Aplicação de LLMs na Sumarização de Documentos

1st Flávio Marcilio de Oliveira
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG
flavio.marcilio@dcc.ufmg.br

2nd Marcos André Gonçalves
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG
https://orcid.org/0000-0002-2075-3363

***Abstract*—Nos últimos anos, o volume de dados textuais disponíveis digitalmente tem crescido exponencialmente, gerando uma demanda crescente por ferramentas capazes de processar, compreender e sintetizar essas informações de maneira eficiente e precisa. A Sumarização Automática de Textos destaca-se como uma técnica indispensável para facilitar o acesso e a compreensão de grandes quantidades de dados. Este estudo revisa as abordagens mais recentes que empregam Grandes Modelos de Linguagem (LLMs) na sumarização automática de textos, abordando tanto técnicas extrativas quanto abstrativas. A abordagem extrativa seleciona e concatena as sentenças mais importantes do documento, enquanto a abordagem abstrativa gera novas sentenças que transmitem de maneira concisa as informações mais significativas. Ambas as técnicas podem ser aplicadas na sumarização de um único documento ou de múltiplos textos relacionados a um mesmo tema. A pesquisa foi conduzida através de uma revisão bibliográfica sistemática dos trabalhos mais relevantes na área de sumarização automática e LLMs. Foram utilizados critérios de inclusão e exclusão para selecionar artigos publicados em periódicos ou conferências entre 2019 e 2024, além de alguns artigos anteriores devido à sua relevância. Os modelos de sumarização extrativa são amplamente utilizados devido à sua simplicidade e eficiência, mas enfrentam limitações em termos de coerência e consistência factual. Modelos de sumarização abstrativa, por outro lado, geram resumos mais fluentes e gramaticais, mas enfrentam desafios relacionados à utilização do contexto completo e à "maldição do meio". Abordagens híbridas, que combinam técnicas extrativas e abstrativas, mostram-se promissoras para superar essas limitações. A sumarização automática de textos com LLMs possui grande relevância para o avanço de tecnologias de NLP e para áreas práticas que demandam a análise e interpretação de grandes volumes de dados textuais. Este estudo contribui para o entendimento do papel dos LLMs na sumarização automática, promovendo o desenvolvimento de sistemas mais robustos e eficientes. Futuras pesquisas devem focar na melhoria da coerência, precisão factual e relevância dos resumos gerados.**

***Index Terms*—LLM; Sumarização Automática de Textos; Sumarização Extrativa; Sumarização Abstrativa;**

I. Introdução

Nos últimos anos, o volume de dados textuais disponíveis digitalmente tem crescido de forma exponencial, gerando uma demanda crescente por ferramentas capazes de processar, compreender e sintetizar essas informações de maneira eficiente e precisa. Nesse cenário, a Sumarização Automática de Textos destaca-se como uma técnica indispensável para facilitar o acesso e a compreensão de grandes quantidades de dados. Seu objetivo principal é criar uma versão concisa de um texto original, preservando sua essência e os pontos mais relevantes.

A sumarização automática pode ser realizada por meio de duas abordagens principais: **extrativa** e **abstrativa**. Na abordagem extrativa, são selecionadas e concatenadas as sentenças mais importantes do documento, com base em critérios de relevância. Já a abordagem abstrativa gera novas sentenças que não estão presentes no texto original, mas que transmitem de maneira concisa as informações mais significativas. Por essa razão, a abordagem abstrativa é mais complexa e desafiadora. Ambas as técnicas podem ser aplicadas na sumarização de um único documento ou de múltiplos textos relacionados a um mesmo tema.

Nos últimos anos, a Sumarização Automática de Textos tem atraído crescente atenção, com um aumento substancial nas pesquisas publicadas sobre o tema. Esse avanço tem sido impulsionado especialmente pelo progresso no campo do Processamento de Linguagem Natural (NLP, na sigla em inglês), particularmente com o desenvolvimento de Grandes Modelos de Linguagem (LLMs, como BERT, GPT, BART, entre outros modelos baseados em transformadores). Esses modelos tornaram a sumarização automática uma tarefa ainda mais essencial.

Diante desse panorama, esta pesquisa tem como principal objetivo revisar os estudos mais recentes sobre o tema, com foco específico nas abordagens que empregam Grandes Modelos de Linguagem. Apesar do sucesso dos LLMs em diversas tarefas de NLP, desafios persistem na Sumarização Automática de Textos, incluindo a capacidade de garantir coerência, precisão factual e relevância nos resumos gerados. Com esta análise, espera-se contribuir para o entendimento do papel dos LLMs na sumarização automática, promovendo o desenvolvimento de sistemas mais robustos e eficientes nessa área.

Acreditamos que este estudo possui grande relevância não apenas para o avanço de tecnologias de NLP, mas também para áreas práticas que demandam a análise e interpretação de grandes volumes de dados textuais, como jornalismo, direito e ciência da informação. A Sumarização Automática de Textos tem o potencial de transformar a produtividade e a tomada de decisão em diversos setores, ao possibilitar que informações essenciais sejam acessadas de forma mais rápida, clara e eficiente.

Este trabalho está estruturado da seguinte maneira: na **Seção II**, apresentamos os fundamentos teóricos e os conceitos principais relacionados à Sumarização Automática de Textos, com

ênfase na aplicação dos LLMs nesse campo. Na **Seção III**, detalhamos a metodologia de pesquisa adotada, descrevendo os procedimentos e critérios utilizados no estudo. A **Seção IV** aborda a análise e discussão dos resultados obtidos, fornecendo uma visão crítica sobre os achados. Por fim, a **Seção V** encerra o estudo com uma síntese das conclusões, além de apresentar recomendações e direções para futuras pesquisas nessa área em constante evolução.

II. Referencial Teórico

Nesta seção apresentaremos as bases teóricas e os principais conceitos que fundamentam os estudos e pesquisas no campo da Sumarização Automática de Textos com a aplicação de LLMs. Para isso, esta seção será dividida em três partes principais: (a) uma visão geral sobre o desenvolvimento da sumarização automática de textos, (b) uma introdução aos Grandes Modelos de Linguagem e (c) abordagens para avaliação da qualidade dos resumos gerados.

A. Sumarização Automática de Textos

A Sumarização Automática de Textos é o processo de gerar uma versão reduzida de um documento original, preservando as informações mais relevantes. A pesquisa nesse campo teve início na década de 1950, com o trabalho pioneiro de Lunh [19]. Nesta pesquisa, Luhn propôs uma técnica de criação automática de resumos para artigos científicos utilizando informações estatísticas obtidas com a frequência e distribuição de palavras num texto para calcular a significância relativa. O método foi aplicado primeiro às palavras e depois às frases. As frases com maior significância foram extraídas para criar um resumo automático. A partir deste estudo, outros pesquisadores buscaram desenvolver novas técnicas para a sumarização automática de textos tentando sempre meios de aprimorar os resumos gerados. Entretanto, os resumos produzidos careciam de profundidade semântica, coerência e expressividade [10]. Esses desafios, observados nos primeiros modelos de sumarização automática, começaram a ser superados com o advento dos modelos de aprendizado profundo. Esses modelos são capazes de compreender semânticas, produzindo, assim, resumos com maior coerência e expressividade. Neste novo cenário, vários pesquisadores adotaram uma categorização técnica para os métodos utilizados na sumarização: métodos que adotam uma abordagem "extrativa" [23; 34] e métodos que adotam uma abordagem "abstrativa" [8; 17].

Sumarização Extrativa

Este método consiste em selecionar sentenças ou fragmentos do texto original que representam o conteúdo essencial. A sumarização extrativa não modifica o texto selecionado, resultando em resumos que mantêm o estilo e a redação originais. Por sua simplicidade e menor necessidade de processamento semântico, essa abordagem foi amplamente adotada em pesquisas iniciais de sumarização automática [27; 24].

Sumarização Abstrativa

Diferentemente da extrativa, a sumarização abstrativa procura reescrever as ideias centrais do texto em novas palavras e estruturas, criando uma versão mais condensada e que não necessariamente replica sentenças inteiras do documento original. Esse método, mais avançado e semelhante à forma como os humanos sumarizam, exige uma compreensão profunda do texto e foi impulsionado pelo avanço de modelos de linguagem complexos [31].

B. Grandes Modelos de Linguagem - LLMs

Com a evolução dos modelos de linguagem baseados em arquiteturas de redes neurais profundas, especialmente os modelos transformers, como BERT, T5, GPT e seus sucessores [36; 4], os LLMs se destacaram como uma ferramenta eficaz para tarefas complexas de processamento de linguagem natural, incluindo a sumarização automática.

Arquitetura Transformer

A arquitetura transformer revolucionou o campo do NLP ao permitir que modelos aprendessem e aplicassem dependências complexas entre palavras sem a necessidade de processamento sequencial, como nos métodos de redes recorrentes. Esse avanço possibilitou a criação dos grandes modelos de linguagem treinados com bilhões de parâmetros em grandes volumes de dados textuais. Modelos como BERT [4] e GPT [28; 1] utilizam técnicas de pré-treinamento, o que aumenta sua capacidade de compreensão e geração de texto de alta qualidade.

Sumarização com LLMs

Grandes modelos de linguagem podem realizar sumarização através de duas abordagens principais: fine-tuning e few-shot learning. O fine-tuning ajusta o modelo para a tarefa específica de sumarização, enquanto o few-shot learning permite que o modelo produza resumos sem ajuste adicional, utilizando apenas um conjunto mínimo de exemplos. O modelo T5 (Text-To-Text Transfer Transformer) [29] é um exemplo de LLM que oferece flexibilidade na aplicação para múltiplas tarefas de NLP, incluindo sumarização.

C. Avaliação de Resumos

Avaliar a qualidade dos resumos gerados é uma etapa crucial e desafiadora. Tradicionalmente, as métricas de avaliação incluem métodos quantitativos e qualitativos.

Métricas Quantitativas

As métricas ROUGE (Recall-Oriented Understudy for Gisting Evaluation), desenvolvidas por [16], são amplamente utilizadas para avaliar a similaridade entre o resumo gerado e um ou mais resumos de referência. As variantes ROUGE-1, ROUGE-2 e ROUGE-L calculam, respectivamente, a precisão e o recall de n-gramas, bigramas e subsequências. Embora eficazes, essas métricas têm limitações, pois se baseiam apenas em sobreposição lexical e não avaliam a coerência ou a precisão semântica dos resumos [22].

Métricas Qualitativas e Métricas Baseadas em LLMs

Mais recentemente, a aplicação de LLMs como avaliadores automáticos, utilizando métricas baseadas em similaridade semântica e coerência, tem se mostrado promissora. Modelos como BERTScore [38] avaliam similaridade semântica de maneira mais refinada, levando em conta contextos e relações

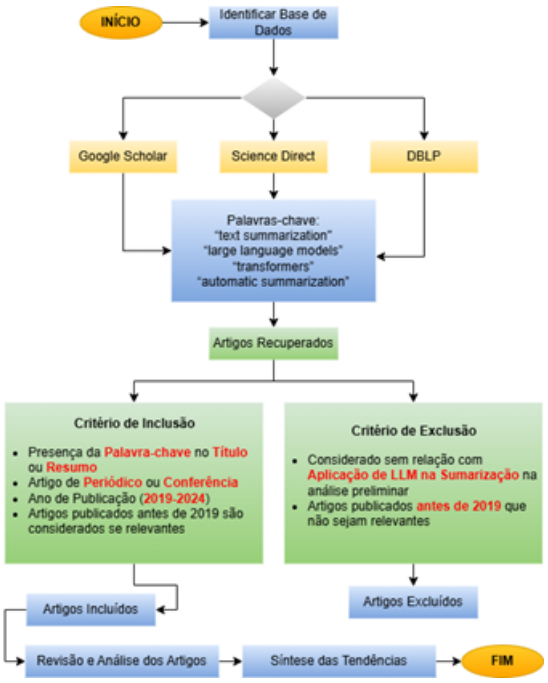

Fig. 1. Metodologia de pesquisa (Fonte: Adaptado de [35])

de palavras. Além disso, abordagens de avaliação humana continuam essenciais para garantir que o modelo não incorra em viés ou problemas de precisão factual.

III. Metodologia de Pesquisa

Este capítulo detalha o processo metodológico adotado para a realização da pesquisa sobre sumarização de textos com o uso de LLMs. A pesquisa, de caráter exploratório e descritivo, será conduzida através de uma revisão bibliográfica sistemática dos trabalhos mais relevantes na área de sumarização automática e LLMs. Conforme ilustrado na Fig. 1, o processo de pesquisa começa com a identificação das bases de dados relevantes para artigos recentes sobre sumarização automática de textos utilizando LLMs. Os artigos foram obtidos de três fontes principais: Google Scholar (scholar.google.com), Science Direct (sciencedirect.com) e a base de dados DBLP (https://dblp.org/). Utilizando várias palavras-chaves como critério de busca para recuperar os artigos relacionados, incluindo ”text summarization”, ”large language models”, ”transformers” e ”automatic summarization”. Definimos critérios de inclusão e exclusão para os artigos recuperados como apresentado na Fig. 1. Selecionamos artigos que continha a palavra-chave no título ou no resumo e que foram publicados em periódicos ou conferências. Priorizamos artigos recentes (2019-2024) que abordam a sumarização automática de textos utilizando LLMs, mas também incluímos artigos publicados antes de 2019 devido a sua relevância para o tema e, também, por fundamentar as pesquisas recentes. Após essa seleção, os artigos foram revisados e analisados. Finalmente, os resultados das analises foram apresentados e discutidos ao longo do estudo.

IV. Análises, Resultados e Discussões

O conteúdo desta seção revela a análise e os resultados da pesquisa, bem como discute pontos importantes destacados dos trabalhos recentes em sumarização extrativa, abstrativa e híbrida, bem como dos métodos de avaliação dos resumos gerados.

A. Sumarização Extrativa

Os modelos de sumarização extrativa, amplamente utilizados devido à sua simplicidade e eficiência, geralmente selecionam frases do texto original para compor o resumo. Esse processo pode ser dividido em três etapas principais: (1) calcular a importância de cada frase, (2) classificar as frases com base em sua importância e (3) selecionar as frases mais bem classificadas para formar o resumo [10]. Esses métodos têm se tornado predominantes no domínio da sumarização de texto e são aplicados em diversos campos. Pesquisas recentes buscam continuamente melhorar a qualidade dos resumos gerados.

Apesar dos avanços, esses modelos enfrentam limitações significativas. O processo de seleção de frases pode levar à escolha de sentenças muito distantes entre si, prejudicando a continuidade do texto. Além disso, relações causais, de transição, progressivas ou de justaposição entre sentenças consecutivas podem ser desarticuladas, comprometendo não apenas a coerência, mas também a consistência factual dos resumos gerados [9]. Em textos longos, resumos com baixa proporção de frases tendem a ser menos representativos. Outro problema recorrente é a inclusão de frases que dependem de contexto adicional para serem compreendidas, como aquelas que contêm pronomes (”isso”, ”aqueles”) ou conectores (”também”), o que afeta negativamente a qualidade dos resumos [21].

No estudo de [21], foi proposto um algoritmo para sumarização extrativa de transcrições de palestras, baseado na geração de embeddings utilizando o modelo BERT em frases tokenizadas. Esses embeddings foram agrupados por meio do algoritmo K-Means, permitindo que frases próximas ao centróide fossem selecionadas como candidatas ao resumo. Para lidar com textos de qualidade variável, o autor combinou diferentes técnicas de tokenização e, para aumentar a representatividade do resumo, incluiu múltiplas frases de clusters próximos ao centróide, adicionando contexto e melhorando a qualidade da saída.

Quanto às palavras sem contexto, soluções como a remoção de frases contendo pronomes ou o uso do NLTK para substituir pronomes por palavras específicas foram exploradas. No entanto, tais abordagens enfrentaram dificuldades devido à necessidade de identificar referências contextuais em frases anteriores, o que comprometeu a qualidade dos resumos.

Buscando superar as limitações dos modelos extrativistas, [37] desenvolveram o **DISCOBERT**, um modelo de sumarização neural consciente do discurso. Esse modelo utiliza dois grafos orientados para capturar dependências de longo alcance entre as unidades discursivas do texto. Os grafos são construídos a partir de árvores da Teoria da Estrutura

Retórica (Rhetorical Structure Theory, RST) e menções de correferência, que são codificadas utilizando Redes Convolucionais de Grafos (RCG).

No DISCOBERT, a Unidade de Discurso Elementar (UDE), uma unidade subfrásica derivada da RST, é adotada como unidade mínima de seleção em vez de frases inteiras. Isso reduz redundâncias e melhora a precisão da sumarização. O grafo RST é construído a partir das árvores de análise RST sobre as UDEs, enquanto o grafo de correferência conecta entidades e clusters de menções ao longo do documento. A navegação pelas correferências permite explorar interações entre conceitos ou eventos relacionados, enriquecendo o contexto do resumo.

Embora a segmentação baseada em UDEs possa impactar a fluidez gramatical de algumas saídas, a árvore de dependência RST garante relações retóricas consistentes entre as unidades. Regras de pós-processamento simples, mas eficazes, ajudam a corrigir eventuais incoerências. Experimentos mostram que o DISCOBERT supera métodos baseados em BERT em benchmarks populares de sumarização, obtendo resultados significativamente superiores.

Entretanto, análises de erros revelaram que problemas na resolução de dependências RST e falhas na análise sintática contribuíram para a perda de coerência e gramaticalidade em algumas saídas. A classificação incorreta das relações retóricas e a dependência de regras manuais ainda representam desafios importantes para o aprimoramento do modelo.

Para abordar esses desafios, [9] propuseram uma nova arquitetura que combina um módulo extrativo e um discriminador de coerência. O discriminador de coerência é treinado online utilizando os vetores de sentenças extraídos do texto de entrada, melhorando sua capacidade de avaliar se as sentenças estão ordenadas de forma coerente. Paralelamente, as pontuações geradas pelo discriminador são otimizadas durante o treinamento do módulo extrativo.

Duas estratégias foram introduzidas para tornar as sentenças extraídas diferenciáveis: uma baseada em modelos de conversão pré-treinados e outra em uma matriz de conversão (MAT), ambas projetadas para mesclar representações de sentenças. O modelo utiliza o BERT como base para a representação textual, aproveitando o vetor do token [CLS] para representar cada frase. Uma camada adicional de transformer foi incorporada acima do vetor [CLS] para atribuir uma pontuação de importância a cada frase.

Para aumentar a robustez do modelo, foi implementado um mecanismo de embaralhamento de frases, criando exemplos incoerentes que servem como referência para o treinamento. Nesse processo, frases que não seguem logicamente as anteriores no texto original são rotuladas como ”incoerentes”, enquanto aquelas que mantêm a continuidade são classificadas como ”coerentes”.

O discriminador de coerência, peça central do modelo, avalia se uma sequência de sentenças é coerente em nível frasal. Ele utiliza camadas de transformer que recebem os vetores de representação gerados pelo BERT ([CLS]) como entrada e produzem pontuações que medem a coerência de cada frase em relação à anterior.

O processo de funcionamento da arquitetura inicia com o processamento, do texto de entrada, pelo BERT, gerando vetores [CLS] para cada frase. Em seguida, esses vetores seguem três caminhos:

- Passam por um codificador transformer para gerar pontuações de importância para cada frase.
- São avaliados pelo discriminador de coerência, que calcula pontuações de coerência para cada frase.
- São combinados com um vetor binário, que determina se uma frase será selecionada ou não, utilizando as técnicas Gumbel-Softmax e Top-K aplicadas às pontuações de importância.

Ainda nesse contexto, [25] apresentam o **DeepExtract**, um modelo que integra técnicas baseadas em semântica para a análise e sumarização de documentos complexos de maneira eficaz. O DeepExtract é construído em torno de uma estrutura hierárquica dinâmica que categoriza frases e seções não apenas com base na sua posição física no texto, mas também em seu significado contextual e temático, utilizando embeddings dinâmicos gerados pelo GPT-4.

O modelo adota um sistema de pontuação multifacetado que avalia as frases com base em critérios como coerência, relevância e novidade. Essa abordagem garante que os resumos gerados sejam não apenas concisos, mas também ricos em conteúdo essencial, refletindo com precisão os temas principais e as discussões detalhadas do documento.

A principal inovação do DeepExtract reside na sua capacidade de construir uma árvore hierárquica que organiza os segmentos textuais tanto por aspectos sintáticos quanto por características semânticas. Essa estrutura permite um processamento mais contextualizado, assegurando que os resumos não apenas capturem a essência do texto, mas também mantenham uma representação fiel da sua estrutura e significado.

O processo de sumarização envolve as seguintes etapas:

1) **Pré-processamento e Segmentação:** O documento é segmentado em componentes hierárquicos, como cabeçalhos, parágrafos e frases, agrupados em seções coerentes.
2) **Construção da Árvore Hierárquica Dinâmica:**
   - Segmentos de texto são analisados estruturalmente para identificar componentes como títulos e subtítulos.
   - A segmentação semântica utiliza os embeddings gerados pelo GPT-4 e algoritmos de agrupamento para categorizar frases.
   - Uma árvore hierárquica dinâmica é construída, onde cada nó representa um segmento ou seção. Relações pai-filho são estabelecidas para capturar a organização hierárquica do documento.
3) **Codificação Posicional Hierárquica:** Cada nó da árvore recebe uma codificação posicional detalhada, que melhora a capacidade do modelo de reconhecer e explorar a estrutura inerente de documentos longos.

4) **Cálculo de Importância Contextual:**
   - A relevância semântica de cada nó é avaliada.
   - As pontuações são ajustadas considerando a função do nó na estrutura do documento.
   - Relações de embedding contextual são usadas para aprimorar as pontuações, enquanto ajustes de novidade e redundância garantem diversidade no conteúdo selecionado.
5) **Seleção e Organização:** Os nós com as maiores pontuações de importância contextual são selecionados para compor o resumo. Esses segmentos são extraídos e organizados para assegurar coerência e legibilidade.

Os resumos gerados são avaliados com métricas padrão, como ROUGE e BLEU, e o modelo é refinado com base no feedback dessas avaliações, aprimorando continuamente a qualidade dos resultados.

No entanto, quando a sumarização envolve documentos longos ou múltiplos documentos, o desafio deixa de ser um problema da técnica empregada e passa a ser inerentemente da ferramenta utilizada, ou seja do LLM. Isso se dá pois os LLMs possuem uma limitação fundamental que é o comprimento fixo de sua janela de contexto, por não possuírem memória fora dessa janela [15].

Na busca por soluções para esse problema, [7] propuseram uma adaptação da arquitetura transformer multicamada, incorporando uma camada hierárquica entre as camadas tradicionais do modelo. Essa nova camada visa propagar informações entre blocos sucessivos, permitindo processar textos longos de forma eficiente. A abordagem consiste em aplicar redes transformadoras independentemente a pequenos blocos de texto e utilizar uma Unidade Recorrente Fechada Bidirecional (BiGRU) para compartilhar informações globais entre esses blocos. Essa estratégia preserva a estrutura do modelo pré-treinado, permitindo a transferência de pesos com a adição de poucos parâmetros. Resultados demonstram que o modelo consegue resumir documentos extensos, mantendo a informatividade (avaliada pelo ROUGE-1) e a fluência (avaliada pelo ROUGE-L) dos resumos gerados. Apesar de ainda apresentar influência da posição das frases no texto, o modelo seleciona frases de todo o documento, aproximando-se da distribuição oráculo.

Para sumarizar múltiplos documentos relacionados ao mesmo tema, [13] adotaram uma abordagem de extração-reescrita que utiliza uma função monótona-submodular orientada ao evento principal para seleção de conteúdo. O objetivo principal é gerar resumos sucintos e informativos, enfatizando o evento central enquanto mantém a objetividade. A metodologia combina três etapas principais:

1) **Extração do Evento Principal:** Baseada na estrutura discursiva de Van Dijk (que considera que todas as unidades linguísticas em um documento compartilham uma relação discursiva com a unidade linguística que representa o evento principal) identifica-se o evento central como a unidade linguística com maior relevância no conjunto de documentos. Métodos propostos por Choubey [3] são usados para rotular e selecionar essas unidades discursivas.
2) **Extração de Contexto:** Utiliza um método de sumarização extrativa enviesado pelo evento principal, garantindo a inclusão de informações essenciais para o contexto.
3) **Reescrita:** O conteúdo extraído é reescrito por um LLM ajustado para gerar um texto coeso e fluente.

Essa abordagem demonstrou excelência em métricas objetivas e avaliações humanas, superando métodos convencionais em cobertura, coerência e informatividade. Trabalhos futuros podem explorar aprimoramentos em cada componente, como a extração de eventos principais e a reescrita.

Os modelos de sumarização extrativa possuem vantagens claras, como a capacidade de capturar terminologias precisas diretamente dos textos originais e a menor dependência de grandes volumes de dados de treinamento, tornando-os rápidos e econômicos. Contudo, apresentam limitações em termos de expressividade, frequentemente gerando resumos redundantes, excessivamente longos ou com contradições contextuais. Embora sejam úteis para determinados fins, ainda divergem da qualidade rica e matizada dos resumos produzidos por especialistas humanos, apontando a necessidade de aprimoramentos para reduzir essas lacunas. Nesse sentido, a abordagem abstrativa na sumarização de texto produz resumos com características mais próximas dos resumos produzidos por especialistas humanos.

B. Sumarização Abstrativa

Grandes modelos de linguagem (LLMs) transformaram drasticamente o cenário da sumarização de textos, especificamente na sumarização abstrativa. Essa abordagem abriu um novo paradigma: os resumos gerados por LLMs se tornaram altamente fluentes, gramaticais e relevantes [6].

Mesmo nesse cenário promissor de avanços e conquistas relevantes no campo da sumarização abstrativa, alguns questionamentos básicos ainda permanecem. A questão inicial envolve a janela de contexto dos LLMs: **Será que eles fazem uso adequado de todo o seu contexto?** Buscando compreender esse comportamento, [18] através de experimentos em tarefas de responder perguntas sobre múltiplos documentos e recuperação do tópico principal, descobriram que os LLMs se concentram principalmente no início e no fim da sua janela de contexto. Esse comportamento gera um padrão de desempenho em forma de U, refletindo dificuldade em lidar com informações localizadas no meio da entrada. Tal viés é especialmente preocupante em tarefas de síntese, onde informações cruciais frequentemente estão distribuídas de forma dispersa ao longo do(s) documento(s).

Na tentativa de compreender essa limitação em tarefas de sumarização, os autores [30] conduziram o primeiro estudo sobre o uso de contexto e o viés de posição quando a abordagem abstrativa é empregada. Os autores examinaram seis LLMs (Flan-UL2, Llama-2 7B, Llama-2 13B, Xgen-7B, Mistral-7B, GPT-3.5), dez conjuntos de dados (cinco com tamanho de entrada padrão e cinco com entrada longa) e cinco métricas

de avaliação (ROUGE-2, BERTScore, A3CU, SummaC, GPT-3.5) e, junto a esses experimentos, elaboraram um conjunto de dados de avaliação (MiddleSum) em que informações importantes estão concentradas no meio do contexto, permitindo quantificar automaticamente esse mesmo comportamento em formato de U.

Os resultados dos experimentos mostraram que os LLMs se concentram no conteúdo no início do(s) documento(s) de origem para tirar suas informações. Em relação à janela de contexto, os resultados revelaram uma fraqueza significativa na sumarização abstrativa: os LLMs enfrentam dificuldades em utilizar informações localizadas no meio da janela de contexto, fenômeno que os autores definiram como ”maldição do meio”. Analisando as fontes dos resumos, concluíram que os LLMs tendem a focar em informações no início ou no final do(s) documento(s), enquanto ignoram amplamente o conteúdo intermediário. Além disso, concluíram que a *maldição do meio* é independente do método de decodificação utilizado.

Buscando avaliar a possibilidade de minimizar a *maldição do meio*, os autores compararam métodos alternativos de inferência no MiddleSum: **sumarização hierárquica** e **sumarização incremental**. Embora a inferência hierárquica e incremental tenha se mostrado eficaz em textos científicos – possivelmente devido à divisão estruturada em seções –, tais abordagens prejudicaram a qualidade dos resumos em outros domínios.

Sobre o dimensionamento do comprimento da janela de contexto, os resultados sugerem que, com a atual estrutura de inferência e avaliação dos LLMs, não há necessidade de exceder 4.000 tokens em modelos de código aberto, uma vez que, se por um lado o tamanho maior permite fornecer mais informações ao modelo contribuindo para um resumo mais rico, por outro lado, o custo computacional sobre um contexto mais longo se torna mais desafiador. Em suma, enquanto técnicas para expandir o comprimento de contexto e métodos de inferência alternativos têm potencial, desafios relacionados ao viés de posição e à utilização de informações no meio da entrada persistem, demandando avanços para aprimorar o desempenho dos LLMs em tarefas de sumarização.

Nessa busca por modelos mais eficientes, os autores [11] propõem uma nova abordagem para sumarização abstrativa, utilizando uma rede única baseada em modelo de linguagem transformer pré-treinada. Nesse método, a mesma rede é responsável tanto por codificar a fonte quanto por gerar o resumo, eliminando a necessidade de componentes separados de codificador e decodificador. Isso assegura que todos os parâmetros da rede, incluindo aqueles relacionados à atenção sobre os estados da fonte, sejam pré-treinados antes do ajuste fino.

Durante o ajuste fino, é utilizada uma arquitetura de rede somente de decodificador baseada em Transformer, tratando a sumarização como uma tarefa de modelagem de linguagem. Em cada exemplo, o resumo é anexado ao artigo original, permitindo que a mesma rede processe e gere representações para ambos. De acordo com os autores, essa abordagem oferece diversas vantagens:

1) **Eliminação de redundância:** evita a duplicação de pesos pré-treinados que ocorreria ao usar componentes separados de codificador e decodificador.
2) **Eficiência de parâmetros:** utiliza menos parâmetros em comparação com modelos tradicionais de codificador-decodificador.
3) **Pré-treinamento completo:** garante que todos os pesos, incluindo os responsáveis pela atenção sobre os estados da fonte, sejam completamente pré-treinados.

Experimentos realizados no conjunto de dados **CNN/Daily Mail** demonstram que o Transformer LM pré-treinado supera significativamente os modelos tradicionais de codificador-decodificador pré-treinados, especialmente em cenários com dados limitados. Essa abordagem reduz a complexidade do modelo, melhora a eficiência computacional e oferece ganhos substanciais na qualidade dos resumos gerados.

Os modelos de resumo abstrato geram resumo de uma forma que se assemelha mais aos humanos pela próxima previsão de token. Em comparação com a abordagem extrativa, essa abordagem é caracterizada pela flexibilidade na expressão e taxa de compressão louvável. No entanto, é imperativo reconhecer os desafios inerentes associados ao desenvolvimento desta abordagem. A complexidade da implementação dessa técnica é notavelmente alta, muitas vezes exigindo conjuntos de dados de qualidade superior para um treinamento eficaz. Além disso, a utilização desse método envolve um consumo substancial de recursos computacionais e tempo de treinamento, exigindo um trade-off entre custo e eficiência.

C. Sumarização Híbrida

Embora o resumo extrativo seja eficaz em preservar as informações relevantes do conteúdo original, ele muitas vezes compromete o fluxo natural e a coerência entre as frases do resumo, tornando-o menos fluido em comparação aos resumos produzidos por humanos. Por outro lado, modelos abstrativos, embora sejam eficazes para documentos curtos, suas limitações de memória dificultam a escalabilidade para documentos longos. Por essas razões, explorar uma abordagem híbrida – que combina a extração inicial de frases relevantes com uma etapa subsequente de sumarização abstrativa – pode ser altamente eficiente. Essa estratégia pode ser entendida como uma forma de ”atenção rígida”, reduzindo o contexto a ser processado na fase abstrativa e otimizando os recursos computacionais [26].

Adotando essa perspectiva, [26] propuseram uma arquitetura híbrida composta por dois componentes treináveis:

1) **Modelo extrativo:** Utilizando um codificador hierárquico para gerar representações de frases, esse componente identifica ou classifica as frases mais relevantes na entrada.
2) **Modelo de linguagem transformador:** Condicionado pelas frases extraídas, bem como por parte ou todo o documento de entrada, este modelo gera o resumo final.

Embora este trabalho represente um avanço na geração de resumos abstrativos mais eficazes, o desafio de desenvolver

modelos que respeitem integralmente os fatos do conteúdo original, ao mesmo tempo em que sintetizam resumos criativos, coerentes e concisos, permanece em aberto. Além disso, os modelos de linguagem usados nesta abordagem possuem tamanho moderado em comparação com as capacidades dos modelos mais avançados disponíveis atualmente. Investigar o desempenho de modelos maiores e mais sofisticados nesse contexto seria uma direção promissora para pesquisas futuras.

Buscando superar as limitações na geração de resumos de alta qualidade, [5] propuseram o **HMSumm**, um método híbrido para sumarização de múltiplos documentos, que combina técnicas extrativas e abstrativas. O processo consiste em duas etapas principais:

1) **Resumo extrativo:** Inicialmente, um resumo extrativo é gerado para identificar as informações mais importantes dos documentos de entrada, enquanto gerencia a redundância – um desafio típico na sumarização de múltiplos documentos. Para isso, é empregado o Processo de Ponto Determinante (DPP - *Determinantal Point Process*), um método baseado em otimização empregado na composição de vários documentos que considera a redundância maximizando a diversidade [2]. Uma Rede Submodular Profunda (DSN - *Deep Submodular Network*) é empregada para determinar a qualidade das frases no resumo extrativo e semelhanças baseadas em BERT para calcular a redundância e selecionar frases relevantes. Além de controlar o comprimento do texto de entrada, essa etapa reduz o tempo computacional e preserva as partes mais significativas dos documentos para a etapa subsequente.
2) **Resumo abstrativo:** O resumo extrativo gerado é usado como entrada para os modelos pré-treinados BART e T5, que produzem dois resumos abstrativos. A diversidade de frases em cada resumo é então avaliada, e o resumo mais diversificado é escolhido como o resultado final.

No HMSumm, o modelo BERT é utilizado para representar as frases de maneira contextualizada e sensível ao conteúdo. A partir dessas representações, é construído um grafo onde os vértices correspondem às frases e as arestas indicam o grau de similaridade entre elas. Este grafo é podado para eliminar frases mais longas em favor de suas contrapartes mais curtas e similares, reduzindo a redundância e o custo computacional.

Após a poda, as frases restantes são processadas por uma Rede Submodular Profunda (DSN), que avalia a qualidade das frases e atribui pontuações. As frases com maior pontuação compõem o resumo extrativo inicial. Este, por sua vez, é refinado na etapa abstrativa, garantindo que o resumo final combine precisão e fluidez.

O HMSumm aborda desafios comuns na sumarização de múltiplos documentos, como redundância e custo computacional, utilizando uma abordagem eficiente e integrada que equilibra as forças dos métodos extrativos e abstrativos.

D. Métodos de Avaliação de Resumos

A qualidade dos resumos gerados automaticamente tem sido significativamente aprimorada por modelos de linguagem pré-treinados. No entanto, a tarefa de sumarização automática ainda enfrenta desafios importantes, especialmente relacionados à inconsistência factual entre os resumos e seus documentos de origem. Um dos principais problemas dos métodos atuais é a dificuldade de garantir que os resumos gerados sejam factualmente consistentes com as informações originais.

Nesse contexto, [12] propuseram uma abordagem fracamente supervisionada, baseada no modelo BERT, para verificar a consistência factual no nível da frase e identificar possíveis conflitos entre os documentos originais e os resumos gerados. Devido à ausência de conjuntos de dados específicos para essa tarefa, os autores criaram dados de treinamento por meio de transformações baseadas em regras, inspiradas na análise de erros observados em modelos de sumarização de última geração. O modelo proposto foi treinado para realizar três tarefas principais:

1) Determinar se as frases permanecem factualmente consistentes após as transformações.
2) Identificar trechos dos documentos de origem que sustentam a previsão de consistência.
3) Apontar, se houver, as partes inconsistentes nas frases resumidas.

Dando continuidade aos esforços para avaliar a consistência factual e aproveitando o desempenho notável dos Grandes Modelos de Linguagem (LLMs), como o ChatGPT, [20] exploraram a capacidade desse modelo em uma configuração zero-shot. O objetivo foi avaliar a consistência factual em diferentes tarefas, incluindo:

- **Inferência de implicação (EI):** verificar se uma frase deriva logicamente de outra.
- **Classificação de consistência em resumos:** comparar a consistência factual entre resumos e documentos de origem.
- **Classificação quantitativa de consistência:** julgar a factualidade de afirmações quantitativas.

Os resultados experimentais revelaram os seguintes achados:

1) **Desempenho competitivo:** O ChatGPT demonstrou grande potencial para avaliar a factualidade de resumos em configurações zero-shot, superando métodos de avaliação de última geração em diversos conjuntos de dados testados.
2) **Tendência à semelhança léxica:** Apesar do desempenho impressionante, o modelo apresentou uma inclinação para considerar documentos e afirmações consistentes quando há alta semelhança léxica, ignorando, em alguns casos, a implicação semântica. Além disso, foram observados casos de inferências falsas, expondo limitações em sua capacidade de raciocínio linguístico.
3) **Limitações nos prompts:** Embora as instruções fornecidas (prompts) tenham sido eficazes para guiar o ChatGPT na detecção de inconsistências, os resultados não foram consistentemente aderentes às diretrizes fornecidas, sugerindo que os prompts testados são insuficientes para garantir resultados uniformes.

Essas análises indicam avanços significativos, mas também ressaltam os desafios remanescentes na avaliação de consistência factual por modelos de linguagem avançados.

Dando continuidade à linha de pesquisa sobre inconsistências factuais em resumos, [14] realizaram uma análise detalhada dos benchmarks existentes para avaliação de inconsistência factual e identificaram problemas que comprometem a precisão dessas avaliações. Para abordar essas limitações, os autores propuseram um novo protocolo para a criação de benchmarks de detecção de inconsistências, implementado no **SUMMEDITS**, um benchmark abrangendo 10 domínios. O protocolo envolve a verificação manual da consistência de um conjunto inicial de resumos e a geração de várias versões editadas desses resumos.

O SUMMEDITS apresenta vantagens significativas:

- É 20 vezes mais econômico por amostra em comparação com benchmarks anteriores.
- É altamente reprodutível, alcançando uma concordância entre anotadores de aproximadamente 0,9.

Apesar dessas melhorias, a maioria dos Grandes Modelos de Linguagem (LLMs) apresenta dificuldades no SUMMEDITS, com desempenho próximo ao acaso. O GPT-4, embora seja o modelo de melhor desempenho, ainda está 8% abaixo do nível estimado de desempenho humano, evidenciando as lacunas dos LLMs em raciocínio factual e detecção de inconsistências, conforme também apontado por [20].

Avançando na avaliação da qualidade de resumos em termos de alucinação, emissão de informações e verbosidade, [33] introduziram o **FineSurE** (Fine-Grained Summarization Evaluation). Essa estrutura utiliza LLMs para avaliar resumos com base em três critérios fundamentais:

1) **Fidelidade:** minimização de erros factuais nos resumos.
2) **Integridade:** cobertura da maioria dos fatos importantes.
3) **Concisão:** eliminação de detalhes desnecessários.

O FineSurE adota uma abordagem de duas etapas:

1) **Verificação de fatos:** identifica erros factuais específicos presentes em cada frase resumida.
2) **Alinhamento de fatos-chave:** foca em alinhar cada fato-chave (definido como uma frase concisa que transmite uma única informação importante, com no máximo 2-3 entidades) às frases resumidas das quais são inferidos.

Essas contribuições destacam avanços metodológicos importantes para a avaliação de resumos, mas também reforçam os desafios remanescentes na busca por modelos capazes de gerar resumos factualmente consistentes, completos e concisos.

No estudo conduzido por [32], foi realizada uma análise abrangente para avaliar a estabilidade e a confiabilidade dos Grandes Modelos de Linguagem (LLMs) como avaliadores automáticos de resumos gerados pela abordagem abstrativa. O objetivo era determinar se os LLMs poderiam substituir de forma confiável especialistas humanos nessa tarefa. Embora modelos como ChatGPT e GPT-4 tenham demonstrado desempenho superior em relação a métodos automáticos tradicionais, eles ainda apresentam limitações significativas que os tornam inadequados como substitutos humanos.

Os autores identificaram que os avaliadores baseados em LLMs classificam os sistemas candidatos de maneira sistemática, mas com dependências e limitações específicas, incluindo:

1) **Dificuldade em diferenciar desempenhos próximos:** Os LLMs apresentam dificuldades para comparar resumos de qualidade semelhante, resultando em avaliações menos precisas.
2) **Dependência do candidato:** O alinhamento das avaliações dos LLMs com os julgamentos humanos varia de acordo com o sistema avaliado, o que pode levar a favoritismos ou penalizações injustas.
3) **Dependência da dimensão:** Os LLMs apresentam diferentes graus de eficácia ao avaliar dimensões específicas, como coerência, fluência e factualidade, o que afeta a uniformidade das avaliações.
4) **Correlação reduzida com julgamentos humanos em resumos de alta qualidade:** À medida que a qualidade dos sistemas de sumarização melhora, a capacidade dos LLMs de se alinhar aos julgamentos humanos diminui, como evidenciado por uma métrica de metacorrelação proposta pelos autores.

Essas descobertas indicam que, apesar dos avanços, os LLMs ainda enfrentam desafios para fornecer avaliações confiáveis e consistentes, especialmente em cenários onde a qualidade dos resumos é elevada ou os sistemas candidatos apresentam desempenhos similares. Assim, enquanto os LLMs representam uma ferramenta promissora para complementar avaliações humanas, eles não estão prontos para substituí-las integralmente em tarefas de avaliação de resumos.

V. Conclusões

A aplicação de LLMs na sumarização automática de textos representa um avanço significativo no campo do NLP. Este estudo revisou as abordagens mais recentes, destacando tanto as técnicas extrativas quanto as abstrativas, e explorou as vantagens e limitações de cada uma.

Os modelos de sumarização extrativa, embora eficientes e simples, enfrentam desafios relacionados à coerência e consistência factual dos resumos gerados. Por outro lado, os modelos de sumarização abstrativa, apesar de produzirem resumos mais fluentes e gramaticais, ainda lutam com a utilização completa do contexto e a ”maldição do meio”. Abordagens híbridas, que combinam técnicas extrativas e abstrativas, mostram-se promissoras para superar essas limitações, oferecendo uma síntese mais equilibrada e precisa.

A pesquisa também destacou a importância de métodos robustos de avaliação para garantir a qualidade dos resumos gerados. Métricas tradicionais, como ROUGE, são amplamente utilizadas, mas apresentam limitações. Abordagens mais recentes, que utilizam LLMs para avaliação semântica e coerência, têm se mostrado promissoras, mas ainda enfrentam desafios significativos.

Em suma, a sumarização automática de textos com LLMs possui grande relevância para o avanço de tecnologias de NLP e para áreas práticas que demandam a análise e interpretação de grandes volumes de dados textuais. Este estudo contribui para o entendimento do papel dos LLMs na sumarização automática, promovendo o desenvolvimento de sistemas mais robustos e eficientes. Futuras pesquisas devem focar na melhoria da coerência, precisão factual e relevância dos resumos gerados, além de explorar novas abordagens híbridas e métodos de avaliação mais sofisticados.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] S. Cho, L. Lebanoff, H. Foroosh, and F. Liu, "Improving the similarity measure of determinantal point processes for extractive multi-document summarization," *arXiv preprint arXiv:1906.00072*, 2019.

[3] P. K. Choubey, A. Lee, R. Huang, and L. Wang, "Discourse as a function of event: Profiling discourse structure in news articles around the main event," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[4] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] A. Ghadimi and H. Beigy, "Hybrid multi-document summarization using pre-trained language models," *Expert Systems with Applications*, vol. 192, p. 116292, 2022.

[6] T. Goyal, J. J. Li, and G. Durrett, "News summarization and evaluation in the era of gpt-3," *arXiv preprint arXiv:2209.12356*, 2022.

[7] Q. Grail, J. Perez, and E. Gaussier, "Globalizing bert-based transformer architectures for long document summarization," in *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume*, 2021, pp. 1792–1810.

[8] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Systems with Applications*, vol. 121, pp. 49–65, 2019.

[9] R. Jie, X. Meng, L. Shang, X. Jiang, and Q. Liu, "Enhancing coherence of extractive summarization with multitask learning," *arXiv preprint arXiv:2305.12851*, 2023.

[10] H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan, "A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods," *arXiv preprint arXiv:2403.02901*, 2024.

[11] U. Khandelwal, K. Clark, D. Jurafsky, and L. Kaiser, "Sample efficient text summarization using a single pre-trained transformer," *arXiv preprint arXiv:1905.08836*, 2019.

[12] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," *arXiv preprint arXiv:1910.12840*, 2019.

[13] L. J. Kurisinkel and N. F. Chen, "Llm based multi-document summarization exploiting main-event biased monotone submodular content extraction," *arXiv preprint arXiv:2310.03414*, 2023.

[14] P. Laban, W. Kryściński, D. Agarwal, A. R. Fabbri, C. Xiong, S. Joty, and C.-S. Wu, "Summedits: measuring llm ability at factual reasoning through the lens of summarization," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9662–9676.

[15] Y. Li, "Unlocking context constraints of llms: Enhancing context efficiency of llms with self-information-based content filtering," *arXiv preprint arXiv:2304.12102*, 2023.

[16] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[17] H. Lin and V. Ng, "Abstractive summarization: A survey of the state of the art," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9815–9822.

[18] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

[19] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[20] Z. Luo, Q. Xie, and S. Ananiadou, "Chatgpt as a factual inconsistency evaluator for text summarization," *arXiv preprint arXiv:2303.15621*, 2023.

[21] D. Miller, "Leveraging bert for extractive text summarization on lectures," *arXiv preprint arXiv:1906.04165*, 2019.

[22] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *arXiv preprint arXiv:1808.08745*, 2018.

[23] N. Nazari and M. Mahdavi, "A survey on automatic text summarization," *Journal of AI and Data Mining*, vol. 7, no. 1, pp. 121–135, 2019.

[24] A. Nenkova and K. McKeown, "A survey of text summarization techniques," *Mining text data*, pp. 43–76, 2012.

[25] A. Onan and H. A. Alhumyani, "Deepextract: Semantic-driven extractive text summarization framework using llms and hierarchical positional encoding," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 8, p. 102178, 2024.

[26] J. Pilault, R. Li, S. Subramanian, and C. Pal, "On extractive and abstractive neural document summarization with transformer language models," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 9308–9319.

[27] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.

[28] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[30] M. Ravaut, S. R. Joty, A. Sun, and N. F. Chen, "On context utilization in summarization with large language models," in *Annual Meeting of the Association for Computational Linguistics*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264146949

[31] A. Rush, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[32] C. Shen, L. Cheng, X.-P. Nguyen, Y. You, and L. Bing, "Large language models are not yet human-level evaluators for abstractive summarization," *arXiv preprint arXiv:2305.13091*, 2023.

[33] H. Song, H. Su, I. Shalyminov, J. Cai, and S. Mansour, "Finesure: Fine-grained summarization evaluation using llms," *arXiv preprint arXiv:2407.00908*, 2024.

[34] D. Suleiman and A. Awajan, "Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges," *Mathematical problems in engineering*, vol. 2020, no. 1, p. 9365340, 2020.

[35] A. A. Syed, F. L. Gaol, and T. Matsuo, "A survey of the state-of-the-art models in neural abstractive text summarization," *IEEE Access*, vol. 9, pp. 13 248–13 265, 2021.

[36] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[37] J. Xu, Z. Gan, Y. Cheng, and J. Liu, "Discourse-aware neural extractive text summarization," *arXiv preprint arXiv:1910.14142*, 2019.

[38] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.