitimador em convenções internacionais que proclamam a essencialidade dessa franquia processual, que compõe o próprio estatuto constitucional do direito de defesa, enquanto complexo de princípios e de normas que amparam qualquer acusado em sede de persecução criminal, mesmo que se trate de réu estrangeiro, sem domicílio em território brasileiro, aqui processado por suposta prática de delitos a ele atribuídos. O INTERROGATÓRIO JUDICIAL COMO MEIO DE DEFESA DO RÉU. - Em sede de persecução penal, o interrogatório judicial notadamente após o advento da Lei nº 10.792/2003 - qualifica-se como ato de defesa do réu, que, além de não ser obrigado a responder a qualquer indagação feita pelo magistrado processante, HC 94.016 ED também não pode sofrer qualquer restrição em sua esfera jurídica em virtude do exercício, sempre legítimo, dessa especial prerrogativa. Doutrina. Precedentes. POSSIBILIDADE JURÍDICA DE UM DOS LITISCONSORTES PENAIS PASSIVOS, INVOCANDO A GARANTIA DO 'DUE PROCESS OF LAW', VER ASSEGURADO O SEU DIREITO DE FORMULAR REPERGUNTAS AOS CO-RÉUS, QUANDO DO RESPECTIVO INTERROGATÓRIO JUDICIAL. - Assiste, a cada um dos litisconsortes penais passivos, o direito fundado em cláusulas constitucionais (CF, art. 5º, incisos LIV e LV) - de formular reperguntas aos demais co-réus, que, no entanto, não estão obrigados a respondê-las, em face da prerrogativa contra a auto-incriminação, de que também são titulares. O desrespeito a essa franquia individual do réu, resultante da arbitrária recusa em lhe permitir a formulação de reperguntas, qualifica-se como causa geradora de nulidade processual absoluta, por implicar grave transgressão ao estatuto constitucional do direito de defesa. Doutrina. Precedente do STF." A douta Procuradoria Geral da República, ao deduzir os presentes embargos de declaração, alega, em síntese, o que se segue (fls. 2.221/2.228): "(...) o acórdão incorreu, acidentalmente, em erro material na sua parte dispositiva. Isso acabou por gerar contradição e obscuridade no julgado (...). .... (...) Muito embora toda a consistente fundamentação desenvolvida no voto condutor tenha-se debruçado exclusivamente sobre a questão da realização do interrogatório dos co-réus sem a participação dos defensores do paciente; muito embora em nenhum momento o acórdão desta Egrégia Corte haja feito qualquer consideração sobre quaisquer outros atos processuais praticados no curso da ação penal em epígrafe, a parte dispositiva, por equívoco, acabou determinando a HC 94.016 ED anulação dos atos praticados desde os interrogatórios judiciais dos demais co-réus. (...): .... A interpretação literal do excerto em epígrafe pode levar à equivocada interpretação – não condizente com a fundamentação judicial – de que não foram anulados somente os interrogatórios dos co-réus, mas todos os atos processuais praticados posteriormente aos mesmos. Isso porque, nos termos da legislação vigente à época – recentemente alterada pela Lei nº 11.719/2008 – o interrogatório (que atualmente é o último ato da instrução, nos termos do art. 532 do CPP) era levado a efeito logo no início da instrução criminal, antes da oitiva das testemunhas. Desde modo, ao ler palavra 'desde' na frase 'concedo, de ofício, a ordem de habeas corpus ao ora paciente, para anular, desde os interrogatórios judiciais dos demais co-réus [...] o Processo-Crime nº 2006.61.81.008647-8 [...]', pode-se chegar à conclusão – certamente não pretendida por esta Excelsa Corte – que todos os demais atos processuais realizados posteriormente, tais como as oitivas de testemunhas, deveriam ser renovados. Tanto não foram essas as conseqüências desejadas por esta Suprema Corte que o dispositivo do acórdão, nas linhas seguintes, assume linguagem mais restritiva, determinando tão somente a realização de novos interrogatórios, sem fazer qualquer menção à renovação de outros atos processuais (vide nesse sentido, a parte sublinhada do excerto acima transcrito). E não poderia ser diferente, na medida em que, segundo o princípio da causalidade – ou da conservação do ato processual, na terminologia utilizada por Frederico Marques –, que rege a teoria das nulidades, só se contaminam da invalidade de um determinado ato aqueles outros que com ele guardem alguma relação de necessidade. (...): .... Levando em conta tais reflexões, não haveria sentido algum em se determinar a anulação de depoimentos de testemunhas em que foram perfeitamente asseguradas as garantias do contraditório e da ampla defesa (nesse sentido, veja-se a certidão expedida pela Justiça Federal referente aos autos nº 2006.61.81.008647-8, à fl. 2207 – bem como as informações prestadas pelo juízo de origem à fl. 2203). Não há qualquer liame de HC 94.016 ED causalidade entre estes atos e aqueles atacados pelos impetrantes (interrogatórios). Entender a questão de outro modo seria negar qualquer eficácia aos princípios da efetividade e da instrumentalidade do processo. (...): .... No caso em apreço, mais inusitada ainda seria a compreensão de que devem ser renovados todos os atos processuais cronologicamente subseqüentes ao interrogatório, se levarmos em consideração que, com a já mencionada reformulação do procedimento comum pela Lei nº 11.719/2008, o interrogatório passou a ser o último ato da instrução. Assim, tomando-se o processo em epígrafe, tendo em vista que as leis processuais têm aplicação imediata, seriam primeiramente repetidas todas as numerosas oitivas de testemunhas – que já haviam sido realizadas sob o crivo do contraditório e da ampla defesa, insista-se – para, somente ao final, serem renovados os atos processuais efetivamente nulos." (grifei) A parte embargada, por sua vez, ao manifestar-se sobre os embargos de declaração (fls. 2.284/2.294), sustenta que "(...) não houve erro material, nem obscuridade ou contradição: o processo foi anulado desde os interrogatórios, inclusive. E a ressalva final a como devem ser feitos os novos interrogatórios é lógica, porque a base do 'habeas' era justamente contra a forma com que foram realizados pelo juízo de primeiro grau" (fls. 2.286 – grifei). Submeto, desse modo, ao exame desta colenda Segunda Turma, os presentes embargos de declaração. É o relatório.",

"voto": "O SENHOR MINISTRO CELSO DE MELLO - (Relator): Entendo assistir plena razão à parte ora embargada, eis que não há, no acórdão ora impugnado, qualquer evidência de obscuridade, omissão ou contradição a sanar, circunstância essa que torna processualmente inviável o recurso em exame. Como se sabe, os embargos de declaração destinam-se, precipuamente, a desfazer obscuridades, a afastar contradições e a suprir omissões que eventualmente se registrem no acórdão proferido pelo Tribunal. acórdão de Essa modalidade embargado, viabilizar quando um recursal só permite o reexame utilizada com pronunciamento integrativo-retificador, o específico jurisdicional vocacionado a afastar as do objetivo de caráter situações de obscuridade, omissão ou contradição, e a complementar e esclarecer o conteúdo da decisão proferida. Desse modo, a decisão recorrida - que aprecia, como no caso, com plena determinada pretensão recursal dos embargos jurídico-processual exatidão jurídica de e em - não declaração, dessa toda permite sob modalidade pena de a sua o de inteireza, emprego grave recurso, da via disfunção eis que HC 94.016 ED inocorrentes, em tal situação, os pressupostos que justificariam a sua adequada utilização. Cumpre enfatizar, de outro lado, que não se revelam admissíveis os embargos de declaração, quando a parte embargante - a pretexto de esclarecer uma inexistente situação de obscuridade, omissão ou contradição – vem a utilizá-los com objetivo de infringir o julgado e de viabilizar, assim, um indevido reexame da causa (RTJ 191/694-695, Rel. Min. CELSO DE MELLO, v.g.). É por tal razão que a jurisprudência do Supremo Tribunal Federal, ao versar os aspectos ora mencionados, assim se tem pronunciado: "Embargos declaratórios. Inexistência de omissão, contradição, obscuridade ou dúvida, no acórdão embargado (art. 337 do RISTF). Embargos rejeitados. O que pretenderam os embargantes foi sustentar o desacerto do julgado e obter sua desconstituição. A isso não se prestam, porém, os embargos declaratórios." (RTJ 134/1296, Rel. Min. SYDNEY SANCHES - grifei) "- A jurisprudência do Supremo Tribunal Federal tem ressaltado que os embargos de declaração não se revelam cabíveis, quando, utilizados com a finalidade de sustentar a incorreção do acórdão, objetivam, na realidade, a própria desconstituição do ato decisório proferido pelo Tribunal. Precedentes: RTJ 114/885 RTJ 116/1106 - RTJ 118/714 - RTJ 134/1296." (AI 153.147-AgR-ED/RS, Rel. Min. CELSO DE MELLO) HC 94.016 ED "- O recurso de embargos de declaração não tem cabimento, quando, a pretexto de esclarecer uma inocorrente situação de obscuridade, contradição ou omissão do acórdão, vem a ser utilizado com o objetivo de infringir o julgado." (RE 177.599-ED/DF, Rel. Min. CELSO DE MELLO) "Embargos declaratórios só se destinam a possibilitar a eliminação de obscuridade (...), contradição ou omissão do acórdão embargado (art. 337 do RISTF), não o reconhecimento de erro de julgamento. E como, no caso, é esse reconhecimento que neles se reclama, com a conseqüente reforma do acórdão, ficam eles rejeitados." (RTJ 134/836, Rel. Min. SYDNEY SANCHES - grifei) O exame dos autos evidencia que os presentes embargos declaratórios consideradas revestem-se as razões de nítido expostas pela caráter própria infringente, parte embargante, circunstância essa que, por si só, basta para tornar incabível a espécie recursal ora em análise, consoante adverte o magistério jurisprudencial desta Corte, na linha dos precedentes anteriormente referidos. Impende assinalar, ainda, por necessário, e ao contrário do que sustenta a douta Procuradoria Geral da República, que a nulidade radical que afeta os diversos atos de interrogatório judicial, quando provocada pela conduta do magistrado que arbitrariamente nega, ao réu, o direito – por este titularizado – de HC 94.016 ED formular reperguntas contamina, processo, por aos efeito notadamente demais causal, litisconsortes todos aqueles de os índole penais atos passivos, subseqüentes probatória, do como a inquirição de testemunhas arroladas pelo Ministério Público. É que, no caso, há nítida relação de causa e efeito entre a conduta judicial e as conseqüências geradas no âmbito do processo penal de conhecimento, pois, com a arbitrária recusa do magistrado em permitir, ao réu, a legítima formulação de reperguntas aos demais litisconsortes penais passivos, o ora paciente, por intermédio de seus Advogados, deixou de obter dados informativos que lhe poderiam ser úteis quando da inquirição das testemunhas arroladas pelo Ministério

Fig. 1 Sample legal ruling highlighting the spread of the critical information that should be covered by the summary

"hallucination" (Maynez et al. 2020; Zhao et al. 2020) when the model generates unrelated or false information. Preliminary studies showed that nearly 30% of the outputs generated by state-of-the-art summarization systems suffer from this problem (Cao et al. 2018).

Finally, another technical difficulty is to allow the model to work with more extensive sequences. Generally speaking, the worst problem is to deal with the quadratic complexity when using self-attention. Reformer (Kitaev et al. 2020) uses local self-attention and Locality Sensitive Hashing to deal with distant references. It also uses other tricks to reduce the memory requirements: chunked feed-forward layers, reversible residual layers, and axial position encodings. Longformer (Beltagy et al. 2020) replaces the standard Transformer self-attention combining local windowed attention with a task motivated global attention. Sparse Transformers (Child et al. 2019) deals with the quadratic complexity of self-attention using sparse factorizations of the attention matrix. Sparse Sinkhorn (Tay et al. 2020) splits the input sentences into buckets and uses sinkhorn normalization to sample a permutation matrix that matches the most relevant buckets. Routing Transformers (Roy et al. 2021) uses a sparse routing module based on k-means to replace the standard self-attention. Despite these advances, the applicability of these methods for summarization is still an open problem.

The objectives of this work are (*i*) to deal with more extensive input lengths typical in legal rulings and (*ii*) to reduce the number of hallucinations or unrelated topics in the generated summaries. The main contribution of this article is the proposal of LegalSumm, a summarization approach designed specifically for legal texts. To

deal with long documents, LegalSumm creates different "views" over the documents, capturing chunks of text from different parts covering more content than a single model would be able to. Each one of these chunks is used to generate a candidate summary. To avoid hallucinations, it uses the entailment module to judge how faithful the candidate summaries are with respect to the source text. The entailment module is advantageous because it can capture deep features from the text beyond a fixed set of predefined rules.

Most works (Cao et al. 2018; Mudrakarta et al. 2018) defined a set of rules for filtering the output or generating negative examples. Kryściński et al. (2020) has also used BERT for evaluating the generated summaries, but they created a fixed set of rules to generate the negative examples. This approach requires NER and POS taggers and does not allow the model to learn by itself to distinguish and a correct summary from another containing extraneous topics or terms.

We carried out experiments on the RulingBR dataset (Feijo and Moreira 2018) which consists of 10K real court rulings written in Portuguese. In our evaluation, LegalSumm is compared to internal and external baselines. The results show that LegalSumm can select the most suitable summary among a set of candidates, leading to improved summarization quality (measured by ROUGE scores).

The remainder of this article is organized as follows. Section 2 presents an overview of summarization applied for the legal domain and related work dedicated to mitigating the "hallucination" problem. Section 3 describes the RulingBR dataset used in our experiments. Section 4 gives an overview of LegalSumm and details how training and inference are made. Section 5 describes the experiments designed to check the effectiveness of our proposed method and discusses the results. Finally, Sect. 6 summarizes the main aspects of LegalSumm, showing its suitability as an abstractive method for summarizing legal rulings.

2 Related work

Legal text summarization has been studied for decades. The works on this topic can be split into two categories. The first, called extractive summarization, identifies and ranks the passages with meaningful information until the desired summary is reached. The second, called abstractive summarization, builds an abstract representation and generates a summary often paraphrasing the source text.

2.1 Extractive summarization

Early techniques were able to generate simple headnotes extracting entire text passages (Gelbart and Smith 1991; Moens and Uyttendaele 1997). Their approach used a combination of the search for keywords and particular patterns combined with a weighting mechanism to determine the relevant sections to be included in the headnotes. Galgani et al. (2012) used Ripple Down Rules (Compton and Jansen 1990) to create incremental Knowledge Acquisition. These rules use features such as the

position of the citation and word statistics to decide which sentences should make up the summaries.

Pandya (2019) used the fact that the summary should cover topics addressed by the decision. The author employed the *k*-means algorithm to identify sentences that were central to each cluster. More recently, Zhong et al. (2019) presented an extractive summarizer using a CNN classifier based on the Maximum Marginal Relevance (Carbonell and Goldstein 1998) algorithm. Their work included a module to classify sentences according to their function in classes "Reasoning/Evidential Support" or "Others", following previous work that has explored rhetorical roles of sentences (Grover et al. 2003b, a; Yousfi-Monod et al. 2010) to create an extractive summary.

All these aforementioned methods are extractive. These summaries are usually "safe" because they select entire sentences avoiding introducing information that is not present in the source text.

2.2 Abstractive summarization

When a human generates a summary, they often paraphrase the source text. A summary that resembles human-generated summaries should paraphrase topics and ideas from the source text. Deep learning models were able to encode these main components and generate (decode) text representing these topics. However, while generating the output text, sometimes the decoder produces some information not present in the source text. The applicability of abstractive text summarization for the legal area requires that these models produce faithful summaries. Some recent works have been proposed to try to mitigate this issue.

Cao et al. (2018) proposed building extra relevant information using a set of rules of extracted passages and dependency parsing techniques to identify fact descriptions in the source text. Later, this data is used together with the source to generate the final summaries. Zhao et al. (2020) proposed a re-ranking method from candidate summaries generated by beam search. They extract specific kinds of entities (dates, numbers, sums of money, etc) to provide additional information for their re-ranking strategy. The authors pointed out that this strategy is similar to the one proposed by Falke et al. (2019), which is applied at the sentence level. The difference is that Zhao et al. (2020) use it globally. These methods require an efficient Named Entity Recognition (NER) tool to identify all such entities in both articles and summaries. This approach is harder to apply in legal texts because of the specific expressions and jargon inherent to the domain. Also, reliable trained NER tools may not be available for some languages (like Portuguese). These reasons make these approaches impractical in our context.

Goodrich et al. (2019) proposed using triplets *subject-relation-object* from a Part-Of-Speech (POS) tagger to check facts between the generated summary and the source document. Kryściński et al. (2020) proposed generating entailment relations from unsupervised data. They modified the original data with external tools for paraphrasing, sentence negation, pronoun swap, entity swap, number swap, and noise injection. This approach has the advantage of creating multiple examples from the training data, offering a good option for low-resource languages or small datasets.

However, this approach requires both NER and POS tagger, which can be difficult for low-resource languages or domain-specific areas like the legal domain. Matsumaru et al. (2020) used human evaluations to judge entailment from generated summaries. With this supervised data, the authors trained a binary BERT model to classify the entailment of an article-summary pair. Later, entailment information was used to train their final transformer model. They reported better entailment according to human evaluations, without quantitative improvement. This approach is more expensive as it requires previous human evaluation for building the required training data. Also, legal rulings would require specialized human judges, who are not readily available. Finally, while using just the first three lines from news articles may be enough to yield a good summary, legal rulings require more information for generating the summary.

General abstractive summarization models on their own are not suitable for generating summaries of legal documents. The standard procedure of truncating the start of long sequences may hide vital information that could jeopardize the model's ability to generate the expected summary. Abstractive textual summarization in the Legal domain is a growing field. There are still only a few works dedicated to this topic (Feijo and Moreira 2019; Luijtgaarden 2019). The lack of works may be explained by the differences between Common and Civil Law, which makes it harder to define a general approach suitable for any legal system. Improvement proposals based on annotated data or reliable POS and NER taggers are unfeasible for the legal domain in low-resource languages. These facts make abstractive textual summarization in the Legal field a challenging task.

2.3 Evaluation metrics

The standard evaluation metric for text summarization is the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin 2004). The general idea of this metric is to count the number of overlapping units between one or more reference summaries and the machine-generated summary. It is expected that a high-quality summary should use the same words found in the reference summaries and preferably in the same order.

There are five variants of ROUGE metrics: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. In our experiments, we will be using the two most common variants to evaluate the summarization task: ROUGE-N and ROUGE-L. ROUGE-N measures n-gram units between a candidate and a collection of reference summaries. This metric is recall-oriented because the denominator in Equation 1 is the sum of n-grams in the reference.

$$ROUGE - N = \frac{\sum_{S \in RefSums} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{S \in RefSums} \sum_{gram_n \in S} count(gram_n)} \quad (1)$$

where n stands for the length of the n-gram, $gram_n$, and $count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of

reference summaries. When multiple references are used, the measure retrieves the one with the maximum score.

ROUGE-L measures the Longest Common Subsequence (LCS). Considering a sentence as a sequence of words, the task is to find a sub-sequence (an ordered subset) match between the candidate and one reference. Even if they are separated by other words, words occurring in the same order might indicate a good match between candidate and reference.

$$\begin{aligned} R_{lcs} &= \frac{LCS(X,Y)}{m} \\ P_{lcs} &= \frac{LCS(X,Y)}{n} \\ F_{lcs} &= \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \end{aligned} \tag{2}$$

where: X = reference summary of length m
Y = candidate summary of length n
R_{lcs} = recall
P_{lcs} = precision
β = weight of recall over precision
F_{lcs} = ROUGE-L

Equation 2 shows how ROUGE-L is calculated. First, the LCS between a reference X and a candidate Y is computed. ROUGE-L is the F-measure, *i.e.,* the harmonic mean between recall and precision. One critical distinction between ROUGE-N and ROUGE-L is that the first is recall-based, while the second is F-measure-based. ROUGE-N is not affected by a lengthy candidate. On the other hand, ROUGE-L accounts for both recall and precision and is affected by the length of the candidate.

Despite some criticism about the effectiveness of ROUGE scores (Kryściński et al. 2020) when evaluating the introduction of false facts in abstractive summarization, this metric is still the most used in assessing summarization.

3 Dataset

Each country has its own set of rules and statutes that define the Law. The Brazilian Judicial System uses precedents but is more tied to written statutes. In this work, we will be using the RulingBR (Brazilian Supreme Court Rulings) dataset (Feijo and Moreira 2018). It is composed of around 10K court rulings written in Portuguese, amounting to about 176MB of data. The rulings cover abstract constitutional control, administrative, civil, civil procedural, constitutional, consumer, criminal, criminal procedural, economic, electoral, environmental, financial, fundamental rights, labor, notarial, public international, social security, request of extradition, tax, and urban Law. There are cases from 18 judges.

Each ruling is split into four sections: *summary*, *report*, *vote*, and *judgment*. The summary contains the main topics discussed in the case and how the judges

decided. It addresses the main topics from the other three sections. The summary has an average length of 191 tokens, ranging from 25 to a few hundred tokens. The report section contains a compilation of the main arguments and events that happened during the trial. Its length can vary broadly, from a few dozens up to a few thousand tokens. The vote section contains the judges' positions; they evaluate the arguments and expose their decision. The length of this section can also vary even more than the report. Finally, the judgment section is, in general, short and compiles the outcome as granted or denied. The text of the rulings and their corresponding summaries have an average length of 2,661 and 191 tokens, respectively. The compression ratio (average number of tokens in the summary divided by the average number of tokens in the ruling) is 7%.

The conventional format for the summary section is to have a header that presents representative keywords, terms, and expressions designating the case's fundamental elements; and a second part where the extract of the conclusion of the decision is given (Guimarães 2011). The topics covered by the summary are spread among the report, vote, and judgment. In general, none of these three sections by themselves are enough to summarize the ruling correctly. The summary section will be used as our *ground-truth summary*, and the combination of all other sections will compose the *source text* to be summarized.

4 LegalSumm

As discussed in the Introduction, abstractive summarization might introduce extraneous subjects or facts. This problem may happen because the model has often seen one subject tied to some context or does not have enough training data. When training for summarization, the model learns to generate texts that it has often seen during training. Deep learning usually applies complex models that are limited in length. This represents a problem for applications in the legal area, where long texts are the standard. If the model cannot work with entire documents, the standard practice is to truncate the document at the maximum length. This approach may hide vital context data for the model correctly generating summaries. Our goal here is twofold. We aim to handle long texts and also to minimize the hallucinations generated by the model. We hypothesize that we may teach a model to distinguish if a summary belongs to a given ruling and that this can improve summarization quality.

LegalSumm is suitable for long documents in which the focus of interest is dispersed throughout the text. It requires some structure of the source text to allow the extraction of coherent "views". LegalSumm does not require external taggers and may look at several parts of the rulings. This ability is handy when working with long sequences and when the summary topics are spread through the source text. LegalSumm is not geared towards news and scientific articles because the focus of the summaries from these types of documents is usually either at the beginning (as in news articles) or the end of the document (as in scientific papers).

4.1 Overview

Figure 2 shows our LegalSumm proposal for abstractively generating text summaries from legal decisions.

Two practical difficulties arise from the fact that legal decisions are frequently long. First, complex models often are unable to handle long sequences due to high memory requirements. Second, even when the models can work with such long texts, the attention mechanism becomes too sparse and unable to focus on the relevant topics.

4.1.1 Building the input data

To overcome these practical problems, we propose splitting the text from a ruling into smaller samples, called *chunks*, that are generated according to predefined rules. The set of rules used for generating these views is called *strategy*. Each strategy defines how a chunk is generated from the ruling's text. These chunks are depicted on the left bottom corner of Fig. 2 and they could use the first or the last few paragraphs, or even an arrangement of sentences. Each one may be more biased to emphasize the initial, middle, or final parts of the ruling. This approach allows the model to handle long documents and also to focus on these restricted parts. LegalSumm requires at least two strategies because it needs to select the best summary within the number of candidates. Multiple strategies allow the generation of different "views", allowing our summarization module to generate an independent version of the summaries for each strategy.

Ideally, these chunks should contain coherent parts of the text. An incoherent text could jeopardize the subsequent evaluation step that assesses whether the summary

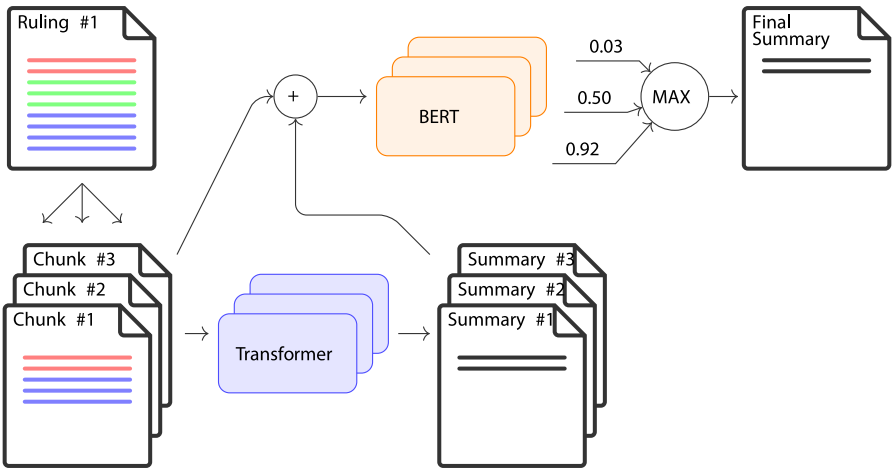


Fig. 2 LegalSumm Overview. The input source document is split into smaller chunks. Each chunk is passed through a Transformer (in light blue), which generates a summary. Each chunk-summary pair is submitted to a set of BERT models (in light orange) that output scores. The summaries with the highest scores are kept

could be inferred from the source text. Thus, the rules for selecting these portions should be preferably defined according to the inner structure of the text (paragraphs or sections). These strategies use the ruling structure (as described in Sect. 3) to split the text.

The strategy adopted here was to split the source text from the RulingBR dataset into sections. The chunk was composed of an arrangement from the three sections: "report", "vote", and "judgment". As these sections may contain more than our operational limit, the input data is always truncated at length 400. This limitation of 400 tokens required creating different strategies for assembling the chunk to be summarized, as discussed in Sect. 4.1.1. Each line in Table 1 represents a strategy for building the chunks. The first strategy defines that the "chunk" is composed using up to 300 tokens from the "report" section, concatenated with up to 100 tokens from the "judgment" section. For strategies 1 to 5, the generated chunks may be smaller than the maximum length allowed. This happens because the length of each section is enforced independently. For example, for strategy 4, if "report" has only 250 tokens, this will be the final input length under this strategy. Strategies 6, 7, and 8 allow more than 400 tokens, requiring posterior truncation at length 400.

4.1.2 Generating candidate summaries

Transformer models (Vaswani et al. 2017) (shown in light blue in Fig. 2) generate independent summaries for each chunk. The Transformer uses an Encoder-Decoder architecture with self-attention, and it can handle sequential data suitable for tasks such as translation and summarization. Having these alternative versions is vital for our framework because it can choose which version the model believes is the most truthful in relation to the source.

4.1.3 Scoring candidate summaries

The candidate summaries are combined with the input chunks (round node with the + sign in Fig. 2) and submitted to BERT (Devlin et al. 2018) models (in light orange) that predict scores for each pair, representing how confident the model is that this summary is related to the chunk. The input for BERT uses the special token

Table 1 Strategies for generating the chunks to be summarized. Each strategy (shown in the rows) uses the concatenation of the referred sections, limited to the corresponding length

#	1*st* Section	Length	2*nd* Section	Length	3*rd* Section	Length
1	report	300	vote	0	judgment	100
2	report	0	vote	300	judgment	100
3	report	150	vote	150	judgment	100
4	report	400	vote	0	judgment	0
5	report	0	vote	400	judgment	0
6	report	400	vote	400	judgment	400
7	vote	400	judgment	400	report	400
8	judgment	400	report	400	vote	400

CLS at the start, the chunk tokens, one special *SEP* token, the predicted summary tokens, another *SEP* token. The score is the output after the softmax of a binary classifier. It is a number between 0 and 1, representing how confident the model is that this summary matches the chunk. The score represents the verisimilitude that this summary was generated from a given ruling, based both on its contents and writing style. There is one score for each input strategy.

After all scores have been generated, the summary with the highest score among the candidates is selected and used as the final output summary for each case. These final summaries are the result of the collection from any of the summarizers.

4.2 Training procedure

Fig. 3 depicts the steps required for training LegalSumm. The training "chunks" are generated following the same description presented in Sect. 4.1.1. These are the inputs for building the entailment data and for training the summarization module. We can see that the two most time-consuming training tasks (steps 2 and 3) can be done in parallel, reducing the total training time. In the following sections, we detail each one of these stages.

4.2.1 Building entailment data

To reduce the number of hallucinations, we train a model to evaluate whether a generated summary could be inferred from the source text. This task is closely related to Recognizing Textual Entailment (RTE) task in Natural Language Processing (NLP). The goal of RTE is to define if one text entails, contradicts, or is neutral concerning another. We use the BERT (Devlin et al. 2018) model for RTE because it showed state-of-the-art results in this task.

To teach the model to distinguish between an appropriate summary and one out-of-the-context, we need to feed it with real and fake examples. A real example is a ground-truth chunk-summary pair from the dataset. A fake example is formed using the same "chunk", but randomly selecting another summary from the dataset. Facts are the original chunk-summary pair, and fake are the ones in which the summary cannot be entailed from the chunk.

LegalSumm generates artificial entailment data following the procedure depicted in Fig. 4. Each example (shown large on the left) comprises the chunk and its corresponding ground-truth summary. From this example, both its summary and chunk

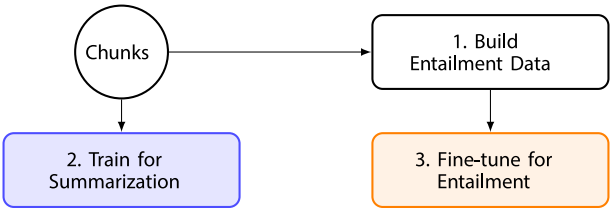


Fig. 3 LegalSumm's training procedure

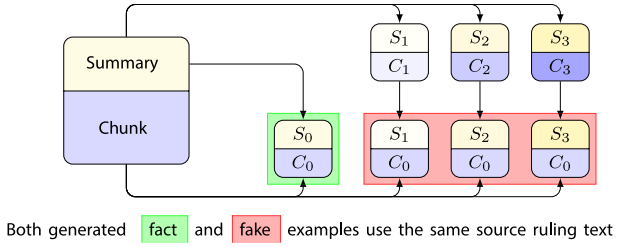


Fig. 4 Building Entailment Data

are used, and the example is tagged as "fact" (*i.e.,* S_0, shown in green). For the fake examples, the intuition is to simulate the effects of hallucinations, *i.e.,* the introduction of extraneous topics close to the ruling's broad subject. To do that, we use the category label given in the dataset that represents the broad subject of the case (*e.g.,* Habeas Corpus, Habeas Data, *Writ* of Mandamus, Appealing in Civil Procedural Law, etc). We randomly select ten rulings belonging to "related" categories (to save space, Fig. 4 depicts only three – S_1, S_2, and S_3). We consider that categories that share word bigrams in their labels are related. For example, a ruling categorized as "Appealing in Labor Procedural Law" is related to another ruling from the same category and to categories such as *"Appealing in* Criminal Law" or "Civil *Procedural Law*" as they share bigrams. Figure 4 uses different background shades to represent these differences from the source ruling. We keep these similar summaries and replace their ruling texts with the original chunk (shown using C_0). Finally, we tag these modified examples as "fake" (generated examples with red backgrounds on the right). With this technique, some fake examples will be more closely related to the source than others. We hypothesize that this method allows distinguishing distant subjects and is even more closely related in incremental degrees of difficulty.

4.2.2 Training summarization module

LegalSumm uses one Transformer (Vaswani et al. 2017) for each strategy because the model needs to learn to summarize chunks that are in the format defined by the strategy (see Table 1). The intuition is that each strategy covers distinct topics from the original ruling. This approach allows each summarizer to be more specialized; and a specialized model can learn more easily to distinguish the relevant parts of the chunks.

To train the summarization module, we used OpenNMT-py (Klein et al. 2017), more specifically, the standard Transformer model (Vaswani et al. 2017) with 6 encoder and 6 decoder layers, embedding and hidden dimension of 512. One model is trained for each input strategy from Table 1. They were trained for 20K steps, using batch size 56, and learning rate 1×10^{-3}. Training took between two and six hours (training time varies because the input size of each strategy has a different length).

We used these trained models to generate the candidate summaries. Some constraints were applied during decoding. We use beam search with size eight, avoid

3-ngram repetition (Paulus et al. 2017), and require a minimum of 25 tokens and a maximum of 256. We chose these limits because they represent the minimum and maximum lengths found in the dataset. The test set contains around 2K examples, and the prediction took 20 minutes for each input strategy.

4.2.3 Fine-tuning for entailment

Training an entailment model from scratch requires the model to learn to "understand" the language and the complex relations from the topics. BERT (Devlin et al. 2018) is a method for pre-training language models that can later be fine-tuned for various NLP tasks. It has been applied to several such tasks, achieving results that outperform the state-of-the-art. Its authors propose a costly pre-training method using a large volume of data so that this model can understand the general topics of the language. This pre-training on general texts allows the model to generate rich representations of the source data. Later, this pre-trained model can be fine-tuned for specific tasks such as RTE.

The authors of BERT released a multilingual model trained on the Wikipedias from 104 languages (Pires et al. 2019). Portuguese is among the languages that compose the multilingual model so that it can be used with the dataset of Brazilian rulings. LegalSumm uses one entailment model for each one of the input strategies. There is no parameter sharing between these models.

To train our entailment module, we used the Hugging Face Transformers library (Wolf et al. 2019). We start with bert-base-multilingual-cased and fine-tune it for entailment. We append on top of the model a classification head for the model to choose if a given chunk-summary pair belongs to the "fact" of the "fake" class. Despite the fact that we converted all text to lowercase, we used the cased version of BERT multilingual because the uncased version also strips diacritics. Using lowercase for all data may jeopardize recognizing proper nouns, entities, and the start of the sentences. Considering that the dataset is not consistent with casing, we opted to use lowercase everywhere. Also, Portuguese uses diacritics and, while the text can be understood without them, removing them introduces noise as some discriminating features are lost – *e.g.,* the distinction between "baby" (*bebê*) and "s/he drinks" (*bebe*) is on the diacritical mark.

The input is the concatenation of a classification token, up to 400 tokens from the chunk, a special SEP token, and up to 110 tokens from the candidate summary. The output layer is just two classes representing how confident the model is that the summary entails the input data. Each fold contains 20% of testing data. From the remaining, we separated 10% for validation. For each example, we generate another ten fake examples using the method shown in Sect. 4. We trained each model for 6K steps until validation loss stopped diminishing, batch size 32, learning rate 2×10^{-5}. Training took between two and six hours for each model, depending on the length of the inputs.

Furthermore, to avoid data leakage when selecting the fake examples, we ensure that these examples come from the training folds. This way, the test fold is protected and not mixed with the training folds.

Unfortunately, the restriction of 400 tokens from the chunk might not provide enough "coverage" of the ruling for the model to correctly entail that this chunk-summary is a valid combination. In this case, as the model does not have enough information, this sample would be noisy, probably jeopardizing the model performance. We address this issue using eight models with different chunking strategies. We expect that the chunking strategy with enough information (*i.e.,* best "coverage") would provide the best summaries. In contrast, the situations in which the model has insufficient information and would not generate the best summaries receive a less confident score.

5 Experimental evaluation

To evaluate the effectiveness of LegalSumm in generating automatic legal summaries, we performed an experimental evaluation using the RulingBR dataset (described in Sect. 3).

The automatically generated summaries were scored using the official ROUGE script, without stemming. The evaluation metrics are described in Sect. 2.3. Additionally, to save space on the tables, we multiply the scores by ten in all experimental runs.

We organized the results of our experimental evaluation into four parts, each designed to answer one of the following questions. Each of these questions is addressed in the next subsections.

1. How useful is the entailment module in LegalSumm?
2. How does LegalSumm compare with baselines for text summarization?
3. How do legal experts rate the quality of the automatically generated legal summaries?
4. How does the number of fake summaries impact the results in LegalSumm?

5.1 How useful is the entailment module in LegalSumm?

Table 2 shows the results of a comparison among LegalSumm and the eight strategies for assembling the input data as shown in Table 1. Recall that these strategies rely only on standard Transformer models. Thus, this comparison works as an internal baseline.

Looking at the scores, we find that strategies 6, 7, and 8 produced the best summaries among the baselines. The advantage of these strategies can be attributed to the fact that they use the maximum number of tokens allowed. With more "features", the summarization module produces better summaries.

Although it is inappropriate to say that ROUGE correctly evaluates entailment, it is safe to consider that a higher ROUGE score might indicate that the summary more closely represents the expected topics. In this case, we assume that if our entailment module can distinguish between "fact" and "fake" chunk-summary pairs, the ROUGE score should show some improvement.

Table 2 Average ROUGE Scores for Summaries Generated by each Strategy using 5-fold cross-validation. The numbers between brackets are the standard deviations across folds. Best results in bold

Model	R1-F	R1-P	R1-R	R2-F	R2-P	R2-R	**RL-F**	RL-P	RL-R
LegalSumm	**43**(±3)	**63**(±1)	**37**(±3)	**27**(±3)	**39**(±3)	**23**(±3)	**35**(±3)	**51**(±1)	**30**(±3)
Transformer 1	37(±2)	55(±2)	31(±2)	20(±3)	30(±2)	17(±3)	29(±2)	43(±2)	25(±2)
Transformer 2	38(±3)	57(±2)	33(±3)	22(±3)	33(±3)	19(±3)	31(±3)	45(±2)	26(±3)
Transformer 3	39(±2)	58(±1)	33(±2)	23(±2)	33(±2)	20(±3)	31(±2)	46(±2)	27(±2)
Transformer 4	36(±2)	56(±1)	31(±2)	20(±2)	31(±1)	17(±2)	28(±2)	44(±1)	24(±2)
Transformer 5	39(±2)	59(±1)	33(±3)	23(±2)	34(±2)	20(±2)	31(±2)	47(±2)	27(±2)
Transformer 6	42(±3)	60(±2)	36(±3)	26(±3)	36(±3)	22(±3)	34(±3)	48(±2)	29(±3)
Transformer 7	42(±3)	60(±2)	36(±3)	26(±3)	36(±3)	**23**(±4)	34(±3)	48(±2)	29(±3)
Transformer 8	42(±2)	60(±2)	36(±3)	26(±3)	36(±2)	22(±3)	34(±2)	48(±2)	29(±3)

To illustrate LegalSumm's ability to select a summary among the candidates, we show a shortened example in Table 3. The original summaries in Portuguese were translated into English so that a broader audience could understand them. The first line represents the ground-truth summary. The following lines represent the candidate summaries generated by each strategy. Imprecise or factual errors are highlighted in yellow. Missing information such as "credit cooperative" are not represented and would lower the score. The "Score" column shows the confidence that this is a summary from the input chunk. We can see that LegalSumm assigns higher scores to the summaries with fewer factual errors.

Summaries 2, 5, and 7 have more invalid information and received the lowest scores. Summaries generated with strategies 1, 4, and 6 contain fewer errors and received the highest scores. In this example, the summary from strategy 6 was chosen.

In an error analysis experiment, we scored each of the 10K summaries generated by each of the eight Transformer models and the summaries picked by LegalSumm using RL-F. If the entailment module made the optimal choice every time (*i.e.,* selected the summary that maximizes RL-F) then the average score for LegalSumm would be ten points higher. The choices made by the entailment module are not optimal because of two main reasons (*i*) the goal of the training procedure is not to maximize ROUGE scores, which are not known at that stage; and (*ii*) if multiple good candidate summaries are available, LegalSumm will be penalized for not choosing the one that follows the same writing style as the ground-truth.

5.2 How does LegalSumm compare with baselines for text summarization?

In this section, we compare LegalSumm with existing text summarization systems, including BertSumExt and BertSumAbs by Liu and Lapata (2019), and BART by Lewis et al. (2020). Additionally, we also compare with the results published by Feijo and Moreira (2019) on the same dataset.

Next, we provide a brief description of these baselines. Notice that BertSumExt, BertSumAbs, and BART cannot readily work on the RulingBR dataset, which is in

Table 3 Candidate summaries are automatically generated by the summarization module and their scores.

Strategy	Score	Summary
Ground Truth	-	labor. union contribution. credit cooperative. equivalent to a banking establishment. absence of the required pre-questioning. overviews 282 and 356 of the stf. interpretation of statutory legislation.
1	0.96	labor law. rural union contribution. proof. absence of pre-questioning. overviews 282 and 356 of the stf. review of facts and evidence. summary no. 279 of the stf. unfeasibility of the extraordinary appeal.
2	0.00	administrative. restructuring of the workers of the brazilian institute for the environment and renewable natural resources - ibama.
3	0.80	labor law. rural union contribution. prescriptive period. absence of pre-questioning. overviews 282 and 356 of the stf. analysis of statutory legislation.
4	0.96	tax law. rural union contribution. absence of pre-questioning. Precedent 282 of the stf. analysis of statutory legislation.
5	0.67	search and seizure action. fiduciary sale in guarantee. civil prison. decree-law n. 911/69. binding summary no. 25 of this court. No pre-questioning. overviews 282 and 356 of the stf. analysis of statutory legislation.
6	**0.99**	labor. wage differences. rural union contribution contract. absence of pre-questioning. overviews 282 and 356 of the stf.
7	0.76	civil procedure. execution. savings. inflationary purges. no pre-questioning. overviews 282 and 356 of the stf. analysis of statutory legislation.
8	0.84	labor law. rural union contribution. compensation. pre-questioning. overviews 282 and 356 of the stf. analysis of statutory legislation.

Factual errors are highlighted. All summaries were shortened to save space and translated into English to be understood by a broader audience

Portuguese. Thus, adaptations were needed. These models were trained using the concatenation of sections "Report", "Vote", and "Judgment", *i.e.,* the same contents seen by LegalSumm.

- **BertSumExt** (Liu and Lapata 2019) is an extractive approach that uses BERT to generate the features for the sentences and to decide which sentences should compose the final summary.
- **BertSumAbs** (Liu and Lapata 2019) is an abstractive model that uses a standard BERT as encoder and trains a decoder with a causal mask from scratch. Because BertSumExt and BertSumAbs were trained using English-only texts, and we need these models to work with Portuguese-only texts, we could not use any fine-tuned checkpoints. Thus, we used the original code shared by the authors and replaced "bert-base-cased" with "bert-base-multilingual-cased" (which includes Portuguese) and fine-tuned it for summarization using the RulingBR dataset. Fine-tuning BertSumExt required four hours, and BertSumAbs required 28 hours, both using one Nvidia Tesla V-100 GPU.
- **BART** (Lewis et al. 2020) is also an encoder-decoder model that uses the same BERT architecture as the encoder. The decoder uses a causal mask with a cross-attention with the encoder. Again, there is no BART model trained using Portuguese texts. Thus, we were required to run the entire pre-training. For this task, we use three million sentences available from Goldhan et al. (2012). BART's "large" configuration was used, and pre-training used one Nvidia Tesla V-100

GPU for 210K steps. Pre-training took about 70 hours. After pre-training, each fold was fine-tuned for summarization. Fine-tuning took about two hours.

- Feijo and Moreira (2019) published results on the RulingBR. In their work, they ran several experiments using both extractive and abstractive models. For the abstractive models, they used standard RNN and Transformer architectures. The minimum summary lengths were set to 100 and 120 tokens, respectively. The Transformer model with a minimum summary length of 100 tokens was the configuration that produced the best scores – and these are the results we reproduce here. The architecture is an ordinary encoder-decoder Transformer model (Vaswani et al. 2017), akin to the one used by the summarization module in LegalSumm. To be comparable with the scores published by Feijo and Moreira (2019), we report results using the same train/test splits rather than using cross-validation.

Table 4 shows the results for the comparison with external baselines. BertSumExt did not produce good summaries in our legal data. We believe this is due to differences between summarizing news articles and legal documents (as discussed in Sect. 1). On the other hand, BertSumAbs and BART produced competitive summaries but required a very long training procedure. Even with these strong baselines, the results showed that LegalSumm outperformed or matched the ROUGE-F scores for all metrics evaluated. Despite the required training of several similar models, we improved ROUGE scores without using larger versions that are expensive to train. We ran paired *t*-tests between the RL-F scores (which is the most important of the metrics we reported) of LegalSumm and all baselines, including our eight Transformers. The results showed statistically significant differences between LegalSumm and all other baselines (p-values $\ll 0.01$ for BertSumExt and BertSumAbs and p-value < 0.03 for BART considering the Bonferroni correction). Our method has the advantage of not requiring additional trained NER or POS tools, and it is suitable even for low-resource languages such as Portuguese.

Table 5 shows that LegalSumm improved or matched all ROUGE F-scores in comparison to the basic encoder-decoder Transformer by Feijo and Moreira (2019). The most significant advantage was on precision scores. LegalSumm can choose safer alternatives avoiding unusual terms or extended sentences that could append wrong information. This behavior positively impacts precision, but it could harm recall because more complete or extended candidates would not be selected. The baseline, however, achieves a higher recall. That happens because it does not have the restriction imposed by the entailment module and thus can generate more terms that are found in the ground-truth summaries. Since LegalSumm's advantage in the precision is greater than its loss in the recall, its F-scores were higher.

Table 4 Average ROUGE scores for summaries generated by LegalSumm and external baselines using 5-fold cross-validation.

Model	R1-F	R1-P	R1-R	R2-F	R2-P	R2-R	**RL-F**	RL-P	RL-R
LegalSumm	**43**(±3)	**63**(±1)	37(±3)	**27**(±3)	**39**(±3)	23(±3)	**35**(±3)	**51**(±1)	30(±3)
BertSumExt	23(±2)	38(±1)	20(±2)	6(±1)	10(±1	5(±1)	16(±1)	28(±1)	14(±1)
BertSumAbs	41(±3)	62(±1)	34(±3)	25(±3)	37(±2)	21(±3)	33(±3)	49(±1)	28(±3)
BART	**43**(±3)	56(±2)	**41**(±3)	25(±4)	32(±4)	**24(±4)**	32(±4)	41(±3)	**31(±4)**

The numbers between brackets are the standard deviations across folds. Best results in bold

5.3 How do legal experts rate the quality of the automatically generated summaries?

To answer this question, we designed an evaluation experiment in which legal experts evaluate randomly selected cases. Our experts were eleven volunteers with a law degree and over ten years of experience working with legal rulings. For each case, the legal experts were presented with (*i*) the full text of the ruling, (*ii*) the ground-truth summary, (*iii*) the summary generated by BertSumAbs, and (*iii*) the summary produced by LegalSumm. This assessment was a time-consuming task that required familiarity with the topic. Therefore, it was restricted to ten cases. We chose BertSumAbs as the comparison model because it obtained the highest RL-F scores among our external baselines. For each of the automatically generated summaries, the legal expert was asked to rate the summary concerning four aspects by assigning a score from one and five. A score of one represents strong disagreement, while five represents strong agreement with each of the following statements.

1. The summary covers important parts of the text.
2. The summary presents a coherent flow.
3. The summary is faithful to the facts and does not introduce extraneous facts.
4. The summary could replace the original (manually created) summary.

The results of the evaluation by the legal experts are shown in Fig. 5. In comparison with BertSumAbs, LegalSumm was better in all four evaluated aspects. Regarding flow coherence and faithfulness, LegalSumm had about twice the number of positive assessments, and, for replaceability, the number was almost three times as high. In LegalSumm, the aspects with the most positive assessments

Table 5 ROUGE Scores on train/test splits.

Approach	R1-F	R1-P	R1-R	R2-F	R2-P	R2-R	**RL-F**	RL-P	RL-R
LegalSumm	**44**	**64**	38	**29**	**42**	25	**36**	**52**	31
Feijo and Moreira (2019)	**44**	49	**46**	27	28	**28**	35	39	**37**

Best results are in bold

were coherence (56% of the answers agreed that the flow was coherent) and coverage (50% of the answers). However, concerning being faithful to the facts and the possibility of using the generated summary as a replacement to the original one, over half of the answers were in disagreement. These negative ratings mean that, despite the improvements, these generated summaries still lack the reliability to be used independently. Nevertheless, LegalSumm's summaries could be used as drafts that require manual checking, thus alleviating the burden on human summarizers.

5.4 Impact of the variation of the number of fake examples

Some points deserve deeper analysis. First, according to Cao et al. (2018), state-of-the-art abstractive summarization models generate around 30% of summaries with one or more false facts. The definition of a false fact is when a triplet (subject + predicate + object) could not find support in the source text. Thus, when applying our entailment classifier over the test set, we expected that the summaries would contain a fair number of "hallucinations", with around 30% of low (< 0.5) scores, indicating that these summaries were considered as fake. However, the histogram shown in Fig. 6 shows that the model classified 60% of cases as "fake". This bias is a consequence of the training data having ten times more fake samples than true samples.

To investigate the influence of generating ten fakes for each original example, we also experimented using five and 15. The number of "replacements" (*i.e.,* when the generated summary is replaced by another with higher confidence) increased with the higher number of fake examples. A ratio of 5 to one generated 1097 replacements, which increased to 1172 and 1501 for ratios of ten to one and 15 to one, respectively. Despite these changes, the results were almost identical with slight

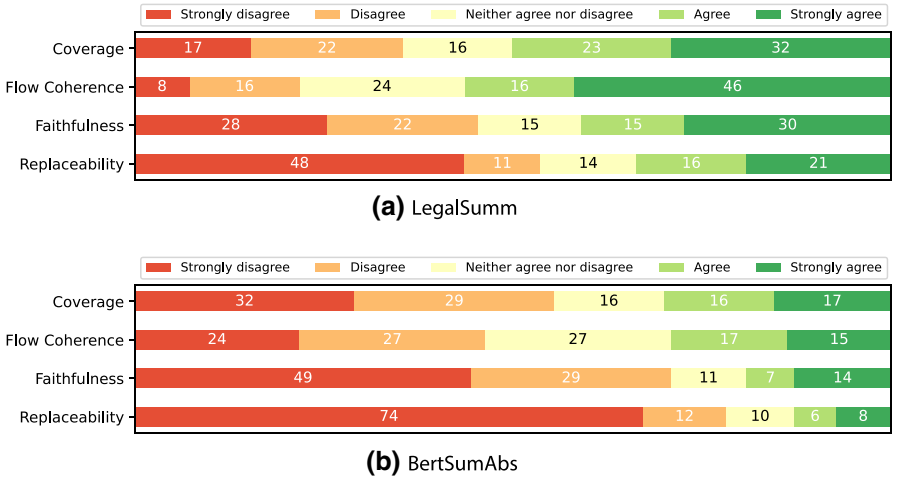


Fig. 5 Scores given by experts to summaries generated by LegalSumm and BertSumAbs to each of the four statements regarding the quality of the summaries

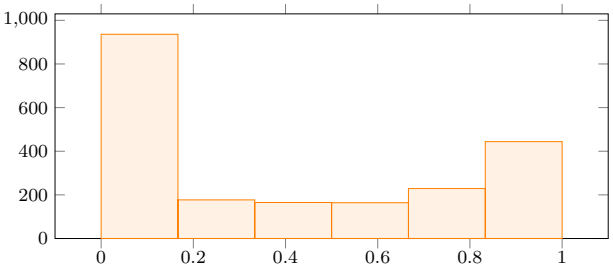


Fig. 6 Histogram of entailment scores for a test fold in one of the chunking strategies

modifications in precision or recall, but not altering the F-score of any strategy. These results show that the entailment component is relatively tolerant to class imbalance and can be used in contexts where the source data has some structure, and the relevant data for the summaries are spread in the text.

5.5 Limitations

Using BERT for assessing the entailment restricts the maximum length supported by LegalSumm. BERT restricts the length up to 512 tokens. As BERT needs to evaluate both the chunk and the candidate summary, the chunk length had to be limited to be at a maximum of 400 tokens. Each summary length increment requires reducing the length of the chunks. Therefore, if the summaries needed to be twice as large, the chunks would be limited to almost this same length. In practice, LegalSumm would not be suitable for summaries longer than 200 tokens. To overcome this limitation, BERT should be replaced by another model that supports more than 512 tokens. Furthermore, as discussed in Sect. 5.1, we observed that the choices among the candidate summaries are not always optimal. Finally, as shown in Sect. 5.3, despite outperforming all baselines, the summaries produced by LegalSumm are not good enough to be used as a replacement of manually generated summaries.

6 Conclusion

This article presented LegalSumm, a method for improving abstractive summarization of legal rulings. LegalSumm works with long documents by splitting the legal document into predefined chunks that the model can handle and generating a candidate summary. Also, it can remove extraneous topics by selecting the most suitable summaries among these candidate summaries. Compared to other methods for avoiding "hallucination", LegalSumm has the advantage of not requiring NER or POS taggers. It improves the ROUGE scores without increasing the size of the summarizer.

We conducted different evaluation experiments. The first experiment demonstrates that LegalSumm can make good choices among the candidate summaries generated by standard Transformer models. In the second experiment, LegalSumm

is compared with summarization baselines including BertSumExt, BertSumAbs (Liu and Lapata 2019), and BART (Lewis et al. 2020), which needed to be adapted to work with Portuguese texts. We also show how LegalSumm compares with the results by Feijo and Moreira (2019) on the same dataset. LegalSumm was able to outperform or match all ROUGE F-scores from these baselines. In our third evaluation, we asked legal experts to assess the automatically generated summaries in terms of coverage, coherence, faithfulness, and replaceability. Finally, we discussed the impact of varying the number of fake examples and the limitations of the proposed method.

Acknowledgements This work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors thank the two anonymous reviewers whose suggestions helped improve and clarify this manuscript.

References

Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer arXiv preprint arXiv:200405150

Cao Z, Wei F, Li W, Li S (2018) Faithful to the original: fact aware neural abstractive summarization. In: thirty-second AAAI conference on artificial intelligence

Carbonell J, Goldstein J (1998) The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, pp 335–336

Chandrasekaran MK, Yasunaga M, Radev D, Freitag D, Kan MY (2019) Overview and results: Cl-scisumm shared task 2019. In: in proceedings of joint workshop on bibliometric-enhanced information retrieval and NLP for digital libraries (BIRNDL 2019)

Child R, Gray S, Radford A, Sutskever I (2019) Generating long sequences with sparse transformers. arXiv preprint arXiv:190410509

Compton P, Jansen R (1990) Knowledge in context: A strategy for expert system maintenance. In: proceedings of the second Australian joint conference on artificial intelligence, Springer-Verlag, Berlin, Heidelberg, AI '88, p 292–306

Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Falke T, Ribeiro LF, Utama PA, Dagan I, Gurevych I (2019) Ranking generated summaries by correctness: an interesting but challenging application for natural language inference. In: proceedings of the 57th annual meeting of the association for computational linguistics, pp 2214–2220

Fan A, Grangier D, Auli M (2018) Controllable abstractive summarization. In: proceedings of the 2nd workshop on neural machine translation and generation, 45–54

Feijo D, Moreira V (2018) Rulingbr: A summarization dataset for legal texts. In: computational processing of the portuguese language (PROPOR 2018), Springer International Publishing, pp 255–264

Feijo D, Moreira V (2019) Summarizing legal rulings: comparative experiments. in: proceedings of the international conference on recent advances in natural language processing (RANLP 2019), pp 313–322

Galgani F, Compton P, Hoffmann A (2012). Combining different summarization techniques for legal text. In: proceedings of the workshop on innovative hybrid approaches to the processing of textual data, Association for Computational Linguistics, pp 115–123

Gelbart D, Smith J (1991) Beyond boolean search: Flexicon, a legal text-based intelligent system. In: proceedings of the 3rd international conference on Artificial intelligence and law, pp 225–234

Goldhan D, Eckart T, Quasthoff U (2012) Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In: proceedings of the 8th international language resources and evaluation (LREC'12)

Goodrich B, Rao V, Liu PJ, Saleh M (2019) Assessing the factual accuracy of generated text. In: proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 166–175

Grover C, Hachey B, Hughson I, Korycinski C (2003a) Automatic summarisation of legal documents. In: proceedings of the 9th international conference on artificial intelligence and law, association for computing machinery, ICAIL '03, p 243–251

Grover C, Hachey B, Korycinski C (2003b). Summarising legal texts: Sentential tense and argumentative roles. In: proceedings of the HLT-NAACL 03 on text summarization workshop-volume 5, association for computational linguistics, pp 33–40

Guimarães JAC (2011) Elaboração de ementas jurisprudenciais: elementos teórico-metodológicos. Série Monografias do CEJ 9

Kitaev N, Kaiser Ł, Levskaya A (2020) Reformer: the efficient transformer arXiv preprint arXiv:2001044451

Klein G, Kim Y, Deng Y, Senellart J, Rush A (2017). OpenNMT: Open-source toolkit for neural machine translation. In: proceedings of ACL 2017, system demonstrations, association for computational linguistics, Vancouver, Canada, pp 67–72

Kryściński W, McCann B, Xiong C, Socher R (2020) Evaluating the factual consistency of abstractive text summarization. In: proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 9332–9346

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, 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, association for computational linguistics, Online, pp 7871–7880, https://doi.org/10.18653/v1/2020.acl-main.703, https://aclanthology.org/2020.acl-main.703

Lin CY (2004) Rouge: a package for automatic evaluation of summaries. In: text summarization branches out, pp 74–81

Liu Y (2019) Fine-tune BERT for extractive summarization. arXiv preprint arXiv:190310318

Liu Y, Lapata M (2019) Fine-tune BERT for extractive summarization. arXiv preprint arXiv:190310318

Luijtgaarden N (2019) Automatic summarization of legal text. Utrecht University Master's thesis

Matsumaru K, Takase S, Okazaki N (2020) Text summarization with pretrained encoders. In: proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 3721–3731

Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. In: proceedings of the 58th annual meeting of the association for computational linguistics, pp 1906–1919

Moens MF, Uyttendaele C (1997) Automatic text structuring and categorization as a first step in summarizing legal cases. Inf Proces Manag 33(6):727–737

Mudrakarta PK, Taly A, Sundararajan M, Dhamdhere K (2018) Did the model understand the question? In: proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 1896–1906

Pandya V (2019) Automatic text summarization of legal cases: A hybrid approach. 5th international conference on advances in computer science and information technology (ACSTY-2019)

Paulus R, Xiong C, Socher R (2017) A deep reinforced model for abstractive summarization. arXiv preprint arXiv:170504304

Pires T, Schlinger E, Garrette D (2019) How multilingual is multilingual BERT? In: proceedings of the 57th annual meeting of the association for computational linguistics, pp 4996–5001

Roy A, Saffar M, Vaswani A, Grangier D (2021) Efficient content-based sparse attention with routing transformers. Transac Assoc Comput Linguistics 9:53–68

See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. In: proceedings of the 55th annual meeting of the association for computational linguistics 1:1073–1083

Tay Y, Bahri D, Yang L, Metzler D, Juan DC (2020) Sparse sinkhorn attention. In: international conference on machine learning, PMLR, pp 9438–9447

Turtle H (1995) Text retrieval in the legal world. Artif Intell and Law 3(1):5–54

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: proceedings of the 31st international conference on neural information processing systems, pp 6000–6010

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, et al. (2019) Huggingface's transformers: state-of-the-art natural language processing. arXiv preprint arXiv:191003771

Yousfi-Monod M, Farzindar A, Lapalme G (2010) Supervised machine learning for summarizing legal documents. In: Canadian conference on artificial intelligence, Springer, pp 51–62

Zhang J, Zhao Y, Saleh M, Liu P (2020) Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: international conference on machine learning, PMLR, pp 11328–11339

Zhang X, Wei F, Zhou M (2019) HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In: proceedings of the 57th annual meeting of the association for computational linguistics, pp 5059–5069

Zhao Z, Cohen SB, Webber B (2020) Reducing quantity hallucinations in abstractive summarization. arXiv preprint arXiv:200913312

Zhong L, Zhong Z, Zhao Z, Wang S, Ashley KD, Grabmair M (2019) Automatic summarization of legal decisions using iterative masking of predictive sentences. In: proceedings of the seventeenth international conference on artificial intelligence and law, ICAIL '19, p 163–172