UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS ANDRÉ SALVINO

ANÁLISE DE TÉCNICAS DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTO SUPERFICIAIS E PROFUNDAS

CAMPINA GRANDE - PB

2019

LUCAS ANDRÉ SALVINO

ANÁLISE DE TÉCNICAS DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTO SUPERFICIAIS E PROFUNDAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Professora Dra. Joseana Macêdo Fechine Régis de Araújo.

CAMPINA GRANDE - PB

2020

S185a Salvino, Lucas André.
Análise de técnicas de sumarização automática de texto superficiais e profundas. / Lucas André Sanvino. – 2019.

12 f.

Orientador: Prof. Dr. Joseana Macêdo Fechine Régis de Araújo.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Sumarização automática. 2. Aprendizagem de máquina. 3. Mineração de dados. I. Araújo, Joseana Macêdo Fechine Régis de. II. Título.

CDU:004.6(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

LUCAS ANDRÉ SALVINO

ANÁLISE DE TÉCNICAS DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTO SUPERFICIAIS E PROFUNDAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professora Dra. Joseana Macêdo Fechine Régis de Araújo
Orientadora – UASC/CEEI/UFCG

Professor Dr. Franklin de Souza Ramalho
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 02 de julho de 2019.

CAMPINA GRANDE - PB

Análise de Técnicas de Sumarização Automática de Texto Superficiais e Profundas

Lucas André Salvino
Unidade Acadêmica de Sistemas e Computação
Centro de Engenharia Elétrica e Informática
Universidade Federal de Campina Grande
Campina Grande
lucas.a.salvino@gmail.com

ABSTRACT

Automatic summarization of texts is the field of information retrieval area, based on textual documents, character, collection and production of relevant phrases. Thus, in a series of activities of searching for relevant data, a scenario of generation of summaries, which, in this scenario, are summaries of texts. There are several techniques for automatic summarization, comprising a linguistic science and the statistical domain. In this article, we discuss the positive and negative aspects of two subgroups of automatic text summarization techniques, such as those that use superficial approaches and those that use a deep approach. A historical study is presented on the main studies in the area of summarization and the functioning of the process of automatic summarization of texts. The challenges and limits in the automatic summarization process are highlighted. In terms of the results in the research, was observed that there is a technique of summarization, since each has a set of benefits and disadvantages. In view of the above, it was verified the intention to intensify the studies in the area of automatic summarization of texts.

KEY WORDS

Automatic summarization, machine learning, text mining.

RESUMO

A sumarização automática de textos é o campo da área de recuperação da informação que, baseado em documentos textuais, caracteriza, coleta e produz frases pertinentes. Assemelhando-se, assim, à atividade humana de selecionar componentes relevantes de um texto, a fim de gerar sumários, que, neste cenário, são resumos de textos. Há inúmeras técnicas para sumarização automática, compreendendo a ciência linguística e o domínio estatístico. Neste artigo, são discutidos os aspectos positivos e negativos de dois subgrupos de técnicas para sumarização automática de textos, as que utilizam abordagens superficiais e as que utilizam abordagem profunda. É apresentado um breve histórico sobre os principais estudos na área de sumarização e abordado o funcionamento do processo de sumarização automática de textos. São destacados os desafios e limites no processo de sumarização automática. Em se tratando dos resultados obtidos na pesquisa, observou-se que não há uma técnica ímpar de sumarização, uma vez que cada técnica possui um conjunto de benefícios e desvantagens. Diante do exposto, verificou-se a necessidade de intensificar os estudos na área de sumarização automática de textos.

PALAVRAS-CHAVE

Sumarização automática, aprendizagem de máquina, mineração de texto.

1 Introdução

A diversidade dos meios de comunicação e o avanço da internet facilitaram o acesso a uma gama de informações, por meio de blogs, sites, redes socais, revistas online, dentre outros. Esses meios facilitam também o acesso a conhecimentos científicos a partir de documentos, texto tais como artigos científicos, teses, dissertações, monografia, e-books, periódicos, entre outros. Em contrapartida, o aumento da quantidade de informações e conhecimentos dificulta uma leitura seletiva destes documentos, o que incentivou o uso de procedimentos automáticos de indexação, classificação e apresentação de informações diretas e sucintas, facilitando a compreensão e demandando menos tempo para a leitura.

Helena et al. [18], afirmam que um procedimento utilizado com a finalidade de extrair textos com informações relevantes dos documentos é denominada sumarização de textos, que gera automaticamente um resumo de um ou mais documentos. Helena et al. [18], afirmam também que algumas técnicas de sumarização automática extraem seus resumos por meio da mineração de texto, implicando na extração automática de palavras-chave.

Neste trabalho, serão analisados dois subgrupos de técnicas de sumarização automática. No primeiro subgrupo, serão analisadas apenas técnicas que fazem uso de abordagens superficiais para criação dos resumos, como o método o método estatístico. No segundo subgrupo, será analisado o desempenho de técnicas que contam com a utilização de abordagem profunda em suas implementações. De acordo com Brandel et al. [21] abordagens superficiais são limitadas apenas à identificação, seleção e exclusão/extração de segmentos textuais enquanto que a abordagem profunda contempla o conhecimento linguístico e extralinguístico associado ao texto de origem, a fim de compor seu possível sumário. Diante das diferenças entre dois modos de

sumarização, pretende-se apresentar as vantagens, desvantagens e as particularidades de cada um desses grupos.

Além destes dois subgrupos alvos deste trabalho, a sumarização automática pode ser subdividida em dois outros subgrupos, que são sumarização abstrativa e sumarização extrativa (LLORET & PALOMAR, [6]). A técnica abstrativa visa obter um resumo coerente, sem redundâncias, podendo, inclusive, gerar novas frases, quando utilizado um componente apropriado para esta finalidade. Trata-se de um resumo mais realista e aprimorado, entretanto, ainda não foi elaborado um algoritmo padrão para este tipo. A técnica extrativista apresenta as frases mais relevantes presentes no artigo original. Entender estes dois conceitos também é importante para ponderar sobre a utilização da melhor técnica para sumarização.

2 Breve histórico sobre os principais estudos na área de sumarização

Desenvolver algoritmos para extração automática de informações de um texto é um dos grandes desafios da computação. Ao longo dos anos, foram desenvolvidos diversos estudos na tentativa de melhorar o desempenho das técnicas de recuperação de informação e alguns destes estudos ainda se mostram de suma importância nos dias atuais. Por exemplo, Salton et al.. [11] apresentaram métodos para estruturação e recuperação de informação e textos de conteúdos heterogêneos em seu artigo "Automatic structuring of text files". Chien et al.. [1] mostraram que é possível extrair palavras-chave de textos utilizando PAT-trees[1], de maneira eficiente. No final da década de 1990, técnicas de sumarização baseadas em machine learning começaram a se popularizar. Turney et al.. [16] publicaram diversos estudos sobre recuperação automática de informação em texto, envolvendo inteligência artificial e extração de palavras-chave. Esses estudos comprovaram que métodos estatísticos para sumarizações automáticas podem ter limitações em determinados aspectos e levantaram diversas indagações sobre as possíveis vantagens e desvantagens acerca do casamento entre aprendizagem de máquina e recuperação da informação.

Na última década, tanto algoritmos para recuperação da informação baseados em inteligência artificial, quanto baseados em grafos e estatísticas demonstravam diversas limitações. Segundo Martins et al. [7], embora no passado parecesse necessário escolher entre uma das duas abordagens supracitadas, os pesquisadores passaram a buscar uma combinação coerente entre as mesmas, investigando amplamente tecnologias híbridas e fazendo uso de metodologias simbólicas e técnicas estatísticas, simultaneamente. Um exemplo desta abordagem híbrida é o sumarizador de eventos de Maybury [22], que utiliza tanto abordagens superficiais como profundas para lidar com mensagens em um simulador de batalha.

Nos anos de 2007, 2008, e 2016 foram desenvolvidos mais 3 estudos de suma importância para a área da sumarização automática de textos. Ercan et al. [3], em 2007, desenvolveu um projeto de pesquisa que demonstrava resultados em sumarizações com aprendizagem de máquina em artigos de jornais. Em 2008, Litvak et al. [5], também utilizando aprendizagem de máquina demonstraram resultados de algoritmos capazes de resumir artigos de páginas de internet. Mais recentemente, em 2016, mais uma vez utilizando aprendizagem de máquina, Thomas et al. [15], desenvolveram pesquisas, as quais demonstravam que algoritmos de sumarização poderiam resumir artigos jornalísticos. E em 2018 Sahoo [24] realizou pesquisas que contribuiram para o campo da sumarização automática abstrativa através de métodos híbridos.

Pode-se notar que, ao longo dos anos, os algoritmos se desenvolveram de modo que utilizassem cada vez mais aprendizagem de máquina. As 3 últimas pesquisas supracitadas envolviam diretamente inteligência artificial e foram capazes de proporcionar resultados relevantes para o mundo acadêmico. Embora o histórico de evolução dos sumarizadores tenha mostrado um progresso relevante na área, essas 3 últimas pesquisas envolviam apenas artigos de jornais e de páginas web. Essa extração, voltada para textos informativos, normalmente é mais fácil do que a sumarização para outros gêneros semanticamente mais complexos.

3 Processo de sumarização automática de textos

O processo de sumarização de texto pode ser classificado em sete categorias baseado nos elementos descritos a seguir.

3.1 Sumarização de texto a partir de um único documento

Para realização deste procedimento, é utilizado um único documento de entrada, assim como, é gerado um único documento de saída. Esta técnica pode ser usada para fazer uma seleção de palavras-chave ou para sumarizar discursos, afirmam Kumar et al. [19].

3.2 Sumarização de texto a partir de múltiplos documentos

Na sumarização feita a partir de múltiplos documentos, são utilizados diversos documentos de entrada e é gerado um único documento de saída. Este procedimento pode ser utilizado para fazer a seleção de palavras-chave fundamentada na relação temporal, para fazer um resumo de diversos documentos ou, até mesmo, para aprimorar a qualidade do conteúdo sumarizado, como fez Min et al. [8], usando as informações compartilhadas por diversos documentos.

[1] É uma estrutura de dados que permite uma pesquisa muito eficiente com pré-processamento.

3.3 Sumarização de texto baseada em consulta

Neste processo de sumarização, é utilizada uma seção específica do documento para selecionar as palavras-chave fundamentais e, a partir destas, fazer a sumarização do documento.

3.4 Resumo de texto extrativo

Nesta técnica de sumarização é feita uma seleção mais ampla das informações e das sentenças do documento de entrada para gerar seu respectivo resumo, sendo utilizadas, para tanto, técnicas estatísticas para selecionar as frases essenciais do documento. Um exemplo de técnica bastante utilizada para realizar sumarizações extrativas é o "TextRank" que, segundo Joshi [14], ranqueia as sentenças mais importantes de acordo com o seu número de aparições no texto.

Dentre os principais métodos extrativos, destacam-se os métodos da palavra-chave, que se baseiam na hipótese de que a ideia principal do texto pode ser sintetizada em palavras-chave, a saber: palavra-chave do título, que considera as palavras que compõem o título para procurar sentenças relevantes no texto; da localização, que considera a posição das sentenças para mensurar sua importância no resumo final; dos marcadores linguísticos, que conta com um dicionário prévio de palavras consideradas relevantes para escolher as melhores sentenças; e o método relacional, que considera termos com a maior dificuldade de dileção[2] para escolher uma sentença relevante.

3.5 Resumo de texto abstrativo

Nesta técnica de sumarização, a máquina assimila a ideia presente nos documentos de entrada para gerar um resumo aprimorado e com sentenças próprias.

3.6 Sumarização do texto baseada em aprendizagem supervisionada

Para o treinamento, é utilizado um *dataset* rotulado. Mirroshandel et al. [9], por exemplo, por meio da utilização de dados rotulados, treinaram o sistema, a fim de categorizar os eventos com as relações temporais.

3.7 Sumarização de texto baseada em aprendizagem não supervisionada

Na utilização desta técnica não existem diretrizes pré-estabelecidas acessíveis na fase do treinamento. Mirroshandel et al. [9], fundamentados no algoritmo Expectation-Maximization (EM), sugeriram uma técnica para extração de relações temporais. Dentre as EM, foram utilizadas técnicas como busca gulosa e programação linear inteira[3] para extrair inconsistência temporal. As técnicas que utilizam EM são embasadas na extração completamente não supervisionada para sumarização de texto.

4 Comparação entre técnicas de sumarização automática superficiais e profundas

Segundo Hutchins [4], a análise do conteúdo de documentos é uma das atividades mais importantes de um sistema de informação No entanto, esse é um trabalho repleto de dificuldades, devido à subjetividade da tarefa. No que diz respeito à análise de conteúdo realizada por humanos, para desenvolver resumos, Hutchins [4] afirma que o leitor lembra somente da ideia central. Mas, identificar qual é a ideia central do texto a ser sumarizado, a fim de preservá-la no sumário correspondente por meio de um algoritmo, ainda é algo desafiador para as técnicas de sumarização atuais. Esse problema está relacionado tanto à forma, quanto ao conteúdo do texto.

Em abordagens superficiais de sumarização, um dos principais problemas encontrados, segundo Martins et al..[7] é a falta de coesão. O problema da falta de coesão textual, que implicará, quase que certamente, na falta de coerência, é quase que inerente aos métodos de sumarização superficial e há poucas soluções atualmente para mitigar o problema, que ainda não garantem a coesão, tampouco a coerência, caracterizando os textos como *non-sequitur*, ou seja, que possuem falácias lógicas.

Em relação às limitações da abordagem profunda, Martins et al.. [7] afirma que o inter-relacionamento entre clareza, abstração, coerência e coesão e outros conceitos que para os seres humanos são intuitivos, são de difícil incorporação e manipulação em um sistema computacional. Estes conceitos remetem ao nível profundo de estruturação e garantia da coerência textual, assim como ao nível superficial de expressão adequada do nível profundo e, portanto, da correta manipulação das informações linguísticas disponíveis para se produzir o texto final.

Para entender melhor estes conceitos e limitações expressos em relação as duas abordagens, seguem dois exemplos de resumos gerados a partir das duas técnicas. Embora as técnicas sejam diferentes, o texto analisado foi o mesmo para facilitar a comparação.

Texto original (trecho de artigo do Wikipédia sobre Julian Assenge) :

"(...)Julian Paul Assange nascido Julian Paul Hawkins; 3 de julho de 1971 é um ativista australiano, programador de computador, jornalista e fundador do site WikiLeaks. Atualmente ele está sob custódia da Polícia Metropolitana de Londres após ser preso em 11 de abril de 2019, sob a acusação de ter violado as condições estabelecidas na sua fiança em 2010. Antes, ele estava refugiado na embaixada do Equador em Londres, vivendo lá como refugiado de 2012 até seu encarceramento em 2019. Assange

[2] Ganhar atenção especial.

3 Técnica de programação utilizada para modelagem dos problemas através de variáveis inteiras (discretas), não contínuas

fundou o site WikiLeaks em 2006 e ganhou atenção internacional em 2010 quando o site publicou uma série de documentos sigilosos do governo dos Estados Unidos que haviam sido vazados por Chelsea Manning (na época chamada Bradley Manning). Entre os vazamentos estavam dados sobre o ataque aéreo a Bagdá em 12 de julho de 2007, os registros de guerra do Afeganistão e do Iraque e o CableGate (novembro de 2010). Após os vazamentos de 2010, autoridades dos Estados Unidos começaram uma investigação criminal sobre o WikiLeaks e pediu apoio a nações aliadas pelo mundo. Em novembro de 2010, a Suécia emitiu um mandado de prisão internacional contra Assange. Ele havia sido interrogado, três meses antes, sob suspeita de agressão e estupro contra uma mulher no país. Assange negou as acusações e afirmou que, caso ele fosse preso em território sueco, ele seria extraditado para os Estados Unidos por ter publicado os documentos do governo americano. Assange se entregou para a polícia do Reino Unido em dezembro de 2010 mas foi libertado dez dias depois após pagamento de fiança. Não tendo sido bem sucedido na contestação do processo de extradição, ele violou os termos da sua fiança em junho de 2012 e fugiu. Foi concedido a ele asilo político na embaixada do Equador em Londres, em agosto de 2012, e lá permaneceu até abril de 2019. Entre 2017 e 2019, Assange deteve cidadania equatoriana. Eventualmente, as autoridades suecas encerraram a investigação no caso de estupro e revogaram seu pedido de prisão europeu ainda em 2017. A polícia de Londres, contudo, afirmou que caso Assange deixasse a embaixada, seria preso imediatamente. Durante as primárias do Partido Democrata para a eleição presidencial nos Estados Unidos em 2016, o WikiLeaks revelou diversos emails da candidata Hillary Clinton do seu servidor privado na época que ela era Secretária de Estado. Os Democratas, junto com analistas e especialistas em cibersegurança, afirmaram que órgãos de inteligência da Rússia haviam hackeado os emails de Hillary e então entregado estas informações para o WikiLeaks; Assange consistentemente negou qualquer associação ou colaboração com o governo russo. Em 27 de julho de 2018, o presidente equatoriano, Lenín Moreno, afirmou que havia iniciado conversas com autoridades britânicas para remover o direito de asilo de Assange. Isso se concretizou em 11 de abril de 2019 e então a polícia de Londres, com a serventia do governo equatoriano, entrou na embaixada do Equador e prendeu Julian Assange. Sua prisão dividiu opiniões pelo mundo, com muitos exortando o acontecimento devido as supostas conexões de Assange com o governo russo para conspirar contra nações ocidentais, enquanto outros afirmam que a prisão dele viola o direito internacional e seria um atentado contra a liberdade de informação. (...)"

4.1 Resumo produzido através de abordagem superficial utilizando aprendizagem de máquina

O SummaryBot [13] é uma ferramenta de sumarização *online* que pode ser acoplada ao site de gerenciamento de turmas Slack. Ao adicionar a extensão do SummaryBot ao Slack, é possível enviar uma mensagem de requisição de resumo e, após alguns instantes, será retornada uma mensagem com o resumo pretendido.

O SummaryBot utiliza, além de todas as técnicas mais conhecidas e difundidas de processamento de linguagem natural, a técnica de blockchain[4] para auxiliar na construção dos resumos. Com o blockchain, o SummaryBot é capaz de aprender e validar diversos resumos remotamente. Após as análises de diversos resumos, a capacidade de gerar resumos novos vai se tornando mais consistente e de melhor aceitação por leitores humanos.

Ao resumir o trecho do artigo do wikipedia sobre Julian Assange, o SummaryBot retornou o resultado apresentado a seguir.

"Julian Paul Assange nascido Julian Paul Hawkins; 3 de julho de 1971) é um ativista australiano, programador de computador, jornalista e fundador do site WikiLeaks.Atualmente ele está sob custódia da Polícia Metropolitana de Londres após ser preso em 11 de abril de 2019, sob a acusação de ter violado as condições estabelecidas na sua fiança em 2010. Assange fundou o site WikiLeaks em 2006 e ganhou atenção internacional em 2010 quando o site publicou uma série de documentos sigilosos do governo dos Estados Unidos que haviam sido vazados por Chelsea Manning (na época chamada Bradley Manning).Entre os vazamentos estavam dados sobre o ataque aéreo a Bagdá em 12 de julho de 2007,[3] os registros de guerra do Afeganistão e do Iraque e o CableGate (novembro de 2010). Após os vazamentos de 2010, autoridades dos Estados Unidos começaram uma investigação criminal sobre o WikiLeaks e pediu apoio a nações aliadas pelo mundo. Em novembro de 2010, a Suécia emitiu um mandado de prisão internacional contra Assange. Assange negou as acusações e afirmou que, caso ele fosse preso em território sueco, ele seria extraditado para os Estados Unidos por ter publicado os documentos do governo americano. A polícia de Londres, contudo, afirmou que caso Assange deixasse a embaixada, seria preso imediatamente. Os Democratas, junto com analistas e especialistas em cibersegurança, afirmaram que órgãos de inteligência da Rússia haviam hackeado os emails de Hillary e então entregado estas informações para o WikiLeaks; Assange consistentemente negou qualquer associação ou colaboração com o governo russo. Isso se concretizou em 11 de abril de 2019 e então a polícia de Londres, com a serventia do governo equatoriano, entrou na embaixada do Equador e prendeu Julian Assange.Sua prisão dividiu opiniões pelo mundo, com muitos exortando o acontecimento devido as supostas conexões de Assange com o governo russo para conspirar contra nações ocidentais, enquanto outros afirmam que a prisão dele viola o direito internacional e seria um atentado contra a liberdade de informação."

Analisando atentamente o resultado obtido, através da sumarização do SummaryBot, é possível perceber que não foi realizada uma sumarização abstrativa. A inteligência artificial em questão se limitou apenas a extrair os segmentos de texto que se mostraram mais relevantes para o resumo. E, por mais que a técnica de sumarização extrativa traga bons resultados em um texto

[4] bases de registros e dados distribuídos e compartilhados

meramente informativo, em resumo de textos do gênero lírico, por exemplo, a falta de abstração de ideias pode acarretar em uma ausência de lógica textual.

Apesar de não abstrair a ideia principal do texto, o SummaryBot conseguiu ranquear de maneira aceitável as sentenças e ordená-las de modo cronológico. Os eventos citados no texto são realmente relevantes para se saber quem é Julian Assenge e quais fatos foram marcantes em sua militância na Web.

No resumo produzido pelo SummaryBot não foram encontrados problemas relevantes de coesão, já que o método utilizado foi o resumo extrativo. Também não foram encontrados problemas significativos de pontuação, já que o Bot extraiu apenas frases completas até o sinal de ponto final. Apesar disto, o SummaryBot não conseguiu ordenar as sentenças com a gradação adequada. Deste modo, as frases citadas pelo resumo parecem ser apenas frases soltas e levemente desconexas. A partir deste fenômeno, a falta de gradação de ideias, torna-se possível observar a falta de coesão citada por Hutchins [4] acerca de resumos gerados por métodos superficiais de sumarização.

Na Figura 1, extraída do banco de dados do SummaryBot, é possível notar com a inteligência artificial classificou a relevância de cada sentença chave.

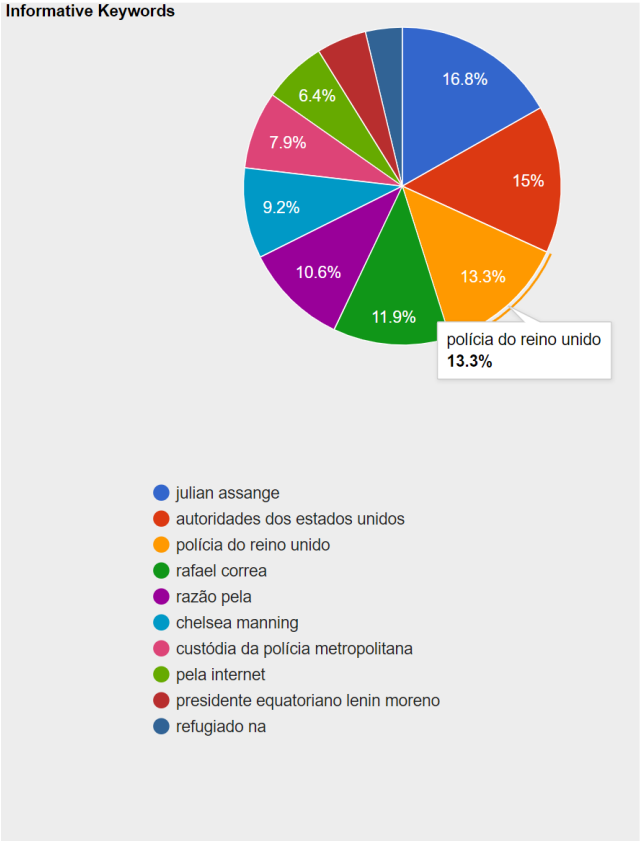

Figura 1: Sentenças chave para o SummaryBot.

4.2 Resumo produzido com abordagem superficial através de métodos estatísticos

O genSim [20] é uma biblioteca para processamento de linguagem natural que possui diversos recursos de processamento de textos, como tokenização, ranqueamento de sentenças e até mesmo sumarizações de trechos de texto. Para produzir o resumo abaixo, foi utilizada a função "sumarizer", que utiliza uma técnica estatística conhecida como TextRank em sua execução. O produto gerado pelo GenSim, ao se inserir o trecho do texto do wikipedia sobre Julian Assenge, é apresentado a seguir.

"Em novembro de 2010, a Suécia emitiu um mandado de prisão internacional contra Assange. Ele havia sido interrogado, três meses antes, sob suspeita de agressão e estupro contra uma mulher no país. Assange negou as acusações e afirmou que, caso ele fosse preso em território sueco, ele seria extraditado para os Estados Unidos por ter publicado os documentos do governo americano. Assange se entregou para a polícia do Reino Unido em dezembro de 2010 mas foi libertado dez dias depois após pagamento de fiança. Em 27 de julho de 2018, o presidente equatoriano, Lenín Moreno, afirmou que havia iniciado conversas com autoridades britânicas para remover o direito de asilo de Assange. Isso se concretizou em 11 de abril de 2019 e então a polícia de Londres, com a serventia do governo equatoriano, entrou na embaixada do Equador e prendeu Julian Assange."

O texto analisado foi um texto expositivo que, segundo muitos autores, é o tipo de texto mais fácil para se fazer um resumo automático. Mesmo para um texto de fácil extração de sentenças, o GenSim foi incapaz de selecionar frases que fossem capazes de introduzir ao leitor as características essenciais do Assange. O GenSim "decidiu" começar seu resumo de maneira não muito lógica devido à provável falta de ranqueamento preciso do sumarizador. Em comparação direta com o SummaryBot, o GenSim não conseguiu extrair frase com a relevância correta e com cronologia correta. Provavelmente, esse fenômeno ocorreu devido à deficiência das heurísticas que ranquearam e validarem as escolhas das sentenças relevantes.

A partir de uma análise de dados rápida utilizando o Keaggle com python, é possível avaliar mais detalhadamente como o GenSim definiu a escolha de sentenças para compor seu resumo. Com a ajuda de outras funções da biblioteca GemSim, foram geradas nuvens de palavras que contêm os termos que mais se destacaram, baseado em frequência, para o sumarizador (Figura 2).

Figura 2: Nuvem de palavras para o resumo Assange utilizando o GemSim

A nuvem de palavras é uma forma de visualização de dados que mostra que as palavras "Governo", "Londres", "Assange" e "WikiLearks" tiveram mais aparições e, portanto, serviram de gatilho para a escolha das sentenças chave do sumarizador automático.

Em uma segunda visualização, é possível observar quais são os termos que tiverem as frequências de aparição mais e elevadas.

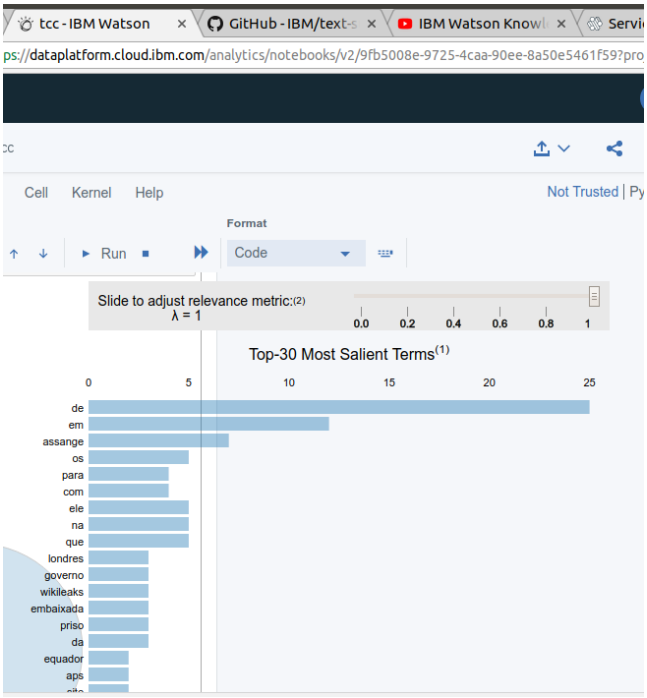

Figura 3: Frequência das palavras para o resumo Assange utilizando o GemSim.

Conforme visualização, apresentada na Figura 3, é possível notar que algumas palavras, que não possuem relevância para o sentido geral do texto, foram destacadas como frequentes. Os termos "de" e "em" apareceram mais frequentemente do que a palavra "Assange", que é o nome do personagem principal do texto. Este é mais um dos desafios de um sumarizador automático: conseguir identificar quais das palavras destacadas como relevantes realmente possuem relevância.

4.3 Sumarização abstrativa através da abordagem profunda

O EazyMind é um sumarizador *online* capaz de utilizar aprendizagem de máquina para realizar resumos abstrativos [2]. Os dois exemplos anteriores foram resumos extrativos de texto,, em que o sumarizador é capaz apenas de ranquear as sentenças mais importantes e remontá-las de maneira cronológica.

A área que trata de resumos abstrativos automáticos ainda é muito obscura. A maioria dos algoritmos criados nesse tipo de abordagem é deficiente e incapaz de criar sentenças completamente lógicas e coesas. A partir de uma análise do resumo gerado pelo EazyMind, é possível constatar quais os vieses criados pelos algoritmos de aprendizagem no momento da geração do resumo produto.

O EazyMind utiliza principalmente o algoritmo seq2seq (algoritmo que trabalha com redes neurais e converte sequências de dados codificados como entrada em sequências decodificadas na saída) em conjunto com a biblioteca TensorFlow e funciona apenas para a língua Inglesa. Devido a este fato, foi realizada uma tradução do texto, utilizado neste artigo, para a obtenção dos resumos em inglês e, após a sumarização, foi realizada uma nova tradução para o português. Ao realizar a tradução, foi tomado o devido cuidado para que os erros cometidos pelo sumarizador não fossem confundidos com erros de tradução. A seguir, é apresentado o resumo obtido através da sumarização abstrativa no EazyMind.

"julian paul assange nascido julian paul hawkins julho é um programador de computador ativista australiano julho está sob a custódia da polícia metropolitana de Londres depois de ser preso em abril sob a acusação de violar a polícia.

Assange, em seguida, conhecido como bradley manning os vazamentos incluídos dados sobre o ataque aéreo em Bagdá e Iraque e Cablegate novembro seguindo o site publicou uma série de documentos furtivos nos governo que tinham sido vazados por chelsea manning.

Mandado de detenção internacional se rendeu três meses sob suspeita de estados por ter publicado documentos do governo nos EUA entregou a polícia do Reino Unido em dezembro a Suécia emitiu um estupro contra uma mulher no país assange negou as acusações e disse que se ele foi preso na Suécia.

O secretário de cibersegurança disse que as agências de inteligência na Rússia hackearam os e-mails de Hillary e então entregaram essas informações para o Wikileaks. De modo consistente, negaram qualquer associação ou colaboração com os e-mails do governo russo.

O presidente da Líbia, Lenin Moreno, disse que iniciou conversações com autoridades britânicas para remover o direito de asilo contra as nações ocidentais, enquanto outros afirmam que sua

prisão viola a lei internacional e seria um ataque à liberdade de informação."

Pode-se observar no primeiro parágrafo, obtido pelo resumo abstrativo, que o sumarizador confundiu o significado dos meses. O EazyMind inseriu a palavra "julho" de maneira desconexa ao longo do parágrafo. No segundo parágrafo, é possível notar que mesmo com alguns erros de sintaxe, o sumarizador conseguiu extrair corretamente a informação de que houve documentos vazados através de Chelsea Manning.

Após o segundo parágrafo, é possível notar que o sumarizador alterna entre erros sintáticos e semânticos ao tentar construir sua abstração. Para conhecedores do texto original é fácil reconhecer a ideia principal do resumo gerado pelo EazyMind, mas para leitores que conhecem pouco sobre o tema, fica difícil identificar a ideia base do resumo.

Uma das principais dificuldades do EazyMind é validar as sentenças geradas semântica e sintaticamente. Talvez a falta de treinamento adequado tenha prejudicado a capacidade de reconhecimento do significado de determinadas palavras e, muito possivelmente, estes erros foram potencializados devido à falta de supervisão no treinamento. Não existe um algoritmo que determine com precisão se uma frase faz sentido ou não. Portanto, não é possível obter uma aprendizagem não supervisionada ideal para sumarizadores abstrativos.

Outro fator importante a destacar é a falta de um banco de dados de palavras da língua inglesa. Talvez um dicionário de sinônimos ajudasse o sumarizador a utilizar a melhor palavra para sentenças específicas sem que o sentido original fosse perdido. O EazyMind é incapaz de utilizar palavras que não estão contidas no texto alvo do resumo e, devido a este fato, o resumo produto será carente de conectivos e conjugações adequados na formulação das sentenças.

5 Análise dos sumarizadores

A partir dos resumos apresentados, foi possível observar que nenhum dos sumarizadores foi capaz de re-escrever o texto com base na sua ideia principal, de maneira coesa e coerente. Todos os três sumarizadores analisados neste artigo, independentemente de utilizarem abordagem superficial ou profunda, realizaram pequenos cortes no texto alvo do resumo e selecionaram as frases que julgaram como sendo mais importantes. Ambos os textos apresentaram problemas de coerência, com ressalvas para o método profundo, que selecionou frases mais conexas do texto e tentou criar suas próprias sentenças, buscando conectar as sentenças através de vírgulas ao invés de pontos. Mas, vale ressaltar que o texto resumido na Seção 4 é meramente informativo e facilitou bastante o trabalho dos sumarizadores. Em um poema, por exemplo, os problemas de coesão e coerência seriam agravados.

6 Desafios e limites no processo de sumarização automática

Diversos conceitos básicos para um humano, do ponto de vista de produção do conhecimento, podem parecer bastante obscuros para uma máquina. Para gerar um bom resumo, de forma automática, faz-se necessário que o sumarizador consiga reconhecer e reproduzir as ideias principais do texto. Nos últimos 70 anos, isso não tem se mostrado uma tarefa fácil.

Segundo Martins et al. [7], existem dois fatores principais que distinguem um sumarizador automático de um sumarizador humano, a saber: a) o domínio do assunto específico, que o sumarizador detém para entender e, portanto, abstrair ou generalizar as informações que lê do texto-fonte; b) o conhecimento prévio que esse possa ter, como experienciador nesse domínio.

Vale salientar, também, que sumários produzidos por humanos remetem ao mesmo texto-fonte, mesmo quando contêm informações diversas ou adicionais, em relação ao texto-fonte. Tais fatos evidenciam a necessidade de se importar um conhecimento previamente adquirido, externo ao texto, para que o resumo possa ser escrito e possa ter mais aceitabilidade.

Mesmo que as considerações sobre a ideia principal do texto sejam bastante relevantes para o entendimento do processo de sumarização, essas são de difícil abstração para um computador, pois contêm um alto nível de subjetividade e requerem uma representação bastante complexa do conhecimento do mundo.

Segundo Martins et al. [7], em geral, os modelos explorados para a sumarização automática se baseiam no conteúdo explícito dos textos-fonte e em suas características estruturais, quando se baseiam em metodologias profundas. Quando contemplam técnicas superficiais, baseiam-se em suas características estruturais.

Para Sampson [12], a sumarização automática é baseada em teorias linguísticas formais. Entendido dessa forma, o sistema computacional deveria simular a inteligência humana, para proporcionar um processamento eficiente da língua. Porém, a grande maioria de pesquisadores entende que a tarefa de simular a inteligência humana para um domínio aberto, no PLN, ainda está fora do alcance.

O processo de sumarização automática, de modo geral, possui problemas e desafios que, independente da técnica utilizada, são recorrentes. Alguns dos principais problemas, segundo Pardo [23], que precisam ser sanados em um sumarizador automático são por descritos a seguir.

7 Problemas da sumarização

(i) Em sumarizações com vários documentos: limitações ou inexistência de técnicas capazes de avaliar a redundância de uma informação; a fragilidade da avaliação automática da época a qual uma informação se refere; a diferença que existe de gramaticalidade entre textos distintos, que pode influenciar no resultado final; a diferença que pode haver na extração de significados de textos escritos em línguas diferentes; e a coreferência entre os textos.

(ii) Em sumarizações com único documento: realizar um resumo de qualidade para um texto em diversas línguas; não existe padrão para identificar palavras-chave de qualidade; e as palavras-chave podem variar em textos referentes ao mesmo assunto.

8 Solução para os problemas da sumarização

Diante das situações apresentadas, é possível observar que não existe uma técnica singular de sumarização automática, que consiga produzir um resumo com qualidade aceitável em todos os casos. Existem diversos cenários, com situações específicas, que exigem do algoritmo habilidades específicas. Otimizar uma técnica específica para tentar sanar estes problemas não vem se mostrando confiável ao longo das últimas décadas.

Como afirmam Martins et al. [7], "houve uma grande evolução no campo da sumarização automática e áreas relacionadas, tais como interpretação ou geração textual. Entretanto, ainda há muitos problemas que precisam ser solucionados, para que a sumarização automática de textos seja plenamente realizada, desde a extração de textos até a condensação de conteúdo.". A partir de diversas análises do estado da arte, é possível afirmar que apesar dos avanços na área, a sumarização automática de textos ainda é um campo de estudo que precisa se consolidar. Cada método de sumarização apresenta tanto méritos quanto dificuldades.

Para tentar resolver os problemas supracitados, o ideal é recorrer a técnicas híbridas de sumarização. Técnicas baseadas em abordagens profundas ainda são incapazes de abstrair a ideia principal de um texto de maneira satisfatória. Esses vieses são prejudiciais para o resumo produto da execução do algoritmo. Para mitigar ou corrigir cada um desses vieses, é recomendável que se desenvolva heurísticas com abordagens superficiais para que se possa identificar com maior precisão quais partes do texto merecem maior atenção.

REFERÊNCIAS

[1] CHIEN, L. F. PAT Tree-Based Keyword Extraction for Chinese Information Retrieval. Taipei, Taiwan, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.2087&rep=rep1&type=pdf.
[2] Eazymind, http://eazymind.herokuapp.com/arabic_sum.
[3] ERCAN, G., CICEKLI, I., "Using lexical chains for keyword extraction, Information Processing & Management," vol. 43 (6), 2007, pp. 1705- 1714

[4] HUTCHINS, J., "Summarization: Some Problems andMethods," Proc. Informatics 9: Meaning—The Fron-tier of Informatics, K.P. Jones, ed., Aslib, London, 1987

[5] LITVAK, M., LAST M., "Graph-based keyword extraction for singledocument summarization," in: Proceedings of the workshop on Multisource Multilingual Information Extraction and Summarization, ACL, 2008, pp. 17-24.

[6] LLORET, E., PALOMAR, M. Text summarisation in progress: a literature review, artificial Intelligence Review, vol. 37(1), 2012, pp.1-41.

[7] MARTINS, C. B., et al.. Introdução a sumarização automática de textos. Disponível em:<http://conteudo.icmc.usp.br/pessoas/taspardo/RTDC00201-CMartinsEtAl.pdf>. Acesso em 25 Mai. 2019.

[8] MIN, Z. L., CHEW, Y. K., TAN, L. "Exploiting category-specific information for multi-document summarization," in Proceedings of COLING, ACL, 2012, pp. 2093–2108.

[9] MIRROSHANDEL, S. A., GHASSEM-SANI, G. "Towards unsupervised learning of temporal relations between events," Journal of Artificial Intelligence Research.

[10] QASSEMA, L. M. A., et al. Automatic Arabic Summarization: A survey of methodologies and systems. Procedia Computer Science 117 (2017) 10–18. , https://www.sciencedirect.com/science/article/pii/S1877050917321452.

[11] SALTON, G., BUCKLEY, C., ALLAN, J. Automatic structuring of text files, http://cajun.cs.nott.ac.uk/wiley/journals/epobetan/pdf/volume5/issue1/ep056gs.pdf.>, Acesso em 10 Abr. 2019.

[12] SAMPSON, G. (1987). Probabilistic models of analysis. In G. Leech, R. Garsude and G. Sampson (eds.), The computational analysis of English, pp. 16-29. Longman. Harrow.

[13] Summarizebot. Summary. https://www.summarizebot.com/api/ce576cc2dcb7443298ce5252c97e344d.html

[14] Joshi, Prateek. An Introduction to Text Summarization using the TextRank Algorithm,https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/.

[15] THOMAS, J. R., BHARTI, S. K., BABU, K. S., "Automatic keyword extraction for text summarization in e-newspapers," in: Proceedings of the International Conference on Informatics and Analytics, ACM, 2016, pp. 86-93.

[16] TURNEY, P. Learning to Extract Keyphrases from Text. Canada. Disponível em:<http://cogprints.org/1802/5/ERB-1057.pdf>, Acesso em 5 Mai. 2019.

[17] WITTEN, n achei no artigo I. H., et al... KEA: Practical Automatic Keyphrase Extrantion. Hamilton, New Zealand. Saskatoon, Canada. NY, USA.Disponível em:<https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-GWP-etal-KEAPractical.pdf.

[18] Helena, Lucia, et al A Sumarização Automática de Textos: Principais Características e Metodologias*, http://conteudo.icmc.usp.br/pessoas/taspardo/JAIA2003-RinoPardo.pdf

[19] Kumar, Automatic Keyword Extraction for Text Summarization: A Survey,https://arxiv.org/ftp/arxiv/papers/1704/1704.03242.pdf

[20] Gensim, https://radimrehurek.com/gensim.

[21] Brandel, Camila Martins , INTRODUÇÃO À SUMARIZAÇÃO AUTOMÁTICA, http://conteudo.icmc.usp.br/pessoas/taspardo/RTDC00201-CMartinsEtAl.pdf.

[22] Maybury, M. (1993). Automated Event Summarization Techniques. In: Seminar Report of Summarizing Text for Intelligent Communication Seminar. Dagstuhl, Germany.

[23] Pardo, Alexandre Thiago Salgueiro, Sumarização Automática: Principais Conceitos e Sistemas para o Português Brasileiro, http://conteudo.icmc.usp.br/pessoas/taspardo/NILCTR0804-Pardo.pdf .

[24] Sahoo, Deepak et al. , Hybrid Approach To Abstractive Summarization , https://reader.elsevier.com/reader/sd/pii/S1877050918307701?token=D5C553A7347250B70F21B59A1CCDDE0D84808E1CC2D69D9F925FFDEBABADC5FEE85DDB0C5A4D1435CCC04AEBFBDD0BF8.