

Capítulo 22

Sumarização Automática

*Jackson Wilke da Cruz Souza
Paula Christina Figueira Cardoso
Crysttian Arantes Paixão*

Publicado em: 13/03/2024

22.1 Introdução

Talvez resumir seja uma das tarefas mais comuns em nosso cotidiano. Não são raras as vezes que nos pegamos reportando algo a alguém: quando compartilhamos sobre o último filme que assistimos, ou o livro que lemos, ou quem sabe quando contamos para alguém sobre aquela história ou notícia que vimos ou vivenciamos.

Nessas tarefas, partimos do princípio que não conseguiremos reproduzir exatamente o que aconteceu, mas faremos o máximo de esforço para transmitir uma mensagem o mais próximo do original, valorizando o **conteúdo** mesmo em detrimento da **forma**. Assim, podemos optar por reestruturar o fato acontecido ou vivenciado dentro da narrativa sintetizada: selecionamos o que julgamos ser o mais importante do evento; lançamos mão de inversões dos fatos, prevendo melhorar a compreensão do ouvinte/leitor da narrativa; ou ainda trocamos as palavras e construções sintáticas originais por outras semelhantes ou sinônimas.

Outra consideração importante nesse processo é levar em conta o tempo ou limite (textual) que temos à disposição para a construção da narrativa que foi resumida. Saber disso de antemão ditará o quão detalhista ou generalista precisaremos ser na síntese dos fatos, tendo em vista que quanto mais tempo e espaço tivermos, mais próximo à realidade estará a narrativa sintetizada.

Ao chegar nesse ponto do nosso exemplo, é necessário nos questionar sobre algo bastante relevante: o quanto das percepções pessoais e autorais ficam na narrativa quando comparado ao evento real? O fato de haver liberdade estilística na produção da síntese, não significa que a narrativa deva veicular conteúdo falso, já que ela precisa representar o evento real. Apesar disso, é possível, em alguns casos, acrescentar e evidenciar nossas avaliações e perspectivas. O esforço do autor é pensar em diferentes operações textuais que incidem sobre o efeito de sentido de verossimilhança ao conteúdo, como a paráfrase ou ainda a cópia.

Do ponto de vista linguístico, os estudos sobre as operações associadas à transformação de texto-fonte extenso em um texto-alvo sintetizado remetem ao conceito de **reescrita**. Para Fiad (2013), esse conceito é tido como o trabalho que é feito por um autor (com ou sem auxílio de um mediador) por meio de operações na linguagem que abrangem diferentes



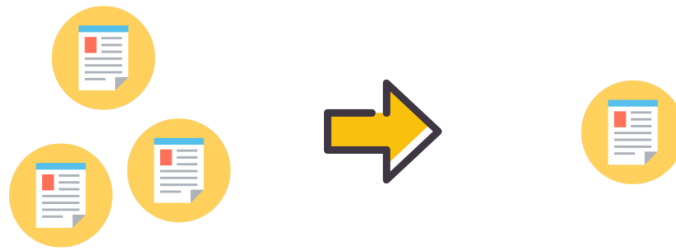
aspectos do texto conduzindo-o à sua modificação. Assim, podemos dizer que o processo de resumir algo está compreendido na reescrita.

Já do ponto de vista computacional, essa atividade que nos é tão comum e feita de maneira, por vezes, intuitiva, pode ser realizada automaticamente. Em processamento de linguagem natural (PLN), essa tarefa é conhecida como **sumarização automática** (doravante, SA). A SA pode ser definida como a produção automatizada de versões reduzidas de outros textos, resultando em sumários (ou “resumos”). Rino; Pardo (2003) elucidam algumas premissas para a SA, a saber:

1. Existência de um texto que deve ser condensado, tido como *texto-fonte*;
2. O texto-fonte dispõe de (i) uma ideia ou tópico principal, (ii) um conjunto de informações relacionadas entre si, (iii) um propósito comunicacional que organiza e relaciona as informações entre si e (iv) uma narrativa visando a informatividade de maneira coerente;
3. Identificação do conteúdo relevante no texto-fonte e construção de uma nova narrativa coerente no sumário, preservando a ideia inicial.

A partir dessas premissas, apresentamos um modelo de representação do processo de SA, na Figura 22.1.

Figura 22.1: Processo genérico de Sumarização Automática



Na Figura 22.1, tem-se que SA se inicia na consulta de um ou mais textos-fonte, que será(ão) submetido(s) a determinada análise para fazer a verificação do que é mais ou menos importante. Quando o processo de sumarização se baseia em apenas um texto-fonte, diz-se que a sumarização é **monodocumento**; quando baseado em dois ou mais textos-fonte, **multidocumento**. Como resultado, é elaborado o sumário, caracterizado por ser uma versão menor do(s) texto(s)-fonte. Mais adiante, detalharemos esses e outros aspectos importantes sobre a SA, como quantidade de textos-fonte a resumir e de línguas nesse processo, além de tendências e perspectivas metodológicas.

Com relação às tendências de **abordagens da SA**, ao longo do tempo, percebemos que as pesquisas em SA do começo dos anos 70 até o início da década de 90 tinham como motivação a recuperação de informação em textos. Por conta disso, os métodos utilizados para a sumarização nesse período se baseiam quase que exclusivamente em informações na superfície textual, como a ocorrência de palavras-chave no título e no corpo do texto. No começo dos anos 90, percebeu-se que os métodos utilizados anteriormente não conseguiam dar conta de problemas linguísticos mais complexos, como resolução de

correferência (Capítulo 12). Assim, métodos linguisticamente motivados começaram a ser implementados em sistemas de SA.

Já nas duas primeiras décadas dos anos 2000, notou-se uma tendência sócio-comportamental que também influenciou a SA: a democratização da Web. Por conta disso, nesse período, a sumarização se popularizou devido ao fato de haver um grande volume de informação produzida e circulante em relação ao pouco tempo que as pessoas tinham disponível para consumí-la. Mais recentemente, vemos a influência dos *Large Language Models* (LLMs) (Capítulo 15) em PLN. Com relação à SA, a implementação desse tipo de abordagem apontou que é possível produzir sumários mais coerentes, coesos, mesmo em contextos que exigem manobras de reescrita dos textos.

Em PLN há diferentes motivações que fomentam interseções com a SA. Diversas aplicações e subáreas, como compreensão de textos, recuperação da informação, indexação, mineração de textos e sistemas de pergunta-resposta, partem de texto(s)-fonte que necessitam ter seus conteúdos classificados entre mais e menos importante e, a partir disso, propor uma nova versão do texto, ou em um tamanho menor ou um outro conteúdo. Além disso, os sumários também podem beneficiar pessoas que precisam ler biografias ou coleções de documentos em que seja necessário ter acesso à informação de maneira resumida.

Ao longo deste capítulo queremos introduzir conceitos importantes para a SA, bem como propor um material que você possa utilizar na construção de seus primeiros sumários. Você irá perceber que em alguns momentos remontamos a história da área, pois entender as limitações que sistemas e métodos apresentaram ao longo da trajetória da SA nos permite entender quais os desafios atuais impostos ora pelas tecnologias desenvolvidas, ora pela própria língua em relação aos sistemas computacionais. Ainda, apresentamos exemplos de algoritmos introdutórios e simples para que você possa praticar a sumarização e, quem sabe, poder aplicá-la em outros cenários.

22.2 Por trás do processo

22.2.1 Sobre conceitos

Antes de discutir métodos e sistemas de geração e avaliação da SA, é necessário conhecer alguns conceitos importantes. Para tanto, destacamos as seguintes características da SA: (i) função, (ii) público-alvo, (iii) tipo, (iv) abordagem, (v) fonte, e (vi) idioma.

A versão sumarizada do texto-fonte é tida como sumário, que pode ser classificado como informativo, indicativo ou crítico, a depender da **função** comunicativa que exerce (Mani; Maybury, 1999). Os sumários informativos são aqueles que apresentam a informação original de maneira a preservá-la, a ponto de dispensar a leitura do texto-fonte. Os sumários indicativos são orientações genéricas sobre o texto original, fazendo com que o leitor tenha um panorama sobre o conteúdo, sem ter acesso a seus detalhes, como o índice de um livro. Por fim, os sumários críticos são aqueles que permeiam entre a síntese e a avaliação do conteúdo, aproximando-se do gênero resenha.

Quando pensamos no **público-alvo**, os sumários podem ser caracterizados em genéricos ou específicos. Sumários genéricos são aqueles que extraem a informação do texto-fonte levando em consideração apenas critérios técnicos, como importância do conteúdo. Já os sumários específicos, além de considerarem aspectos técnicos também devem levar em conta elementos relacionados ao leitor que afetam a construção do sumário, como o conhecimento



prévio ou o interesse do público-alvo sobre o assunto do texto-fonte.

No que se refere ao **tipo**, os sumários podem ser classificados em extrativos e abstrativos (do inglês *abstracts*). Os sumários extrativos têm a característica de serem compostos por sentenças dos textos-fonte. Nesse sentido, o conteúdo não é submetido a nenhum tipo de edição linguística, mas apenas selecionado a partir de determinados critérios (como tamanho disponível do sumário e a importância do conteúdo, por exemplo). Já os sumários abstrativos são caracterizados por seu conteúdo (parcial ou integral) ter sido submetido a alguma operação linguística de reescrita, como a paráfrase.

Com relação à **abordagem**, em PLN há três grandes vertentes, organizadas em função da quantidade e profundidade de informação linguística utilizada no processo de sumarização. A abordagem profunda utiliza muito conhecimento linguístico na elaboração dos sumários. Nesse caso, podem ser levados em conta aspectos pertinentes à semântica, ao discurso e à pragmática, por exemplo, para tratar determinados fenômenos linguísticos. Já a abordagem superficial utiliza pouco conhecimento linguístico; quando o faz, utiliza majoritariamente informações dos níveis morfológico, morfossintático e sintático. Esse tipo de abordagem se baseia em conhecimento estatístico e empírico, como abordaremos mais detidamente na próxima seção. Por fim, a abordagem híbrida seria a combinação das duas abordagens anteriores.

É importante destacar que há possibilidade de elaborar diferentes formatos de sumários a partir da abordagem escolhida. Os sumários produzidos sob a abordagem superficial tendem a ser extrativos, já que utilizam, na maioria das vezes, técnicas de contagem de *tokens* e/ou identificação de informações que estão na superfície textual. Ao passo que abordagens profundas e híbridas tendem a produzir sumários abstrativos, pois concebem estratégias linguísticas de retextualização que podem resultar em reelaboração da forma com que o conteúdo está expresso no sumário.

Quanto à **fonte**, os sumários podem ser classificados em mono ou multidocumento. Conforme mencionado no início deste capítulo, os sumários monodocumento são resultantes de apenas um único texto que serviu de base para a sumarização. No caso dos sumários elaborados a partir de dois ou mais textos que dissertam sobre o mesmo assunto, tem-se a sumarização multidocumento. Este último tipo de SA, em especial, precisa lidar com fenômenos linguísticos que tornam o processo mais desafiador, como “redundância”, “complementaridade” e “contradição” provenientes de fontes produzidas com estilos, vieses e perspectivas autorais diferentes durante a escrita dos textos¹.

Quanto ao **idioma**, os sumários podem ser produzidos a partir de uma única língua (monolíngue) ou com duas ou mais línguas (multilíngue).

Outro conceito importante na SA é a **taxa de compressão** dos sumários, isto é, quanto de informação será incluída no sumário. Essa quantidade é estabelecida em função da taxa de compressão, que é a razão entre o tamanho do sumário e o tamanho do texto-fonte (Mani, 2001). Um sumário com taxa de compressão de 70% apresenta tamanho equivalente a 30% do tamanho do texto-fonte, medido ou em função da quantidade de palavras ou de sentenças. Assim, se o texto-fonte possuir 1000 palavras e deseja-se construir um sumário com uma taxa de compressão de 0,7, o sumário terá 300 palavras. No caso da SA multidocumento, a taxa de compressão pode ser definida em relação ao maior texto ou à soma total de palavras ou de sentenças de todos os textos. Porém, em um cenário de

¹Sugerimos a leitura de (Souza; Felippo; Pardo, 2011), (Souza; Felippo, 2018), (Souza, 2019) e (Silva; Di Felippo, 2014) que se aprofundam nos referidos fenômenos linguísticos.



sumarização abstrativa, é preciso ponderar que nem sempre os valores estabelecidos nas taxas de compressão farão segmentações de sentenças de maneira devida. Nesse sentido, com relação à informatividade, é preciso avaliar o conteúdo que já foi escolhido para compor o sumário, caso a taxa de compressão pretendida exceda a quantidade de palavras estimada para o sumário.

22.2.2 Sobre o processo de sumarização

A seguir, apresentamos uma possível análise de um sumário retirado do *corpus* CSTNews² (Cardoso et al., 2011). Os textos do Quadro 22.1 serviram de fonte de informação para a elaboração de sumários extrativos multidocumento do Quadro 22.2.

Quadro 22.1: Textos-fonte extraídos do *corpus* CSTNews

Texto-fonte 1	Texto-fonte 2
O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador –Cahe descartou pancreatite ou úlcera. “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”. Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.	BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. “Agora está estável. Mesmo com esta melhora, ele continuará internado”, disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão). Cahe reforçou que Maradona ainda tem problemas. “Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave”, afirma, em entrevista ao diário La Nación. No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão. Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino. Sua filha, Dalma, foi ao estádio assistir ao jogo.

Os textos resumidos apresentam uma taxa de compressão de 70% do total de palavras do conjunto de textos-fonte. Assim, o humano deveria selecionar sentenças inteiras para compor

²O *corpus* CSTNews é um repositório de textos jornalísticos publicados no ano de 2007, originalmente escritos em português. Ele está organizado em 50 conjuntos de textos, em que cada conjunto possui de dois a três textos sobre um mesmo assunto. Cada conjunto de textos inclui anotações linguísticas (como morfosintática, semântica e discursiva, por exemplo), além de apresentar sumários humanos e automáticos. Esse repositório serviu de base para muitas pesquisas em SA em língua portuguesa, e está disponível em: <http://nilc.icmc.usp.br/CSTNews/>.



seu sumário; caso a próxima sentença ultrapassasse muito o tamanho pré-estabelecido, permitia-se ter um sumário com menos palavras³.

Quadro 22.2: Sumários extraídos do *corpus* CSTNews

Sumário humano multidocumento	Sumário automático multidocumento
BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. <Texto-fonte 2> Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”. <Texto-fonte 1> Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>	Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”. <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>

Os textos-fonte do Quadro 22.1 noticiam o estado de saúde do ex-jogador de futebol Maradona e sobre seu processo de internação para tratar de um quadro de hepatite aguda. Apesar de serem textos jornalísticos, percebe-se que eles apresentam estilos de escrita diferentes, como o uso mais extensivo do discurso direto; ou ainda a construção narrativa dando ora mais detalhes, ora menos.

Os textos dos sumários extrativos também apresentam particularidades entre os sumarizadores: o humano, das 5 sentenças que ele escolhe para elaborar o sumário, 3 foram extraídas do Texto-fonte 1, admitindo que o sumarizador humano pode ter preferência por alguma fonte que julgue ser de sua confiança; ao passo que o sistema computacional, das 3 sentenças escolhidas, apenas 1 advém do Texto-fonte 2. Diante disso, é possível inferir que o primeiro texto-fonte apresenta o conteúdo de maneira a ganhar a predileção de ambos os sumarizadores, seja pelo estilo, composição ou qualidade da informação.

Apesar de tanto o sumarizador humano quanto o automático terem escolhido sentenças em comum, algo a se destacar é o fato de os textos sumarizados apresentarem estruturas narrativas discretamente diferentes. No sumário automático, a sentença inicial é apenas a segunda do texto sumarizado por um humano. Assim, é possível inferir que na sumarização manual elegeu-se uma sentença para abrir o texto que pudesse introduzir melhor o assunto, antes de chegar ao cerne da informação.

Também é possível identificar que os textos sumarizados, em relação aos textos-fonte, mantêm uma ordenação da informação: o início do texto apresenta informações mais importantes em relação ao seu desfecho. Nesse caso, os sumarizadores precisavam identificar

³Para saber mais detalhes sobre a criação de mais sumários humanos para o *corpus* CSTNews, consulte (Dias et al., 2014).



as informações mais relevantes dos textos-fonte e, em seguida, refletir como esse conteúdo se organizaria temporalmente no próprio sumário.

A identificação de conteúdo relevante não é uma tarefa trivial. Desde o advento da SA, um dos maiores desafios para as pesquisas foi desenvolver maneiras de identificar a informação tida como a mais relevante no(s) texto(s)-fonte. Conforme se superavam alguns desafios, outros foram identificados e passaram a pautar muitos estudos, como a sumarização multilíngue e abstrativa. Nos exemplos que trabalharemos nas seções seguintes, detalharemos uma das maneiras disponíveis para seleção automática de conteúdo em textos-fonte.

22.3 Avaliação na SA

Usualmente, a avaliação de desempenho dos sistemas de SA é medida por meio da análise de seus sumários, levando em consideração os critérios de **informatividade** e **qualidade** dos textos produzidos. A informatividade geralmente é calculada de forma automática e consiste em verificar quanto da informação relevante dos textos-fonte é preservada no sumário automático. A avaliação da qualidade, por sua vez, é realizada por humanos, pois o foco reside na análise de aspectos relacionados à gramaticalidade, coesão e coerência, foco e clareza referencial, para os quais ainda não há uma forma de medir automaticamente⁴.

Para medir a informatividade, a métrica de avaliação amplamente utilizada é a ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Essa medida compara automaticamente a quantidade de n-gramas (conjunto de palavras em sequência) em comum entre um sumário automático e um ou mais sumários de referência, ou seja, produzido por um humano. Porém, em cenários em que não houver sumários de referência, o texto-fonte pode ser adotado como critério para avaliação (Louis; Nenkova, 2013). A ROUGE possui quatro variantes principais, descritas a seguir:

- a) ROUGE-N (*ROUGE-Ngram*): Avalia a sobreposição de n-gramas entre o sumário automático e o sumário de referência. A ROUGE-1, por exemplo, considera a sobreposição de unigramas (palavras individuais), enquanto o ROUGE-2 considera a sobreposição de bigramas (pares de palavras consecutivas), e assim por diante.
- b) ROUGE-L (*ROUGE-Longest Common Subsequence*): Mede a similaridade baseada na sequência de palavras comuns mais longa entre o sumário automático e o texto de referência. A ROUGE-L não exige correspondência consecutiva, mas sim a mesma ordem das palavras tanto no sumário automático quanto no sumário de referência.
- c) ROUGE-W (*ROUGE-Word*): Calcula as correspondências consecutivas mais longas entre o sumário automático e o sumário de referência.
- d) ROUGE-S (*ROUGE-Skip-gram*): Essa métrica considera pares de palavras com um certo número de palavras intermediárias “puladas”. Ela é útil para capturar dependências de longa distância entre palavras.

O resultado ROUGE é dado em termos de precisão (Equação 22.1), revocação (Equação 22.2) e medida-f (Equação 22.3), e possui correlação com a avaliação humana

⁴Veja mais sobre Avaliação de sistemas de PLN no Capítulo 14.



(Lin, 2004). A precisão (P) expressa a proporção de n-gramas coincidentes entre os sumários automático e de referência em relação ao número de n-gramas do sumário automático. A revocação (R) representa a proporção de n-gramas coincidentes entre os sumários automático e de referência em relação ao número de n-gramas do sumário de referência. Tais medidas são complementares e, por isso, costuma-se utilizar a medida-f (F) que representa a média harmônica entre precisão e revocação. Como precisão e revocação são inversamente relacionadas, uma tende a diminuir quando a outra sofre um aumento.

$$\text{Precisão} = \frac{\text{Número de n-gramas em comum com o sumário automático}}{\text{Número de n-gramas do sumário automático}} \quad (22.1)$$

$$\text{Revocação} = \frac{\text{Número de n-gramas em comum com o sumário automático}}{\text{Número de n-gramas do sumário referência}} \quad (22.2)$$

$$\text{Medida f} = \frac{2 * P * C}{P + C} \quad (22.3)$$

Para exemplificar, considere o seguinte fragmento como texto-fonte “o gato estava embaixo da cama” e a sentença candidata “o gato foi encontrado embaixo da cama”. Sem considerar os procedimentos de pré-processamento, as palavras em comum das duas sentenças são [o, gato, embaixo, da, cama]; então, os valores para revocação, precisão e medida-f considerando a Rouge-1 são respectivamente: 0,83, 0,71 e 0,76.

Por ser rápida, barata e não sujeita à subjetividade, a ROUGE é uma das medidas mais populares para avaliar sumários (p.ex. (Cardoso, 2014), (Gambhir; Gupta, 2017), (Aliguliyev et al., 2019), (Parida; Motlicek, 2019)). A correlação dessa medida com o julgamento humano aumenta quando se utiliza vários sumários de referência. No entanto, a ROUGE aborda somente a capacidade de seleção de conteúdo dos sistemas, ignorando aspectos de qualidade linguística como coerência e gramaticalidade.

Nesse sentido, em 2005, a DUC (*Document Understanding Conference*)⁵ introduziu cinco propriedades linguísticas para avaliar a qualidade dos sumários automáticos e que não utiliza sumários de referência, a saber⁶:

- (i) **gramaticalidade:** que diz respeito à ausência de erros de ortografia, pontuação e sintaxe;
- (ii) **não redundância:** que se refere à ausência de informações repetidas;
- (iii) **clareza referencial:** que diz respeito à clara identificação dos componentes da superfície textual que fazem remissão a outro(s) elemento(s) do sumário;
- (iv) **foco:** se refere ao fato de que as informações de uma sentença devem se relacionar com as informações do restante do sumário;
- (v) **estrutura e coerência:** que diz respeito à organização do sumário considerando sua textualidade.

⁵Em 2008, a DUC passou a se chamar TAC (*Text Analysis Conference*). Disponível em: <https://tac.nist.gov/>.

⁶Original: <https://www-nlpir.nist.gov/projects/duc/duc2006/quality-questions.txt>.



Para avaliar de acordo com os critérios estabelecidos pela DUC, coleta-se a opinião de um grupo de juízes sobre um mesmo sumário e calcula-se a média para cada critério julgado. Cada anotador atribui uma nota que varia de 1 (muito ruim) a 5 (muito bom). Vale destacar que, embora esse tipo de avaliação não dependa do uso de sumário de referência, ela pode beneficiar sumários automáticos que sejam bastante diferentes dos sumários de referência. Provavelmente, esses sumários automáticos teriam notas muito baixas pela ROUGE, mas, ainda assim, poderiam ser considerados informativos e coerentes. Por essa razão, alguns trabalhos de SA utilizaram tais medidas avaliativas, além da ROUGE (Ermakova; Cossu; Mothe, 2019; Pitler; Louis; Nenkova, 2010).

Outro método que também pode ser utilizado para avaliação é o da pirâmide (Nenkova; Passonneau, 2004). Tal método necessita de um conjunto de sumários de referência, dos quais se extraem manualmente “unidades de conteúdo do sumário” (SCUs - *Summary Content Units*) que são usadas na avaliação dos sumários automáticos. A intuição é que SCUs mencionadas somente em um sumário humano têm menor relevância comparadas àquelas citadas em múltiplos resumos humanos. Cada SCU recebe um peso que é igual ao número de sumários de referência nela contidos.

A pirâmide é formada após a anotação de SCUs: no topo, ficam as unidades que aparecem em mais de um sumário, portanto, são as SCUs com melhores notas; na base, ficam as unidades que aparecem em poucos sumários, portanto, possuem notas mais baixas. O sumário automático ideal deve conter o subconjunto de SCUs das posições mais altas da pirâmide. Apesar de a construção da pirâmide ser muito trabalhosa, essa é uma forma de avaliar que incorpora a variedade do julgamento humano.

Conforme Luhn (1958), a produção de sumários é uma tarefa intelectual e que sofre influência da familiaridade com o assunto, atitude e disposição do produtor. O autor também sugere que a produção de sumários de referência pode depender dos interesses de quem o produz, dos interesses dos leitores e da importância subjetiva que ele atribui às informações textuais. Assim, algumas ponderações podem ser feitas a respeito disso: os humanos divergem na escolha de informações importantes; e existe a possibilidade de que a mesma pessoa, ao resumir novamente um texto-fonte, crie um sumário totalmente diferente do anterior (Hailu; Yu; Fantaye, 2020; Luhn, 1958).

Além disso, no cenário atual em que a grande área de geração de texto em linguagem natural evolui rapidamente, críticas surgem a medidas de avaliação que dependem estritamente da sobreposição de unidades de n-gramas entre sumários de referência e candidatos, o que não é adequado para medir a qualidade de sumários abstrativos. Assim, têm surgido frentes de trabalhos que buscam criar protocolos de avaliação desde as métricas aplicadas até a qualidade dos conjuntos de dados utilizados.

Alguns autores apontam que a maioria das métricas baseadas na sobreposição de vocabulário não são adequadas para avaliação com base na comparação com um texto completo. Diversos conjuntos de dados de sumarização foram lançados nos últimos anos com pouca discussão sobre sua qualidade, indicando ser necessário buscar formas automáticas para avaliação de métricas que não requer nenhuma anotação humana⁷.

⁷Sobre essas novas frentes, indicamos a leitura de (Ermakova; Cossu; Mothe, 2019), (Bommasani; Cardie, 2020), (Hailu; Yu; Fantaye, 2020), (Zhu; Bhat, 2020) e (Fabbri et al., 2021).



22.4 Aplicação

De maneira geral, os sumários são produzidos a partir de três etapas⁸, conforme a Figura 22.2.

Figura 22.2: Etapas do processo de sumarização



Fonte: Adaptado de (Sparck-Jones, 1998).

Na primeira etapa, conhecida como **análise**, os textos-fonte passam por um processo de interpretação, resultando em uma representação formal deles. Na fase seguinte, tida como **transformação**, o sumário é elaborado a partir do ranqueamento de segmentos dos textos-fonte em função de algum critério que indique a relevância, sendo selecionados os segmentos com maiores pontuações até que se atinja a taxa de compressão desejada. Na etapa de **síntese**, gera-se, então, o sumário em língua natural a partir do conteúdo selecionado. Nessa etapa podem ser utilizados métodos de tratamento de correferência, fusão, linearização, justaposição e ordenação de sentenças. Essas três fases contribuem umas com as outras, de modo que alguns métodos que, ocorrem na síntese também poderiam estar na fase de transformação, e vice-versa.

A partir dessa arquitetura genérica, apresentamos nesta seção dois algoritmos com estratégias distintas de seleção de conteúdo para gerar os sumários. É possível que para compreender algumas questões que levantaremos aqui sejam necessárias informações prévias. Pensando nisso, elaboramos o Apêndice D, em que discutimos aspectos relevantes ao pré-processamento dos textos para SA e que complementa o assunto abordado nesta seção. Ademais, para complementar a leitura, disponibilizamos um notebook⁹ virtual com os algoritmos¹⁰ apresentados neste capítulo. Sugerimos que a leitura desta seção seja acompanhada desse material suplementar.

22.4.1 Algoritmo baseado na frequência das palavras

Uma das unidades mínimas de análise linguística é a palavra, compreendida dentro do nível morfológico; quando combinadas entre si, é possível construir outros objetos de análise, como a sentença. Para a sumarização de textos escritos, as palavras e as sentenças podem

⁸Para mais detalhes sobre as três etapas dessa arquitetura genérica, indicamos (Jones, 1993) e (Mani, 2001).

⁹https://colab.research.google.com/drive/1qGfpxQJgnJy_UKKiHpKQSrACiFRwjyeP?usp=sharing Os resultados podem ser um pouco diferentes dos que são apresentados aqui, devido aos testes realizados durante a escrita do texto.

¹⁰*Grosso modo*, algoritmo pode ser definido como um conjunto de regras, compreendendo uma sequência finita de ações visando a resolução de alguma tarefa/problema.



ser o início do processo de identificação de conteúdo relevante. Os métodos clássicos de identificação de segmentos relevantes avaliam a importância de cada sentença em um texto-fonte com base em seu peso e, em seguida, selecionam aquelas com pesos mais elevados (acima de um limite mínimo) para formar o resumo (Baxendale, 1958; Edmundson, 1969; Luhn, 1958). As propostas mais recentes utilizam *word embeddings* (Hailu; Yu; Fantaye, 2020) e modelos de redes neurais profundas, como BERT (Liu; Lapata, 2019).

Nesta seção, apresentamos uma estratégia simples baseada na frequência de palavras para selecionar sentenças relevantes do texto-fonte para compor os sumários. Nessa perspectiva, quanto mais uma palavra se repete no texto, mais relevante ela é. Propusemos um algoritmo geral, organizado em sete passos, que seleciona as sentenças com maior número de palavras relevantes para compor um possível sumário. Vamos aplicar o algoritmo ao Texto-fonte 1 do Quadro 22.1 com o objetivo de determinar um *ranking* de sentenças para a composição do sumário.

Quadro 22.3: Algoritmo geral

Pré-processamento

Passo 1: Normalização do texto

Passo 2: Remoção de caracteres de pontuação

Passo 3: Remoção de *stopwords*

Geração do sumário

Passo 4: Cálculo das frequências das palavras que compõe o texto pré-processado

Passo 5: Determinação do peso de cada palavra

Passo 6: Cálculo do peso de cada sentença

Passo 7: Seleção das sentenças com maiores pesos até atingir a taxa de compressão

Ao executar o **Passo 1**, todo o texto será transformado para letras minúsculas para fins de padronização. Esse procedimento faz com que tenhamos palavras graficamente distintas a serem consideradas de maneira similar, como “hepatite” e “Hepatite”.

Após a normalização, no **Passo 2**, devem ser removidos os caracteres de pontuação. Essa operação se justifica pelo fato de que uma palavra, acompanhada de uma pontuação, difere-se de outra com a mesma grafia sem pontuação (como “cahe” e “cahe,”). Os caracteres de pontuação podem divergir de um método para outro, a partir de diferentes linguagens de programação ou *software*. Aqui, consideramos a pontuação sendo o ponto final, a vírgula, a sequência de dois hífen (--) e as aspas (” ou “).

Já no **Passo 3**, removemos as *stopwords*. Nesta aplicação, utilizamos a *stoplist* do *Natural Language Toolkit* – NLTK – (Bird, 2006). As palavras removidas foram aquelas que estão compreendidas na *stoplist*, saindo de “o médico pessoal do argentino diego”, por exemplo, e chegando em “médico pessoal argentino diego”, tendo sido removidos os *tokens* “o” e “do”.

É importante destacar que a remoção das *stopwords* se baseia na hipótese da distribuição (Capítulo 10). Assim, na SA com abordagem lexical, acredita-se que após a remoção das *stopwords*, as palavras restantes representam melhor o conteúdo e aquelas que mais se repetem, são relevantes. Assim, no **Passo 4**, deve-se calcular a frequência de ocorrência dos *tokens*. Os cinco termos mais frequentes, em ordem alfabética do Texto-fonte 1 estão listados na Tabela 22.1.



Tabela 22.1: Cinco termos mais frequentes do Texto-fonte 1

Termo	Frequência	Peso
maradona	5	$5/5 = 1,0$
cahe	4	$4/5 = 0,8$
hepatite	3	$3/5 = 0,6$
internado	3	$3/5 = 0,6$
recaída	3	$3/5 = 0,6$

A partir da frequência das palavras, no **Passo 5** conseguimos determinar diferentes pesos para os *tokens* com base na frequência relativa. O cálculo é definido pela divisão da frequência de cada termo pela maior frequência calculada. No caso, o termo “maradona” é o que mais ocorre no texto, com frequência 5. Assim, todos os termos do texto devem ser divididos por 5, determinando o peso de cada um deles. Ressalta-se que essa não é a única forma de ponderar a relevância de uma palavra ou sentença em um texto. No exemplo, para fins de simplificação, optamos por considerar a fala do médico como uma única sentença por se tratar de um discurso direto compreendido entre aspas.

Uma vez calculados os pesos de cada uma das palavras, podemos estimar a pontuação das sentenças. Considerando o Texto-fonte 1, no **Passo 6**, definimos a pontuação por meio da soma dos pesos da frequência relativa de cada palavra restante em cada sentença. Na Tabela 22.2, apresenta-se o peso final de cada sentença. Por exemplo, a Sentença 1 tem o peso 5,8, pois a soma dos pesos dos *tokens* (destacados no texto) resulta nesse valor.

Tabela 22.2: Sentenças do texto-fonte com seus respectivos pesos

Sentença	Marcações	Peso
1 O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da Hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.	O (médico - 0,2) (pessoal - 0,2) do (argentino - 0,2) (Diego - 0,2) (Maradona - 1,0), (Alfredo - 0,2) (Cahe - 0,8), (revelou - 0,2) (nesta - 0,2) (segunda-feira - 0,2) que uma (recaída - 0,6) da (Hepatite - 0,6) (aguda - 0,4) de que sofre fo (motivo - 0,2) da (nova - 0,2) (internação - 0,2) do (ex-craque - 0,2).	5,8
2 Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera.	(Maradona - 1,0) (havia - 0,2) (recebido - 0,2) (alta - 0,4) no (último - 0,2) (dia - 0,2) (11 - 0,2), mas (voltou - 0,4) a ser (internado - 0,6) na (sexta-feira - 0,2) e os (boletins - 0,2) (médicos - 0,2) não (especificaram - 0,2) o que se (passava - 0,2) com o (ex-jogador - 0,2) (Cahe - 0,8) (descartou - 0,2) (pancreatite - 0,2) ou (úlcera - 0,2).	6,0
3 “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”.	“(Maradona - 1,0) teve uma (recaída - 0,6) na (hepatite - 0,6) (aguda - 0,4). (Agora - 0,2) está (estável - 0,2). (Apesar - 0,2) de ter (melhorado - 0,2) no (domingo - 0,2), (deverá - 0,2) (continuar - 0,2) (internado - 0,6)”, (dis (Cahe - 0,8), em (declarações - 0,2) ao (jornal - 0,2) “(La - 0,2) (Nación - 0,2)”.	6,6



	Sentença	Marcações	Peso
4	Maradona, 46, desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.	(Maradona - 1,0), (46 - 0,2), (desenvolveu - 0,2) um (hepatite - 0,6) (tóxica - 0,2) por (excesso - 0,2) de (consumo - 0,2) de (álcool - 0,2), o que já o (manteve - 0,2) (internado - 0,6) (durante - 0,2) (13 - 0,2) (dias - 0,2) (antes - 0,2) da (primeira - 0,2) (alta - 0,4).	5,0
5	Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.	(Cahe - 0,8) (disse - 0,4) (ainda - 0,2) que (Maradona - 1,0) não (voltou - 0,4) a (consumir - 0,2) (bebidas - 0,2) (alcoólicas - 0,2) e que as (causas - 0,2) da (recaída - 0,6) estão (sendo - 0,2) (investigadas - 0,2).	4,6

No **Passo 7**, ocorre a seleção de conteúdo relevante. Para fins de exemplificação, a taxa de compressão utilizada foi de 60% em relação ao número de sentenças originais. Apesar disso, destacamos que geralmente a taxa de compressão é calculada em função do número de palavras do texto-fonte. Assim, o sumário poderá ter duas sentenças, conforme se mostra no Quadro 22.4.

Quadro 22.4: Sumário gerado com algoritmo de frequência de palavras

Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera.

“Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”.

A partir desse algoritmo, é importante fazermos algumas considerações:

- **termos com maiores frequências:** a depender do estudo, talvez seja interessante remover os termos que ocorrem em maior frequência, pois podem não determinar exatamente a essência do texto. Suponha que estamos analisando textos com a palavra “covid”. Se o *corpus* for criado a partir desse termo, é quase certo que todos os textos que o compõem devam possuí-lo. O pesquisador deve avaliar se é interessante tê-lo como fator determinante ou inseri-lo em sua *stoplist*;
- **frases curtas versus longas:** avaliando essa comparação, as frases com mais palavras tendem a ter um maior peso, porém podem não ser tão relevantes. Conforme demonstrado na Tabela 22.2, por exemplo, a Frase 1 tem 29 palavras e possui peso 5,8. Já a Frase 2 possui 33 palavras com peso 6 e, a Frase 3 apresenta 27 palavras e o maior peso 6,6. A forma como o algoritmo determina o peso de uma frase deve ser criteriosamente avaliada;
- **stoplist:** a lista de *stopwords* pode interferir no resultado, uma vez que determina quais palavras devem ser removidas, influenciando na frequência e no peso calculado pelo método apresentado. Portanto, os resultados dos métodos de SA podem variar dependendo da lista de *stopwords* utilizada;



- **a ordenação do conteúdo selecionado:** a tarefa de ordenação de sentenças em sumarização multidocumento é difícil quando comparada com a sumarização monodocumento. O conteúdo é extraído de diferentes fontes e, portanto, nenhum documento pode fornecer a ordenação adequada (Barzilay; Elhadad; McKeown, 2001). Como o sumário do Quadro 22.4 é monodocumento, o conteúdo selecionado foi organizado na ordem em que aparece no texto-fonte.

22.4.2 Algoritmo baseado em Transformers

O modelo de *Bidirectional Encoder Representations from Transformers* (BERT) (Capítulo 15) é considerado um dos melhores algoritmos para sumarização da atualidade, apresentando resultados expressivos (Rogers; Kovaleva; Rumshisky, 2021) em diferentes áreas. O algoritmo foi desenvolvido nos laboratórios da Google, a fim de implementar melhorias no algoritmo de busca, permitindo converter as intenções dos usuários (termo de busca) para a codificação adequada contribuindo para otimizar e melhorar a assertividade (LIVESO, 2020). Outra contribuição é a possibilidade do treinamento do algoritmo para atender a um determinado idioma. Algumas empresas direcionaram seus serviços para o treinamento dos idiomas e áreas específicas com demanda, como o português (NEURALMIND, 2020) e o meio empresarial (Julião, 2024). Em suma, o modelo BERT, por meio de técnicas avançadas de PLN, possibilitou uma melhor compreensão da informação textual por parte dos modelos matemáticos, o que por sua vez permitiu melhorias e uma ampla gama de aplicações em diversas áreas.

Diferente de seus antecessores, aqui o algoritmo se baseia em determinar as palavras-chave e variações para atribuir relevância aos trechos de um texto a fim de sumariá-lo. Os antecessores, na maioria, analisavam as palavras individualmente, ou às vezes, considerando os termos ao entorno da palavra-chave requisitada. O BERT mudou essa estratégia, permitindo mapear o texto de forma geral. Dessa forma, o algoritmo possibilita até prever termos a partir de um contexto, ampliando as possibilidades de aplicações.

O processamento dos primeiros algoritmos era fundamentado em processar o texto, independente de o texto estar repetido ou não entre a coleção de documentos, para gerar uma lista das palavras-chave e atribuir pontuação às sentenças. Nesse sentido, até então, o modelo não possuía uma memória, demandando um novo processamento a cada texto a ser analisado. Essa questão foi contornada após a criação do BERT e apresentado por Devlin et al. (2019), sendo implementado por meio da tecnologia *Transformers* (Vaswani et al., 2017a).

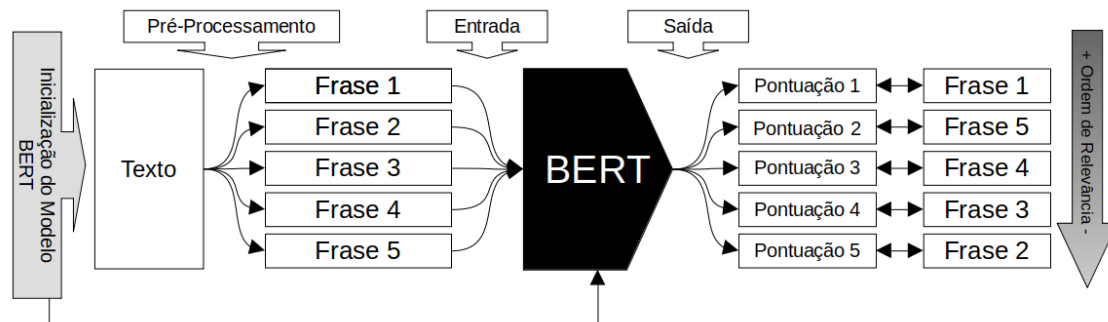
O BERT pode ser aplicado tanto na sumarização abstrativa¹¹ quanto extrativa; porém, aqui, nos deteremos à segunda. O algoritmo é baseado em uma aplicação com treinamento bidirecional. Diferente dos métodos que focam na estrutura composicional das sentenças, que processam as palavras ora da esquerda para direita, ora da direita para esquerda (como os primeiros algoritmos de sumarização, com destaque para o Luhn (1958), Earl (1970) e Edmundson (1969)), no BERT o tratamento ocorre lendo os dados nos dois sentidos durante o treinamento. Nesse sentido, o algoritmo permite que a palavra seja analisada dentro de um “contexto”, levando em consideração o que acontece ao seu redor.

¹¹Quanto à sumarização abstrativa, sugerimos a leitura de (Barros, 2022), (Nascimento, 2023) e (Silva, 2022).



Vaswani et al. (2017a) ressaltam a importância do treinamento bidirecional e demonstram que, uma vez treinado o modelo para o idioma desejado, reduz-se a necessidade de arquiteturas específicas, permitindo uma generalização na solução de desafios, como a escolha do conteúdo mais relevante nos sumários. Portanto, uma vez realizado um pré-treinamento, basta realizar um ajuste fino na rede para atender a casos específicos. No caso da tarefa de SA, o modelo pode ser utilizado para classificar e extrair as sentenças mais relevantes de um texto. Aqui, organizamos o processo em cinco passos, conforme a Figura 22.3.

Figura 22.3: Ilustração do processo simplificado de sumarização do BERT



De acordo com a Figura 22.3, no **Passo 1**: o BERT deve ser inicializado, carregando algumas configurações do modelo já treinado (“Inicialização do Modelo BERT”). No **Passo 2**, o texto a ser fornecido ao BERT passa por um pré-processamento, que compreende segmentação em sentenças e tokenização. O pré-processamento pode incluir outras etapas, como visto no algoritmo apresentado na seção anterior. Após os ajustes iniciais, o texto é submetido ao modelo na camada de codificação (“Entrada”). Ao entrar na rede, os *tokens* passam a compor um vetor numérico e constituem um espaço vetorial¹² para mapear o texto. Já no **Passo 3**, a partir dos vetores no espaço vetorial e do processamento do modelo BERT, em sua camada de decodificação (“Saída”), atribui-se um valor de importância de cada palavra no contexto do texto analisado. No **Passo 4**, com o valor de importância de cada palavra, torna-se possível atribuir um valor (ou “importância”) à sentença que a possui (“Pontuação”). Uma vez definido esse valor para cada sentença, é possível classificá-las em ordem de relevância. Por fim, no **Passo 5**, a partir do nível de compressão informado no início do processamento pelo usuário, o BERT retorna as sentenças para compor o sumário de acordo com a taxa de compressão escolhida. Para exemplificar, são apresentados dois sumários gerados pelo BERT. O Sumário 2 segue a taxa de compressão do algoritmo anterior que é de 60% em relação ao número de sentenças originais e o Sumário 3 coleta uma sentença a mais, isto é, três sentenças.

Caso o sumário produzido não seja condizente com as expectativas, o modelo BERT pode sofrer ajustes finos para aprimorar os resultados. A tecnologia do BERT já fornece um modelo treinado a partir de dois *corpora* volumosos (a saber, *BookCorpus*¹³ e Wikipedia). O treinamento do BERT consiste em fornecer ao algoritmo um *corpus* específico da área

¹²Para um melhor entendimento do espaço vetorial, sugere-se consultar o Capítulo 10.

¹³<https://paperswithcode.com/dataset/bookcorpus>

requisitada, o qual deve ser modificado para se adequar a essa etapa. As adequações previstas são *texto-previsão* - a criação de um conjunto de texto (frases); *previsão da próxima frase*; e *máscara* - modelagem de linguagem mascarada. A adequação texto-previsão consiste em ajustar o texto de entrada e a respectiva frase que deverá ser prevista pelo algoritmo; com isso, parte do treinamento é executado. A segunda parte do treinamento consiste na adaptação máscara em ajustar o texto em *tokens* (usando o WordPiece¹⁴), fazendo a permuta de alguns termos por máscaras (como “####”) e os possíveis valores que podem substituí-las. Após o *corpus* ser ajustado, ele é submetido ao BERT para treinamento e geração do modelo adequado ao caso requisitado. De forma geral, esse ajuste fino ocorre em casos muito específicos, nos quais o texto pode ser de determinados domínios, como jurídico ou médico, nos quais podem ocorrer jargões e termos técnicos que não estão presentes na base geral de treinamento do modelo.

22.4.3 Avaliando os sumários

Até aqui, indicamos de maneira teórico-prática como funcionam os algoritmos de sumarização baseado em frequência de palavras e no modelo BERT. Agora, nosso objetivo é analisar como os sumários produzidos por esses modelos podem diferir entre si a partir de um texto-fonte com cinco sentenças.

No Quadro 22.5 são apresentados os sumários extrativos gerados pelos métodos Frequência de Palavras e BERT. O texto-fonte é o mesmo do Quadro 22.1.

Quadro 22.5: Sumários gerados a partir dos métodos Luhn e BERT

Método	Texto	Qtd. sentenças	Qtd. palavras
Sumário 1 - Frequência de palavras	Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera. “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”.	2	60

¹⁴Wordpieces são partes de palavras, devendo não confundir-se com morfemas, já que não carregam nenhum significado. Wordpieces podem ser obtidas de maneira empírica e constituem um vocabulário induzido a partir de dados do texto. Para saber mais detalhes sobre esse conceito, confira o Capítulo 2 e o Capítulo 4. Neste capítulo utilizamos o pacote WordPiece proposto pela Google, e disponível em: <https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt>.



Sumário 2 - BERT	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona, 46, desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.	2	54
Sumário 3 - BERT - 3 sentenças	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera. Maradona teve uma recaída na hepatite aguda.	3	69

De maneira geral, observa-se que a taxa de compressão pode impactar diretamente a informatividade do sumário. Todos os três sumários gerados apresentam informações necessárias à compreensão do evento (no caso, o problema de saúde de Maradona e a recaída de seu quadro clínico). Apesar disso, destaca-se que, apesar de a informatividade ser um conceito, por vezes, subjetivo, ele está associado ao objetivo do sumário (se ele deve ser informativo, indicativo ou avaliativo, por exemplo).

Conforme já mencionado, é importante planejar a organização do conteúdo selecionado para que o texto final faça sentido e satisfaça a necessidade do usuário. Por exemplo, no Sumário 1, embora a segunda sentença tenha maior pontuação que a primeira, se ela iniciar o texto final, o leitor perderá o contexto de que Cahe é o médico de Maradona.

Num cenário em que comparamos os sumários com o texto-fonte, notamos que todos selecionaram sentenças do início e do meio do conteúdo original. Possivelmente essa escolha se baseou no fato de as sentenças que finalizam textos jornalísticos apresentarem detalhes sobre o tópico principal, enquanto as primeiras sentenças apresentarem sumariamente o assunto.

Além de aspectos linguístico-textuais que devem ser observados nos sumários, como visto, há parâmetros estatísticos que podem ser utilizados durante a avaliação. Para tanto, os três sumários foram avaliados usando a medida ROUGE, mais especificamente ROUGE-1, ROUGE-2 e ROUGE-L. Os valores obtidos para cada sumário são apresentados na Tabela 22.3.

Tabela 22.3: Resultado da avaliação pelo método ROUGE

Algoritmos	Rouge 1			Rouge 2			Rouge 3		
	R	P	F	R	P	F	R	P	F
Frequência de Palavras	0.447	1.00	0.618	0.443	1.00	0.614	0.447	1.00	0.618



Algoritmos	Rouge 1			Rouge 2			Rouge 3		
	R	P	F	R	P	F	R	P	F
Bert - 20%	0.394	1.00	0.566	0.390	1.00	0.561	0.394	1.00	0.566
Bert - 3 sentenças	0.519	1.00	0.683	0.516	1.00	0.681	0.619	1.00	0.683

Uma característica do texto jornalístico é que as informações localizadas no início do texto expressam o fato principal de uma notícia (Canavilhas, 2012), por isso, são selecionadas para compor o sumário. Tal consideração pode ser diferente para outros gêneros textuais. O resultado da métrica ROUGE, na comparação de sumários automáticos extrativos com seu texto-fonte (referência), justifica a precisão ser 1 para todos os casos. Se a comparação fosse em relação a um sumário humano com alguma reescrita, o resultado seria totalmente diferente.

Observa-se que os valores da medida-f não são muito diferentes entre si. Cada um dos sumários traz um contexto que pode ser importante para um tipo de leitor, mas não para outro: o Sumário 1 recupera a opinião do médico e o Sumário 2 apresenta o breve histórico da saúde de Maradona. Ao reduzir a taxa de compressão, o Sumário 3 foi mais abrangente que os demais, resultando em uma melhor avaliação em termos de medida-f.

22.5 A Sumarização automática e a língua portuguesa

Ao analisar as publicações registradas em diferentes repositórios, percebe-se que a SA aplicada ao português brasileiro acompanhou as tendências metodológicas e de aplicações da literatura internacional.

Com relação a alguns *recursos* desenvolvidos, destacamos o TeMário (Pardo; Rino, 2003) por ser um *corpus* de textos jornalísticos acompanhados de seus respectivos sumários, e que inspirou metodologicamente outros *corpora* que foram elaborados para fins de sumarização, como o CSTNews (Cardoso et al., 2011). Este último é tido como padrão ouro por apresentar diferentes camadas de anotação linguística, além de sumários humanos que servem de referência para o processo automatizado. Para apoiar pesquisas de SA de opiniões baseadas em aspectos, há o *corpus* OpiSums-PT¹⁵ (López et al., 2015), que contém vários resumos extrativos e abstrativos de opiniões escritas em português brasileiro, nos quais cada resumo é derivado da análise de 10 opiniões sobre diferentes produtos.

Já com relação a algumas ferramentas, sublinhamos a importância da ferramenta CSTTool (Aleixo; Pardo, 2008c), que utiliza o modelo discursivo *Cross-document Structure Theory* (Radev, 2000)¹⁶ para solucionar desafios linguísticos da sumarização multidocumento. Outra ferramenta tida como precursora por muitos estudiosos da área é o Gistsumm (Balage Filho; Pardo; Nunes, 2007; Pardo, 2002), que gera sumários extrativos a partir da identificação do tópico principal dos textos, utilizando métodos superficiais (como palavras-chave).

Cabe destacar que nos últimos anos, com o avanço de modelos de língua e abordagens mais robustas em PLN, a SA passou a apresentar sumários potencialmente com mais qualidade linguística. Nesse sentido, foram encontrados trabalhos que utilizam LLMs

¹⁵<https://sites.icmc.usp.br/taspardo/sucinto/files/OpiSums-PT.zip>

¹⁶Para saber mais detalhes sobre o modelo discursivo CST, confira o Capítulo 11.



(Barros, 2022; Paiola, 2022), apresentando respostas às questões ora não bem resolvidas por abordagens utilizadas no início da SA. Além disso, observaram-se estudos que aplicam a SA em outros domínios para além do jornalístico, como o jurídico (Feltrin; Vianna; Silva, 2023) e códigos de programação (Pontes; Oliveira; Assis Boldt, 2022). Aplicações baseadas na dependência de língua e/ou domínio seriam bastante custosas do ponto de vista do PLN, o que pode, em partes, ser resolvido com os LLMs.

Muitos outros trabalhos foram e vêm sendo desenvolvidos em SA, já que se trata de uma área ainda em desenvolvimento. Nesse sentido, indicamos que conheçam os trabalhos capitaneados pelo Núcleo Interinstitucional de Linguística Computacional (NILC)¹⁷, que deu suporte a muitas pesquisas¹⁸ de diferentes níveis de formação (graduação e pós-graduação). Tais trabalhos utilizaram distintas abordagens que vigoravam como o estado da arte, como SA mono e multidocumento, mono e multilíngue, ora com abordagem mais linguística, ora com abordagem mais estatística. Outra indicação nossa é o Sistema de Informação e Banco de Dados, da Universidade Federal de Campina Grande (SINBAD), que também abriga pesquisas recentes sobre SA. Por fim, outro grupo de destaque é o Grupo de pesquisa em sistemas inteligentes (GSI), da Universidade Tecnológica Federal do Paraná, que tem feito estudos em SA multidocumento e com modelos computacionais mais recentes, como o BERT.

22.6 Considerações finais

Neste capítulo, apresentamos um esboço sobre uma das áreas que tem evoluído no PLN, a SA. Como visto, o processo de automatização de sumários (extrativos ou abstrativos) não é uma tarefa simples. As dificuldades tendem a ser decorrentes de informações faltantes nos textos. Essa falta de dados exige ter-se algum conhecimento prévio do que é informado, de modo que se possa fazer associação com outros fatos relevantes, o que tende a gerar um melhor entendimento do sumário.

Ainda, detalhamos o funcionamento de dois algoritmos de sumarização com abordagens distintas. O algoritmo baseado em frequência tem como essência a simplicidade de realizar a análise a partir da contagem das palavras e estabelecer a sua relação com o todo (frequência relativa). Para fechar, demonstramos a aplicação de um algoritmo baseado em BERT, que marca uma nova etapa no processamento textual. Baseado em *Transformers*, o BERT como outros algoritmos, utilizam do poder computacional de redes neurais de aprendizado profundo e dos novos hardwares (com destaque ao processamento em placas gráficas) para executar as suas ações no processamento.

De acordo com o panorama que fizemos, chegamos a esta conclusão com o sentimento que muitos desafios encontrados no surgimento da área de SA foram superados. Porém, alguns ainda persistem, como a seleção de conteúdo relevante e personalizado, avaliação do sumário automático sem a necessidade de compará-lo com o sumário produzido por humanos, geração de *abstract* semelhante a um sumário produzido por humano) e quais outros poderão ser superados *na* e *com* a sumarização.

É possível que alguns desses tópicos possam ser trabalhados com abordagens mais recentes

¹⁷<https://sites.google.com/view/nilc-usp/?pli=1>

¹⁸Indicamos acesso à página do projeto Sucinto, que abrigou diferentes pesquisas na área de SA ligadas ao NILC. Disponível em: <https://sites.icmc.usp.br/taspardo/sucinto/index.html>.



em SA, como a utilização de LLMs. Do ponto de vista do marco histórico, consideramos que o *Transformers* deve ser o limiar entre o antes e o depois, quando se tratar de Sumarização, muito em decorrência dos ganhos e da qualidade dos resultados, apesar de ainda não conseguirmos explicar completamente como o conteúdo pode ser recuperado.

Uma das tarefas que não pode ser negligenciada no processo de sumarização é a avaliação do produto gerado. Assim, pontuamos sobre os métodos de avaliação, com destaque para a ROUGE, além de uma abordagem mais linguística segundo a proposta da TAC/DUC. Cabe ressaltar que aqui também residem outros desafios à SA, já que aplicações de sumarização estão sendo implementadas em conteúdo gerado pelo usuário da Web (como avaliações de produtos e publicações em redes sociais). Nesse caso, concepções linguísticas sobre gênero e suporte textual deverão ser levadas em consideração para que os sumários gerados possam ser informativos, coerentes e coesos, para além de uma perspectiva normativista da língua.

Uma das motivações mais utilizadas para as pesquisas iniciais em SA era a falta de tempo que os usuários (especialmente na internet) tinham frente à quantidade de conteúdo que era constantemente disponibilizado. Atualmente, essa motivação continua viável, mas deve ser acolhida no interior da seguinte reflexão: material gerado por “usuários-autores” pode conter ainda mais enviesamento e incorrer em informações distorcidas ou falsas, podendo resultar em desinformação. Nesse caso, vislumbramos que a SA não poderá se afastar de pesquisas que abordem a língua de uma perspectiva funcional-descritiva, mas que também dialoguem com soluções de identificação de conteúdo falso e desinformação.

Diante dessas ponderações, podemos concordar que não esgotamos o tópico de sumarização apenas neste capítulo, especialmente em um contexto em que há uma tendência nas pesquisas em PLN em utilizarem LLMs. Em decorrência da constante evolução da qualidade dos sumários, talvez torne-se necessário a criação de métodos mais apurados que permitam identificar diferenças de estilos de escrita e a identificação de conteúdo relevante. Esse processo de evolução das tecnologias permite imaginar os avanços que podem ocorrer por conta de trabalhos com redes neurais generativas. Fato é que todo esse processo demandará melhorias ou desenvolvimento de novos algoritmos, em especial dos voltados para a língua portuguesa.

Referências

- ALEIXO, P.; PARDO, T. A. S. **Uma Ferramenta Semi-automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST**. [s.l.] Instituto de Ciências Matemáticas e de Computação, 2008.
- ALIGULIYEV, R. M. et al. COSUM: Text summarization based on clustering and optimization. **Expert Systems**, v. 36, n. 1, p. e12340, 2019.
- BALAGE FILHO, P. P.; PARDO, T. A. S.; NUNES, M. DAS G. V. **Sumarização automática de textos científicos: Estudo de caso com o sistema gistsumm**. [s.l.] Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2007.
- BARROS, T. S. **Um modelo BERT para sumarização extrativa de textos em documentos da Polícia Federal**. mathesis—[s.l.] (Mestrado em Ciências da Computação) - Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande, 2022.
- BARZILAY, R.; ELHADAD, N.; MCKEOWN, K. **Sentence ordering in**



- multidocument summarization**. Proceedings of the first international conference on Human language technology research. **Anais...**2001.
- BAXENDALE, P. B. Machine-made index for technical literature—an experiment. **IBM Journal of research and development**, v. 2, n. 4, p. 354–361, 1958.
- BIRD, S. **NLTK: the natural language toolkit**. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. **Anais...**2006.
- BOMMASANI, R.; CARDIE, C. **Intrinsic evaluation of summarization datasets**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais...**2020.
- CANAVILHAS, J. Webjornalismo: Da pirâmide invertida à pirâmide deitada. **Aprender**, n. 32, p. 58–65, 2012.
- CARDOSO, P. C. F. et al. **CSTNews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese**. Proceedings of the 3rd RST Brazilian Meeting. **Anais...**2011.
- CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo**. tese de doutorado—[s.l.] (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014.
- DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. (J. Burstein, C. Doran, T. Solorio, Eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. **Anais...**Minneapolis, MN, USA: Association for Computational Linguistics, 2019. Disponível em: <<https://doi.org/10.18653/v1/n19-1423>>
- DIAS, M. S. et al. **A qualitative analysis of a corpus of opinion summaries based on aspects**. Proceedings of the 1st Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish. **Anais...**2014.
- EARL, L. L. Experiments in automatic extracting and indexing. **Information Storage and Retrieval**, v. 6, n. 4, p. 313–330, 1970.
- EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM (JACM)**, v. 16, n. 2, p. 264–285, 1969.
- ERMAKOVA, L.; COSSU, J. V.; MOTHE, J. A survey on evaluation of summarization methods. **Information processing & management**, v. 56, n. 5, p. 1794–1814, 2019.
- FABBRI, A. R. et al. SummEval: Re-evaluating Summarization Evaluation. **Transactions of the Association for Computational Linguistics**, v. 9, p. 391–409, 2021.
- FELTRIN, G. R.; VIANNA, D.; SILVA, A. DA. **Um Estudo Sobre Métricas de Avaliação para Sumarização de Acórdãos**. Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados. **Anais...**SBC, 2023.
- FIAD, R. S. Reescrita, dialogismo e etnografia. **Linguagem em (Dis) curso**, v. 13, p. 463–480, 2013.
- GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. **Artificial Intelligence Review**, v. 47, p. 1–66, 2017.
- HAILU, T. T.; YU, J.; FANTAYE, T. G. A framework for word embedding based automatic text summarization and evaluation. **Information**, v. 11, n. 2, p. 78, 2020.
- JONES, K. S. What might be in a summary? **Information retrieval**, v. 93, n. 1, p. 9–26, 1993.



- JULIÃO, A. **Algoritmo do Google: Veja o impacto que tem no SEO**. Disponível em: <<https://blog.ajestrategia.com.br/algoritmo-do-google-veja-o-impacto-que-tem-no-seo/>>.
- LIN, C.-Y. **ROUGE: A Package for Automatic Evaluation of Summaries**. Text Summarization Branches Out. **Anais...** Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <<https://aclanthology.org/W04-1013>>
- LIU, Y.; LAPATA, M. Text summarization with pretrained encoders. **arXiv preprint arXiv:1908.08345**, 2019.
- LIVESO. **O que é BERT? - O mais recente algoritmo da Google**. Disponível em: <<https://liveseo.com.br/seo/o-que-e-bert-o-mais-recente-algoritmo-da-google/#:~:text=Bem%2C%20o%20BERT%2C%20de%20maneira,respostas%20possíveis%20para%20seus%20usuários>>.
- LÓPEZ, R. et al. **A qualitative analysis of a corpus of opinion summaries based on aspects**. Proceedings of the 9th Linguistic Annotation Workshop. **Anais...**2015.
- LOUIS, A.; NENKOVA, A. Automatically assessing machine summary content without a gold standard. **Computational Linguistics**, v. 39, n. 2, p. 267–300, 2013.
- LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of research and development**, v. 2, n. 2, p. 159–165, 1958.
- MANI, I. Automatic Summarization. **John Benjamins Publishing Company**, v. 2, p. 399–408, 2001.
- MANI, I.; MAYBURY, M. T. **Advances in automatic text summarization**. [s.l.] MIT press, 1999.
- NASCIMENTO, D. N. C. R. DO. **Sumarização de artigos científicos em português no domínio da saúde**. mathesis—[s.l.] (Mestrado em Informática) - Programa de Pós-Graduação em Informática da PUC-Rio, Rio de Janeiro, 2023.
- NENKOVA, A.; PASSONNEAU, R. J. **Evaluating content selection in summarization: The pyramid method**. Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004. **Anais...**2004.
- NEURALMIND. **NeuralMind disponibiliza modelo BERT do Google em português**. **Neuralmind blog**. Disponível em: <<https://neuralmind.ai/2020/01/26/neuralmind-disponibiliza-modelo-bert-inteligencia-artificial-do-google-em-portugues/>>.
- PAIOLA, P. H. **Sumarização abstrativa de textos em português utilizando aprendizado de máquina**. mathesis—[s.l.] (Mestrado em Ciências da Computação) - Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual Paulista Júlio de Mesquita Filho, 2022.
- PARDO, T. A. S. **Gistsumm: Um sumarizador automático baseado na ideia principal de textos**. [s.l.] Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional, Universidade de São Paulo, 2002.
- PARDO, T. A. S.; RINO, L. H. M. **TeMário: Um corpus para sumarização automática de textos**. [s.l.] Série de Relatórios Técnicos da Universidade de São Carlos, 2003.
- PARIDA, S.; MOTLICEK, P. **Abstract text summarization: A low resource challenge**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais...**2019.
- PITLER, E.; LOUIS, A.; NENKOVA, A. **Automatic evaluation of linguistic quality**



- in multi-document summarization.** Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. **Anais...**2010.
- PONTES, L. B. L.; OLIVEIRA, H. T. A. DE; ASSIS BOLDT, F. DE. **Avaliação de Modelos Neurais para Sumarização de Código-fonte.** Anais do XLIX Seminário Integrado de Software e Hardware. **Anais...**SBC, 2022.
- RADEV, D. R. **A common theory of information fusion from multiple text sources step one: cross-document structure.** 1st SIGdial workshop on Discourse and Dialogue. **Anais...**2000.
- RINO, L. H. M.; PARDO, T. A. S. **A Sumarização Automática de textos: principais características e metodologias.** Anais do XXIII Congresso da Sociedade Brasileira de Computação. **Anais...**2003.
- ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A primer in BERTology: What we know about how BERT works. **Transactions of the Association for Computational Linguistics**, v. 8, p. 842–866, 2021.
- SILVA, J. F. F. **Estratégias para sumarização de documentos.** mathesis—[s.l.] (Mestrado em Engenharia informática) - Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, 2022.
- SILVA, N. L. DA; DI FELIPPO, A. **Descrição e Análise do Fenômeno da Contradição para a Sumarização Automática Multidocumento.** [s.l.] Série de Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional, 2014.
- SOUZA, J. W. DA C. **Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento.** tese de doutorado—[s.l.] (Doutorado em Linguística) - Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, 2019.
- SOUZA, J. W. DA C.; FELIPPO, A. D. Caracterização da complementaridade temporal: subsídios para sumarização automática multidocumento. **Alfa: Revista de Linguística (São José do Rio Preto)**, v. 62, p. 125–150, 2018.
- SOUZA, J. W. DA C.; FELIPPO, A. D.; PARDO, T. A. S. **Investigação da Identificação da Redundância na Sumarização Multidocumento.** Anais do III Student Workshop on Information and Human Language Technology. **Anais...**2011.
- SPARCK-JONES, K. **Automatic Summarizing: Factors and Directions.** In Mani, I. And Maybury, M., editors, **Advances in Automatic Text Summarization.** MIT Press, 1998.
- VASWANI, A. et al. **Attention is All you Need.** (I. Guyon et al., Eds.)Advances in Neural Information Processing Systems. **Anais...**Curran Associates, Inc., 2017. Disponível em: <<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>>
- ZHU, W.; BHAT, S. GRUEN for evaluating linguistic quality of generated text. **arXiv preprint arXiv:2010.02498**, 2020.

