Sumarização multi-documento para o português com modelos BERT

Multi-document summarization for Portuguese using BERT

Giordano Pydd Berwanger *, Gustavo Henrique Paetzold †

RESUMO

Em decorrência da quantia abundante de informações que temos disponível na Internet, encontrar formas de sintetizar conteúdos para obter apenas o cerne de um documento é essencial para auxiliar no processamento de grandes volumes de dados. A sumarização automática consiste na tarefa de gerar automaticamente versões condensadas de textos fontes denominadas sumários, que são textos expressos em forma reduzida mantendo apenas as partes necessárias sem perder o sentido nem o contexto original do texto. Mesmo com avanços substanciais realizados nessa área na língua inglesa, a escassez de estudos e esforços para pesquisa e desenvolvimento de sumarizadores multi-documento com ênfase na língua portuguesa são aspectos que motivam o presente trabalho. Com a ascensão dos modelos BERT e seus excelentes resultados, nos últimos anos tem crescido muito a sua utilização em várias tarefas de PLN, incluindo a sumarização de documentos. A aplicação dos modelos bidirecionais explorados no BERT possibilitam o desenvolvimento de sumarizadores abstrativos que aprendem com as relações contextuais entre as palavras, gerando assim, sumarizadores mais precisos e estruturados. O presente trabalho visa explorar o uso de modelos BERT aplicados a tarefa de sumarização multi-documento na língua portuguesa e investigar o comportamento e os resultados obtidos ao utilizar algumas ferramentas e técnicas de simplificação textual e sumarização.

Palavras-chave: Modelos BERT Sumarização automática Sumarização multi-documento

ABSTRACT

Due to the abundant amount of information that we have available on the Internet, finding ways to synthesize content to get just the heart of a document is essential to assist in processing large volumes of data. Automatic summarization is the task of automatically generating condensed versions of source texts called summaries, which are texts expressed in reduced form keeping only the necessary parts without losing the meaning or the original context of the text. Even with substantial advances made in this area in the English language, the scarcity of studies and efforts for research and development of multi-document summarizers with an emphasis on the Portuguese language are aspects that motivate this work. With the rise of BERT models and their excellent results, in recent years their use in various PLN tasks, including document summarization, has grown a lot. The application of the bidirectional models explored in BERT enables the development of abstract summaries that learn from the contextual relationships between words, thus generating more precise and structured summaries. This work aims to explore the use of BERT models applied to the task of multi-document summarization in Portuguese and investigate the behavior and results obtained when using some tools and techniques of textual simplification and summarization.

Palavras-chave: BERT Templates Auto-Summary Multi-Document Summary.

* Bacharel em Engenharia de Computação; giordanoberwanger@alunos.utfpr.edu.br.
† Ph.D em Linguística Computacional; ghpaetzold@outlook.com.

1 INTRODUÇÃO

De 2017 para 2019, a Internet cresceu aproximadamente 63%, alcançando a marca de 4,4 Zetabytes de informação, com previsão de superar os 175 Zetabytes até 2025 (REINSEL; GANTZ; RYDNING, 2018). Em decorrência deste ritmo acelerado de crescimento que a Internet vem obtendo nos últimos anos, enfrentamos um grande acúmulo de informação textual, que vem tornando tarefas como pesquisa e consulta cada vez mais árduas e custosas, criando um fenômeno que passou a ser descrito como uma "sobrecarga de informação".

Nesse cenário, uma das soluções úteis para atenuar esse problema é condensação e redução do volume de informação textual disponível. A sumarização clássica (ou monodocumento) tem o intuito de gerar uma síntese a partir de um único documento, enquanto a sumarização automática multi-documento (SAM) é mais adequada ao contexto pois consiste na produção automática de um único sumário a partir de um grupo de textos sobre um mesmo tópico (MANI, 2001).

A sumarização é de fato, uma tarefa bastante comum e presente em nosso cotidiano e entre suas variadas aplicações, podemos citar ela sendo muito utilizada em resumos de publicações científicas, síntese de notícias e até sinopse de filmes e seriados.

Com a escassez de trabalhos relacionados e de poucos recursos destinados a pesquisas na área de sumarização na língua portuguesa, este trabalho justifica a sua importância, já que é uma área que desperta cada vez mais interesse, principalmente para empresas como o Google e Bing, que estão constantemente reforçando a qualidade de seus buscadores a fim de sempre retornar os melhores resultados nas buscas.

A sumarização multi-documento é outro ponto de interesse, já que tem sua importância justificada por ser uma abordagem com mais aplicações no mercado, dado a vasta quantia de documentos que cresce cada dia mais na Internet. Além disso, este trabalho também tem o objetivo de explorar a eficácia de modelos BERT, que tem beneficiado muito a área de PLN por conta do amadurecimento das redes neurais recorrentes. Tal modelo será aplicado na criação de sumarizadores multi-documento para a língua Portuguesa e será comparado com outras estratégias distintas em avaliações tanto automáticas quanto manuais.

2 MATERIAIS E MÉTODOS

A partir de uma base de dados com diversos textos jornalísticos de diferentes fontes, serão implementados sumarizadores multi-documento extrativos e abstrativos utilizados para gerar sumários, que posteriormente serão avaliados de duas formas: qualitativamente e quantitativamente.

A base de dados utilizada foi o córpus de textos jornalísticos em português do Brasil denominado CSTNews(ALEIXO; PARDO, 2008) que é uma base de dados muito renomada e bem estruturada na língua portuguesa, contando com textos-fonte de temas e tamanhos variados e utilizando sumários de referência desenvolvidos por linguistas profissionais. O *dataset* fpo desenvolvido pelo NILC (Núcleo interinstitucional de linguística computacional) para fins de sumarização. Contendo uma coleção de 50 textos, onde cada coleção aborda temas distintos contendo em média, 3 documentos de diferentes fontes.

As fontes dos textos foram os jornais *online*: Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo e as notícias são subdivididas em 6 temáticas, conforme apresentado na Tabela 1.

Tabela 1 – Distribuição de textos da base de dados

Tema	Qt. de documentos	Qt. de sentenças	Qt. de palavras
Mundo	54	913	21006
Cotidiano	49	1027	19833
Política	39	785	16055
Esportes	38	740	13531
Dinheiro	3	46	1136
Ciência	2	23	587
Total	185	3534	72148

Fonte: Autoria própria (2021).

2.1 PROCESSAMENTO

O processamento dos textos presentes no córpus acontece paralelamente de duas formas: gerando os sumários extrativos e abstrativos a partir de um conjunto de uma das coleções disponíveis no *dataset*. Ao todo são gerados 4 sumários extrativos utilizando os seguintes algoritmos: algoritmo de Luhn, algoritmo TextRank, método TF-IDF e o algoritmo de Earl.

Os algoritmos foram implementados na linguagem Python utilizando recursos das bibliotecas NLTK, *spacy*, *heapq*, *math*, *summa*, *numpy* para sua implementação e funções e métodos da biblioteca ROUGE para avaliar a qualidade da informatividade dos sumários. Tais sumários servirão de base para comparar com o desempenho do sumário abstrativo desenvolvido utilizando modelos BERT que é paralelamente processado para posteriormente ser avaliado.

2.2 MÉTODOS AVALIATIVOS

Como este trabalho tem a intenção de avaliar a qualidade dos sumários, as sentenças selecionadas e a organização estrutural do resumo, os métodos avaliativos utilizados serão de natureza intrínseca. Eles tem como tarefa fundamental identificar e avaliar os sumários automaticamente gerados em função dos critérios expostos

por (DANG, 2005) na DUC (*Document Understanding Conferences*), que é a divisão de sumarização automática na TAC (*Text Analysis Conference*).

2.2.1 AVALIAÇÃO DA QUALIDADE LINGUÍSTICA

A avaliação da qualidade linguística de um sumário tem como objetivo verificar a redundância na gramática e qualidade estrutural do texto gerado automaticamente. No contexto da DUC'05, Dani propôs que a qualidade linguística dos sumários pode ser avaliada em função dos seguintes critérios (DANG, 2005):

- **Gramaticalidade:** Através deste critério é verificado a ortografia, pontuação, sintaxe e formatação a fim de encontrar erros que prejudiquem a legibilidade do texto;
- **Não-redundância:** Nesse tópico é necessário avaliar a repetição de sentenças, fatos e nomes pois um sumário de qualidade não deve conter informações repetitivas.
- **Clareza referencial:** O sumário deve fornecer a identificação clara das siglas que referenciam uma pessoa, entidade ou organização.
- **Foco:** O sumário também deve conter um foco temático bem definido, de forma que ele possa ser identificável através de sentenças que contenham informações inter-relacionadas com a temática proposta.
- **Estrutura e coerência:** Sobre este atributo, o sumário deve possuir uma organização adequada, bem informativa e estruturada coerentemente com o tópico do texto.

Os sumários serão avaliados de forma manual e individual por humanos de acordo com os critérios acima e classificados com uma das notas segundo a Tabela 2.

Tabela 2 – Parâmetros de avaliação da qualidade linguística

Nota Atribuída	1	2	3	4	5
Descrição	Péssimo	Ruim	Regular	Bom	Excelente

Fonte: Autoria própria (2021).

2.2.2 AVALIAÇÃO DA INFORMATIVIDADE DO TEXTO

A avaliação da informatividade consiste em identificar quanto de informação relevante dos textos-fonte o sumário gerado automaticamente incorpora. Utilizando o *framework* ROUGE, os sumários gerados são avaliados automaticamente em comparação com os sumários de referência a fim de se calcular a informatividade por meio da coocorrência de n-gramas entre os sumários (LIN; HOVY, 2003).

Serão utilizados tanto o método ROUGE-1 quanto o ROUGE-2 para avaliação dos sumários pois ao avaliar tais medidas com uma granularidade menor, é possível avaliar a fluência dos sumários com relação ao resumo de referência. Os resultados são fornecidos em termos da precisão (P), cobertura (C) e da medida-F (F), que variam de 0 a 1. Quanto maior o valor obtido, mais informativo é o sumário com relação ao resumo utilizado como referência.

3 RESULTADOS

Os resultados preliminares referentes a avaliação da qualidade linguística e informatividade dos sumários, gerados utilizando o *dataset* foram obtidos para as seguintes implementações: Algoritmo de Luhn, TextRank, algoritmo de Earl e o método TF-IDF.

A avaliação foi aplicada pelo autor do texto em 10 coleções de documentos do banco de dados, o que corresponde a 28 textos-fonte que aplicados aos 4 algoritmos citados, geraram 40 sumários diferentes para serem avaliados seguindo os parâmetros apresentados na Tabela 2.

3.1 QUALIDADE LINGUÍSTICA

A 3 apresenta os resultados obtidos na avaliação da qualidade linguística dos algoritmos, onde temos que o Algoritmo clássico elaborado por Luhn obteve os melhores resultados, sendo considerado o melhor dos sumários implementados no quesito da qualidade linguística.

Tabela 3 – Resultados da qualidade linguística dos métodos avaliados

Critérios	Método			
	Luhn	TextRank	Earl	TF-IDF
Gramaticalidade	**4.6**	**4.6**	4.4	4.3
Não-redundância	**3.9**	2.6	3.5	3
Clareza referencial	**4.6**	4.5	4.3	3.6
Foco	**4.2**	4	3.9	3.2
Estrutura e coerência	**4**	3.6	**4**	3.3

Fonte: Autoria própria (2021).

3.2 INFORMATIVIDADE DO TEXTO

Para avaliação da informatividade do texto foram utilizadas as métricas ROUGE-1 e ROUGE-2, que tem o papel de avaliar os sumários gerados automaticamente de forma comparativa com os sumários considerados ideais para tal conjunto de textos. Essa avaliação é fornecida em termos da Precisão (P), cobertura (C) e medida-f (F). A Tabela 4 apresenta a média aritmética dos valores ao aplicar tais métodos ao mesmo conjunto de 10 coleções utilizado na avaliação da qualidade linguística.

Tabela 4 – Resultados preliminares da avaliação de informatividade

Método	ROUGE 1			ROUGE 2		
	P	C	F	P	C	F
Luhn	0.4243	0.3890	0.3961	0.2237	0.2012	0.2067
TextRank	0.3635	0.5675	**0.4196**	0.2111	0.3297	**0.2433**
Earl	0.3873	0.3625	0.3662	0.1781	0.1645	0.1673
TF-IDF	0.3154	0.4100	0.3368	0.1012	0.1383	0.1109

Fonte: Autoria própria (2021).

4 CONCLUSÕES

A partir dos sumários obtidos aplicando-se cada um dos métodos nos textos fontes disponíveis no *dataset*, concluímos que:

- **Método de Luhn:** Para esta abordagem, obtivemos boa qualidade linguística e informativa em seus sumários, porém apresentando problemas estruturais em alguns resumos obtidos. Além disso, os sumários apresentaram boa clareza referencial e mantiveram altas notas de gramaticalidade.
- ***TextRank:*** Apresentando altas taxas de informatividade, esta abordagem se destacou obtendo as maiores medidas-F nos métodos ROUGE 1 e 2. Possuindo também boa gramaticalidade e clareza referencial, este método apresenta muito potencial a ser explorado mas deixa a desejar no quesito não redundância, visto que os sumários resultantes acabaram por vezes utilizando duas sentenças muito semelhantes, caracterizando-se como uma repetição que prejudicou a qualidade do resumo.
- **Método de Earl:** Com resultados próximos a abordagem de Luhn, os sumários obtidos através deste método apresentavam índices de foco próximos ou abaixo da média e problemas de não-redundância e repetição de informações no decorrer do texto. Mas, apesar destes erros, seus resumos ainda possuem boa clareza referencial e altos índices de gramaticalidade.
- **Método TF-IDF:** Através da avaliação automática pelo método ROUGE, foi identificado que o método TF-IDF possui pontos críticos na organização e estruturação de seus sumários, além de seus resumos apresentarem um alto índice de dispersão do foco do texto, o que caracteriza uma dificuldade em encontrar adequadamente quais as palavras-chave que descrevem a temática central do texto-fonte.

REFERÊNCIAS

ALEIXO, Priscila; PARDO, T. A. S. CSTNews: um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory). **Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP**, 2008.

DANG, H. T. Overview of DUC 2005. **Proceedings of the Document Understanding Conference**, 2005.

LIN, Chin-Yew; HOVY, Eduard. Automatic evaluation of summaries using n-gram co-occurrence statistics. **Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1**, Association for Computational Linguistics, 2003. Disponível em: .

MANI, I. Summarization Evaluation: An Overview. In the Proceedings of the Workshop on Automatic Summarization. **Pittsburgh, Pennsylvania**, 2001.

REINSEL, David; GANTZ, John; RYDNING, John. Data Age 2025: The Digitization of the World From Edge to Core. **IDC White Paper**, Seagate, 2018. Disponível em: .