UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Rafaella Ferreira do Vale

Análise comparativa de métodos de simplificação de sentenças para sumarização de textos

Recife

2017

Rafaella Ferreira do Vale

Análise comparativa de métodos de simplificação de sentenças para sumarização de textos

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Ciência da Computação

Orientador: Prof. Dr. Rafael Ferreira Leite de Mello

Recife

2017

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por Rafaella Ferreira do Vale às 14 horas do dia 25 de agosto de 2017, no Auditório do CEAGRI-02 – Sala 07, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **Análise comparativa de métodos de simplificação de sentenças para sumarização de textos**, orientado por Rafael Ferreira Leite de Mello e aprovado pela seguinte banca examinadora:

Rafael Ferreira Leite de Mello
DEINFO/UFRPE

Rafael Dueire Lins
DEINFO/UFRPE

Adenilton José da Silva
DEINFO/UFRPE

Agradecimentos

Agradeço aos meus professores do ensino médio, que acreditaram em mim e ainda influenciam na minha visão do mundo. Não esquecerei da participação que tiveram na minha vida durante esta fase. Agradeço também aos meus professores da graduação. Em especial, ao professor da disciplina de Cálculo II, Renato Teixeira, que viu potencial em mim e me impulsionou a desenvolvê-lo; ao professor Wilson de Oliveira, que muito contribuiu na minha formação acadêmica; ao professor Adenilton Silva, que confiou na minha capacidade e me encorajou a tomar decisões importantes para o meu futuro; e por último mas não menos importante, ao professor Rafael Ferreira, que me aceitou como orientanda mesmo no momento tardio em que o procurei, pela sua paciência e por também acreditar em mim.

Resumo

A sumarização automática de textos é uma ferramenta valiosa para lidar com grande quantidade de informação, sendo útil para filtrar conteúdo relevante com esforço humano reduzido. Entretanto, abordagens extrativas de sumarização, onde o sumário é composto pela cópia integral de um conjunto de frases do texto original, possuem limitações, podendo não capturar a informatividade de um texto. Uma possível estratégia para melhorar o conteúdo dos sumários gerados é usar simplificação de sentenças. Esta pesquisa tem como foco a aplicação da tarefa de simplificação de sentenças como parte do pré-processamento da sumarização extrativa de textos, bem como comparar diferentes métodos de simplificação para este propósito. Como objetivo, procura-se descobrir se, na realização de sumarização extrativa, a inclusão do processo de simplificação de sentenças é favorável à informatividade dos resumos obtidos como resultado. Além disso, o trabalho visa investigar o comportamento das técnicas de simplificação mais adequadas para esse tipo de aplicação. Para isso, uma análise comparativa de três métodos de simplificação de sentenças é efetuada utilizando 15 métodos de sumarização aplicados a um *corpus* de 1038 textos de notícias na língua inglesa. Como referência, a sumarização também é executada no corpus não simplificado e, adicionalmente, no *corpus* com *stop words* filtradas. Os resultados de *recall*, precisão e *F-score* em geral não exibem vantagem dos métodos de simplificação na aplicação dos métodos de sumarização individualmente, porém sugerem que métodos que consideram propriedades linguísticas e preservação gramatical obtêm melhor desempenho do que aqueles que não têm essas características.

Palavras-chave: processamento de linguagem natural, simplificação de sentenças, sumarização extrativa de textos.

Abstract

Automatic text summarization is possibly a valuable tool in extracting information from the Internet and digital libraries nowadays, proving itself useful to filter relevant content with reduced human effort. Nevertheless, extractive summarization approaches have limitations, possibly not fully capturing the informativeness of a text. A potential strategy to tackle this problem is to use sentence simplification. This work focuses on the application of the sentence simplification task as a preprocessing step for extractive text summarization, and on comparing different methods of simplification for such purpose. This work seeks to answer the question of whether sentence simplification increases the informativeness of extractive summaries. Furthermore, if considered valid, this work aims to point at the most adequate simplification techniques for such an application. A comparative analysis between three sentence simplification methods is performed using 15 summarization approaches applied to a corpus of 1038 news texts in the English language. The summarization process is also performed directly on the original corpus and, additionally, on the corpus with stop words removed. The results obtained for recall, precision and F-score showed no advantage in simplifying sentences when applying summarization methods individually. However, they suggest that the methods that take into account linguistic features and grammaticality obtain better performance than the other ones.

Keywords: natural language processing, sentence simplification, extractive text summarization.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

PLN	Processamento de linguagem natural
LC	Linguagem controlada
PBMT	*Phrase-based machine translation*
POS	*Part-of-speech*
ROUGE	*Recall-oriented understudy for gisting evaluation*
BLEU	*Bilingual evaluation understudy*
SVC	Sujeito, verbo e complemento (método de compressão de sentenças)
RT	*Reluctant Trimmer* (método de compressão de sentenças)
ILP	*Integer linear programming*

Sumário

1 Introdução

O grande volume de dados disponíveis na Internet e em bibliotecas digitais trouxe o desafio de manipular o que é mais pertinente, descartando o que não tem utilidade informativa. No contexto da Internet, por exemplo, podemos encontrar uma abundância de textos em notícias, *blogs*, fóruns, entre outros. Em outras circunstâncias, há livros, artigos científicos, documentos, etc., e o empenho humano torna-se fundamental para lidar com a expansiva quantidade de informação (GUPTA; LEHAL, 2010).

Segundo Gupta e Lehal (2010), a sumarização extrativa de textos é uma técnica que possibilita a filtragem de conteúdo relevante, além de diminuir o esforço humano para a produção de resumos. Contudo, este tipo de procedimento possui limitações, entre elas na forma de decidir quais sentenças melhor condensam o assunto tratado em um texto, podendo comprometer a qualidade dos resumos produzidos. Diversas estratégias apresentadas na literatura tratam desse problema (GUPTA; LEHAL, 2010; DAS; MARTINS, 2007).

Uma opção é a simplificação de textos. Particularmente, a simplificação de sentenças é estudada para auxiliar na compreensão para certos grupos de leitores, como não nativos de uma língua ou pessoas com problemas cognitivos (SIDDHARTHAN, 2014). Entretanto, no setor industrial, é possível encontrar instâncias de uso da simplificação de sentenças na sintetização de manuais técnicos, por exemplo (O'BRIEN, 2003; SIDDHARTHAN, 2014). Esta é uma tarefa que possui escopo abrangente, apontando para novas possibilidades de aplicação, como em assitência à sumarização de textos.

Neste trabalho, métodos de simplificação de sentença são comparados quantitativamente quanto ao seu desempenho ao usar diversas abordagens de sumarização extrativa. Com este estudo, pretende-se descobrir se existe proveito a ser tirado da aplicação de simplificação como um passo anterior à sumarização de textos. Além disso, pretende-se analisar as características dos métodos que trazem melhores resultados, indicando haver melhor condensação de conteúdo.

1.1 Pergunta de pesquisa

Este trabalho busca responder a seguinte pergunta: a utilização de técnicas de simplificação de sentenças traz melhorias à informatividade dos resumos produzidos por sumarização extrativa de textos?

Obstáculos que surgiram na área de sumarização extrativa não foram completamente solucionados. Uma qualidade importante em um resumo, além da fluência linguística, é a alta informatividade sobre o texto original. Neste momento, a fluência não está no escopo

deste trabalho, mas com a simplificação de sentenças supõe-se que pode haver benefícios na concentração de conteúdo informativo em textos sumarizados. Adicionalmente, espera-se que o produto deste trabalho sirva como guia para auxiliar tentativas futuras de atacar o problema da sumarização extrativa.

1.2 Objetivos

Nesta seção estão expostos os objetivos almejados com a realização deste trabalho.

1.2.1 Objetivo geral

Efetuar uma análise comparativa quanto à preservação de informatividade de métodos de simplificação de sentenças aplicados como pré-processamento na tarefa de sumarização de textos.

1.2.2 Objetivos específicos

1. Aferir a adequação da estratégia de simplificação como parte do pré-processamento de sumarizadores automáticos.
2. Inferir as propriedades de abordagens de simplificação de sentenças que produzem melhores resultados de sumarização.

1.3 Organização do trabalho

O restante deste trabalho está disposto da seguinte maneira: o Capítulo 2 apresenta aspectos teóricos envolvidos em processamento de linguagem natural e ligados ao tema deste trabalho; no Capítulo 3, são discutidos outros trabalhos que têm interseção de ideias com este, principalmente em se tratando de abordagens envolvendo simplificação de sentenças; o Capítulo 4 detalha a metodologia utilizada para trabalhar com as estratégias de simplificação de sentenças, aplicá-las aos sistemas de sumarização de textos e realizar análisa qualitativa dos resultados; o Capítulo 5 examina os resultados obtidos por cada sistema de simplificação em diferentes métodos de sumarização, discutindo e comparando as diferentes características dos sistemas diante das estatísticas observadas; o Capítulo 6 expõe as considerações finais e expectativas para o futuro da pesquisa.

2 Fundamentação teórica

Neste capítulo, a área de processamento de linguages naturais é contextualizada. Também são apresentadas noções básicas necessárias para o embasamento desta pesquisa, que trata de simplificação de sentenças e sumarização de textos.

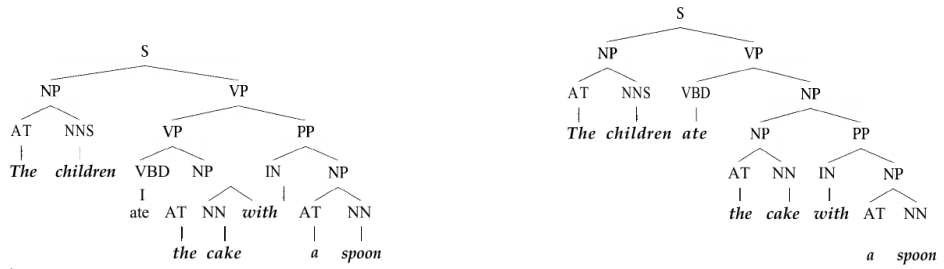
2.1 Processamento de linguagem natural

A área de processamento de linguagens naturais (PLN) "explora como computadores podem ser usados para entender e manipular texto ou fala em linguagem natural para fazer coisas úteis" ((CHOWDHURY, 2003), tradução nossa) a partir deles. PLN compreende uma grande variedade de estratégias que lidam com esse assunto, envolvendo ortografia, gramática, semântica e outras disciplinas linguísticas, e também fundamentando-se em áreas da matemáticas, como probabilidade e estatística, entre outros tópicos de outras áreas. Este trabalho lidará com textos, portanto PLN será apresentada somente do ponto de vista de linguagens naturais em forma textual.

Sistemas computacionais que manipulam textos em linguagens naturais a fim de buscar informações ou produzir novos conteúdos possuem um arcabouço de recursos no campo de processamento de linguagem natural. A partir do texto não estruturado, através de análise sintática, conhecida em inglês como *parsing*, pode-se tentar determinar possíveis estruturas gramaticais de sentenças para se obter informações procuradas. Múltiplas representações sintáticas para uma mesma sentença ocorrem devido à ambiguidade inerente a linguagens naturais. Como é observado por Manning e Schütze (1999), quando a complexidade gramatical cresce, a tendência é que mais ambiguidades sejam geradas. Por esse motivo, é ideal que tais sistemas não dependam unicamente da análise sintática. A Figura 1 exibe um exemplo de duas árvores de dependências possíveis após a análise sintática de uma sentença ambígua.

O problema de desambiguação pode envolver outras atividades de PLN, como escolhas baseadas em sentido ou categoria de vocábulos, estrutura sintática e escopo semântico. Devido à natureza dinâmica da linguagem natural, sistemas baseados exclusivamente em regras também não conseguem ter um comportamento satisfatório diante desse obstáculo. Também é mostrado que estratégias estatísticas de PLN são uma boa forma de abordar esse problema, pois são capazes de trabalhar com relacionamentos entre termos e outras informações extraídas dos textos, fornecendo bom desempenho e generalização (MANNING; SCHÜTZE, 1999). Normalmente, métodos de PLN trabalham em cima de coleções de textos chamadas *corpus* (*corpora*, no plural), processando esses textos em vez de manifestações da linguagem em qualquer tipo de situação.

Figura 1 – Exemplo de ambiguidade sintática



(a) Árvore sintática passa a ideia de que a colher foi um instrumento usado para comer o bolo.

(b) Com esta estrutura sintática, a árvore sintática sugere que a colher acompanha o bolo.

Fonte: Manning e Schütze (1999)

Métodos estatísticos para PLN incluem uma gama de atividades, muitas delas exploradas por Manning e Schütze (1999). Simples estatísticas de frequência, por exemplo, podem ser utilizadas de diversas formas, sendo algumas elas para:

a) indicar as palavras mais comuns (ou raras);

b) indicar as categorias de palavras mais comuns (ou raras);

c) estabelecer uma relação entre a quantidade de palavras mais frequentes e a quantidade de palavras mais raras;

d) estabelecer uma relação entre a frequência de uma palavra e a sua quantidade de significados;

e) estabelecer uma relação entre palavras frequentes e as distâncias entre repetições das mesmas.

Esse tipo de análise pode não ser eficaz o suficiente para boa parte dos problemas, pois algumas classes de palavras como artigos, preposições, conjunções, etc. não trazem informação suficiente sobre o que é tratado no texto. Esse conjunto de palavras é denominado *stop words*, e é comum que seja eliminado de cada documento da base de dados durante a fase de pré-processamento.

Palavras ou conjuntos de palavras de interesse para o problema podem ser identificadas de forma mais eficiente se um sistema que usa métodos estatísticos para PLN implementar alguma estratégia para a busca de colocações. Uma colocação é uma expressão formada por duas ou mais palavras e é usada comumente em um idioma (MANNING; SCHÜTZE, 1999). Exemplos de colocações são “música clássica” e “dar uma chance”, em português, e *“fast food”* e *“give a presentation”*, em inglês. Processamentos desse tipo são úteis para atividades como extração de informações e sumarização de textos, pois colocações em linguagem natural são de uso frequente e, assim, ideias importantes do domínio do problema são preservadas.

Diversas outras técnicas estão envolvidas no processamento de linguagem natural em diversos níveis de complexidade e aplicação, esta podendo se estender a palavras, sentenças, textos completos ou vários documentos. Nas subseções a seguir serão introduzidas atividades importantes de PLN que são realizadas frequentemente na preparação do texto para outras atividades de escopo mais voltado para aplicações no mundo real. Sendo assim, as tarefas de simplificação e sumarização de textos serão apresentadas subsequentemente em suas próprias seções.

2.1.1 Segmentação textual

Para trabalhar com processamento de textos, é necessário delimitar palavras ou conjuntos de palavras potencialmente importantes e, muitas vezes, sentenças também devem ser identificadas individualmente. Esse tipo de tratamento é efetuado através da segmentação textual. Os exemplos de pedaços de textos mencionados são categorizados como segmentação de palavras e segmentação de sentenças.

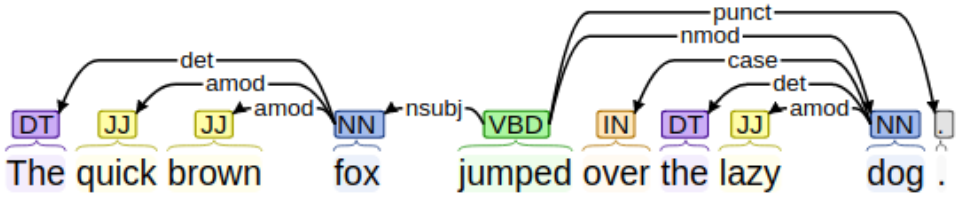
A segmentação de palavras é conhecida em inglês como *tokenization*, de modo que as unidades segmentadas são chamadas de *tokens*. Esse processo precede qualquer tipo de análise de uma linguagem natural, pois a princípio um texto puro (não processado) não apresenta informações linguísticas suficientes para aplicação da maioria das técnicas de PLN (MIKHEEV, 2003). O procedimento de segmentação de palavras pode ter um critério tão simples quanto o uso de espaços em branco como delimitadores, entretanto distinções importantes como pontuações ou entidades (por exemplo, reconhecer "*United States*" como um *token*) podem ser perdidas. Sendo assim, as unidades de segmentação nesse caso não são somente palavras, logo métodos mais avançados geralmente são necessários.

Quando a segmentação ocorre no nível de sentenças, a mesma é chamada *sentence splitting* ou quebra de sentenças. Apesar de ser mais simples que a segmentação de palavras, considerando que sinais de pontuação indicam o fim de uma sentença, é preciso eliminar ambiguidades que ocorrem em domínios específicos, como pontos em números decimais, siglas ou abreviações (MIKHEEV, 2003). A quebra de sentenças, assim como a segmentação de palavras, é uma atividade de pré-processamento usada antes de diversas aplicações de PLN, e a complexidade da técnica utilizada deve ser adaptada de acordo com a aplicação e seu domínio.

2.1.2 *Part-of-speech tagging*

Esta atividade consiste na categorização de palavras já segmentadas de acordo com suas determinadas classes gramaticais (*parts-of-speech* ou POS *tag*). Seu uso é predominante durante o pré-processamento, auxiliando a aplicação de outras tarefas como análise sintática, recuperação de informação e na anotação de bases de dados (VOUTILAINEN, 2003). Na

Figura 2 – *Part-of-speech tagging* e análise de dependências de uma sentença usando a ferramenta Stanford CoreNLP.



Fonte: http://corenlp.run/

Figura 2 encontra-se um exemplo de sentença cujos *tokens* estão etiquetados por POS *tags* como DT (artigo), JJ (adjetivo) e NN (substantivo). Essa informação é então usada para a criação da árvore de dependência, também presente na imagem através das relações entre os *tokens*. Pode-se ver, por exemplo, que a relação "*nsubj*" ocorre entre o verbo "*jumped*" e o substantivo "*fox*", indicando que o último é o núcleo do primeiro.

2.2 Simplificação de sentenças

A simplificação de textos tem o objetivo de reduzir a complexidade dos mesmos sem suprimir o conteúdo original e seu significado (SIDDHARTHAN, 2014). Sendo um termo abrangente, pode incluir metodologias com o objetivo de gerar textos com diversos graus de redução de complexidade, de forma totalmente manual, manual com sugestões de uma ferramenta ou totalmente automática. As transformações do texto podem ser de natureza léxica ou sintática.

A simplificação léxica normalmente se encarrega de executar substituições léxicas (como na Tabela 3) ou expansões de termos com explicações do seu significado, em geral buscando reduzir a complexidade para um certo tipo de leitor (SIDDHARTHAN, 2014).

Figura 3 – Exemplo de simplificação por substituição léxica

Once upon a midnight **dreary**, while I pondered, weak and **weary**

Once upon a midnight **sad**, while I pondered, weak and **tired**

Fonte: a autora

Por sua vez, a simplificação sintática busca reduzir a complexidade estrutural de sentenças, o que pode ser conseguido de diversas formas, como pode ser visto na Tabela 1. Ambos podem ser combinados de acordo com o escopo da aplicação. Neste trabalho opta-se por usar apenas a simplificação sintática, escolha que é justificada pela finalidade da sumarização de textos. O intuito é preservar o conteúdo principal, portanto, embora

Tabela 1 – Exemplos de simplificações de sentenças.

Original	Simplificada	Natureza da simplificação
I have read the books that you bought last week.	I have read the books. You bought the books last week.	Desagregação de orações
The infected rabid fox eventually dies, but a simple scratch can spread the virus to other animals or people.	The infected rabid fox eventually dies. A simple scratch can spread the virus to other animals or people.	Eliminação de conjunção e desagregação de orações
After the raid took place on Saturday around 8:00 pm, Ibrahim took a group of journalists to the site of the house.	After the raid took place, Ibrahim took a group of journalists to the site of the house.	Eliminação de expressão temporal

Fonte – Bawakid e Oussalah (2011)

modificar expressões para uma representação equivalente não necessariamente resulte na perda da essência do documento, este trabalho procura evitar distanciamento da forma como os termos do texto são apresentados. Além disso, a prioridade é investigar a influência das transformações sintáticas de sentenças nos resumos resultantes.

Em termos de aplicabilidade, como é sintetizado por Siddharthan (2014), sistemas de simplificação podem ser direcionados a grupos específicos de acordo com suas habilidades de leitura, assim como podem servir para aperfeiçoar os processos de sumarização de textos, extração de informação, entre outras tarefas de PLN. A seguir serão mencionadas algumas abordagens que requerem o uso de *software* para auxílio na simplificação.

Uma possível metodologia de simplificação de linguagem natural é a utilização de um conjunto de regras para linguagens naturais controladas (LCs), que visam gerar linguagem mais simples e menos ambígua. O'Brien (2003) propõe as seguintes categorias de regras de LCs:

a) léxicas: definem como selecionar palavras com base, por exemplo, na frequência de uso no vocabulário popular, entre outros fatores;

b) sintáticas: definem como a sintaxe é empregada, controlando, por exemplo, o uso de pessoa gramatical, os tempos verbais admitidos ou se a voz passiva é permitida;

c) textuais, subdivididas em regras de *estrutura textual* e *regras pragmáticas*:

   – regras de estrutura textual: estabelecem restrições sobre a estrutura do texto ou sobre as informações contidas no mesmo;

   – regras pragmáticas: usam artifícios que delimitam o propósito do texto para causar uma resposta específica no leitor.

Um conjunto de regras de LCs serve de diretrizes específicas para o tipo de texto,

sua intenção e seu público-alvo, a fim de guiar o escritor na elaboração do texto. Em particular, regras de LCs são mais comumente usadas com ferramentas para analisar o texto e dar sugestões de modificações. Apesar de não ser o único modo de simplificar sentenças, o uso de regras predefinidas é ideia frequentemente utilizada em sistemas de simplificação automática.

Outra abordagem que tem ganho evidência é a de tradução automática monolíngue (*monolingual machine translation*). A intenção dessa abordagem é considerar que a simplificação é como uma tradução de uma língua em forma mais complexa para a mesma língua, porém em forma mais simples (SIDDHARTHAN, 2014). Um exemplo que adota essa ideia é a tradução automática baseada em frases (originalmente *phrase-based machine translation* ou PBMT), que procura alinhar sequências de palavras sem levar em conta a sintaxe, e calcula a probabilidade de uma sequência (ou frase) ser traduzida como outra sequência.

2.3 Sumarização automática de textos

Mani (1999) define sumarização de textos como "o processo de destilar a informação mais importante de uma fonte (ou várias fontes) para produzir uma versão resumida para um usuário (ou usuários) e atividade (ou atividades) em particular" (tradução nossa). Pelo fato de existir necessidade de adequação do conteúdo e do formato do texto sumarizado para torná-lo compatível a um certo público-alvo, supõe-se que essas restrições possibilitam a tarefa de automatização do método por procedimentos computacionais. Isto é, à sumarização de textos estão atrelados fatores que limitam o alcance do texto para servir a um propósito específico, portanto é natural assumir que esse processo é praticável com interferência humana reduzida ou inexistente.

Os métodos mais antigos de sumarização automática se baseavam em características do texto, como frequência de palavras, posições de sentenças, o emprego de palavras sugestivas sobre o texto, etc (DAS; MARTINS, 2007; GUPTA; LEHAL, 2010). Já as abordagens modernas, além de métodos estatísticos, contam com modelos de aprendizagem de máquina. Das e Martins (2007) resumem as principais estratégias clássicas e modernas encontradas na literatura. Entre as técnicas modernas estão: modelo Naïve-Bayes, árvores de decisão, modelos ocultos de Markov (HMM), modelos log-linear e redes neurais artificiais.

Gupta e Lehal (2010) classificam a sumarização de textos como extrativa ou abstrativa. Essas categorias são definidas da seguinte forma:

a) extrativa: partes do texto, como sentenças ou parágrafos, são filtradas por meio de certas características para compor o resumo, algumas delas introduzidas na seção 4.3;

b) abstrativa: um novo texto é gerado a partir da compreensão dos conceitos do

texto original, usando métodos linguísticos para interpretação e reformulação de forma mais breve.

Com o uso de simplificação de sentenças, a sumarização de textos geralmente se encontra entre as classes extrativa e abstrativa. Métodos de simplificação mais voltados a aprendizagem de máquina, por exemplo, podem se aproximar de geração de linguagem, posicionando o tipo de sumarização que utiliza tais métodos mais próximo da classe abstrativa.

Esta pesquisa lida com sumarização extrativa. Para guiar a decisão de um sistema de sumarização extrativo, um critério de pontuação que atribui um peso às sentenças é utilizado. As sentenças que pontuam mais são selecionadas para o resumo resultante. A seção 4.3 discorrerá sobre os métodos de pontuação utilizados neste trabalho.

3 Trabalhos relacionados

Este capítulo é subdividido em seções de acordo com a finalidade de trabalhos que têm relação com este, os quais são descritos brevemente. Estes estudos, em geral, têm escopo voltado para a proposição de novos métodos de simplificação de textos. Em vez de avaliar o desempenho de métodos no que concerne à qualidade das simplificações produzidas, procurou-se investigar o resultado da combinação dos mesmos com abordagens de sumarização extrativa.

Os trabalhos descritos na seção 3.2 aplicam simplificação de sentenças à sumarização de textos. Como diferencial, este trabalho tem o intuito de comparar os diferentes métodos de simplificação precedendo a sumarização extrativa de textos, tomando como referência resumos dos textos originais escritos por humanos. Serão aproveitados os dois métodos apresentados por Angrosh, Nomoto e Siddharthan (2014), com introdução de uma estratégia de compressão baseada em relações gramaticais. O objetivo é mostrar, a partir da análise comparativa, se há evidências de vantagens associadas à realização de simplificação posteriormente à sumarização de textos, e no caso positivo, que propriedades dos métodos podem ter influência nesse resultado. Neste contexto, este trabalho se contrasta ao que é encontrado na literatura. Das pesquisas descritas a seguir, as que combinam simplificação de sentenças à sumarização automática não fornecem um direcionamento sobre técnicas adequadas que preservam informatividade dos textos.

3.1 Trabalhos direcionados a um público-alvo

Uma das possíveis metas da simplificação de sentenças é tornar a leitura menos complexa para pessoas com dificuldades por diversas razões, como transtornos relacionados à linguagem ou idade, ou para acomodação a um determinado contexto. Adotando o último objetivo, Daelemans, Höthker e Sang (2004) apresentaram dois métodos de simplificação de sentenças para a tarefa de legendagem automática em inglês e holandês. O primeiro é um método de aprendizagem de máquina, cujo modelo aprende de transcrições de programas de televisão. As ações de cópia, remoção e substituição de palavras são executadas durante a simplificação de uma sentença, levando em conta também a taxa de compressão da legenda em relação à sentença transcrita original. O segundo método utiliza uma abordagem baseada em regras de remoção de componentes de sentenças, novamente conformando-se à taxa de compressão. O primeiro modelo obteve desempenho acima do modelo de regras, porém abaixo do esperado de acordo com o sistema usado como referência para comparação. As duas estratégias foram combinadas, mas trouxeram desempenho parecido com o do método baseado em regras.

Por sua vez, a abordagem de Siddharthan (2006) possui características adequadas para a redução de complexidade da leitura. O estudo buscou explorar um problema que não havia sido considerado por trabalhos anteriores de simplificação textual. O sistema é dividido em três fases, começando pela análise, que funciona como uma etapa de pré-processamento. Depois disso, na fase de transformação são aplicadas regras de simplificação sintática construídas manualmente que tratam de conjunções, orações iniciadas por pronomes relativos e apostos. É adicionada uma etapa de pós-processamento que tem como meta algo não trabalhado até esses estudos, que é a coesão conjuntiva e de pronomes referentes a termos citados anteriormente.

Nunes et al. (2013) têm como objetivo explícito a diminuição da complexidade de textos de acordo com o público-alvo e seu nível de aprendizagem. Na abordagem utilizada são realizadas substituições de palavras por outras mais populares que se adequam ao contexto ou nível de aprendizagem desejado.

Angrosh, Nomoto e Siddharthan (2014) descreveram dois sistemas de simplificação textual. O primeiro é direcionado à tradução automática combinada com regras definidas manualmente, aplicando transformações em árvores de dependência. Sua construção teve como base o sistema proposto por Siddharthan (2011), que apenas usava regras pré-estabelecidas. Já o segundo escolhe uma de várias versões comprimidas de uma sentença de acordo com restrições estabelecidas. Ambos os sistemas, incluindo a composição dos dois, foram testados com falantes não nativos da língua inglesa, indicando que houve aumento de compreensão por parte de leitores menos habilidosos.

O sistema de Ferrés, Marimon e Saggion (2015) encadeou um simplificador léxico e outro sintático baseado em regras, gerando simplificações de sentenças com base em anotações feitas por análises precedentes.

Alguns trabalhos seguem uma abordagem vinda da tradução automática, entretanto considerando a simplificação de sentenças como um problema monolíngue. Seguindo essa linha, o modelo baseado em árvores sintáticas de Zhu, Bernhard e Gurevych (2010) foi proposto como uma técnica de simplificação de âmbito sintático e léxico, usando aprendizagem de máquina e aplicando uma série de operações à árvore de uma sentença. As operações de desagregação e de remoção servem o mesmo propósito que a desagregação e a compressão de sentenças empregadas por Bawakid e Oussalah (2011). Partes da sentença podem também ser reordenadas ou substituídas, podendo o último caso se aplicar a palavras ou trechos de sentenças. O modelo proposto é uma adaptação de modelos probabilísticos de tradução para aplicação em simplificação de sentenças, de modo que o resultado, em vez de uma tradução, é uma simplificação da sentença original.

De forma semelhante, Coster e Kauchak (2011) propuseram um esquema inspirado em um problema de tradução automática, tendo a operação de inserção como uma de suas distinções. Ainda no paradigma em questão, o modelo de Wubben, van den Bosch e

Krahmer (2012) opera usando tradução automática baseada em frases, contrastando com tradução (ou simplificação) de palavras individuais. A saída do modelo é escolhida a partir da dissimilaridade entre simplificações candidatas consideradas melhores e a sentença original, a fim de obter maior número de simplificações para a sentença e melhor qualidade no resultado.

Narayan e Gardent (2014) exploram uma abordagem híbrida formada por um modelo probabilístico para desagregação e remoção e um modelo de tradução monolíngue baseado em frases para substituição, reordenação, fluência e adequação gramatical. O trabalho tem como diferencial o fato de usar uma representação semântica em vez de árvore de dependências, favorecendo a ocorrência de transformações que dependem da semântica.

Os estudos de Štajner, Béchara e Saggion (2015) revelaram que, na simplificação de sentenças como um problema de tradução monolíngue, a quantidade de exemplos de treinamento não afeta o desempenho do sistema. Além disso, os experimentos realizados indicam que a escolha de exemplos de similaridade moderada contribui para o alcance de resultados gramaticalmente corretos que preservam o significado original e atingem um bom nível de simplificação.

3.2 Trabalhos direcionados à sumarização de textos

O ponto central deste trabalho é o uso de simplificação de sentenças como um processo que possibilite a obtenção de melhores resultados gerados por sumarização de texto. Compartilhando esse objetivo, Jing (2000) introduziu um sistema para reduzir sentenças, buscando resumos mais concisos e que expressassem bem as ideias centrais do texto original. Durante o funcionamento, o sistema procura respeitar as regras gramaticais e manter partes de sentenças que são mais importantes no contexto através de uma heurística. Um corpus que consiste de sentenças reduzidas por profissionais é usado para indicar as probabilidades de que fragmentos da sentença sejam eliminados, sendo mais um dos recursos usados para a tomada de decisão. Na época da publicação do trabalho, a maioria dos sumarizadores extraía sentenças importantes sem alterações, tornando o seu foco na redução de sentenças um diferencial.

Vanderwende et al. (2007) estenderam um sistema de sumarização extrativa ao anexar componentes para, entre outras atividades, a simplificação de sentenças. O método de simplificação é voltado para a compressão, com a finalidade de reduzir o resumo gerado. Os resultados produzidos por sumarização sugerem que a compressão de sentenças pode comprometer a gramática ou ocasionar a perda de conteúdo relevante ao assunto, mas em outros casos também é responsável por remover conteúdo redundante e encurtar o texto do resumo.

Bawakid e Oussalah (2011) desenvolveram um esquema de sumarização automática que contém um módulo de simplificação de sentenças composto pelos processos de desagregação e compressão de sentenças. No primeiro, sentenças são separadas em sentenças mais simples, quando necessário, segundo regras predefinidas. O segundo processo implementa regras que removem partes não essenciais para o entendimento das sentenças. Para cada sentença e sua versão simplificada, uma pontuação é agregada levando em consideração os conceitos envolvidos, a relação entre os conceitos e as sentenças e certas características associadas às sentenças. No módulo de sumarização, as sentenças simplificadas com pontuações superiores são escolhidas para o resumo.

Assim como os autores mencionados antes, Finegan-Dollak e Radev (2016) visam o problema de sumarização textual. A metodologia empregada faz uso de operações de inserção, remoção e reordenação de partes de sentenças, assim como ajustes para garantir concordância gramatical. Além disso, os autores dão ênfase à desagregação de sentenças sem a substituição de entidades por pronomes. Isso permite que haja informação suficiente para que seja possível avaliar a importância das sentenças no sumarizador extrativo.

3.3 Trabalhos direcionados a outras atividades de PLN

Outras atividades de PLN podem ser beneficiadas pela simplificação prévia de sentenças. Apresentando uma combinação dos métodos de regras sintáticas de simplificação e aprendizagem de máquina, Vickrey e Koller (2008) formularam um modelo cujo escopo é voltado à tarefa de rotulagem de papel semântico em PLN, que consiste em identificar os papéis semânticos associados a um verbo, em geral representados por um agente e um alvo. Os autores propuseram um modelo probabilístico que aprende, dado um conjunto de regras de transformação, a selecionar a melhor simplificação gerada por uma sequência de aplicações de um subconjunto das regras.

Lima et al. (2014) trazem uma representação de sentenças baseada em grafos — resultando em um grafo de dependências em vez de uma árvore — em um modelo que faz uso de um conjunto de transformações para simplificar sentenças. O objetivo final do modelo é obter uma representação com o máximo de informações relevantes para que se possa extrair relacionamentos entre entidades com melhor desempenho.

Che et al. (2015), abordando a tarefa de análise de sentimentos na língua chinesa, apresentam ideias para usar compressão de sentenças para aprimorar essa tarefa. Os autores tratam de análise de sentimentos baseada em aspectos, que reconhece não apenas a polaridade e a intensidade do texto, mas também o aspecto da opinião. Na arquitetura do sistema, a sentença passa por um pré-processamento que envolve compressão antes de ser dada como entrada para o analisador de sentimentos.

4 Metodologia

A seguir são apresentados os recursos, medidas de avaliação para análise quantitativa dos resultados e por último, as abordagens de execução dos experimentos, onde são expostos os métodos de simplificação de sentenças e as variações de sumarização extrativa empregados na pesquisa.

4.1 Configuração experimental

As próximas subseções descrevem a base de dados utilizada para avaliação dos resumos e as ferramentas de implementação e avaliação dos sistemas.

4.1.1 Base de dados

O *corpus* utilizado para análise das estratégias é formado por 1038 textos na língua inglesa que consistem em notícias do CNN[1] versando sobre negócios, política, justiça, saúde, estilo de vida e opinião. O mesmo é um subconjunto do *corpus* de 3000 notícias usado por Batista et al. (2016), decisão feita devido a limitações de tempo para execução dos experimentos. Cada notícia possui um resumo de até quatro sentenças produzidas por humanos. Um resumo é tido como *gold standard*, o que significa que é um resumo de referência para avaliação das técnicas de sumarização. Uma variação do *corpus* original com *stop words* filtradas foi incluída como referência para comparação dos resultados.

4.1.2 Ferramentas e ambiente de programação

As implementações necessárias (filtragem de *stop words* do *corpus* original e técnica de simplificação SVC, apresentada a seguir) foram desenvolvidos na linguagem de programação Java, utilizando a ferramenta Stanford CoreNLP[2]. Dentre os métodos apresentados a seguir, o RegenT[3] e o Reluctant Trimmer[4] possuem ferramenta própria para simplificação de sentenças.

4.1.3 Método de avaliação

A avaliação dos resumos foi realizada com a ferramenta ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (LIN, 2004), que inclui um conjunto de medidas para

[1] http://edition.cnn.com/
[2] https://stanfordnlp.github.io/CoreNLP/
[3] http://homepages.abdn.ac.uk/cgi-bin/cgiwrap/csc323/RegenT/demo.cgi
[4] http://www.quantmedia.org/coling2014/

análise automática de resumos de acordo com resumos de referência. A medida usada neste trabalho foi o ROUGE-N, que consiste em estatísticas de co-ocorrência baseadas em *n-grams*. Um *n-gram* é uma sequência de n palavras de um texto. Neste caso foi utilizado N = 1, portanto avaliou-se a interseção entre *unigrams* do resumo de referência e com um resumo de sistema.

O ROUGE-N computa a razão entre *n-grams* de um resumo candidato que também estão no resumo de referência em relação ao total de *n-grams* do resumo de referência. Esta medida é chamada *recall*, e informa o quanto há sobreposição entre os resumos de referência e de sistema. A fórmula para calcular o *recall* de um resumo de sistema é

$$Recall = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{4.1}$$

onde n é o tamanho do *n-gram*, $gram_n$ corresponde a um *n-gram* e $Count_{match}(gram_n)$ é o total de interseções de *n-grams* entre um resumo candidato e um resumo de referência. No caso de $n = 1$, os *n-grams* são palavras.

Além do *recall*, também é importante obter informação sobre o percentual de *n-grams* relevantes selecionados. Para o cálculo de precisão, leva-se em consideração a quantidade total de *n-grams* do resumo candidato, como indica a fórmula

$$Precision = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{SystemSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{4.2}$$

Esta medida é a mesma usada pelo método de avaliação de tradução automática BLEU ou *Bilingual Evaluation Understudy* (PAPINENI et al., 2002), que também se baseia em co-ocorrência de *n-grams*.

O *F-score*, que combina precisão e *recall*, funcionando como uma média harmônica ponderada das duas medidas, é calculado da seguinte maneira:

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{4.3}$$

onde β é um valor positivo.

Em posse dos resumos gerados automaticamente e seus respectivos resumos de referência, a ferramenta ROUGE é usada para computar as três medidas para cada resumo associado, avaliando a sobreposição de *unigrams* entre os textos. As médias de cada medida são, então, calculadas para os métodos de sumarização, de modo que cada um possui valores associados aos métodos de simplificação e ao *corpus* original.

4.2 Métodos de simplificação de sentenças

Este trabalho tem o intuito de comparar o desempenho de métodos de sumarização extrativa que fazem uso de diferentes estratégias de simplificação de sentenças. Como

referência para a comparação, faz-se uso do *corpus* original e uma versão do mesmo com *stop words* filtradas. A seguir, são apresentados os métodos de simplificação aplicados neste trabalho.

4.2.1 Filtragem de sujeito, verbo e complemento (SVC)

Tendo em vista o requisito de preservação do conteúdo original no texto resumido por sumarização automática, este método, referido neste trabalho como SVC, é proposto para extrair das sentenças o assunto tratado e seus agentes. Os alvos de tal filtragem são o sujeito, o verbo identificado como principal e o complemento da sentença, compreendido pelas seguintes relações gramaticais usadas por Marneffe e Manning (2008):

a) *copula*: relação entre verbo de ligação e seu complemento;

b) *conjunct*: relação que conecta dois elementos por uma conjunção (mantida para abranger casos em que há múltiplos complementos);

c) *nominal subject*: substantivo ou frase nominal que é o sujeito de uma oração;

d) *nominal passive subject*: o mesmo que *nominal subject* para orações na voz passiva;

e) *clausal subject*: uma oração que também é sujeito de outra oração;

f) *clausal passive subject*: uma oração que é sujeito de oração na voz passiva;

g) *direct object*: substantivo ou frase nominal que é objeto de um verbo;

h) *indirect object*: substantivo ou frase nominal que é objeto de um verbo;

i) *clausal complement*: oração dependente cujo sujeito funciona como o objeto de um verbo de outra oração ou um adjetivo;

j) *open clausal complement*: predicado ou oração que complementa um verbo ou adjetivo e não possui sujeito;

k) *nominal modifier*: substantivo que serve de adjunto para outros substantivos ou orações predicativas.

Vale notar que a definição de complemento usada neste trabalho é um hiperônimo para termos que participam das relações gramaticais listadas acima, não sendo uma definição gramatical oficial. O SVC se baseia na hipótese de que a ideia principal que uma sentença transmite tem como peças chave, usualmente, os três elementos anteriormente citados. Somente a representação de árvore de dependências é usada como recurso para filtrar esses elementos, sem uso de técnicas de extração de informação baseadas em semântica.

4.2.2 RegenT

O método RegenT foi introduzido pela primeira vez no trabalho de Siddharthan (2011). O autor propõe um sistema que usa a representação de dependências do *parser* da ferramenta Stanford CoreNLP para realizar transformações em orações coordenadas e subordinadas, apostos e orações na voz passiva. Dada uma sentença como entrada, o *parser* gera a representação de dependências e, a partir da mesma, são inferidas transformações que podem remover ou inserir relações de dependência para reformular a sentença. Contudo, tais transformações podem não ser suficientes para garantir adequação sintática. Um passo adicional, então, realiza a correção de concordância gramatical e relações de dependência interrompidas.

Neste trabalho, uma versão modificada do RegenT descrita por Angrosh, Nomoto e Siddharthan (2014) é utilizada. A modificação agrega dois novos conjuntos de regras de transformação (MANDYA; SIDDHARTHAN, 2014; SIDDHARTHAN; MANDYA, 2014) aprendidas a partir de *corpora* alinhados de sentenças em inglês e suas reformulações simplificadas. O sistema aproveita a estrutura da abordagem de tradução automática de Ding e Palmer (2005), que se baseia em gramáticas síncronas de dependência. Esse tipo de gramática codifica substituições em árvores de dependência, mapeando sub-árvores da sentença original em sub-árvores da sentença alvo, como sugere a Figura 4, onde sub-árvores equivalentes em sentido têm a mesma coloração. Tais sub-árvores são chamadas de árvores elementares, e o processo de mapeamento é uma transdução.

Figura 4 – Exemplo de transdução de árvores elementares

Original: The girl kissed her kitty cat.
Alvo: The girl gave a kiss to her cat.

Fonte: adaptado de Ding e Palmer (2005)

A aprendizagem das regras, seguindo o modelo proposto por Ding e Palmer (2005), compara árvores de dependências de sentenças alinhadas e infere inserções e remoções necessárias para transformar a sentença original na sentença alvo. As correções finais realizadas no RegenT descrito por Siddharthan (2011) são mantidas.

4.2.3 Reluctant Trimmer (RT)

O método Reluctant Trimmer proposto por Angrosh, Nomoto e Siddharthan (2014) é uma técnica de compressão que também usa a representação de dependências das sentenças. A estratégia é "relutante" pois procura fazer o mínimo de modificações possível no texto na intenção de não perder o significado original.

Figura 5 – Conversão de representação de dependências para estrutura de reticulado pelo método RT

(a) Árvore de dependências da sentença "*2009 detention of American hikers by Iran*".

(b) Estrutura de reticulado da sentença com TDPs representados pelas linhas destacadas.

Fonte: Angrosh, Nomoto e Siddharthan (2014)

Figura 6 – Sentença truncada em reticulado e seleção de compressão de sentença por ILP

(a) Sequência de palavras contíguas passando pela raiz.

(b) Busca do melhor candidato com ILP.

Fonte: Angrosh, Nomoto e Siddharthan (2014)

Partindo da árvore de dependências (Figura 5a), o método transforma a representação em uma forma de reticulado (Figura 5b), gerando várias possibilidades de truncamento da sentença. As k melhores candidatas são escolhidas e, em seguida, o sistema faz otimização usando programação linear inteira (*integer linear programming* ou ILP) para selecionar uma candidata (Figura 6b), satisfazendo certas restrições impostas que levam em consideração propriedades locais (de sentenças) e globais (do texto). Na Figura 6a está um exemplo de truncamento da sentença "*2009 detention of American hikers*", produzindo a sentença compactada "2009 detention of hikers".

4.3 Métodos de pontuação para sumarização extrativa

Na sumarização extrativa, as sentenças de um resumo de sistema são selecionadas de acordo as pontuações mais altas de acordo com algum critério. Neste trabalho, 15

métodos de pontuação segundo Ferreira et al. (2013) são usados para análise dos sistemas, cada um gerando resumos compostos por quatro sentenças.

4.3.1 Pontuação baseada em palavras

Sentenças podem ser pontuadas de acordo com as palavras que a compõem. Os seguintes critérios estão incluídos nesta pesquisa:

a) frequência de palavras (WORDFREQ): parte da hipótese de que palavras mais frequentes têm mais chance de indicar o assunto do documento, logo elevam a pontuação de uma sentença;

b) *term frequency–inverse document frequency* (TFIDF): diminui a relevância de termos muito frequentes ao calcular a frequência de um termo (tf) em uma sentença e a quantidade de sentenças em que o termo ocorre (idf), onde a pontuação é computada através da Equação 4.4 (N é o número de sentenças);

$$\text{TFIDF}(s) = N \times \frac{\log(1 + \text{tf})}{\log(\text{df})} \tag{4.4}$$

c) letras maiúsculas (UPPER): eleva a pontuação de sentenças contendo palavras capitalizadas, de acordo com a Equação 4.6;

$$\text{R}(s) = \frac{\#\text{ de palavras capitalizadas em } s}{\#\text{ de palavras em } s} \tag{4.5}$$

$$\text{UPPER(s)} = \frac{\text{R}(s)}{\max(\text{R})} \tag{4.6}$$

d) nomes próprios (PROPER): caso específico do UPPER em que substantivos próprios são considerados indicadores de relevância;

e) co-ocorrência de palavras (CONGRAM): utiliza *n-grams* para favorecer a seleção de sentenças que contêm termos que ocorrem juntos com frequência ao longo do texto;

f) similaridade léxica (LEXSIM): considera mais relevantes sentenças que possuem um forte vínculo semântico com outras sentenças, seja pelo uso de palavras sinônimas ou pelo uso de palavras com outro tipo de relação semântica.

4.3.2 Pontuação baseada em sentenças

a) expressões conectivas (CUE): aponta as sentenças candidatas ao resumo buscando expressões como "*therefore*", "*in addition to this*" ou "*on the other hand*" (segundo a Equação 4.7), que consistem em elementos de coesão textual e podem introduzir ideias importantes;

$$\text{CUE}(s) = \frac{\#\text{ de expressões conectivas em } s}{\#\text{ de expressões conectivas no texto}} \tag{4.7}$$

b) inclusão de dados numéricos (NUMDATA): supõe que dados numéricos são importantes no contexto do documento, favorecendo sentenças com esse tipo de informação;

c) tamanho de sentença (SLENGTH): penaliza sentenças muito curtas ou muito longas em quantidade de palavras;

$$\text{SLENGTH}(s) = \text{tamanho de } s \times \text{tamanho médio de sentenças do texto} \quad (4.8)$$

d) posição de sentença no texto (SPOSTEXT): prioriza sentenças em certas posições do texto (como o começo ou fim);

e) centralidade de sentença 1 (SCENTR): utiliza a sobreposição de termos entre sentenças como indício da importância de sentenças segundo a Equação 4.9;

$$\text{SCENTR}(s) = \frac{\text{termos em } s \cap \text{termos em outras sentenças}}{\text{termos em } s \cup \text{termos em outras sentenças}} \quad (4.9)$$

f) centralidade de sentença 2 (BLEU): mesmo objetivo do SCENTR, porém utilizando o método BLEU (PAPINENI et al., 2002), que foi proposto para avaliar o quão próxima uma sentença traduzida por tradução automática é da original;

g) semelhança com título (RESTITLE): seleciona sentenças com maior interseção de palavras com o título, fundamentando-se na ideia de que essas sentenças transmitem mais conteúdo tratado no texto que outras sentenças menos semelhantes ao título.

$$\text{RESTITLE}(s) = \frac{\#\text{ de palavras do título em } s}{\#\text{ de palavras no título}} \quad (4.10)$$

4.3.3 Pontuação baseada em grafos

a) *bushy path* (BPATH): neste método, uma sentença é vértice de um grafo e seu *bushy path* é a quantidade de arestas conectando-a a outras sentenças, indicando a existência de referências entre essas sentenças;

b) similaridade agregada (AGGSIM): similar ao BPATH, porém calcula a soma dos pesos das arestas, que representam a similaridade entre as sentenças.

5 Resultados e discussão

Os resultados obtidos pelos três métodos de simplificação avaliados em comparação com a aplicação de sumarização ao *corpus* original (Ref) e ao mesmo com filtragem de *stop words* (STOP) encontram-se na Tabela 2 (*recall*), na Tabela 3 (precisão) e na Tabela 4 (*F-score*). Em todos os casos, Ref alcançou melhor *recall* e *F-score*, portanto não há melhoria a ser observada somente pela aplicação de simplificação de sentenças precedendo a sumarização extrativa.

A abordagem RT conseguiu melhor *recall* entre métodos de simplificação. Sendo um método de compressão, era esperado que este sistema induzisse a seleção de sentenças mais relevantes para o resumo, entretanto nota-se que seu desempenho não foi superior ao sistema de referência. Quanto à precisão, o RT atingiu as melhores marcas entre todos os métodos, sugerindo que o critério de compressão teve alguma vantagem em comparação com o Ref.

Não se pode dizer que a remoção de frases de pouca informação ou a desagregação de sentenças, ambas realizadas pelo RegenT, trouxeram melhorias na sumarização, porém é possível que estas características tenham influenciado no ocasional ganho em precisão em relação ao sistema de referência Ref. A desagregação de sentenças, em particular, pode separar uma informação de pouca utilidade de outra que tenha conteúdo importante ou ter o efeito adverso de descentralizar uma ideia relevante, diminuindo sua importância. Em alguns casos, o RegenT conseguiu precisão superior ao método de referência.

Também era esperado que a filtragem realizada pelos sistemas STOP e SVC tivesse impacto no aumento da sobreposição de termos relevantes com os resumos de referência. Ambos os métodos mostraram o oposto, sendo mais ineficientes que o Ref em todas as estratégias de sumarização abordadas, refutando a hipótese que motivou o uso do SVC. Outro ponto negativo de ambos os métodos está na apresentação final do resumo, que não é apropriada para a leitura por humanos. Em questão de desempenho, a filtragem de *stop words* ainda mostrou-se melhor do que apenas manter sujeito, verbo e complemento, indicando que houve perda de indicadores de relevância nas sentenças simplificadas por SVC.

Observa-se que os métodos STOP e SVC, que realizam filtragens sem levar em conta aspectos linguísticos, mostraram-se ineficientes. Os métodos RegenT e Reluctant Trimmer, que conseguiram precisão superior ao Ref, fazem uso de regras mais complexas de simplificação e procuram manter gramaticalidade. Por esse motivo, é possível que a perda de informações seja minimizada. Ainda assim, não houve um aumento notável de desempenho em comparação à aplicação de sumarização no *corpus* não simplificado.

Tabela 2 – Média do *recall* e respectivos valores de desvio padrão para cada método de pontuação aplicado a cada método de simplificação.

Método de pontuação	*Recall*				
	Ref	RegenT	RT	STOP	SVC
PROPER	0.482 (0.180)	0.396 (0.153)	0.460 (0.163)	0.262 (0.145)	0.178 (0.086)
WORDFREQ	0.486 (0.176)	0.449 (0.154)	0.475 (0.163)	0.302 (0.142)	0.185 (0.089)
TFIDF	0.519 (0.178)	0.472 (0.152)	0.501 (0.161)	0.309 (0.143)	0.219 (0.089)
RESTITLE	0.390 (0.167)	0.251 (0.119)	0.340 (0.150)	0.154 (0.109)	0.140 (0.081)
CONGRAM	0.455 (0.174)	0.394 (0.148)	0.425 (0.157)	0.265 (0.136)	0.161 (0.086)
UPPER	0.482 (0.177)	0.403 (0.152)	0.461 (0.163)	0.264 (0.145)	0.181 (0.086)
SCENTR	0.246 (0.144)	0.204 (0.120)	0.223 (0.129)	0.136 (0.105)	0.117 (0.075)
CUE	0.390 (0.167)	0.251 (0.119)	0.340 (0.150)	0.154 (0.109)	0.140 (0.081)
LEXSIM	0.508 (0.178)	0.451 (0.162)	0.480 (0.165)	0.300 (0.145)	0.207 (0.089)
SPOSTEXT	0.445 (0.200)	0.316 (0.162)	0.401 (0.183)	0.215 (0.144)	0.153 (0.086)
AGGSIM	0.437 (0.155)	0.320 (0.142)	0.403 (0.148)	0.214 (0.132)	0.157 (0.075)
BLEU	0.331 (0.212)	0.285 (0.176)	0.298 (0.202)	0.155 (0.146)	0.132 (0.087)
NUMDATA	0.458 (0.182)	0.355 (0.150)	0.418 (0.163)	0.220 (0.136)	0.158 (0.085)
SLENGTH	0.489 (0.173)	0.444 (0.151)	0.470 (0.156)	0.292 (0.146)	0.207 (0.088)
BPATH	0.436 (0.157)	0.317 (0.136)	0.393 (0.142)	0.207 (0.132)	0.153 (0.078)

Fonte – a autora

Tabela 3 – Média da precisão e respectivos valores de desvio padrão para cada método de pontuação aplicado a cada método de simplificação.

Método de pontuação	Precisão				
	Ref	RegenT	RT	STOP	SVC
PROPER	0.368 (0.160)	0.364 (0.148)	0.372 (0.154)	0.265 (0.155)	0.286 (0.149)
WORDFREQ	0.368 (0.181)	0.358 (0.156)	0.369 (0.172)	0.288 (0.158)	0.257 (0.143)
TFIDF	0.392 (0.180)	0.383 (0.157)	0.392 (0.167)	0.286 (0.161)	0.303 (0.149)
RESTITLE	0.345 (0.142)	0.316 (0.132)	0.340 (0.141)	0.233 (0.145)	0.283 (0.151)
CONGRAM	0.356 (0.174)	0.360 (0.157)	0.367 (0.169)	0.265 (0.158)	0.272 (0.150)
UPPER	0.365 (0.159)	0.359 (0.146)	0.370 (0.157)	0.265 (0.156)	0.278 (0.145)
SCENTR	0.300 (0.114)	0.309 (0.130)	0.305 (0.118)	0.215 (0.141)	0.252 (0.140)
CUE	0.345 (0.142)	0.316 (0.132)	0.340 (0.141)	0.233 (0.145)	0.283 (0.151)
LEXSIM	0.384 (0.169)	0.387 (0.153)	0.388 (0.162)	0.289 (0.160)	0.303 (0.150)
SPOSTEXT	0.384 (0.157)	0.396 (0.168)	0.395 (0.156)	0.293 (0.173)	0.306 (0.156)
AGGSIM	0.343 (0.155)	0.371 (0.134)	0.353 (0.125)	0.247 (0.142)	0.282 (0.132)
BLEU	0.353 (0.180)	0.364 (0.195)	0.367 (0.186)	0.275 (0.217)	0.268 (0.175)
NUMDATA	0.364 (0.154)	0.362 (0.145)	0.366 (0.149)	0.264 (0.161)	0.297 (0.154)
SLENGTH	0.361 (0.163)	0.349 (0.149)	0.363 (0.164)	0.267 (0.156)	0.268 (0.141)
BPATH	0.347 (0.129)	0.369 (0.132)	0.351 (0.127)	0.247 (0.147)	0.276 (0.132)

Fonte – a autora

Foi efetuado um teste estatístico *t-test* com 95% de confiança para comparar o resultado de maior *recall*, alcançado por Ref usando TFIDF, com o segundo melhor resultado, alcançado por RT. Desta forma, pode-se constatar que o método não simplificado Ref é significativamente melhor que o RT. Isso implica que o Ref comporta mais termos relevantes que estão presentes nos resumos *gold standard*. A precisão e o *F-score* dos dois métodos no TFIDF foi equivalente.

Tabela 4 – Média do *F-score* e respectivos valores de desvio padrão para cada método de pontuação aplicado a cada método de simplificação.

Método de pontuação	*F-score*				
	Ref	RegenT	RT	STOP	SVC
PROPER	0.407 (0.155)	0.370 (0.136)	0.402 (0.144)	0.257 (0.141)	0.214 (0.102)
WORDFREQ	0.409 (0.169)	0.389 (0.144)	0.405 (0.157)	0.288 (0.141)	0.210 (0.104)
TFIDF	0.436 (0.166)	0.413 (0.141)	0.429 (0.152)	0.291 (0.144)	0.249 (0.105)
RESTITLE	0.356 (0.139)	0.269 (0.111)	0.329 (0.131)	0.179 (0.116)	0.183 (0.099)
CONGRAM	0.389 (0.162)	0.364 (0.138)	0.382 (0.151)	0.259 (0.138)	0.195 (0.101)
UPPER	0.405 (0.154)	0.370 (0.135)	0.401 (0.147)	0.258 (0.142)	0.214 (0.101)
SCENTR	0.257 (0.118)	0.232 (0.112)	0.244 (0.112)	0.160 (0.113)	0.155 (0.092)
CUE	0.356 (0.139)	0.269 (0.111)	0.329 (0.131)	0.179 (0.116)	0.183 (0.099)
LEXSIM	0.427 (0.160)	0.404 (0.142)	0.419 (0.148)	0.288 (0.143)	0.241 (0.106)
SPOSTEXT	0.400 (0.159)	0.338 (0.146)	0.384 (0.152)	0.239 (0.147)	0.199 (0.103)
AGGSIM	0.375 (0.124)	0.331 (0.122)	0.367 (0.122)	0.223 (0.130)	0.197 (0.089)
BLEU	0.317 (0.183)	0.297 (0.168)	0.302 (0.184)	0.184 (0.162)	0.169 (0.108)
NUMDATA	0.396 (0.153)	0.347 (0.131)	0.381 (0.142)	0.234 (0.139)	0.201 (0.103)
SLENGTH	0.406 (0.157)	0.382 (0.138)	0.400 (0.149)	0.273 (0.142)	0.228 (0.103)
BPATH	0.377 (0.126)	0.329 (0.119)	0.361 (0.119)	0.219 (0.131)	0.192 (0.091)

Fonte – a autora

Realizando a mesma análise para os melhores resultados de precisão, conseguidos pelos métodos RegenT e RT usando o critério SPOSTEXT, verificou-se que não há diferença significativa entre os dois. Inclusive, ao estender este teste ao método Ref, ainda não há diferença significativa neste quesito. Consequentemente, nenhum desses métodos se destaca em relação ao outro no que se refere ao percentual de termos relevantes dos resumos usando o método SPOSTEXT. No entanto, o Ref consegue *recall* superior aos dois, seguido do RT, como foi mostrado pelo teste.

O *t-test* comparando os melhores valores de *F-score* (método LEXSIM) revela que, neste caso, o Ref e o RT são equivalentes, mas superiores ao RegenT. Com isso, infere-se que não há diferença de acurácia entre o Ref e o RT para este método. Entretanto, houve desempenho significativamente superior do Ref quanto ao *recall*, sendo o RT o segundo melhor. A equivalência dos *F-scores* desses métodos foi influenciada pela equivalência em precisão aferida pelo teste.

Em síntese, os experimentos realizados não exibem melhora direta da informatividade de resumos extrativos quando a simplificação de sentenças é usada. Embora em certos resultados haja aparente dominância numérica de métodos como RegenT e RT em relação ao *corpus* não simplificado, foi visto através de *t-test* que não houve diferença estatisticamente significante que sustentasse a hipótese de que haveria melhoria nos resultados.

6 Conclusão

Diante dos resultados obtidos nesta pesquisa, infere-se que houve perda de informatividade dos resumos extrativos devido à aplicação das estratégias de simplificação de sentenças. Nenhum dos métodos de simplificação analisados conseguiu atingir o desempenho do *corpus* original após a efetuação de sumarização extrativa. Entretanto, levando em consideração apenas a sumarização dos textos simplificados, viu-se que o desempenho do método Reluctant Trimmer foi superior, seguido do método RegenT.

Uma das expectativas na realização deste trabalho era que a simplificação de sentenças faria com que, no processo de sumarização extrativa de textos, houvesse maior sobreposição de ideias relevantes entre o resumo gerado e um resumo ideal. Esta ideia foi motivada pelo fato de simplificação de sentenças ter, entre seus objetivos, a compactação de uma sentença capturando ao máximo a ideia central original. Com os experimentos apresentados neste trabalho, ao contrário do esperado, não se pode concluir que a simplificação de sentenças beneficia a sumarização. Porém, percebe-se a partir dos resultados alcançados que propriedades linguísticas aparentam ser um fator importante na seleção de sentenças para o resumo de sistema, visto que os métodos de compressão SVC e STOP, que ignoram tais propriedades, obtiveram uma queda de desempenho considerável.

Como foi visto no Capítulo 2, existem técnicas de tradução automática monolíngues que fazem reformulações de sentenças, portanto visam exatidão gramatical. Tendo isso em vista, algumas hipóteses podem ser levantadas quanto ao emprego de abordagens de aprendizagem de máquina para simplificação de sentenças. No entanto, apenas considerando o desempenho individual dos métodos de pontuação de sentenças, ainda há incertezas sobre possíveis benefícios de tais técnicas. Além disso, em abordagens de reformulação também pode-se esperar que a interseção com o texto original seja parcialmente reduzida, comprometendo o resumo final. Tais questões não fazem parte do escopo deste trabalho, portanto permanecem em aberto.

Quanto à pergunta que este trabalho se propôs a responder, viu-se que, para os métodos de simplificação de sentenças analisados, não houve melhoria na sumarização extrativa. Em todas as abordagens de sumarização utilizadas, o método de referência (não simplificado) obteve desempenho superior ou equivalente aos melhores métodos de simplificação, de acordo com testes estatísticos. É importante explicitar que as abordagens de sumarização não foram combinadas neste trabalho, o que pode ter exercido influência na conclusão tirada dos resultados.

6.1 Trabalhos futuros

Nesta versão da pesquisa, 1038 documentos do *corpus* utilizado foram considerados. Para análises futuras, seria válido utilizar o *corpus* completo de 3000 documentos. Adicionalmente, visto que os métodos RegenT e Reluctant Trimmer foram apresentados em um sistema unificado de simplificação, a composição de ambos poderia fornecer informações úteis para análise. Para promover melhor discernimento dos resultados, o emprego de variações de combinações de critérios de pontuação de sentença pode mostrar-se proveitoso. Por fim, a adição de uma abordagem de aprendizagem de máquina seria relevante, considerando o crescente uso desse tipo de técnica e seu comportamento de conservação de traços linguísticos.

Referências

ANGROSH, M.; NOMOTO, T.; SIDDHARTHAN, A. Lexico-syntactic text simplification and compression with typed dependencies. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. p. 1996–2006. Citado 4 vezes nas páginas 20, 21, 27 e 28.

BATISTA, J.; LINS, R. D.; LIMA, R.; SIMSKE, S. J.; RISS, M. Towards cohesive extractive summarization through anaphoric expression resolution. In: ACM. *Proceedings of the 2016 ACM Symposium on Document Engineering*. Vienna, Austria, 2016. p. 201–204. Citado na página 24.

BAWAKID, A.; OUSSALAH, M. Sentences simplification for automatic summarization. In: *2011 IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS)*. London, UK: IEEE, 2011. p. 59–64. Citado 3 vezes nas páginas 17, 21 e 23.

CHE, W.; ZHAO, Y.; GUO, H.; SU, Z.; LIU, T. Sentence Compression for Aspect-Based Sentiment Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 23, p. 2111–2124, dez. 2015. Citado na página 23.

CHOWDHURY, G. G. Natural language processing. *Annual review of information science and technology*, v. 37, p. 51–89, 2003. Citado na página 13.

COSTER, W.; KAUCHAK, D. Simple English Wikipedia: A New Text Simplification Task. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Portland, Oregon: Association for Computational Linguistics, 2011. p. 665–669. Citado na página 21.

DAELEMANS, W.; HÖTHKER, A.; SANG, E. F. T. K. Automatic Sentence Simplification for Subtitling in Dutch and English. In: *4th International Conference on Language Resources and Evaluation*. Lisbon: European Language Resources Association, 2004. Citado na página 20.

DAS, D.; MARTINS, A. F. T. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, v. 4, p. 192–195, 2007. Citado 2 vezes nas páginas 11 e 18.

DING, Y.; PALMER, M. Machine translation using probabilistic synchronous dependency insertion grammars. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 541–548. Citado na página 27.

FERREIRA, R.; CABRAL, L. de S.; LINS, R. D.; SILVA, G. P. e; FREITAS, F.; CAVALCANTI, G. D.; LIMA, R.; SIMSKE, S. J.; FAVARO, L. Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, Elsevier, v. 40, n. 14, p. 5755–5764, 2013. Citado na página 29.

FERRÉS, D.; MARIMON, M.; SAGGION, H. A Web-based Text Simplification System for English. *Procesamiento del Lenguaje Natural*, v. 55, p. 191–194, 2015. Citado na página 21.

FINEGAN-DOLLAK, C.; RADEV, D. R. Sentence simplification, compression, and disaggregation for summarization of sophisticated documents. *Journal of the Association for Information Science and Technology*, v. 67, n. 10, p. 2437–2453, out. 2016. Citado na página 23.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence*, Academy Publisher, PO Box 40 Oulu 90571 Finland, v. 2, n. 3, p. 258–268, 2010. Citado 2 vezes nas páginas 11 e 18.

JING, H. Sentence Reduction for Automatic Text Summarization. In: *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Seattle, Washington: Association for Computational Linguistics, 2000. p. 310–315. Citado na página 22.

LIMA, R. J.; BATISTA, J.; FERREIRA, R.; FREITAS, F.; LINS, R. D.; SIMSKE, S.; RISS, M. Transforming Graph-based Sentence Representations to Alleviate Overfitting in Relation Extraction. In: *Proceedings of the 2014 ACM Symposium on Document Engineering*. Fort Collins, Colorado, USA: ACM, 2014. p. 53–62. Citado na página 23.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain: Association for Computational Linguistics, 2004. v. 8. Citado na página 24.

MANDYA, A. A.; SIDDHARTHAN, A. Text simplification using synchronous dependency grammars: Generalising automatically harvested rules. In: *8th International Natural Language Generation Conference*. Philadelphia, Pennsylvania, U.S.A.: Association for Computational Linguistics, 2014. Citado na página 27.

MANI, I. *Advances in Automatic Text Summarization*. Cambridge, MA, USA: MIT Press, 1999. ISBN 0262133598. Citado na página 18.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN 0-262-13360-1. Citado 2 vezes nas páginas 13 e 14.

MARNEFFE, M.-C. D.; MANNING, C. D. *Stanford typed dependencies manual*. 2008. Citado na página 26.

MIKHEEV, A. Text segmentation. In: *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003. cap. 10, p. 201–218. ISBN 0198238827. Citado na página 15.

NARAYAN, S.; GARDENT, C. Hybrid Simplification using Deep Semantics and Machine Translation. In: *The 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, United States: ACL, 2014. p. 435–445. Citado na página 22.

NUNES, B. P.; KAWASE, R.; SIEHNDEL, P.; CASANOVA, M. A.; DIETZE, S. As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts. In: *2013 IEEE 13th International Conference on Advanced Learning Technologies*. Beijing, China: IEEE, 2013. p. 128–132. Citado na página 21.

O'BRIEN, S. Controlling controlled English. *Proceedings of EAMT-CLAW*, v. 3, p. 105–114, 2003. Citado 2 vezes nas páginas 11 e 17.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002. p. 311–318. Citado 2 vezes nas páginas 25 e 30.

SIDDHARTHAN, A. Syntactic Simplification and Text Cohesion. *Research on Language and Computation*, v. 4, p. 77–109, 2006. Citado na página 21.

SIDDHARTHAN, A. Text simplification using typed dependencies: a comparison of the robustness of different generation strategies. In: *Proceedings of the 13th European Workshop on Natural Language Generation*. Nancy, France: Association for Computational Linguistics, 2011. p. 2–11. Citado 2 vezes nas páginas 21 e 27.

SIDDHARTHAN, A. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, v. 165, p. 259–298, 2014. Citado 4 vezes nas páginas 11, 16, 17 e 18.

SIDDHARTHAN, A.; MANDYA, A. A. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. Gothenburg, Sweden: Association for Computational Linguistics, 2014. Citado na página 27.

ŠTAJNER, S.; BÉCHARA, H.; SAGGION, H. A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Beijing, China: Association for Computational Linguistics, 2015. p. 823–828. Citado na página 22.

VANDERWENDE, L.; SUZUKI, H.; BROCKETT, C.; NENKOVA, A. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, v. 43, p. 1606–1618, 2007. Citado na página 22.

VICKREY, D.; KOLLER, D. Sentence Simplification for Semantic Role Labeling. In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008. p. 344–352. Citado na página 23.

VOUTILAINEN, A. Part-of-speech tagging. In: *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003. cap. 11, p. 219–232. ISBN 0198238827. Citado na página 15.

WUBBEN, S.; van den Bosch, A.; KRAHMER, E. Sentence Simplification by Monolingual Machine Translation. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Jeju Island, Korea: Association for Computational Linguistics, 2012. (ACL '12), p. 1015–1024. Citado na página 22.

ZHU, Z.; BERNHARD, D.; GUREVYCH, I. A Monolingual Tree-based Translation Model for Sentence Simplification. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China: Association for Computational Linguistics, 2010. (COLING '10), p. 1353–1361. Citado na página 21.