LUCAS MUCIDA COSTA

ISIM: PROPOSTA DE UMA NOVA MÉTRICA PARA SIMPLIFICAÇÃO DE SENTENÇAS EM LINGUAGEM NATURAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Doctor Scientiae*.

Orientador: Alcione de Paiva Oliveira

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
C837i
2024

Costa, Lucas Mucida, 1987-
ISiM: proposta de uma nova métrica para simplificação de sentenças em linguagem natural / Lucas Mucida Costa. – Viçosa, MG, 2024.
1 tese eletrônica (84 f.): il. (algumas color.).

Orientador: Alcione de Paiva Oliveira.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Informática, 2024.
Referências bibliográficas: f. 79-84.
DOI: https://doi.org/10.47328/ufvbbt.2024.593
Modo de acesso: World Wide Web.

1. Inteligência artificial. 2. Processamento de linguagem natural (Computação). I. Oliveira, Alcione de Paiva, 1962-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22. ed. 006.35

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

LUCAS MUCIDA COSTA

ISIM: PROPOSTA DE UMA NOVA MÉTRICA PARA SIMPLIFICAÇÃO DE SENTENÇAS EM LINGUAGEM NATURAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Doctor Scientiae*.

APROVADA: 03 de Junho de 2024.

Assentimento:

gov.br
Documento assinado digitalmente
ALCIONE DE PAIVA OLIVEIRA
Data: 23/09/2024 09:03:35-0300
Verifique em https://validar.iti.gov.br

Alcione de Paiva Oliveira
Orientador

gov.br
Documento assinado digitalmente
LUCAS MUCIDA COSTA
Data: 17/09/2024 15:51:34-0300
Verifique em https://validar.iti.gov.br

Lucas Mucida Costa

Agradecimentos

Agradeço primeiramente a Deus por me prover capacidade e iluminar o caminho independente das dificuldades. Agradeço também à Universidade Federal de Viçosa por toda a estrutura e seu corpo docente, ao Departamento de Informática pelo excelente programa de pós graduação, à Diretoria de Tecnologia da Informação pelo apoio e incentivo à minha evolução como profissional e à todos os funcionários envolvidos. Agradecimento especial ao meu orientador Alcione e ao meu colega de pesquisa Maurilio pelo apoio durante todos estes anos especialmente desafiadores de pesquisa, em meio a pandemia e ainda conciliando estudos e trabalho. Agradeço também aos amigos Ivo, Luiz Paulo e Rafael pelos conselhos, e aos meus familiares, em especial minha mãe que sempre acreditou em mim, e à minha esposa Bruna pelo apoio incondicional. Dedico esse trabalho ao meu filho Bento, que seja motivo de orgulho para ele.

Resumo

COSTA, Lucas Mucida, D.Sc., Universidade Federal de Viçosa, junho de 2024. **ISiM: Proposta de uma métrica para simplificação de sentenças em linguagem natural.**. Orientador: Alcione de Paiva Oliveira.

Em uma sociedade complexa, a habilidade de simplificar textos pode ser bastante útil. Uma comunicação clara, concisa e de fácil compreensão são características bem-vindas na interação entre pessoas. Em virtude dessa necessidade, pesquisas voltadas ao desenvolvimento de modelos capazes de produzir textos mais simples importantes, e a busca por *corpus* adequados para treinar e aperfeiçoar esses modelos é um campo de pesquisa ativo. No entanto, para cumprirmos essa exigência, é necessário que possamos desenvolver métricas que possibilitem verificar o quanto uma sentença é mais simples que outra com significado similar. Nesta pesquisa, desenvolvemos uma métrica de simplificação de textos para a área de Processamento de Linguagem Natural (PLN), denominada ISiM. A métrica proposta supera as limitações das métricas existentes, oferecendo uma abordagem rápida, simples, livre de intervenção humana e independente da língua contribuindo na avaliação da qualidade da simplificação textual. Além disso, ISiM se demonstrou eficiente na criação e no refinamento de *corpora* de pares de sentenças complexo/simples, sendo essa uma contribuição para as pesquisas na área. Também, foi criado nesta pesquisa, um modelo gerador de textos simplificados, utilizando para *fine tunning* um *corpus* otimizado pela métrica ISiM. Durante os experimentos, a métrica demonstrou sua eficácia em diversas aplicações, como sua velocidade ao gerar resultados em poucos segundos, obtendo uma taxa de acerto de 96,94% ao ser testada em um *corpus* existente de pares de frase complexo/simples, 77,5% de acerto ao confrontada com um formulário respondido por humanos, e também superando outros modelos de geração de frases simplificadas da literatura. Além disso, a pesquisa destaca a relevância social da simplificação de textos, especialmente em um contexto como o do Brasil, onde o analfabetismo funcional atinge mais de 62 milhões de pessoas, sendo um desafio significativo a ser superado. A dificuldade de compreensão de textos complexos devido à deficiências na educação da população mostra o quanto ainda precisamos melhorar nosso sistema de ensino, e reforça a importância de desenvolver ferramentas como a ISiM para ajudar a tornar a informação mais acessível e compreensível para todos.

Palavras-chave: Inteligência Artificial; Processamento de Linguagem Natural; Simplificação de Texto; Métrica; ISiM; Mucimples.

Abstract

COSTA, Lucas Mucida, D.Sc., Universidade Federa de Viçosa, June 2024. **ISiM: Proposal of a Metric for Sentence Simplification in Natural Language.**. Advisor: Alcione de Paiva Oliveira.

In a complex society, the ability to simplify texts can be quite useful. Clear, concise, and easily understandable communication is highly valued in interactions between people. Due to this necessity, research focused on the development of models capable of producing simpler texts is important, and the search for suitable *corpora* to train and improve these models is an active field of research. However, to meet this demand, it is necessary to develop metrics that allow us to verify how much simpler one sentence is compared to another with a similar meaning. In this research, we developed a text simplification metric for the field of Natural Language Processing (NLP), named ISiM. The proposed metric overcomes the limitations of existing metrics, offering a quick, simple, language-independent, and human-intervention-free approach, contributing to the evaluation of the quality of text simplification. Additionally, ISiM proved to be efficient in creating and refining *corpora* of complex/simple sentence pairs, making it a valuable contribution to research in the area. Moreover, in this research, a simplified text generator model was created, using a *corpus* optimized by the ISiM metric for fine-tuning. During the experiments, the metric demonstrated its effectiveness in various applications, such as its speed in generating results within seconds, achieving an accuracy rate of 96.94% when tested on an existing *corpus* of complex/simple sentence pairs, and 77.5% accuracy when compared with a human-answered form, also surpassing other simplified sentence generation models from the literature. Furthermore, the research highlights the social relevance of text simplification, especially in a context like Brazil, where functional illiteracy affects more than 62 million people, representing a significant challenge to be overcome. The difficulty in understanding complex texts due to deficiencies in the population's education shows how much we still need to improve our education system and reinforces the importance of developing tools like ISiM to help make information more accessible and comprehensible for everyone.

Keywords: Artificial Intelligence; Natural Language Processing; Text Simplification; Metric; ISiM; Mucimples

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Em um cenário mundial marcado pela globalização e interconexão, a capacidade de simplificar textos emerge como uma ferramenta útil para sustentar uma comunicação tanto eficiente quanto inclusiva. Textos com elevado grau de complexidade não apenas apresentam barreiras à compreensão, mas têm o potencial de marginalizar segmentos substanciais da população confrontados com obstáculos linguísticos ou cognitivos.

A prática da simplificação vai além da mera redução textual ou substituição por termos mais usuais. Visa a apresentação de informações de forma acessível, preservando sua essência e intenção comunicativa. Siddharthan destaca (SIDDHARTHAN, 2014) que a simplificação textual é um processo meticuloso cujo objetivo é atenuar a complexidade linguística, garantindo, no entanto, que o núcleo informativo e a intenção do texto original se mantenham íntegros e transparentes.

A relevância da simplificação textual estende-se além da mera eficácia comunicativa em um contexto global. A simplificação potencializa a disseminação da informação, favorecendo sua compreensão por um espectro demográfico mais amplo. Para ilustrar, ao simplificar um texto, este pode se tornar mais acessível para crianças em etapas iniciais de aprendizagem (KAUCHAK, 2013), bem como para indivíduos com uma formação educacional básica (GASPERIN et al., 2009). Esta prática é também essencial para falantes não nativos que não dominem plenamente o idioma em questão (PAETZOLD; SPECIA, 2016). Ademais, a simplificação beneficia indivíduos com desafios linguísticos específicos, como aqueles com afasia, dislexia ou autismo (DEVLIN; UNTHANK, 2006; RELLO et al., 2013; EVANS; ORASAN; DORNESCU, 2014). Portanto, a simplificação textual serve como um vetor de inclusão, facilitando a compreensão em variados contextos.

Para entender os desafios associados à leitura e interpretação de textos no Brasil, algumas estatísticas do Instituto Brasileiro de Geografia e Estatística (IBGE) são reveladoras. Segundo este órgão, aproximadamente 6,6% da população enfrenta barreiras significativas em relação à leitura e escrita de textos (IBGE, 2019). Contudo, os desafios não se limitam estritamente à decodificação das palavras. Uma proporção considerável de cidadãos, embora capazes de ler, enfrenta adversidades ao tentar interpretar e compreender profundamente o conteúdo lido. Em uma pesquisa conduzida pelo Indicador de Analfabetismo Funcional (*INAF*) em 2018, foi identificado que cerca de 30% da população brasileira se enquadra na categoria de analfabetos funcionais (INAF, 2018). Agravando a situação, a mesma pesquisa destaca que 64% da população apresenta um nível de alfabetização apenas rudimentar, colocando-os em risco de mal interpretar ou não capturar nuances em textos mais complexos. Portanto, pesquisas voltadas ao desenvolvimento de modelos capazes de produzir textos mais simples importantes, e a busca por *corpus* ade-

quados para treinar e aperfeiçoar esses modelos é um campo de pesquisa ativo.

A área atrai o interesse e dedicação de muitos pesquisadores em diversos países, e vários estudos para criação de modelos voltados para essa tarefa tem sido desenvolvidos ((MARTIN et al., 2020), (CLIVE; CAO; REI, 2022), (XU et al., 2016), (OMELIANCHUK; RAHEJA; SKURZHANSKYI, 2021), (MARTIN et al., 2019), (ZHAO et al., 2018), (DONG et al., 2019), (ZHANG; LAPATA, 2017)), onde podemos ver vários tipos de abordagem, como utilização de paráfrases ((MARTIN et al., 2020), (ZHAO et al., 2018)), modelos estatísticos ((XU et al., 2016)) e aprendizagem por reforço ((ZHANG; LAPATA, 2017)).

O domínio da simplificação textual possui métricas que são reconhecidas por sua relevância e aplicabilidade. Dentre estas, o BLEU (PAPINENI et al., 2002) e o SARI (XU et al., 2016) sobressaem devido à sua constante presença em estudos da área. Contudo, é importante ressaltar que essas métricas originaram-se de tarefas específicas, como tradução automática e geração de resumos. Outras métricas, como o Índice Flesch, Gunning Fog e Coleman-Liau, (TALBURT, 1986; GUNNING, 1969; COLEMAN; LIAU, 1975) derivam de estudos direcionados a alunos de instituições educacionais específicas nos Estados Unidos, o que indica uma calibração particular para este tipo de amostragem, potencialmente limitando sua universalidade.

Neste contexto, apresentamos a métrica ISiM (*Independent Simplification Metric*). Sua introdução é importante para o escopo desta pesquisa, uma vez que oferece a possibilidade de avaliação da complexidade textual. Argumentamos que a ISiM não apenas oferece uma perspectiva nova, mas também tem o potencial de aprimorar e refinar tanto a simplificação textual quanto *corpora* de simplificação, incrementando a precisão e eficácia dessas bases.

A métrica ISiM foi apresentada no artigo (MUCIDA; OLIVEIRA; POSSI, 2022). Neste trabalho, apresentamos sua composição e os desfechos observados após sua implementação. Realizamos uma avaliação rigorosa da ISiM utilizando dois *corpora* paralelos, reconhecidos por seu uso em avaliações de modelos de simplificação textual.

Em um artigo subsequente, ampliamos a análise e a contextualização da métrica ISiM, com a intenção de apresentar outras possibilidades para a aplicação da métrica. Em uma das iniciativas, concentramo-nos na transição de *corpora* de paráfrases para *corpora* de simplificação, ampliando sua aplicabilidade. A fim de avaliar essa transição, concebemos uma avaliação prática: um questionário online contendo 40 pares de sentenças que foi respondido por 140 pessoas de localidades distintas. Além disso, ao final do experimento colhemos informações sobre o grau de escolaridade e de especialização dos participantes, para fazer uma análise também nesse sentido. Neste questionário solicitávamos que os respondentes, sob total anonimato, selecionassem obrigatoriamente a sentença que percebiam como sendo mais simples dentre as duas.

Adicionalmente, focamos em aprimorar *corpora* de simplificação já consolidados, otimizando sua eficiência, sobretudo quando avaliados pela métrica SARI. Em uma investigação específica, ponderamos a viabilidade de empregar a ISiM como guia na seleção de palavras subsequentes durante a geração textual por meio do GPT2.

Finalizando esta sequência de experimentos, desenvolvemos um novo modelo generativo com o intuito de gerar sentenças intrinsecamente mais simples do que suas versões originais. Este modelo foi comparado com outros modelos da literatura científica voltados para a simplificação textual.

Cabe destacar que, em contraste com diversos estudos que utilizam métricas específicas (e ocasionalmente menos eficazes) em seus modelos, destacamos a ISiM por sua precisão e eficiência. Isso se alinha à proposta de proporcionar uma métrica computacionalmente ágil por realizar apenas poucos cálculos numéricos, desprovida de dependências da língua e sem a necessidade de intervenção humana.

Ao avançar na leitura deste trabalho, notaremos as diversas facetas da métrica ISiM. Desde a integração do modelo GPT2 na produção de textos mais simplificados, passando pela reformulação e aperfeiçoamento de *corpora*, até a validação da métrica auxiliada pelo ChatGPT. Nossa expectativa é que, ao concluir esta análise, se destaque a versatilidade da ISiM e sua contribuição no campo da simplificação textual.

O principal objetivo deste trabalho é desenvolver e validar uma nova métrica de simplificação textual, denominada ISiM, que seja independente de frases ou anotações geradas por seres humanos e independente de semântica e sintaxe específicas de uma língua.A hipótese central desta pesquisa é que a frequência das palavras e o tamanho da sentença é capaz de medir simplificação de texto de forma a auxiliar no aprimoramento de corpus paralelos voltados para para a tarefa de simplificação de texto. Testamos esta hipótese nas línguas portuguesa e inglesa, mas postulamos que ela pode ser aplicada em línguas que usam o alfabeto latino, sistema de escrita alfabético e estrutura de frase similar.

A métrica ISiM procura contornar as deficiências das métricas anteriores ao focar em características intrínsecas dos textos, como a frequência das palavras e o tamanho das sentenças, sem depender de referências específicas. Precisão, neste contexto, refere-se à capacidade da métrica de refletir a simplicidade e clareza do texto de forma consistente com as percepções humanas de simplificação. A *confiabilidade* refere-se à consistência dos resultados da métrica ao avaliar diferentes conjuntos de dados simplificados. A *versatilidade* reflete a capacidade da métrica em também criar e modificar *corpora*, disponibilizando material para futuras pesquisas em simplificação de texto.

Para avaliar estas características, a métrica ISiM será comparada com métricas existentes através de experimentos que incluem:

- Avaliação da correlação entre as pontuações da métrica ISiM e as avaliações humanas da qualidade dos textos simplificados.
- Avaliação de *corpora* modificados pela métrica ISiM e reavaliados pela métrica SARI, a mais utilizada na literatura.
- Utilização de *corpus* modificado pela métrica na codificação de um modelo de geração de frases simplificadas, independente de intervenção humana e da língua utilizada.
- Avaliação da capacidade da métrica ISiM de criar novos *corpora* de pares de frases complexo/simples para utilização futura em modelos de simplificação.

Esses experimentos visam demonstrar que a métrica ISiM pode medir a simplificação textual de maneira tão eficaz quanto, ou melhor do que, as métricas dependentes de referências, e se pode ser usada para criar *corpora* para treinamento de modelos de simplificação de texto.

Os resultados alcançados neste estudo evidenciam a hipótese acima, destacando-se em dois aspectos. O primeiro diz respeito ao desenvolvimento de uma nova métrica, simples e eficaz, capaz de quantificar a simplicidade relativa entre duas sentenças. Essa métrica foi usada na elaboração de um modelo de geração de textos simplificados, superando em desempenho os modelos previamente disponíveis na literatura especializada. A segunda frente de contribuição se refere ao papel da métrica na criação e aprimoramento de *corpora* paralelos destinados à simplificação linguística. Dada a escassez desses *corpora* na literatura, a capacidade da métrica de gerar novos *corpora* a partir de paráfrases ou refinar os já existentes é importante, ampliando as possibilidades de pesquisa e desenvolvimento no campo da simplificação textual.

Este trabalho está estruturado da seguinte maneira: o capítulo subsequente aborda o referencial teórico que fundamenta a simplificação de textos, oferecendo uma base sólida para a compreensão dos conceitos envolvidos. O terceiro capítulo dedica-se à revisão de literatura, destacando as contribuições mais significativas na área. No quarto capítulo, introduzimos a métrica ISiM, detalhando as variações exploradas até a definição da versão final. O quinto capítulo descreve exaustivamente os experimentos realizados com a métrica ISiM, incluindo a criação e o refinamento de *corpora*, bem como sua aplicação em modelos de geração de texto próprios e a adaptação de modelos existentes, como o ChatGPT. Por último concluímos o trabalho.

2 Referencial Teórico

Para entendermos melhor o que é uma simplificação de texto, devemos introduzir o conceito de similaridade semântica, pois esta irá garantir que a versão simplificada não perde o sentido em relação à versão original.

2.1 Similaridade Semântica

Quando perguntamos a alguém se dois símbolos são parecidos, devemos primeiro inserir um contexto para que ela possa dar uma resposta correta. Por exemplo, dois alimentos podem ser semelhantes no sabor mas totalmente diferentes na aparência, ou semelhantes na aparência, mas totalmente distintos no preço, duas palavras podem ser semelhantes na escrita mas diferentes no significado, etc. Ou seja, para podermos comparar duas coias precisamos primeiramente estabelecer em qual contexto elas estão inseridas.

Para ajudar na formulação de um contexto temos o uso de duas bases, os corpora (MCENERY, 2012) e as ontologias(NOY; MCGUINNESS et al., 2001). O *corpus* seria uma enorme coleção de texto estruturados que podem ser lidos por uma máquina. Ele também contém evidências informais sobre a relação entre os objetos. Por exemplo, um *corpus* pode nos informar que o par de palavras *(piscina, verão)*, tem mais proximidade semântica do que o par *(piscina, dentista)*, pois é mais comum que a palavra piscina esteja associada à palavra verão numa sentença do que à dentista. Essa co-ocorrência é uma informação importante para uma análise semântica determinar o grau de relação entre dois objetos, pois podemos considerar que dois objetos que estão semanticamente próximos tendem a estar correlacionados em uma sentença (HARISPE et al., 2015). A ontologia, como o próprio nome sugere, engloba o conhecimento a respeito dos objetos em estudo, contribuindo para o desenvolvimento de um compreensível, completo e compartilhável sistema de categorias, rótulos e relações que representam o mundo real de maneira objetiva (KIM; STOREY, 2011). Assim, um par de palavras dentro de uma ontologia estarão ligadas por algo em comum entre elas, por exemplo, as palavras *gato* e *rato* podem estar associadas dentro de uma ontologia como *gato naoGostaDe rato*. Em termos computacionais podemos dizer que a ontologia inclui definições interpretáveis por máquinas de conceitos básicos dentro de um domínio e relações entre eles (NOY; MCGUINNESS et al., 2001).

A similaridade semântica tenta responder à seguinte pergunta: o objeto ‘A’ é próximo ou distante do objeto ‘B’?; onde por distância entendemos o quão similares dentro de um determinado contexto são estes objetos. A similaridade semântica pode ser descrita como uma medida usada para computar a similaridade entre dois conceitos dentro de uma

ontologia (BANU; FATIMA; KHAN, 2015).

O próprio conceito de medir a similaridade entre dois objetos não é trivial. Em (GOODMAN, 1972), o autor diz que dois objetos possuem um número infinito de propriedades iguais e também infintas propriedades diferentes, como por exemplo dois objetos quaisquer que possuem a semelhança de terem mais de um metro, também são semelhantes por terem mais de 0.9999 metros, terem mais de 0.9998 metros, terem mais de 0.9997 metros, etc. Para que possamos ter a possibilidade de medir de um certo grau de similaridade, precisamos selecionar algumas características ou pesá-las de forma diferente. Em (HARISPE et al., 2015) vemos que a comparação entre dois objetos não deve ser feita através de suas propriedades físicas/lexicais, mas sim através do entendimento que o agente que está analisando tem sobre eles, visto que tal representação do objeto (através do seu entendimento) não possui um número infinito mas sim limitado de características.

Mais afundo na similaridade semântica temos a implicação textual e a paráfrase. A implicação textual pode ser definida como a inferência de algo a partir de uma sentença, por exemplo, a hipótese H: *'o menino estava com fome'* pode ser confirmada ao lermos o trecho T: *'o menino comeu todo o hambúrguer'*, ou seja, T implica H. Já a paráfrase representa uma nova informação do sentido de um texto utilizando-se outras palavras, ou seja, reescrever um texto sem que ele perca o sentido original. Segundo (ANDROUTSOPOULOS; MALAKASIOTIS, 2010), podemos inclusive expressar em termos lógicos ($\models$) a implicação textual e a paráfrase. Tendo a representação lógica de dois elementos textuais (T e H) dados por Φ_T e Φ_H, temos que T e H possuem implicação textual entre si caso $(\Phi_T \wedge B) \models \Phi_H$, onde B contém postulados correspondentes ao conhecimento humano em relação àquele assunto. Para as paráfrases se temos dois fragmentos de T, T_1 e T_2, com a representação lógica sendo Φ_1 e Φ_2, temos que T_1 é paráfrase de T_2 se $(\Phi_1 \wedge B) \models \Phi_2$ e $(\Phi_2 \wedge B) \models \Phi_1$.

Para resumir e complementar o que foi explicado acima, iremos listar agora algumas características da similaridade semântica contribuem para o entendimento e a execução de simplificação de texto:

- Manutenção do Significado Central: o principal objetivo da simplificação de um texto é tornar o conteúdo mais acessível, sem perder o significado original. A similaridade semântica é quem assegura que, mesmo após a simplificação, o texto simplificado mantenha as informações centrais do texto original. Isso é crucial para evitar perda de informações importantes.

- Seleção Adequada de Sinônimos: uma prática comum na simplificação de textos é a troca de palavras complexas por outras mais simples. A similaridade semântica ajuda a identificar sinônimos que não apenas são mais fáceis de entender, mas também mantêm o significado original naquele contexto.

- Adaptação Contextual e Personalização: a similaridade semântica permite personalizar o texto simplificado ao público-alvo, pois ela consegue captar o significado original e a forma como ele se relaciona com diferentes contextos e públicos, sendo possível personalizar a simplificação para atender às necessidades específicas de diferentes usuários.

- Ambiguidade e Precisão: a simplificação de texto pode acabar causando ambiguidades que são inerentes à linguagem. A análise semântica pode ajudar a resolver essas ambiguidades no texto original, garantindo que a versão simplificada seja clara e precisa.

2.2 Simplificação de Texto

Segundo (SIDDHARTHAN, 2014) a simplificação de textos pode ser descrita de duas maneiras: a mais restrita, refere-se à redução da complexidade linguística de um texto, preservando suas informações e significados originais. Já em um contexto um pouco mais amplo, abrange também a simplificação conceitual, que visa simplificar tanto o conteúdo quanto a forma do texto, confrontando redundância e clareza para realçar aspectos cruciais, e também utilizando a sumarização de texto, que exclui informações secundárias ou irrelevantes.

Simplificar um texto não é uma tarefa fácil. De acordo com (MAX, 2006), as dificuldades surgem da complexidade inerente da linguagem natural, que requer por exemplo, a desambiguação do sentido das palavras e da estrutura sintática. O autor ainda acrescenta que seria improvável que um sistema pudesse simplificar um texto mantendo-o o mais próximo possível do seu sentido original, a menos que fosse pós editado por um profissional.

Sobre a composição e compreensão de uma sentença, em (SMITH et al., 1989), o autor mostra que os símbolos utilizados para a comunicação humana são divididos em duas categorias: dificuldade semântica e complexidade sintática. Uma descrição dessas características pode ser encontrada em (SCARTON; ALUÍSIO, 2010), onde o autor classifica estas duas propriedades da seguinte forma:

- **Dificuldade semântica**: seria a frequência de uso cotidiano de cada palavra contida na sentença. Segundo (BORMUTH, 1966), inferir o significado de um texto inclui a probabilidade da pessoa identificar a palavra num contexto, ou seja, a baixa frequência de uma palavra pode interferir no processo lógico de inferência do texto e tornar a sentença incompreensível.

- **Complexidade sintática**: neste caso o tamanho da sentença é um influenciador na complexidade sintática. A tendência é de que quanto maior a sentença mais estru-

turas sintáticas são acrescentadas na sentença e mais difícil fica sua interpretação. (STENNER, 1996) chama a atenção para o fato de que algumas vezes, quando reduzimos o tamanho da sentença, ela se torna mais difícil de interpretar. O autor mostra que a melhor fórmula para ligar o tamanho de um texto com sua complexidade é utilizando o *log* do tamanho médio das sentenças (inspiração para nossa métrica).

Ainda segundo (SCARTON; ALUÍSIO, 2010), as unidades semânticas variam em familiaridade e as estruturas sintáticas variam em complexidade. A compreensibilidade ou dificuldade de uma mensagem é governada em grande parte pela familiaridade das unidades semânticas e pela complexidade das estruturas sintáticas usadas na construção da mensagem, que neste trabalho são abordadas na medição da frequência que a palavra aparece em um *corpus* (afeta a compreensibilidade) e o tamanho da sentença (afeta a complexidade sintática). Porém, como visto acima, nem sempre reduzir uma sentença implica necessariamente em uma sentença mais simples.

Existem várias formas de se abordar uma simplificação de texto. Iremos listar abaixo algumas das mais utilizadas na literatura, como estudados e exemplificados nos seguintes *surveys* (SIDDHARTHAN, 2014; SHARDLOW, 2014) e em outros trabalhos (ZHANG; LAPATA, 2017; CROSSLEY; ALLEN; MCNAMARA, 2011; SCARTON; ALUÍSIO, 2010)

- **Simplificação Lexical:** Esta abordagem foca na substituição de palavras complexas e ambíguas por sinônimos mais simples e comumente utilizados no contexto em que se encontra. O objetivo é reduzir a complexidade lexical sem alterar o significado do texto. Esta é a uma das abordagens utilizada neste trabalho.
- **Simplificação Sintática:** Envolve a reestruturação de estruturas complexas em estruturas mais simples, o que inclui a divisão ou redução de sentenças longas e complexas em sentenças mais curtas e simples. Também aborda a simplificação de construções gramaticais complexas e a reorganização de sentenças para obter uma ordem mais clara. A redução do tamanho da sentença é também uma abordagem utilizada neste trabalho.
- **Simplificação Semântica:** Esta abordagem se concentra em tornar o conteúdo e as ideias de um texto mais acessíveis. Diferente da simplificação lexical ou sintática vistas anteriormente, que focam na substituição de palavras e na reestruturação de sentenças, a simplificação semântica usa uma abordagem de compreensão e reformulação de conceitos complexos em termos mais simples.
- **Simplificação Baseada em Regras:** Exemplificado no trabalho Coh-Matrix-Port apresentada no capítulo a seguir, essa abordagem utiliza um conjunto de regras para

guiar o processo de simplificação te texto, regras estas que podem ser baseadas em gramática, sintaxe, ou outros aspectos linguísticos.

- **Uso de Paráfrases:** Pode ser vista como uma união de algumas abordagens vistas acima. Ela inclui a reescrita de textos ou sentenças a fim de torná-los mais simples, sem perder o significado original. Isso pode envolver a reestruturação de sentenças, substituição de palavras e a eliminação de redundâncias, entre outras alterações. Em nosso trabalho utilizamos pares de paráfrases para montar corpora de pares de frases complexo-simples.

- **Sumarização:** Embora essa abordagem possa ser vista como algo não estritamente relacionado a simplificação, o resumo de textos pode ser usado para reduzir a quantidade de informação, focando nas mais importantes e omitindo detalhes menos relevantes. Apesar de não ser necessariamente uma simplificação, várias métricas voltadas para este fim, como Simple-QE e Sum-QE, utilizam este conceito para chegar em uma sentença menos complexa.

Um trabalho muito importante para a construção da nossa métrica pode ser encontrado em (STENNER, 1996). Aqui o autor afirma a teoria sobre a base da nossa métrica citada acima: frequência de palavras e tamanho da sentença. Ele diz que, em relação ao componente semântico, a probabilidade de uma pessoa encontrar uma palavra no contexto é o que mais pesa ao inferir seu significado (BORMUTH, 1966), sendo esta a base para a chamada "teoria da exposição"(MILLER; GILDEA, 1987; STENNER; III; BURDICK, 1983), que explica como o vocabulário receptivo se desenvolve. Nela, no contexto do processamento de linguagem natural (PLN), está relacionada à ideia de que a compreensão do significado de uma palavra é influenciada pela frequência e pelo contexto em que essa palavra é encontrada. Esta teoria pode ser vista como parte do campo mais amplo da semântica distribucional, que se baseia na hipótese de que palavras que aparecem em contextos semelhantes tendem a ter significados semelhantes. Detalhando um pouco mais sobre estes aspectos citados acima sobre essa importante teoria, temos os seguintes pontos; Frequência de Exposição: A teoria sugere que a frequência com que uma pessoa encontra uma palavra em diferentes contextos, influencia em quão robusta será sua compreensão do significado dessa palavra. A exposição repetida ajuda a firmar o conhecimento sobre como e quando a palavra é usada; Contexto de Uso: Como citado mais acima, o significado de uma palavra é parcialmente derivado do conjunto de contextos em que ela aparece. Por exemplo, a palavra "pena", que pode ser associada a uma parte do corpo de uma ave ou ao sentimento sobre alguém, dependendo do contexto em que é usada; Aprendizado Baseado em Exemplo: Não faz parte direta da nossa métrica, mas é um aspecto importante da teoria da exposição, que também se alinha com a ideia de que aprendemos o significado das palavras e o uso da linguagem através de exemplos concretos e não abstratos. Isso é particularmente evidente na forma como as crianças aprendem a linguagem.

Para reforçar essa teoria, em (STENNER, 1996) foram analisados mais de 50 componentes semânticos, e eles chegaram a conclusão que o que mais interfere na compreensão do sentido da sentença é a frequência do uso de cada palavra em um contexto.

Existem ainda dois outros estudos importantes relacionados à familiaridade e raridade das palavras (KLARE et al., 1963; CARROLL; DAVIES, 1971) que reforçam a ideia da importância de termos palavras mais comumente usadas nas sentenças para fins de simplificação, reforçando ainda mais a base da nossa métrica e seu fundamento.

Também segundo (STENNER, 1996):

> "Conhecer a frequência das palavras como elas são usadas na comunicação escrita e oral fornece a melhor maneira de inferir a probabilidade de que uma palavra será encontrada e, assim, tornar-se parte do vocabulário receptivo de um indivíduo."

2.3 Modelo base de aprendizagem

Conforme mencionado em discussões anteriores, a geração de texto desprovida de um modelo que possa qualificar de maneira eficaz a conexão semântica entre os conteúdos textuais não é a abordagem mais adequada. De acordo com (ROCHA; CARDOSO, 2018), uma estrutura representativa que pode servir como fundamento para os modelos contemporâneos de análise semântica está ilustrada na Figura 1.

No modelo proposto, as sentenças que estão sob análise, comumente referenciadas como T e H, mas que, no contexto de paráfrases, poderiam ser igualmente designadas como T1 e T2, são inicialmente submetidas a um processo de codificação, onde são transformadas em *tokens*. Este método de codificação, que se destaca por sua vasta aplicação, é representado na parte inferior da figura, especificamente nas seções denominadas *Sentence T* e *Sentence H*.

Após este passo inicial de codificação, os *tokens* obtidos são então mapeados para vetores, um processo conhecido como *word embedding*. Um exemplo notável desse tipo de mecanismo é o *Word2Vec* (RONG, 2014). Em tal representação vetorial, uma característica fascinante é que palavras com semelhanças semânticas tendem a ser localizadas de maneira próxima umas das outras no espaço vetorial, conforme elucidado por (LEVY; GOLDBERG, 2014). Após esse mapeamento, esses vetores específicos são canalizados para a subsequente fase do processo, a codificação da sentença. Vale ressaltar que existem variadas técnicas para executar essa codificação, e uma análise mais profunda sobre essas técnicas será apresentada posteriormente.

Os vetores resultantes, que possuem dimensões pré-definidas, são então utilizados como entradas para uma rede neural de arquitetura profunda. Nesta fase crucial, na ca-

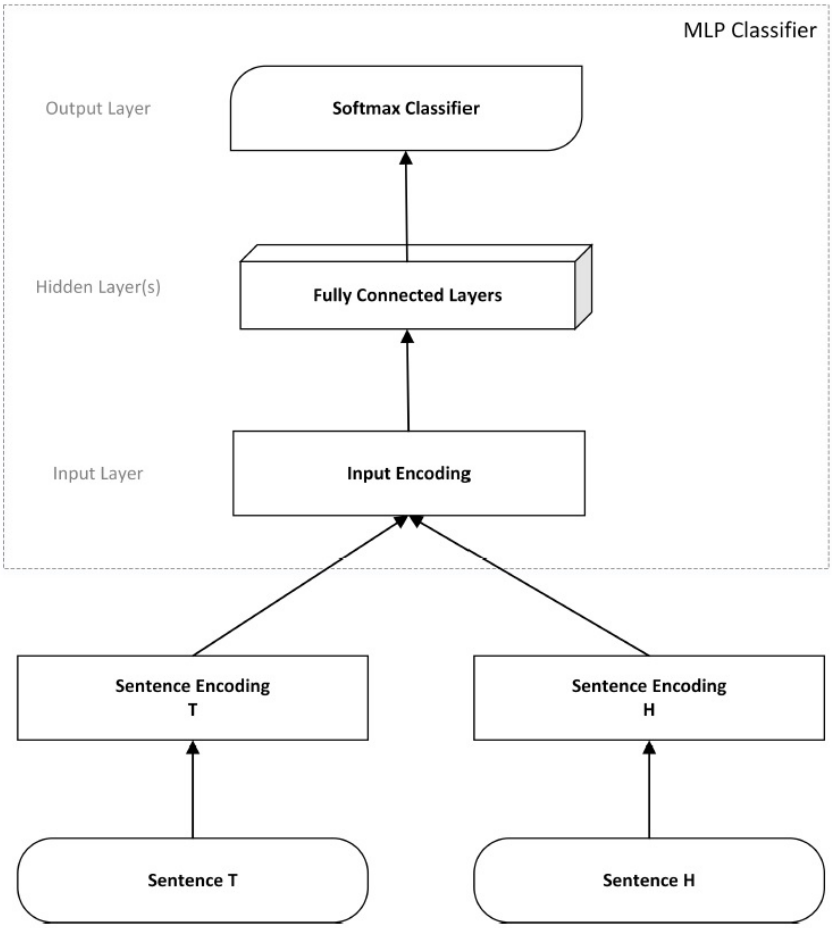

Figura 1 – Modelo base de processamento de sentenças em redes neurais. Fonte: (ROCHA; CARDOSO, 2018)

mada final da rede, funções específicas são empregadas, como é o caso da *Softmax*, cujo objetivo primordial é discernir e predizer a relação semântica entre as sentenças analisadas. Cabe destacar que a natureza dessas relações semânticas pode variar consideravelmente, dependendo especificamente do *corpus* utilizado para treinamento e avaliação. Para ilustrar, no trabalho apresentado por (FONSECA et al., 2016), as relações são avaliadas em uma escala numérica de 1 a 5. Nessa escala, o valor 1 indica uma clara ausência de correlação entre as sentenças, enquanto o valor 5 sugere uma equivalência semântica perfeita entre elas. Outro exemplo de interesse é o *corpus ASSIN*, no qual as relações são classificadas de forma mais categórica, dividindo-se em: ausência de correlação, implicação e paráfrase.

2.4 Redes neurais profundas

Com a ascensão das redes neurais profundas, o campo de análise de linguagem natural testemunhou uma mudança de paradigma. Este avanço tecnológico não apenas possibilitou a utilização de redes neurais em si, mas também acomodou a inclusão de *corpus* de treinamento de grandes dimensões, proporcionando um refinamento sem prece-

dentes nos modelos.

No estudo detalhado em (BOWMAN et al., 2015), são apresentadas três arquiteturas distintas: *Sum of Words*, *Recorrent Neural Network* (RNN) e *Long Short Term Memory Recorrent Neural Network* (LSTM RNN). O mecanismo fundamental em todas essas arquiteturas consiste em mapear cada sentença individualmente para um espaço vetorial multidimensional. Após este mapeamento, os vetores resultantes são amalgamados, resultando em um vetor compósito. Este vetor compósito, por sua vez, serve como insumo primário para a camada de entrada da rede neural. Em uma análise mais detida, quando nos referimos à RNN e à LSTM RNN, há uma etapa adicional: antes de serem integrados à rede neural, os vetores passam por essas arquiteturas específicas, garantindo uma pré-processamento mais rico.

O trabalho de (ROCHA; CARDOSO, 2018) entra em nuances técnicas deste processo. Quando temos uma sentença S que abriga uma sequência de n palavras $(w_1, ..., w_n)$, o valor associado a S é deduzido através da equação $\vec{S} = \sum_{k=1}^{N} \vec{e}(w_k)$, onde $\vec{e}(w_k)$ é a representação vetorial da palavra w_k. No cenário em que as arquiteturas empregam redes neurais recorrentes, o processamento de uma sequência de n palavras $(w_1, \ ... \ , w_n)$ resulta na produção de uma sequência correspondente de n representações $(\overrightarrow{h_1}, \ ... \ , \overrightarrow{h_n})$. Esta sequência é definida pela relação $\overrightarrow{h_n} = \overrightarrow{RNN}(\vec{e}(w_1), \ ..., \ \vec{e}(w_n)$. A culminância deste processo é a representação da sentença pela última camada, explicitada como $\overrightarrow{h_n}$.

Finalmente, os dados consolidados na Tabela 1 extraídos de (BOWMAN et al., 2015) oferecem uma visão dos resultados. Eles destacam que, quando comparada às outras, a arquitetura LSTM manifesta superioridade em termos de desempenho.

Sentence Model	**Train**	**Test**
100d Sum of Words	79.3	75.3
RNN	73.1	72.2
LSTM RNN	**84.8**	**77.6**

Tabela 1 – Resultados dos testes em (BOWMAN et al., 2015).

2.5 Sequence to sequence

No contexto apresentado anteriormente, as redes *LSTM* (Figura 2) (HOCHREITER; SCHMIDHUBER, 1997) se destacam como fundamentais para a evolução da Processamento de Linguagem Natural. Essa evolução é notavelmente evidenciada na maneira como a similaridade textual é gerada e medida. A principal virtude das *LSTM* nesse âmbito é sua capacidade de aprender eficientemente representações de texto no formato de *word embedding*. Isso é exemplificado pelas abordagens como *Word2Vec* (RONG, 2014) e GLOVE (PENNINGTON; SOCHER; MANNING, 2014), que se beneficiaram imensamente dessa estrutura neural.

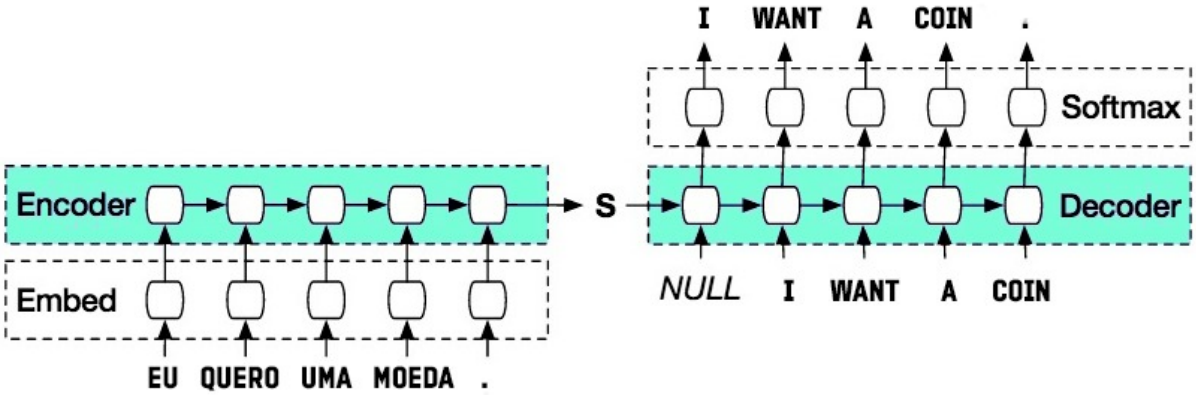

Figura 2 – Modelo sequence to sequence.

A contribuição das *LSTM* não se limitou apenas à representação do texto. Elas também serviram como alicerce para o avanço subsequente na pesquisa de PLN: a adoção dos modelos *sequence to sequence* (SUTSKEVER; VINYALS; LE, 2014). Esses modelos, com sua arquitetura *encoder / decoder*, foram inicialmente concebidos para enfrentar desafios de tradução de textos. No entanto, à medida que a pesquisa avançou, ficou evidente sua versatilidade. Conforme apontado em (CELIKYILMAZ; CLARK; GAO, 2020), esses modelos provaram ser extremamente eficazes não apenas no domínio da tradução, mas também em áreas diversificadas da PLN, como a geração de textos, solidificando sua relevância no campo.

2.6 Encoder - Decoder

Os modelos *sequence to sequence* representaram um marco significativo no domínio do Processamento de Linguagem Natural. Especificamente, os modelos *encoder / decoder* (Figura 3) emergiram como estruturas fundamentais neste paradigma. No entanto, à medida que a aplicação destes modelos cresceu em complexidade e abrangência, certas limitações tornaram-se evidentes.

A solução para algumas dessas limitações veio com a introdução dos modelos de atenção, conforme proposto por (BAHDANAU; CHO; BENGIO, 2014) em 2014. Esta proposta abordou dois problemas principais identificados nos modelos anteriores. O primeiro desafio era o fenômeno do esquecimento, no qual informações vitais de iterações anteriores poderiam ser suplantadas à medida que o modelo avançava em suas iterações de aprendizagem.

O segundo desafio, por outro lado, residia na abordagem global que os modelos *sequence to sequence* adotavam. Esta abordagem, ao tratar a sentença em sua totalidade, muitas vezes negligenciava a importância de *tokens* individuais. Por exemplo, na sentença "No dia de hoje, meu time de futebol ganhou uma partida por 3x1", ao analisar o *token*

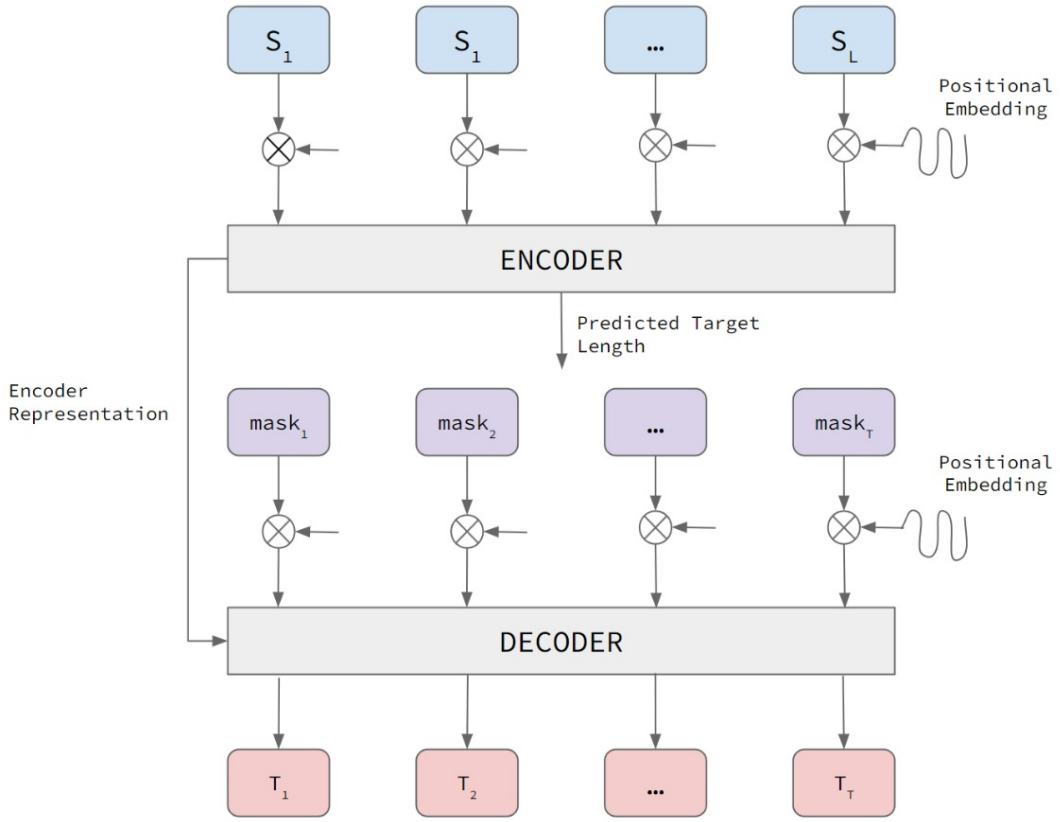

Figura 3 – Modelo genérico de Encoder Decoder. Fonte: Adaptado de (BABU et al., 2021).

"time", *tokens* como "dias"e "hoje"poderiam não oferecer contribuições significativas para o contexto em que "time"está inserido.

Neste cenário, o modelo com componente de atenção emergiu como uma estratégia crucial. Dotado da capacidade de discernir a relevância de palavras ou *tokens* em determinados contextos, otimizou o processo de aprendizagem da rede. Esta eficácia não se limita à teoria, mas é corroborada na prática, como evidenciado em (HU, 2019), demonstrando sua relevância na PLN contemporânea e será melhor explicado abaixo.

2.7 Modelos de atenção

Os modelos de atenção, que surgiram como uma solução robusta para as limitações dos modelos *encoder/decoder*, possuem uma particularidade intrínseca em seu funcionamento. Eles são capazes de criar uma variável de contexto, denotada como C_i, para cada *token* emitido pelo *decoder* (SUTSKEVER; VINYALS; LE, 2014). Este aspecto é notável na medida em que não se limita à última variável latente produzida pelo codificador, mas sim engloba todas as variáveis latentes geradas durante o processo.

Para ilustrar de forma mais específica, consideremos as variáveis latentes produzidas pelo codificador, representadas por $\overrightarrow{h_j}$. Estas variáveis são produzidas para cada entrada na sequência, totalizando N valores. Dessa forma, a variável de contexto C_i, as-

sociada à i-ésima saída do decodificador, é computada considerando todas essas variáveis latentes. Assim, apresentamos o mecanismo pelo qual o contexto C_i é determinado, fornecendo uma visão mais clara sobre a eficiência e precisão destes modelos.

$$C_i = \sum_{i=1}^{N} a_{\mathrm{i,j}} h_j$$

onde $a_{\mathrm{i,j}}$ é um peso entre zero e um indicando o tanto que o decodificador deve prestar atenção na j-ésima variável latente produzida pela codificador enquanto produz a i-ésima saída y_i. Os valores de $a_{\mathrm{i,j}}$ são aprendidos durante o treinamento da rede e são calculados de acordo com (CHOROWSKI et al., 2015) como:

$$a_{\mathrm{i,j}} = \frac{exp(e_{\mathrm{i,j}})}{\sum_{k}^{N} exp(e_{\mathrm{i,k}})}$$

onde os valores de $e_{\mathrm{i,j}}$ são produzidos por uma RNA a que recebe como entrada a variável latente $S_{\mathrm{i\text{-}1}}$, que é a variável fornecida como entrada para a recorrência produzir a saída y_i. e a variável latente $\overrightarrow{h_j}$ produzida pelo codificador. A expressão fica:

$$e_{\mathrm{i,j}} = a(S_{\mathrm{i\text{-}1}}, \overrightarrow{h_j})$$

A Figura 4 abaixo extraída de (GOODFELLOW; BENGIO; COURVILLE, 2016) mostra a representação gráfica do modelo básico de atenção.

Em (HU, 2019) vemos várias formas diferentes de modelos de atenção que podem ser utilizados e que possuem resultados diferentes dependendo da aplicação.

2.8 Transformers

Por último temos a arquitetura de *Transformers* (VASWANI et al., 2017) que propôs o mecanismo denominado por eles de *Self Attention*. O Transformer consiste em uma pilha de *encoders* que processam entradas de qualquer tamanho e um outro conjunto de *decoders* para produzir as sentenças. Ao contrário das arquiteturas RNN e LSTM, nos transformes a sequência de dados (no nosso caso palavras de um texto) pode ser passado de forma paralela, o que torna o treinamento muito mais rápido. Isso é possível por causa da inserção e uma informação a mais no *embedding* das palavras, que é o *embedding* da posição delas na sentença, que possibilitou que várias palavras fossem processadas ao mesmo tempo sem perder o contexto de posicionamento de cada palavra. Este contexto é extremamente importante pois a posição da palavra na sentença pode dizer muitos obre ela, por exemplo:

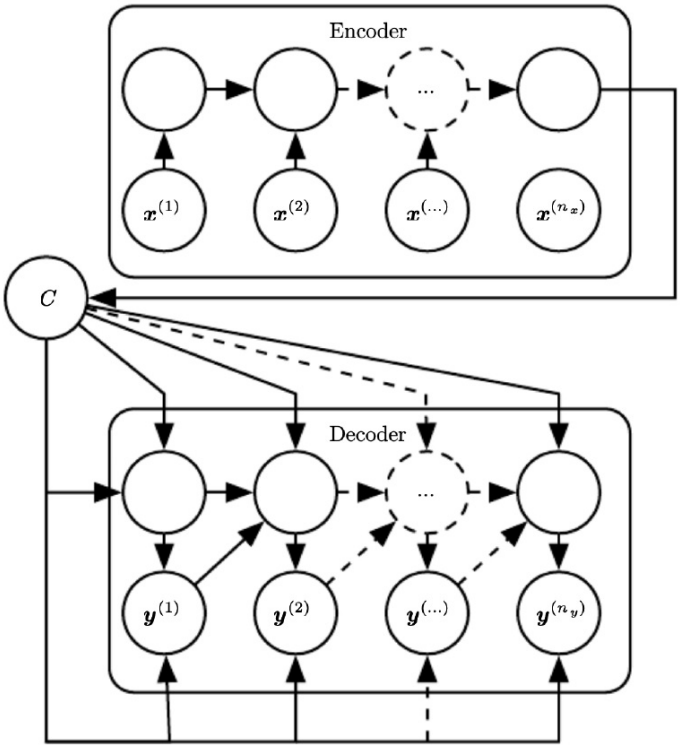

Figura 4 – Modelo utilizando o conceito de atenção. Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016)

- O **cachorro** é o melhor amigo daquele homem.
- O melhor amigo daquele homem é um **cachorro**.

Repare que a palavra cachorro pode ter um significado pejorativo na última sentença e a diferença para a sentença anterior é apenas a posição onde ela se encontra.

Dentro da arquitetura dos *transformers* ainda podemos encontrar algumas diferenças significativas para os modelos de atenção anteriores, que são as camadas de *Mult-Head Attention* nos *Encoders* e a de *Masked Mult-Head Attention* nos *Decoders*.

O primeiro, segundo (VASWANI et al., 2017), ao invés de um único treinamento para criar o vetor de atenção para cada palavra da sentença, realiza vários treinamentos diferentes e finaliza com o vetor de atenção sendo a média destes treinamentos, evitando também que o vetor preste muita atenção na própria palavra que está sendo analisada no momento, o que é inútil para o treinamento. Após esta etapa toda essa informação é jogada para uma camada de rede neural, também organizada para ser paralelizável, e de lá a informação final do *encoder* é enviada ao *decoder*.

Já o *Masked Mult-Head Attention*, segundo o autor, irá ocultar as palavras ainda não analisadas no *decoder*, para que o treinamento ocorra de forma mais eficaz. Por exemplo, se o problema em questão for a tradução de um texto de português para inglês, por exemplo de "Eu gosto de comer pão"para "*I like to eat bread*", durante o processo de

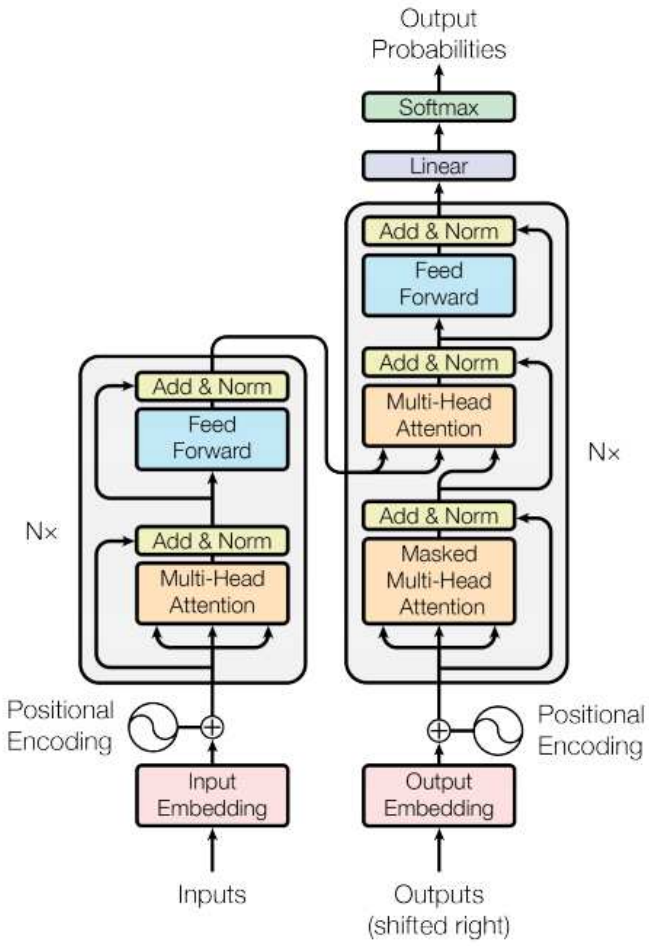

Figura 5 – Modelo de transformers. Fonte: (VASWANI et al., 2017).

aprendizagem da tradução o *decoder* não pode saber a sentença completa, pois assim ele estaria apenas "cuspindo"as palavras corretas para a tradução. O que acontece de fato é que a sentença em inglês tem as próximas palavras ocultadas pela camada *Masked Mult-Head Attention* para que este possa ir aprendendo aos poucos qual a próxima palavra correta da sentença, por exemplo o *decoder* em um momento da tradução sabe apenas *"I like"*, e as palavras *"to"*, *"eat"* e *"bread"* tem valor igual a zero no vetor de atenção, o que é o mesmo que dizer neste momento que elas ainda não existem.

Ainda dentro do *decoder*, a próxima etapa é o que o autor chama de *Encoder-Decoder Attention* onde acontece o mapeamento da sentença do *encoder* e do *decoder*, ou seja, a fase onde se aprende a relação de um com o outro. Voltando ao exemplo de tradução, seria como se nessa etapa o *Transformer* estivesse aprendendo como as palavras em português se relacionam com as de inglês para poder fazer a melhor tradução das sentenças. Após esse mapeamento, o resultado é enviado para mais uma rede neural linear e uma camada *softmax* que irá dizer quais as probabilidades de cada palavra serem a correta para a sequência da sentença.

2.9 Transfer Learning

O conceito de *Transfer Learning* consolidou-se como um pilar central no avanço contemporâneo dos modelos de PLN, promovendo uma revolução nas abordagens e técnicas empregadas (RAFFEL et al., 2019). Modelos de destaque, como ELMo (PETERS et al., 2018), BERT (DEVLIN et al., 2018), RoBERTa (LIU et al., 2019) e XLNet (YANG et al., 2019), incorporam e capitalizam esse método em suas respectivas propostas. A ideia de transferência de conhecimento tem sido estudada há muito tempo na área de aprendizado de máquina e em outras áreas relacionadas (ANDRYCHOWICZ et al., 2016; GANIN; LEMPITSKY, 2015; HOWARD; RUDER, 2018). O *Transfer Learning* foi popularizado na década de 2000, com o crescimento da complexidade dos modelos de aprendizado de máquina e a necessidade de compartilhar conhecimento entre diferentes tarefas. Desde então, muitos artigos e estudos têm sido publicados sobre o tema, incluindo aplicações em diversos campos, como visão computacional, processamento de linguagem natural e reconhecimento de fala.

Essencialmente, *Transfer Learning* emprega uma estratégia em que um modelo é pré-treinado em vastos conjuntos de texto não rotulados por meio de aprendizagem auto supervisionada. Posteriormente, esse modelo pré-treinado é refinado ou adaptado para um conjunto de dados-alvo específico e geralmente mais limitado. Este refinamento tem demonstrado consistentemente uma eficácia superior em comparação ao treinamento de um modelo exclusivamente no conjunto de dados alvo.

No contexto deste estudo, a escolha recai sobre o T5, um modelo considerado uma referência atual em *Transfer Learning* (conforme ilustrado na Figura 6) (RAFFEL et al., 2019). O T5 se destaca por sua versatilidade, funcionando como uma estrutura *text-to-text*. Isso significa que é capaz de empregar um único conjunto de modelo, função de perda e hiperparâmetros para uma variedade de tarefas em PLN. Esta gama de tarefas abrange desde tradução de texto e resumo de documentos até resposta a perguntas e classificação de textos. Uma característica adicional que diferencia o T5 é seu método inovador para geração de texto, denominado *fill in the blank*. Este método distingue-se das técnicas empregadas em modelos como GPT-2 (BROWN et al., 2020), que são amplamente reconhecidos por suas capacidades de geração de texto. Enquanto estes últimos enfocam na previsão da palavra subsequente em um texto, o T5 se concentra em identificar e inserir palavras que completam uma sentença em um espaço designado. Este espaço, notavelmente, inclui um indicador numérico que orienta o modelo sobre quantas palavras devem ser inseridas para completar adequadamente a sentença.

A métrica criada neste trabalho utilizou estas definições acima para buscar uma forma simples de se comparar sentenças e analisar a complexidade entre elas. Utilizando o tamanho da sentença e a frequência da utilização das palavras no contexto (linguagem) atual, os resultados apresentados ao final mostrarão que partir deste princípio nos deu

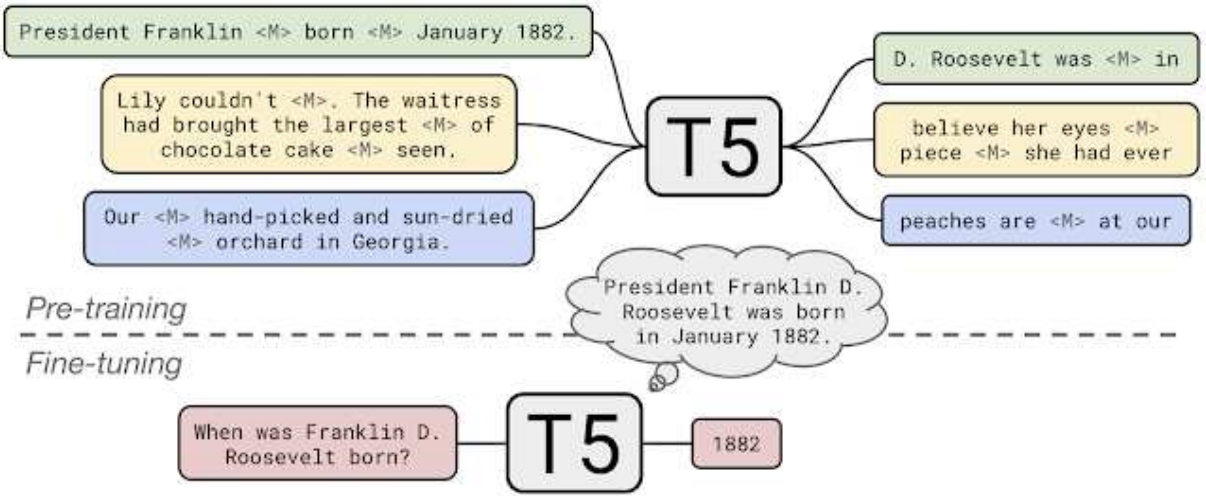

Figura 6 – Modelo de Transfer Learning T5. Fonte: (RAFFEL et al., 2019)

bons resultados apesar da simplicidade da métrica.

3 Trabalhos Relacionados

Nesta seção apresentamos alguns estudos fundamentais sobre métricas que figuram na literatura acadêmica, elucidando as vantagens e desvantagens de cada um. Inicialmente, é pertinente destacar o posicionamento articulado por (ALVA-MANCHEGO; SCARTON; SPECIA, 2021) em relação à necessidade de se investir esforços na criação e refinamento de métricas mais eficazes para simplificação textual.

Os referidos autores oferecem uma análise crítica acerca da adequação e precisão de métricas automáticas amplamente conhecidas, como o BLEU e SARI. Tais métricas, embora detenham sua relevância, apresentam limitações substanciais, muitas vezes resultando em avaliações que divergem significativamente das avaliações realizadas por especialistas humanos. Na investigação conduzida, identificam-se três problemas preponderantes:

Uma notável discordância entre os resultados produzidos pelas métricas automáticas e as avaliações manuais realizadas por humanos; A incapacidade de tais métricas em distinguir e categorizar variados tipos de erros; A inadequação de certas métricas automáticas quando aplicadas a tarefas específicas no campo da linguística computacional. Em acréscimo, os autores sublinham a urgência na formulação de métodos avaliativos mais robustos, que sejam inclusivos e que contemplem perspectivas humanas, ressaltando que abordagens automatizadas, por mais avançadas que sejam, não devem ser consideradas como substitutas infalíveis da avaliação humana. O artigo em questão propõe uma reflexão profunda sobre a eficácia das métricas automáticas atuais e enfatiza a importância de inovação na área, visando avaliar o desempenho de modelos de processamento de linguagem natural com maior acurácia.

Ao concluirmos esta seção, aspiramos que os leitores obtenham uma compreensão ampliada e clara a respeito da operacionalização das métricas discutidas, bem como a urgente necessidade de se criar uma métrica para aprimorar a simplificação textual automatizada.

3.1 BLEU

A métrica BLEU (*Bilingual Evaluation Understudy*), conforme apresentada em (PAPINENI et al., 2002), é extensivamente aplicada no campo da tradução automática de textos e, posteriormente, foi incorporada ao domínio da simplificação textual. Sua metodologia requer a introdução de uma sentença de referência e a saída gerada pelo modelo em questão. Com base nesses dados, a métrica executa cálculos precisos para avaliar a qualidade da tradução. No contexto de nosso estudo, a análise foca em determinar

a simplicidade da sentença gerada em contraste com a sentença original e a referência fornecida.

É pertinente observar que a BLEU adota uma abordagem conservadora em sua avaliação. Em outras palavras, divergências substanciais entre a saída gerada e a sentença de referência acarretam em penalizações no escore final. Esta característica intrínseca da métrica tem origem em sua concepção inicial voltada para a tradução automática. No universo da tradução, a fidelidade ao texto-fonte e a correspondência próxima ao conteúdo de referência são indicativos de uma tradução eficaz. Entretanto, esta inclinação conservadora, que na tradução é vista como uma virtude por preservar a integridade gramatical e semântica da sentença, pode manifestar-se como um obstáculo no campo da simplificação textual. Em situações onde modificações substanciais são necessárias para alcançar uma simplificação adequada, a métrica BLEU, ao penalizar tais alterações, pode não refletir adequadamente o sucesso do processo de simplificação.

Para entendermos o básico do cálculo da pontuação da sentença feita pelo BLUE, vamos analisar a seguinte tradução de alemão para inglês:

- ***Entrada***: Ich bin zur Zeit nicht im Buro
- ***Referência***: I am currently out of the office
- ***Saída***: I am currently not in the office

Na sentença resultante, identificamos um conjunto de 7 termos. Estes serão submetidos a uma análise com base no n-grama, utilizando-se da sentença de referência para tal comparação. Se adotarmos a abordagem de um unigrama no contexto do n-grama, a análise se desenrolará da seguinte maneira:

$$p_1 = \frac{1+1+1+0+0+1+1}{7} = \frac{5}{7} = 0.71 \tag{3.1}$$

onde unidades designadas pelo número 1 indicam os termos da sentença resultante que encontraram correspondência nos termos da sentença de referência, enquanto o número 0 denota aqueles que não apresentaram tal coincidência. Ao analisarmos um exemplo simples, como a sentença "I am", nota-se que a pontuação atribuída seria de 100%, o que claramente não é uma avaliação precisa do contexto. A fim de contornar tais inconsistências e proporcionar uma análise mais robusta, o BLEU incorpora uma formulação matemática específica que penaliza sentenças excessivamente concisas (*Brevity Penalty (BP)*:

$$BP = 1, sec > r \tag{3.2}$$

$$BP = e^{1-r/c}, sec \leq r \tag{3.3}$$

onde c representa a extensão da sentença gerada e r indica a dimensão da sentença de referência. Em 3.4, é apresentada a formulação completa do BLEU, que integra tanto o fator de penalização quanto a agregação dos valores obtidos por meio das comparações n-grama entre a sentença produzida e sua respectiva referência. A partir dessa formulação, obtemos o que é denominado Bleu Score (BS).

$$BS = BP \cdot e^{\left(\frac{1}{n}\sum_{n=1}^{N} P_n\right)} \tag{3.4}$$

onde N é o maior n-grama utilizado.

3.2 iBLEU

O iBLEU (SUN; ZHOU, 2012) apresenta-se como uma adaptação da métrica BLEU, destinada especificamente para a criação de paráfrases. Esse método incorpora a técnica do *pivot* de tradução. Basicamente, ele inicia com um texto no idioma X, converte-o para o idioma Y e, subsequentemente, reverterá essa tradução para o idioma X. A esperança é que o texto reconvertido para o idioma X funcione como uma paráfrase do conteúdo inicial.

Ao contrário do BLEU tradicional, o iBLEU integra em suas operações a sentença original, daí a origem do prefixo "i"em iBLEU, que remete ao conceito de *input aware*, ou conscientização da entrada. Sua fórmula considera não apenas as características originais da métrica BLEU, mas também adaptações específicas para abordar a paráfrase, e é articulada da seguinte maneira:

$$iBLEU(s, r_s, c) = \alpha BLEU - (1 - \alpha)BLEU(c, s) \tag{3.5}$$

onde a variável s denota a sentença original, enquanto r_s é a sentença de referência. Por sua vez, c representa a sentença candidata. O coeficiente α serve como um parâmetro de equilíbrio, governando a relação entre a adequação e a dissimilaridade. Valores menores para α implicam em uma punição mais severa para a autoparáfrase, e modificações em α possibilitam a geração de distintos tipos de paráfrases.

3.3 FKBLEU

O FKBLEU (XU et al., 2016) é uma métrica que integra duas abordagens distintas. A primeira delas é uma conhecida fórmula designada como *Flesch-Kincaid Index*(KINCAID et al., 1975) (FK). Este índice avalia a legibilidade de um texto com base

no seu número total de palavras, sílabas e sentenças. A formulação do FK é apresentada a seguir:

$$FK = 0.39x\left(\frac{\#words}{\#sentences}\right) + 11.8x\left(\frac{\#syllables}{\#words}\right) - 15.59 \tag{3.6}$$

O FK foi adaptado de modo a avaliar sentenças individuais e para incluir tokens de pontuação, tratando-os como se fossem uma sílaba. Esse ajuste evita que elementos de pontuação sejam descartados sem critério. O FK, por natureza, avalia a legibilidade sob a premissa de que o texto já se encontra bem estruturado. Ao amalgamar as características do FK com as do BLEU, obtém-se a formulação do FKBLEU, expressa conforme a fórmula a seguir:"

$$FKBLEU = iBLEU(I,R,O)xFKdiff(I,O) \tag{3.7}$$

$$FKdiff = sigmoid(FKO) - FK(I) \tag{3.8}$$

onde, na equação, um valor mais elevado de FKBLEU indica uma simplificação mais eficaz da sentença. É evidente que esta métrica representa uma adaptação do BLEU com o objetivo de avaliar mais adequadamente a simplificação textual, considerando que o BLEU foi originalmente projetado para avaliar a qualidade das traduções e não especificamente a simplificação de textos.

3.4 SARI

O SARI (**S**ystem Output **A**gainst **R**eferences and **I**nput Sentence) (XU et al., 2016) avalia a eficácia das palavras que foram introduzidas, removidas ou retidas em comparação com a sentença original e a referência. Este indicador examina três operações fundamentais e as quantifica considerando o *recall* ($r(n)$), que se refere ao número de palavras reutilizadas da sentença de referência, e a precisão do n-grama ($p(n)$), detalhadas subsequentemente.

3.4.1 Adição

Aqui são pontuadas palavras na sentença emitida pelo modelo (*output* (O)) que não estavam na entrada (*input* (I)) mas apareciam em alguma das sentenças de referência (*reference* (R)), tal que $(O \cap \bar{I} \cap R)$. As equações podem ser vistas abaixo:

$$P_{add}(n) = \frac{\sum_{g \in O} min(\#_g(O \cap \bar{I}), \#_g(R)}{\sum_{g \in O} \#_g(O \cap \bar{I})} \tag{3.9}$$

$$r_{add}(n) = \frac{\sum_{g \in O} min(\#_g(O \cap \bar{I}), \#_g(R)}{\sum_{g \in O} \#_g(R \cap \bar{I})} \tag{3.10}$$

onde $\#_g$ é um indicador binário da ocorrência de um n-grama g em um determinado exemplo e:

$$\#_g(O \cup \bar{I}) = max(\#_g(O) - \#_g(I), O) \tag{3.11}$$

$$\#_g(R \cup \bar{I}) = max(\#_g(R) - \#_g(I), O) \tag{3.12}$$

O SARI adota uma abordagem distinta quando compara as palavras: ele examina palavras que não estavam presentes no *input* mas que surgiram na *reference*, atribuindo assim valor às modificações. Esta característica contrasta com a métrica BLEU, que limitava sua verificação à correspondência entre palavras do *output* e da *reference*. Dada a natureza intrínseca da simplificação textual, que frequentemente demanda alterações significativas no conteúdo, a métrica SARI demonstra ser mais apropriada e eficaz para avaliar a eficiência dessas transformações.

3.4.2 Manutenção

Neste contexto, o SARI se concentra nas palavras que, originárias da sentença de referência, foram preservadas na sentença resultante, atribuindo-lhes uma pontuação. As fórmulas utilizadas para avaliar e pontuar tais palavras são apresentadas a seguir:

$$P_{keep}(n) = \frac{\sum_{g \in I} min(\#_g(I \cap O), \#_g(I \cap R')}{\sum_{g \in O} \#_g(I \cap O)} \tag{3.13}$$

$$r_{keep}(n) = \frac{\sum_{g \in I} min(\#_g(I \cap O), \#_g(I \cap R')}{\sum_{g \in O} \#_g(I \cap R')} \tag{3.14}$$

Onde R' marca as contagens de n-gramas sobre R com frações, por exemplo, se um unigrama ocorre em duas do total de r referências, então sua contagem é ponderada por 2/r no cálculo de precisão e do *recall*

$$\#_g(I \cap O) = min(\#_g(I), \#_g(O)) \tag{3.15}$$

$$\#_g(I \cap R') = min(\#_g(I), \#_g(R)/r) \tag{3.16}$$

3.4.3 Remoção

Nesta operação, o SARI atribui pontuação às palavras que foram eliminadas, considerando apenas a precisão e não o *recall*. A razão para isso é que a remoção excessiva pode comprometer a clareza do texto, enquanto a não remoção não tem um impacto tão

significativo. As fórmulas utilizadas para calcular a pontuação relacionada à remoção de palavras são apresentadas a seguir:

$$P_{del}(n) = \frac{\sum_{g \in I} min(\#_g(I \cap \bar{O}), \#_g(I \cap \bar{R}')}{\sum_{g \in O} \#_g(I \cap \bar{O})} \tag{3.17}$$

$$\#_g(I \cap \bar{O}) = max(\#_g(I) - \#_g(O), O) \tag{3.18}$$

$$\#_g(I \cap \bar{R}') = max(\#_g(I) - \#_g(R)/r, O) \tag{3.19}$$

A métrica SARI, junto com a BLEU, é mais utilizada em comparações de modelos em estudos atuais (MADDELA et al., 2022; GRABAR; SAGGION, 2022; BLINOVA et al., 2023; VADLAMANNATI; ŞAHIN, 2023; ALKALDI; INKPEN, 2023; LI; SHARDLOW, 2024). Comparando-a com a métrica BLEU citada anteriormente, podemos dizer que SARI é mais receptiva a mudanças, pois ela é capaz de medir se uma adição ou remoção de palavra na sentença foi boa ou ruim, comparando com as sentenças de referências do *corpus*. Por ser mais tolerante a alterações, ela é mais utilizada do que a BLEU que penaliza mudanças e tenta ser mais conservativa, uma característica intrínseca da tradução de textos (tarefa para qual ela foi originalmente projetada).

3.5 Simple-QE

O *Simple-QE* (KRIZ; APIDIANAKI; CALLISON-BURCH, 2020) é uma métrica baseada em BERT (DEVLIN et al., 2019) derivada do *Sum-QE* (XENOULEAS et al., 2019), que é uma métrica para medir a qualidade de resumos de textos. A sigla *QE* vem do inglês *quality estimation*, que significa estimativa de qualidade, logo esta métrica propõem uma estimativa de qualidade do resumo de textos. Outro aspecto importante dessa métrica é que ela não depende das sentenças de referência, diminuindo a necessidade de trabalho humano para fazer a estimativa final da qualidade da simplificação gerada.

Nessa métrica, os autores focaram em três importantes aspectos das sentenças:

- ***Fluência***: diz o quão bem formado é o texto gerado
- ***Adequação***: diz o quanto de significado do texto original foi preservado
- ***Complexidade***: diz o quão simples é a sentença gerada

Simple-QE adaptou o modelo M-3 do *Sum-QE* inserindo também a sentença original na concatenação dos *embeddings* para passá-los para a camada densa da rede, como podemos ver na parte esquerda da Figura 7.

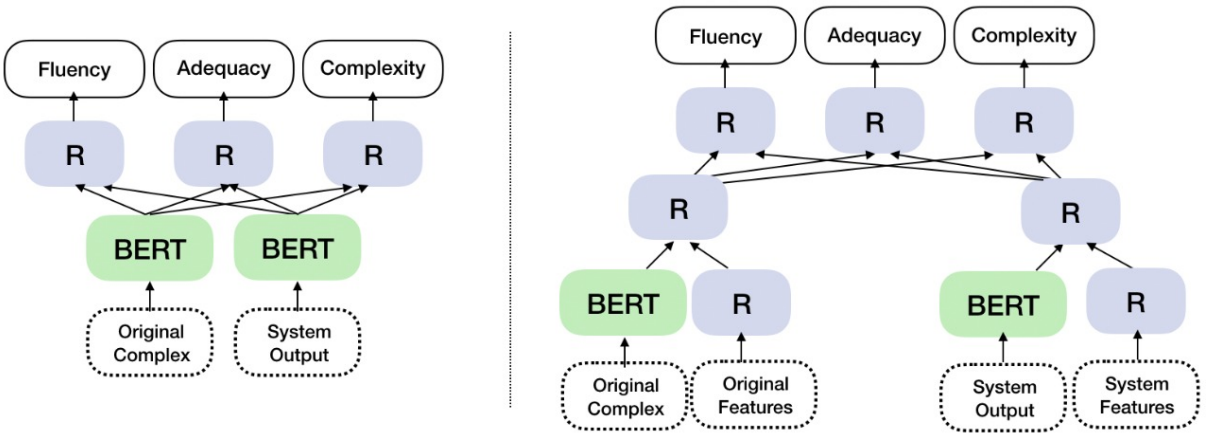

Figura 7 – A arquitetura Simple-QE, com e sem adição de *features* correspondentes à sentença original e à saída do sistema. R denota uma camada de regressão. Fonte: (KRIZ; APIDIANAKI; CALLISON-BURCH, 2020)

Na parte direita da figura, vemos uma tentativa dos autores de inserir *features* como o comprimento médio das palavras de uma uma sentença em caracteres e em sílabas, a frequência do unigrama, o comprimento da sentença e a complexidade em altura da análise sintática da sentença. Segundo os autores esta inserção de *features* não colaborou significativamente com a melhoria da qualidade da métrica.

O autor comparou o trabalho com 3 outras métricas

- BLEU
- SARI:
- Três métricas baseadas em BERT:
  - BERT LM: mascara um *token* e mede a probabilidade da palavra aparecer (mede fluência)
  - BERT SIM: converte o *input* e o *output* para um vetor BERT e calcula a similaridade de coseno (mede adequação)
  - SUM-QE

Segundo os resultados do artigo, o *Simple-QE* obteve melhores resultados na simplificação comparado aos outros trabalhos. O autor diz que seu trabalho também apresenta melhores resultados em comparação aos outros quando comparamos as pontuações dadas às sentenças de saída em relação ao que seres humanos julgam ser mais simples, isto é, se aproxima da avaliação feita por humanos.

3.6 Lexile

O Lexile framework[1] (LENNON; BURDICK, 2004) é um software pago utilizado em grande escala em livros e escolas, principalmente nos Estados Unidos. Baseado na proposta de (SMITH et al., 1989), o software consiste de dois componentes: a medida Lexile e a escala Lexile. O primeiro mede a habilidade do leitor ou a dificuldade de um texto. Já o segundo é uma escala de domínio da leitura variando de 200L (leitores iniciantes) até 1700L (leitores avançados). Os autores acreditam que, aplicando essas medidas, encontrariam livros ideais para cada tipo de leitor medido por sua escala, ou seja, o framework seria capaz de pontuar um leitor e, através dessa pontuação, escolher livros (também pontuados pelo Lexile) que seriam melhor compreendidos pelo leitor no que se diz respeito à complexidade do texto ali contido. Alguns artigos, como (KRASHEN, 2002), vão contra essa ideia pois ela poderia impedir que pessoas progredissem suas habilidades de leitura ao serem sempre direcionados a ler o que está dentro de sua atual capacidade.

O Lexile, um sistema que avalia a habilidade de leitura, oferece uma variedade de medidas, a maioria das quais baseia-se em testes padronizados de leitura aplicados em ambientes educacionais, onde os resultados são registrados e associados ao sistema Lexile. Um exemplo notável é a Scholar Reading Inventory, que envolve a prática de apresentar um texto com uma palavra omitida, desafiando o estudante a escolher a alternativa que melhor completa o trecho. A avaliação das escolhas de palavras é considerada cuidadosamente, levando em consideração diversos fatores, entre eles, a etapa educacional dos alunos submetidos ao teste, a porcentagem de estudantes que selecionaram a resposta correta, a porcentagem que optou por alternativas incorretas, o número total de alunos avaliados e a complexidade do texto original.

As medidas Lexile que refletem a complexidade textual se baseiam em dois principais fatores: a frequência das palavras e o tamanho das sentenças. Da mesma forma que nas métricas abordadas em nosso estudo, sentenças extensas com palavras menos frequentes tendem a receber uma pontuação mais elevada no sistema Lexile, refletindo uma maior complexidade textual. No entanto, vale ressaltar que, uma vez que o Lexile é um serviço pago, os detalhes específicos sobre a forma como esses fatores são atualmente incorporados aos cálculos não são divulgados. Além disso, a estrutura exata desse novo framework permanece incerta.

A única pista disponível sobre os fatores que podem ser levados em consideração no novo framework Lexile está em um estudo de 1989 conduzido por (SMITH et al., 1989), onde ele descreve o processo de desenvolvimento da métrica. O autor baseou-se em dados de um teste aplicado a três mil alunos americanos, que consistia em apresentar um texto com uma lacuna a ser preenchida com uma das opções disponíveis. Usando regressão linear, Smith desenvolveu as equações fundamentais que sustentam o sistema Lexile.

[1] https://lexile.com/

Essa abordagem inicial resultou na seguinte equação de escala Lexile:

$$(9.82247 * LMSL) - (2.14634 * MLWF) - constant = TheoreticalLogit \quad (3.20)$$

onde **LMSL** signfifica *Log of the Mean Sentence Lenght* (log da média do tamanho da sentença) e **MLWF** significa *Mean of the Log Word Frequency* (média do log da frequência da palavra).

Como dito anteriormente, este framework, de acordo com os autores, se assemelha à nossa abordagem. Porém o trabalho final deles, por ser um produto pago, não foi divulgado e muito menos publicado em alguma revista científica, para que a comunidade tivesse acessoa às fórmulas utilizadas. Além disso, ele foi construído com base na língua inglesa e com base nos estudantes americanos, o que dificulta a aplicação em outros países. Outro ponto que devemos lembrar é que a métrica proposta em nosso artigo não leva em consideração nenhum teste realizado com pessoas, ele é totalmente voltado ao *corpus* da linguagem, o que pode torná-lo ajustável para outras línguas no futuro. Levando tudo isso em conta, nossa métrica possui contribuição evidente e, além disso, pode ser utilizada por outros pesquisadores para futuros trabalhos e melhorias no cenário científico da NLP.

3.7 Coh-Metrix-Port

Coh-Metrix-Port (SCARTON; ALUÍSIO, 2010) é uma adaptação da ferramenta *Coh-Metrix* (CROSSLEY et al., 2007) que tem como finalidade extrair e analisar as características de um texto para avaliar a dificuldade ou facilidade de compreensão de sua leitura, usando vários níveis de análise linguística: léxico, sintático, discursivo e conceitual. Em seu desenvolvimento foram analisadas 40 métricas, como por exemplo número de palavras, sílabas por palavras, incidência de pronomes e verbos, números de adjetivos, entre outras. A autora utiliza pares de *corpus* complexos e simples para poder fazer a análise das sentenças. Por exemplo, ela utilizou o *corpus* "CH Ciência hoje" junto com o *corpus* "CHC Ciência hoje para crianças" Os dois corpora que contêm os mesmos artigos, porém o segundo é escrito de forma simplificada para que crianças também possam ler. Ao final da análise métricas utilizou-se um classificador binário "simples - complex" para classificar *corpora* de texto jornalístico e científico utilizados no trabalho.

A autora mostra neste trabalho que as métricas mais discriminativas foram as básicas, como contagem de palavras e sílabas, frequência das palavras e também incidência de substantivos e verbos. Isto evidencia que estas métricas básicas são o melhor sinal para construirmos uma boa métrica, tanto que em suas conclusões ela mostra que a métrica *Coh-Metrix-Port* tem melhores resultados com o reforço da métrica *Flesch* (FLESCH,

1948) que considera apenas as métricas básicas. Vale destacar que, dentre as características analisadas nesta métrica, todas as relacionadas com o uso de um parser como sentenças com vários níveis de subordinação, cláusulas embutidas, sentenças na voz passiva entre outros, não foram computadas e foram reservadas para trabalhos futuros.

A autora diz em suas conclusões que este seria um trabalho inicial e que realizaria melhorias a partir de sua publicação em 2010, mas não foram encontradas mais publicações sobre esta métrica após esta data, apenas o site[1] que fornece as propriedades da sentença que você queira analisar, dentre as várias possibilidades analisadas pela autora.

Em resumo, temos o seguinte:

- Métricas vindas dos campos de tradução e resumo:
  - Tradução: BLEU e SARI necessitam de frases de referência.
  - Resumo: Simple-QE baseado em BERT e necessita de anotações humanas.
- Análise de características das sentenças: Coh-Metrix-Port, que é dependente da sintaxe e semântica da língua portuguesa.
- Métrica baseadas em tamanho de sentença e frequência da palavra: Lexile, que é uma ferramenta paga e não publicada.

A Tabela 2 abaixo deixa mais claro a comparação entre as métricas citadas neste capítulo e algumas de suas características:

	ISiM	**BLEU**	**SARI**	**Simple-QE**	**Coh-Metrix-Port**	**Laxile**
Baseada em tradução de texto		x	x			
Baseada em resumo resumo				x		
Baseada em simplificação de texto	x				x	x
Necessida de frases de referência		x	x			
Necessida de frases anotadas				x		
Dependente de sintaxe				x	x	
Dependente de semântica				x	x	
Criada sob uma língua específica					x	x
Analisa frequência da palavra	x					x
Analisa tamanho da frase	x	x	x	x	x	x
Gratuito	x	x	x	x	x	

Tabela 2 – Comparativo entre as métricas e algumas de suas características

[1] http://143.107.183.175:22680/analyze

4 Métrica ISiM

Nesta seção, apresentamos o desenvolvimento de três versões da métrica proposta, explorando as variadas alternativas consideradas durante a investigação e elucidando as justificativas para sua proposta e eventual descarte, se pertinente. No caso, a métrica 1 foi a mais bem sucedida e a métrica utilizada para o resto das contribuições deste artigo. As métricas 2 e 3 foram utilizadas como teste e não superaram a performance da versão 1. Antes de seguir, uma consideração importante sobre nossa métrica é que ela pressupõe que as sentenças dadas são sempre bem construídas, ou seja, não é função da métrica analisar erros semânticos ou sintáticos.

Nossa métrica foi fundamentada em duas estruturas linguísticas: o comprimento da sentença e a frequência de uso das palavras em um *corpus*. Esta abordagem baseou-se em (STENNER, 1996), onde é afirmado que, no que tange ao componente semântico, a probabilidade de uma pessoa encontrar uma palavra em um contexto é o que mais influencia ao inferir seu significado (BORMUTH, 1966). Essa é a base da chamada"teoria da exposição", que explica como o vocabulário receptivo ou auditivo se desenvolve (MILLER; GILDEA, 1987; STENNER; III; BURDICK, 1983). Outros dois estudos significativos relacionados à familiaridade e raridade das palavras (KLARE et al., 1963; CARROLL; DAVIES, 1971) reforçam essa perspectiva.

Foram criadas sentenças de forma empírica, com o objetivo de abranger uma ampla variedade de cenários linguísticos, incluindo: a reconstrução completa da sentença utilizando palavras de menor frequência; a substituição pontual de um termo por outro de maior complexidade; a introdução de um termo intrincado em uma sentença simples, sem modificar significativamente sua semântica; e a condensação de sentenças para reduzir sua extensão lexical. Essas sentenças, consistentemente aplicadas em várias métricas, possibilitaram uma análise minuciosa de seus méritos e limitações individuais, além de fornecerem *insights* valiosos para potenciais aprimoramentos. A seguir, apresentamos exemplos das sentenças utilizadas nos experimentos:

- "Deglutir o batráqui", "Engolir o sapo"; "Eu estou muito triste", "Eu estou muito sorumbático"; "Eu não sei o que significa sorumbático"; "Estou indo ao mercado para comprar seis ovos", "Irei comprar ovos".

4.1 Métrica 1

A métrica proposta baseou-se apenas na frequência das palavras da sentença e no seu comprimento. A equação básica é a seguinte:

$$\frac{1}{n^{1.19}} \left(\sum_{i=1}^{n} \log_{10} \left(max \left(freq \left(w_i \right), 1 \right) \right) \right) \tag{4.1}$$

onde:

- $\boldsymbol{n}$ é o número de palavras contidas na sentença. É inversamente proporcional à soma do log e também elevado para 1,19 para penalizar sentenças mais longas. Este valor foi calculado de forma epírica e será detalhado mais à frente.
- w_i é a i-ésima palavra da sentença.
- $freq(w_i)$ é a frequência com que a i-ésima palavra aparece no *corpus wordfreq*.
- $\boldsymbol{max}$ é uma função que garante que palavras não encontradas no *corpus* não sejam contadas, garantindo sempre o valor 1.
- log_{10} para evitar que a função cresça muito para palavras com frequência muito alta no *corpus*.

Como a função *zipf frequência* já traz o valor em uma base logarítmica, a função pode ser reescrita da seguinte forma:

$$\frac{1}{n^{1.19}} \left(\sum_{i=1}^{n} freq \left(w_i \right) \right) \tag{4.2}$$

A determinação do peso relativo à potência da variável n foi realizada por meio de uma análise empírica, considerando um intervalo de valores entre 1 e 1.5, pois valores fora dessa faixa foram testados e não deram bons resultados, portanto nos concentramos nessa faixa para tentar um ajuste fino. Após extensas avaliações comparativas, o valor ideal foi inicialmente determinado como 1.4, mas, posteriormente com comparações em *corpus* com maior número de sentenças, que também serão demonstrados mais à frente, o valor final ficou em 1.19. Quaisquer valores significativamente maiores ou menores deste parâmetro resultaram em resultados atípicos e, portanto, foram prontamente descartados.

As Tabelas 3 e 4 exibem os resultados provenientes dessa métrica inicial, levando em consideração variações na penalização referente ao comprimento das sentenças. Notavelmente, a segunda tabela foi construída após a exclusão das *stop-words* contidas nas sentenças antes de realizar os cálculos. De acordo com essa métrica, pontuações mais baixas indicam uma maior complexidade na sentença. Adicionalmente, os valores entre

	1.19	**1.3**	**1.4**	**1.5**
Deglutir o batráquio	0,4551 (1)	0,4074 (1)	0,3597 (1)	0,3119 (1)
Engolir o sapo	0,5793 (3)	0,5315 (5)	0,4838 (6)	0,4361 (6)
Eu estou muito triste	0,6367 (7)	0,5765 (7)	0,5163 (7)	0,4561 (7)
Eu estou muito sorumbático	0,5844 (4)	0,5242 (4)	0,4640 (4)	0,4038 (4)
Eu não sei o que significa sorumbático	0,6166 (6)	0,5321 (6)	0,4476 (3)	0,3631 (3)
Estou indo ao mercado para comprar seis ovos	0,5855 (5)	0,4952 (2)	0,4049 (2)	0,3146 (2)
Irei comprar ovos	0,5690 (2)	0,5213 (3)	0,4736 (5)	0,4259 (5)

Tabela 3 – Resultados para a métrica 1 com *stop words*. Entre parênteses está a colocação da frase, da mais complexa para a mais simples.

parênteses representam a classificação relativa de cada sentença dentro da configuração específica.

Observa-se que, apesar dos resultados coerentes obtidos nas avaliações, houve situações em que sentenças mais extensas, mas ainda de complexidade moderada, foram excessivamente penalizadas. Como exemplo, temos a sentença “Estou indo ao mercado para comprar seis ovos”. Nesse caso, a métrica atribuiu uma penalização desproporcional na sua pontuação de simplicidade devido à extensão da sentença, mesmo que seu conteúdo seja relativamente simples.

No entanto, ao reduzir de maneira significativa para 1.1 o peso relativo atribuído ao comprimento da sentença nas análises, ocorreram situações em que sentenças mais curtas, como “Engolir o sapo”, foram avaliadas desfavoravelmente em comparação com “Eu estou muito sorumbático”. Isso se deve à predominância de palavras de alta frequência na última sentença, o que a métrica considerou como indicativo de menor complexidade, ignorando nuances semânticas e contextuais como a complexidade da palavra "sorumbático".

Essas observações ressaltam a necessidade de encontrar um equilíbrio na configuração da métrica, de forma a não prejudicar injustamente sentenças mais longas, mas ainda de complexidade moderada, enquanto avalia adequadamente a complexidade de sentenças curtas que podem conter termos menos frequentes e, portanto, mais complexos. É um desafio constante no desenvolvimento de métricas de avaliação de complexidade textual.

Neste contexto específico, a remoção das chamadas *stop words* gerou uma distribuição de resultados que não se revelou tão nítida quanto esperado, pois os resultados acabaram sendo piores. A curiosidade reside na variação das pontuações ao manipular o peso da variável *n*. Quando usamos um valor de 1.2, a sentença “Estou indo ao mercado para comprar seis ovos” foi avaliada como mais simples em comparação a “Irei comprar ovos”. Surpreendentemente, com pesos de 1.4 e 1.5, a sentença “Eu estou muito sorumbático” foi categorizada como uma das expressões mais simples. Essas discrepâncias demonstram a sensibilidade da métrica a ajustes na configuração, tornando essencial uma

	1.2	**1.3**	**1.4**	**1.5**
Deglutir o batráquio	0,1516 (1)	0,1215 (1)	0,0914 (1)	0,0613 (1)
Engolir o sapo	0,4421 (2)	0,4120 (2)	0,3819 (2)	0,3518 (3)
Eu estou muito triste	0,6567 (7)	0,6266 (7)	0,5965 (7)	0,5664 (7)
Eu estou muito sorumbático	0,5325 (3)	0,5024 (4)	0,4723 (5)	0,4422 (6)
Eu não sei o que significa sorumbático	0,5735 (5)	0,5133 (5)	0,4531 (4)	0,3929 (4)
Estou indo ao mercado para comprar seis ovos	0,5738 (6)	0,4960 (3)	0,4182 (3)	0,3404 (2)
Irei comprar ovos	0,5690 (4)	06266 (6)	0,4736 (6)	0,4259 (5)

Tabela 4 – Resultados para a métrica 1 sem *stop words*.

análise criteriosa.

Notamos que utilizar a métrica em comparações de sentenças de contextos diferentes não era o ideal. Nossa proposta seguiu para o caminho de comparar pares de sentenças de mesmo contexto, como por exemplo paráfrases. Os resultados obtidos a seguir já demonstram que este foi o caminho correto a seguir, pois a métrica passou a obter resultados muito mais coerentes.

Buscando ampliar nossa base de exemplos e testes, decidimos aplicar a métrica ISiM em dois corpora reconhecidos que contêm pares de sentenças complexas e simples. Esses corpora selecionados são o SimPA (SCARTON; PAETZOLD; SPECIA, 2018), que engloba 2200 pares de sentenças, e o *TurkCorpus* (XU et al., 2016), com um total de 352 pares. Vale ressaltar que o TurkCorpus oferece, adicionalmente, 8 sentenças de referência para cada caso avaliado, enriquecendo nossa análise comparativa.

No intuito de aprimorar a métrica proposta, seguimos um procedimento metódico e abrangente para avaliar sua eficácia. Inicialmente, calculamos as pontuações associadas a todas as sentenças contidas nos dois corpora previamente mencionados. Nossas previsões iniciais indicavam que a sentença original receberia uma pontuação quantitativamente menor, enquanto uma pontuação superior seria atribuída à sentença que, segundo o *corpus*, é classificada como mais simplificada. No entanto, nossos resultados iniciais apontaram para uma diversidade de cenários e desafios a serem superados na busca pela configuração ideal da métrica.

Posteriormente à implementação e execução do algoritmo, conduzimos uma análise quantitativa para determinar o número exato de sentenças que nossa métrica, de fato, classificou como mais simplificadas em relação às suas versões originais. Durante este processo avaliativo, exploramos uma gama diversificada de valores para a potência de n, que, como já abordado, denota o comprimento da sentença em análise. Adicionalmente, realizamos avaliações paralelas, nas quais incluímos e excluímos *stop words*, visando identificar possíveis impactos destas na eficácia de nossa métrica.

A tabela 5, que se encontra a seguir, apresenta, de forma estruturada, os resultados obtidos ao aplicar nossa métrica no *corpus* SimPA, proporcionando uma visão detalhada e clara do desempenho e precisão de nossa proposta.

potência de n	**1.4**	**1.3**	**1.25**	**1.19**	**1.1**
Com stop words	68.45%	70.09%	70.36%	71.55%	71.27%
Sem stop words	71.64%	74.36%	78.45%	**80.27%**	79.36%

Tabela 5 – Porcentagem de acerto da métrica para o *corpus* SimPA variando a potência de n (tamanho da sentença).

Na tabela 6 vemos os resultados da métrica para o TurkCorpus. Note que em ambos os *corpus* os melhores resultados apresentados foram para a mesma potência de n, que foi variado primeiramente no fator 0.10, e ao encontrar o melhor intervalo, passou a variar em 0.01.

	Com stop words	**Sem stop words**
1.4	93.31%	95,82%
1.3	94.15%	96.10%
1.25	95,54%	96.38%
1.19	96.10%	**96.94%**
1.1	96.38%	96.38%

Tabela 6 – Porcentagem de sentenças classificadas corretamente por nossa métrica para o *corpus* TurkCorpus variando a potência de n (comprimento da sentença)

Uma análise desses resultados revela que a métrica que desenvolvemos se distingue por sua simplicidade operacional e estrutural. Sua capacidade de operar de forma autônoma, sem necessidade de intervenção humana, combinada com sua independência linguística, a torna não apenas amplamente aplicável, mas também especialmente eficaz. A eficiência dessa métrica se manifesta através de um tempo de processamento rápido, com todos os cálculos sendo concluídos em menos de um segundo, independentemente do *corpus* utilizado.

Focando especificamente nos dados, vemos que, ao avaliarmos o *corpus* SimPA, alcançamos uma taxa de respostas concordância com o *corpus* de 80,27%, onde por concordância entende-se que a métrica avaliou a primeira sentença como compelxa e a segunda como simples, assim como o *corpus* sugeriu. No entanto, o desempenho no TurkCorpus é ainda mais expressivo, com uma taxa de acerto de 96,94%. Esses números, por si só, atestam a robustez e precisão da métrica na avaliação e categorização da complexidade dos textos, o que será reforçado na seção 5 onde o refinamento de corpus mostra que a ISiM melhora a pontuação dos *corpora* testados em relação a métrica SARI.

Além de sua precisão, a métrica oferece uma vantagem adicional ao funcionar de maneira independente de referências ou anotações humanas anteriores. Isso elimina

potenciais viéses e confere à métrica um caráter verdadeiramente objetivo. Para assegurar sua adaptabilidade a contextos linguísticos diversos, é possível ajustar facilmente o *corpus* de frequência de palavras de acordo com o idioma desejado, reforçando ainda mais sua versatilidade e aplicabilidade.

4.2 Métrica 2

Nesta seção, iremos aprofundar em uma métrica que se fundamenta primariamente em uma abordagem exponencial, baseando-se no produtório normalizado das frequências das palavras. Uma característica dessa abordagem reside na forma como ela administra penalidades a sentenças que incorporam palavras de frequência extremamente baixa. Ao adotar esse mecanismo, sentenças que contenham termos raramente utilizados são submetidas a penalizações consideravelmente acentuadas.

É importante destacar que essa função difere da métrica anterior em um aspecto crucial: enquanto a primeira métrica abordava diretamente o comprimento da sentença, esta não o faz. Contudo, não se deve inferir que o tamanho da sentença seja uma variável irrelevante. Pelo contrário, devido ao mecanismo multiplicativo adotado, o número de termos em uma sentença influencia intrinsecamente o resultado final. Em outras palavras, à medida que multiplicamos mais termos, o produto resultante é afetado de forma proporcional, tornando o comprimento da sentença uma variável implicitamente considerada no cálculo.

Delimitando a configuração desta segunda métrica, sua função é vista da seguinte maneira:

$$e^{(\prod_{i=1}^{n} \frac{max\left(freq\left(w_i\right),1\right)}{m})} \tag{4.3}$$

onde:

- $\boldsymbol{e}$ é a constante de Euler para manter a função exponencial
- $\boldsymbol{m}$ é a maior frequência encontrada dentre todas as palavras da sentença
- $\boldsymbol{n}$ é o número de palavras na sentença
- $\boldsymbol{i}$ é a i-ésima palavra da sentença

Os resultados listados abaixo mostram, entre parênteses a classificação da sentença mais complexa (1) para a mais simples (7):

É interessante observar, ao considerar a presença de stop-words, como a sentença "Deglutir o batráquio" foi classificada como a mais simplificada, surpreendentemente até

	Com stop words
Deglutir o batráquio	1.240295 (7)
Engolir o sapo	1,086586 (6)
Eu estou muito triste	1,036703 (5)
Eu estou muito sorumbático	1,000092 (2)
Eu não sei o que significa sorumbático	1,000048 (1)
Estou indo ao mercado para comprar seis ovos	1,000519 (3)
Irei comprar ovos	1,004910 (4)

Tabela 7 – Resultados para a métrica 2 com *stop words*.

	Sem stop words
Deglutir o batráquio	1.0000000000066 (3)
Engolir o sapo	1,0000000011745 (4)
Eu estou muito triste	1,0003060641245 (6)
Eu estou muito sorumbático	1,0000007826532 (5)
Eu não sei o que significa sorumbático	1,0000000000019 (2)
Estou indo ao mercado para comprar seis ovos	1,0000000000001 (1)
Irei comprar ovos	1,0049101072120 (7)

Tabela 8 – Resultados para a métrica 2 sem *stop words*.

mais do que sua contraparte, “Engolir o sapo”. Este fenômeno específico pode ser atribuído à estrutura de frequência das palavras em questão. Para ilustrar, ao analisar as frequências, observamos que “deglutir” e “batráquio” apresentam frequências de 13 e 139, respectivamente. Em contrapartida, “engolir” e “sapo” possuem frequências mais elevadas, marcando 287 e 7958, respectivamente. Estas disparidades nas frequências fornecem um vislumbre sobre como a função está realmente ponderando o contexto no qual as palavras estão inseridas. Na construção “Deglutir o batráquio”, percebemos que ambas as palavras compartilham um grau de raridade similar, o que as alinha a um contexto onde o vocabulário é rebuscado. Por outro lado, a expressão “Engolir o sapo” exibe uma diferença mais pronunciada entre as frequências dos termos, sugerindo à função que uma palavra de caráter mais complexo foi inserida em um contexto onde palavras mais simples predominam. Com isso, somos levados a inferir uma característica importante desta métrica: sua sensibilidade ao contexto. Quando confrontada com uma sentença composta exclusivamente por palavras de similar complexidade, seja ela elevada ou reduzida, essa métrica tende a atribuir uma pontuação mais elevada (mais simples). Entretanto, sentenças que reúnem termos de distintas frequências são vistas de forma diferente, recebendo, muitas vezes, uma pontuação reduzida devido à mistura heterogênea de frequências (mais complexas).

Fizemos ainda um teste comparativo entre a métrica 1 e a métrica 2 utilizando o *corpus* TurkCorpus com 359 e 2000 pares de sentenças. Os resultados estão nas Tabela 9 e 10 abaixo:

	# sentenças	# remoções	Pontuação
Métrica 1	359	13	**96.37%**
Métrica 2	359	20	94.42%

Tabela 9 – Comparação das métrica 1 e 2 ao analisar o TurkCorpus de 359 pares de frases originais, mostrando o número de pares de sentenças removidos do *corpus* e a porcentagem de concordância com o mesmo.

	# sentenças	# remoções	Pontuação
Métrica 1	2000	112	**94.40%**
Métrica 2	2000	132	93.40%

Tabela 10 – Comparação das métrica 1 e 2 ao analisar o TurkCorpus de 2000 pares de frases originais, mostrando o número de pares de sentenças removidos do *corpus* e a porcentagem de concordância com o mesmo.

Note que, em ambos os casos, a métrica 1 teve um melhor desempenho quando analisamos a quantidade de frases complexo/simples em que ela se alinha com o TurkCorpus. Além disso, devido a peculiaridade da métrica 2 pontuar de forma alta sentenças com várias palavras de baixa frequência, a métrica 1 se apresentou mais relevante ao que pretendíamos construir para a simplificação de texto. Portanto, esta métrica 2 não foi utilizada no restante do trabalho.

4.3 Métrica 3

A métrica que apresentamos agora para análise retoma e expande a função de somatório previamente observada na Equação 4.1 da Métrica 1. No início, nossa principal motivação era unir o conceito de contexto, que se destacou como um elemento essencial na Métrica 2, com o objetivo de aprimorar a eficácia da primeira métrica que concebemos.

Uma exemplificação clara dessa abordagem pode ser encontrada na análise da sentença “Eu estou muito sorumbático."Nesse contexto, a palavra “sorumbático” com uma frequência notavelmente baixa, assume o papel de predicativo do sujeito. Essa particularidade sintática confere um grau de importância significativo à palavra, já que, embora seja uma expressão menos comum, torna-se fundamental para a compreensão da mensagem global da sentença. Como resultado, muitos leitores que não estejam familiarizados com o significado exato de “sorumbático” podem enfrentar dificuldades ao tentar decifrar a ideia central transmitida pela sentença.

Por outro lado, ao considerarmos a construção “Eu não sei o que significa sorumbático” percebemos que “sorumbático” desempenha o papel de um adjunto adnominal, e sua presença não é tão essencial para a compreensão do núcleo da mensagem. Essa dife-

renciação entre os papéis sintáticos das palavras destaca a importância de nosso trabalho na avaliação da complexidade textual.

Sem dúvida, após essas reflexões e análises iniciais, nosso objetivo foi delineado com maior precisão: pretendíamos atribuir pesos específicos aos papéis sintáticos desempenhados por palavras nas sentenças. Acreditávamos que, ao incorporar esses pesos às frequências das palavras, seríamos capazes de desenvolver uma métrica mais sofisticada e precisa, permitindo o ajuste da relevância de determinados termos com base no contexto em que estivessem inseridos. Isso resultaria em uma avaliação mais precisa do grau de simplicidade ou complexidade de uma sentença específica.

Para avaliar a eficácia dessa abordagem, utilizamos um *corpus* de tags que abrange cerca de 1,1 milhão de palavras em português brasileiro, provenientes de artigos de jornal.

Um *corpus* de tags é caracterizado por ser um vasto repositório de recursos textuais, nos quais palavras e sentenças são minuciosamente anotadas com base em suas respectivas categorizações, conforme definido em uma ou várias ontologias predefinidas.

Contudo, observamos que a análise sintática fornecida pelo MAC-MORPHO frequentemente não correspondia à análise sintática correta. Um exemplo emblemático foi o caso da palavra "sorumbático" que, independentemente do contexto em que estava inserida, era consistentemente categorizada apenas como um adjetivo. Não somente esse mais alguns outros exemplos colhidos empiricamente foram analisados por dois profissionais da área de Letras, que reafirmaram o erro das análises por parte da biblioteca. Essa discrepância levantou desafios iniciais que precisavam ser superados em nosso processo de desenvolvimento e aprimoramento da métrica.

Diante dessa persistente inconsistência, tomamos a decisão de redirecionar nossos esforços para a biblioteca FLORESTA (AFONSO, 2004), com a esperança de obter uma análise sintática mais precisa das sentenças. No entanto, mais uma vez em várias instâncias, a biblioteca interpretou incorretamente a sintaxe das palavras, e essa imprecisão também se refletiu na análise da palavra "sorumbático" em diversos contextos. Importante ressaltar que aqui também as análises passaram por uma revisão rigorosa realizada pelos especialistas em linguística, que corroboraram as deficiências identificadas nas saídas geradas pelas bibliotecas.

Continuando o desenvolvimento desta métrica, voltamos nossa atenção para a criação de árvores de dependência sintática, visando auxiliar na avaliação do grau de simplicidade de uma sentença específica. Nossa premissa era que uma árvore de dependência sintática mais rasa poderia indicar uma menor complexidade na sentença analisada, tornando-a mais acessível. Essa abordagem representava uma tentativa de superar os desafios anteriormente enfrentados na busca por uma análise sintática confiável.

As bibliotecas previamente mencionadas apresentaram uma desvantagem notável

que se revelou incompatível com a nossa abordagem no âmbito deste projeto. Essa desvantagem consistia na necessidade de um processo manual e laborioso de construção da árvore sintática, uma demanda que não coadunava com a direção que almejávamos seguir em nossa pesquisa. A diferença crucial que tornou o SpaCy mais atrativo foi a sua capacidade de gerar árvores sintáticas de forma automatizada.

Entretanto, mesmo com essa vantagem inerente, o SpaCy também se demonstrou propenso a imprecisões em suas análises. Um exemplo notável envolveu a classificação da palavra "triste" como um pronome, quando, de fato, em contextos como "eu estou muito triste", sua natureza gramatical deveria ser discernida claramente como um adjetivo.

Além do analisado acima, utilizar características sintáticas da língua poderiam tirar o aspecto de independência da métrica, o que foi mais um motivo para abandonarmos a métrica 3 e prosseguimos com a métrica 1.

5 Utilizações da Métrica ISiM

Neste capítulo apresentaremos todas as formas de utilização da métrica ISiM, tanto para análise de complexidade de pares de frase, quanto para criação e refinamento de corpora.

5.1 Alteração do Modelo GPT2

Nesta seção em particular, tentamos criar um modelo gerador de sentenças complex/simples para posteriormente utilizarmos a métrica ISiM para validar e melhorar o modelo.

O modelo GPT-2 (RADFORD et al., 2019) é um modelo pré-treinado que usa a arquitetura *Transformer*. O modelo foi treinado em um *corpus* muito grande de aproximadamente 40 GB de dados de texto vindos de 8 milhões de páginas na web. Este modelo possui 1.5 bilhões de parâmetros. Seu objetivo é simples, prever a próxima palavra de uma sentença. Já que o GPT-2 é muito eficiente na sua tarefa de prever a melhor palavra que segue a sentença atual, veio a seguinte proposição: será que se pedirmos o GPT-2 que considere a frequência da palavra como um de seus parâmetros de escolha, ele irá começar a sugerir palavras mais simples, assim gerando sentenças mais simples?

Para verificar essa hipótese, realizamos uma pequena modificação no GPT-2, extraindo o *array* de prováveis *tokens* para completar a sentença, juntamente com suas probabilidades.

Em seguida, o código recursivo da biblioteca textitGPT2LMHeadModel retorna quando a probabilidade da próxima palavra é menor do que um dado limite, ou para quando encontrarmos os *tokens* que representam um fim de sentença. Esta é a parte convencional do código, onde dado um início de sentença, ele gera uma sentença que condiz com aquele contexto.

A modificação do código começa com a inserção da biblioteca *zipf frequency*[1] , que nos dá a frequência de uso de uma palavra em uma determinada linguagem. Também inserimos como uma condição de retorno da recursividade um limite (ajustável) de palavras que a sentença poderia ter no total, caso a sentença atingisse esse limite e não tivesse chegado à um *token* de fim de sentença, a busca recomeçava com uma outra palavra.

O primeiro problema encontrado nessa abordagem era o uso de nomes próprios que possuem uma frequência extremamente baixa. Para contornar isso, foi utilizada a

[1] zipf frequency <https://github.com/LuminosoInsight/wordfreq>

biblioteca *Spacy*[2]que contém uma lista de nomes próprios em uma determinada linguagem. Caso a palavra atual se encontre nessa lista, sua frequência não era verificada.

Até este momento as sentenças estavam sendo geradas normalmente. Utilizamos também o código sem as alterações feitas acima para poder compararmos a simplicidade entre as sentenças geradas nas duas formas. Porém, esbarramos em um problema onde modelos acabavam gerando sentenças totalmente diferentes, mesmo o início delas sendo exatamente igual, tornando assim impossível a comparação da simplicidade entre elas. Por exemplo o trecho inicial sendo “A louça da casa...” poderia ser completada pelo modelo padrão como “precisa ser lavada urgentemente.” e pelo modelo modificado como “é feita de porcelana”, são sentenças distintas. Além disso, o início da sentença é fixo, logo se ele for composto de palavras considerada complexas, isso não seria resolvido pelo modelo, pois ele apenas continuaria a sentença a partir do trecho de sentença inicial.

A abordagem seguinte foi tentar fazer o GPT-2 aprender um prefixo e seguir esse modelo para gerar a sentença, como nos exemplos abaixo, retirados aleatoriamente do *TurkCorpus*, onde ele deve refazer a sentença após ver a tag “Simple":

- "**Complex**: All shareholders are accountable to vote for it. **Simple**:"
- "**Complex**: However the first 15 minutes of each meeting are available for members of the public to voice opinions or raise questions. **Simple**:"

dessa forma, além de fazer o GPT-2 a não continuar uma sentença e sim criar uma a partir do exemplo, seria uma tentativa de faze-lo não fugir do contexto da sentença anterior.

Como se pode ver, ainda são geradas sentenças muito diferente uma da outra em alguns casos, apesar de que agora elas já são baseadas na sentença anterior. Apesar dessa melhoria, não temos como ainda comparar as duas sentenças e dizer qual é a mais simples, já que ainda são visivelmente diferentes.

A próxima tentativa foi fazer um *fine-tuning* no modelo, utilizando o *corpus* Turkcorpus (XU; CALLISON-BURCH; NAPOLES, 2015; XU et al., 2016) de pares de sentença complexo;simples, inserindo no início de cada sentença complexa a tag "complex:"e nas simples a tag "simple:". Apesar dessa alteração, pode-se ver na tabela que ainda assim algumas sentenças continuaram se comportando como no teste anterior, ainda podendo gerar palavras que não se encaixam na sentença original e as vezes bem diferente da sentença gerada pelo modelo modificado, o que impossibilita a comparação.

Após várias tentativas de diversas formas diferente de *fine tuning*, utilização de outros pares de sentenças, vários parâmetros diferentes, não obtivemos os resultados esperados, e pausamos essa fase.

[2] spacy <https://spacy.io/>

	Texto
Complexo	All shareholders are accountable to vote for it.
Simples	No shareholders.
Simples 2	If you vote for it, it will be voted on.
Complexo	However the first 15 minutes of each meeting are available for members of the public to voice opinions or raise questions.
Simples	Members of the public can vote on the agenda of the meeting.
Simples 2	Members of the public can vote on their views on the meeting.
Complexo	Samuel Holberry was sentenced to 4 years imprisonment with hard labour and died in prison at York Castle on June twenty one, aged twenty seven.
Simples	Samuel Holberry was sentenced to 4 years imprisonment with hard labour and died in prison at York Castle on 21 June 1842 aged 27.
Simples 2	The trial of Samuel Holmes was held in the High Court in London on June 20, 1893.

Tabela 11 – Tabela de sentenças geradas pelo GPT-2 modificado, onde "simples 2"é a sentença utilizando a frequência das palavras como parâmetro

	Texto
Complexo	All shareholders are accountable to vote for it.
Simples	The shareholders are accountable to the shareholders.
Simples 2	No shareholders are accountable to vote for it.
Complexo	However the first 15 minutes of each meeting are available for members of the public to voice opinions or raise questions.
Simples	The first 15 minutes of each meeting are available for members of the public to voice or raise questions.
Simples 2	The first 15 minutes of each meeting are available for members of the public to voice opinions or raise questions.
Complexo	Samuel Holberry was sentenced to 4 years imprisonment with hard labour and died in prison at York Castle on June twenty one, aged twenty seven.
Simples	Samuel Holberryberryberry was sentenced to 4 years in prison with no parole.
Simples 2	The first time he was sentenced to 4 years imprisonment with hard labour and died in prison at York Castle on June twenty one.

Tabela 12 – Tabela de sentenças geradas pelo GPT-2 com fine tuning

Um dos maiores benefícios da métrica ISiM é a possibilidade de utilizá-la como um método de criação e refinamento de corpora. Nas próximas seções, demonstraremos como a métrica pode transformar um *corpus* de paráfrases em um *corpus* de pares de sentença complexo/simples, e também como ela é capaz de refinar um *corpus* já existente de pares de sentença complexo/simples, deixando-o ainda mais adequados para treinamento de modelos de simplificação de textos.

5.2 Criação de Corpora

Começaremos explicando como funciona o processo de modificação de corpora ilustrando através de um fluxograma de ações do algoritmo.

A Figura 8 mostra as etapas para que, a partir de um *corpus* contendo paráfrases, obter um *corpus* de pares de sentenças complexo/simples. Existe uma variação onde no passo 3 pode-se entrar com apenas 2 pares de paráfrase e eliminar as combinações, dependendo da qualidade das variações encontradas no *corpus*.

Já a Figura 9 mostra o processo de refinamento de *corpus* que já é composto de pares de sentenças complexo/simples. O processo tem uma etapa a menos pois vários destes corpora possuem também as sentenças de referência (criadas pro seres humanos) em sua linha, então eliminar essas linhas é mais interessante do que reorganizar as colunas de sentença complexa e de sentença simples, pois as referências ficariam com uma relação incorreta.

Note que o processo em ambos os casos é bem simples, e também se mostrou muito rápido. Isso tudo graças ao fato da métrica ISiM ser também simples e rápida. Fazer os cálculos para milhares de pares de sentenças não leva mais de 10 segundos e, de posse da pontuação de cada sentença, a reorganização automática dos pares nos fornece um *corpus* complexo/simples pronto para ser utilizado.

Veja abaixo este método de criação de corpora de forma mais detalhada, analisando também os resultados obtidos em cada uma das abordagens.

5.2.1 Transformando *corpus* de paráfrase em *corpus* de simplificação

A pressuposição de que a métrica ISiM pode ser utilizada para simplificação de textos parte possibilidade aplicá-la a um *corpus* paralelo de paráfrases para transformá-lo em um *corpus* paralelo de sentenças complexo/simples, que pode, posteriormente, ser utilizado para treinar modelos de simplificação.

Para testar esta hipótese, foi utilizado o *corpus* TaPaCo (SCHERRER, 2020) que é composto por paráfrases para 73 línguas diferentes, incluindo Português. Juntamente com ele, foi criado um formulário anônimo online para que pessoas pudessem votar em, dado um par de sentenças, qual delas seria a mais simples. Isso foi feito para verificar se o valor da métrica coincidia com avaliações feita por humanos. Participaram deste experimento 120 pessoas não identificadas. Para a preparação deste formulário, primeiro foi realizada um ajuste no formato do TaPaCo. Ele possui uma linha com várias paráfrases para a mesma sentença, que foram transformados em vários pares de sentenças para cada paráfrase. Foi também realizada uma limpeza no *corpus*, removendo as sentenças contendo menos de 2 palavras para que não comparássemos apenas a simplicidade entre duas palavras.

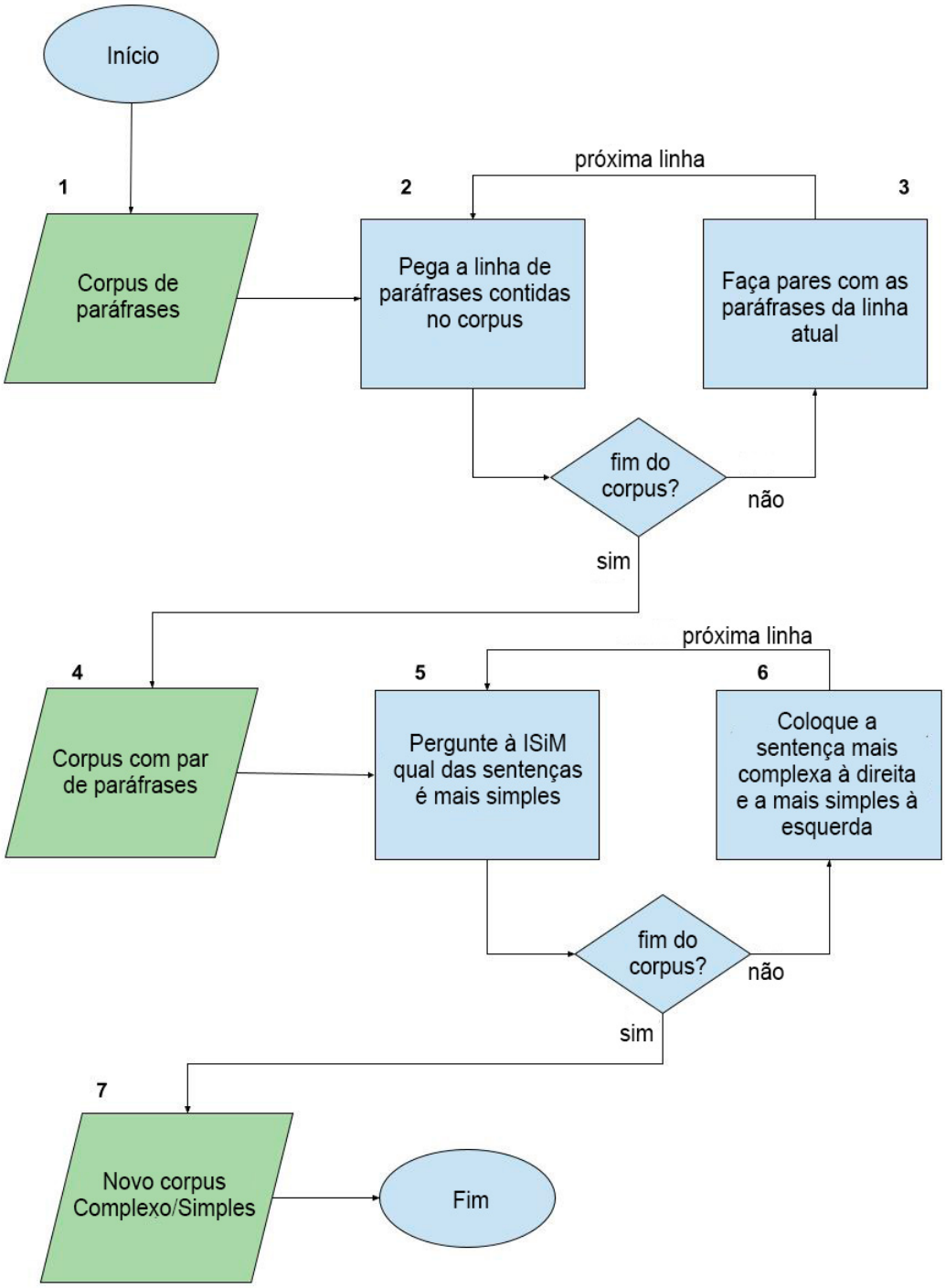

Figura 8 – Processo de criação de *corpus* de sentenças complexo/simples a partir de um *corpus* de paráfrases

Após essa etapa, aplicamos a métrica ISiM em cada um destes pares, obtendo a pontuação de cada sentença, e, se fosse o caso, trocando a primeira sentença de lugar com a segunda para que o par sempre ficasse na ordem complexo/simples.

Após esse rearranjo do *corpus*, foram embaralhadas aleatoriamente os pares de paráfrase resultante, onde as 40 primeiras foram escolhidas para entrar no formulário. No

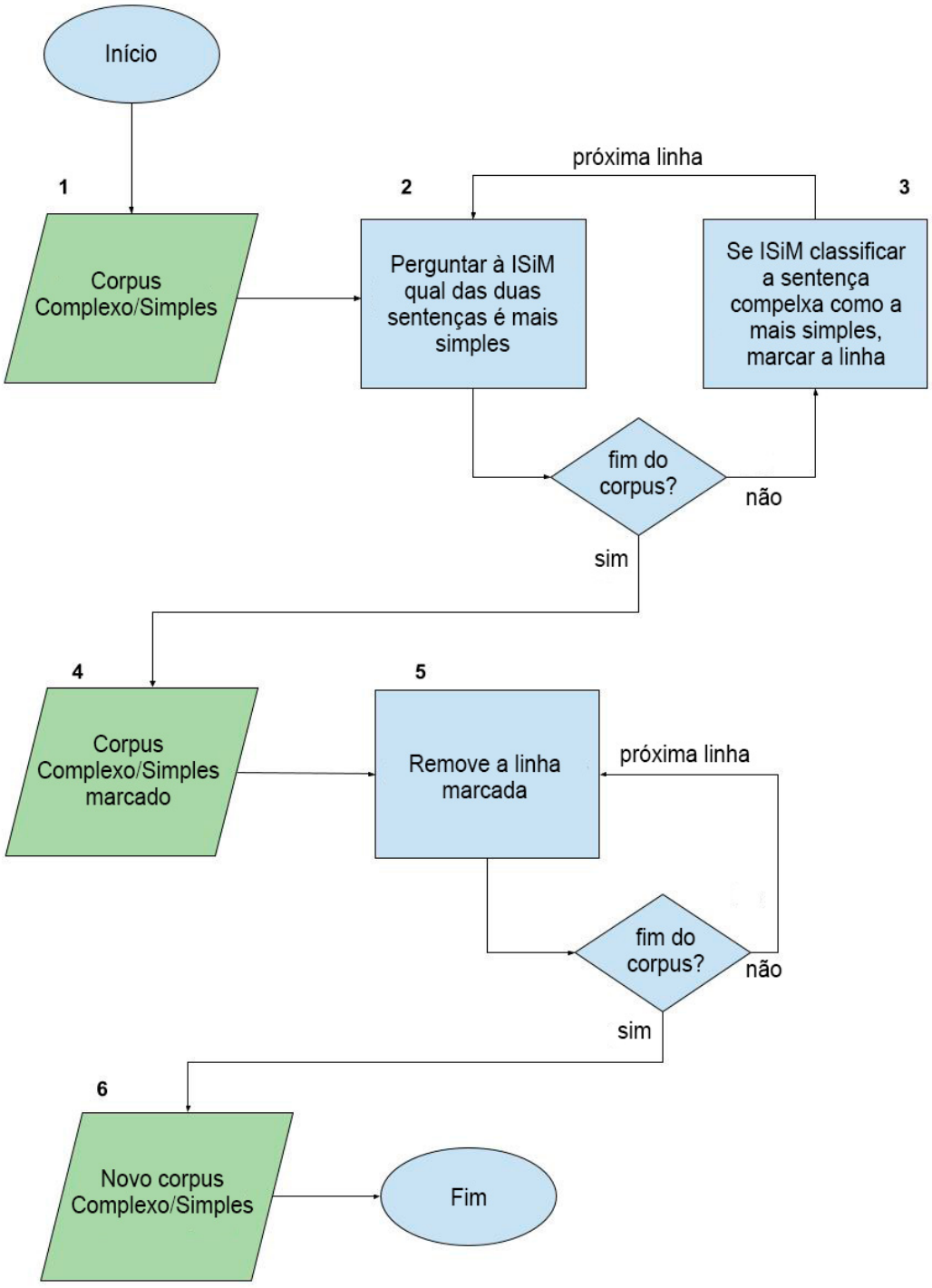

Figura 9 – Processo de refinamento de *corpus* de sentenças complexo/simples

capítulo 7 de material suplementar, na Tabela 26, estão listadas todas os 40 pares de frases utilizados. Este número foi escolhido para que o formulário não ficasse longo demais e as pessoas, devido ao cansaço, começassem a escolher aleatoriamente as respostas, podendo interferir diretamente no resultado final.

A figura 10 mostra um exemplo de como fica a pergunta e as opções de resposta. Foi removida a opção "são equivalentes" para tornar as sentenças dependentes uma da outra

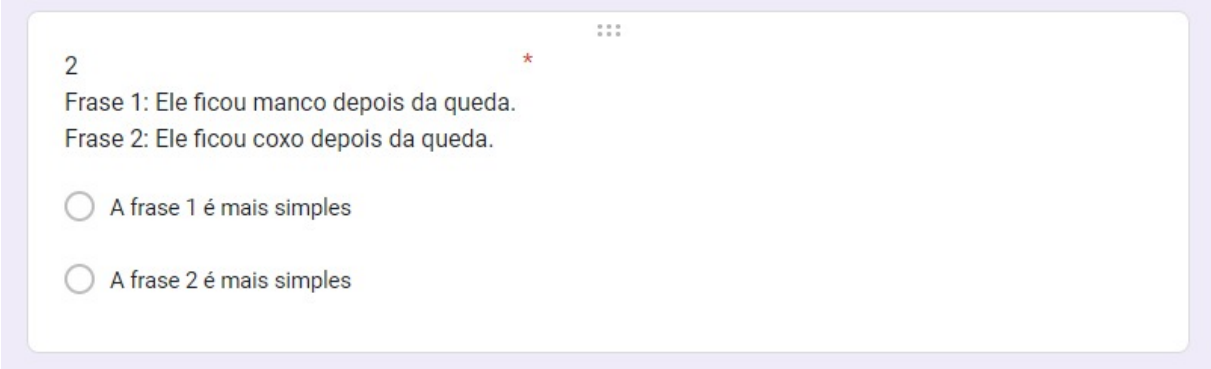

Figura 10 – Exemplo de par de sentença no formulário

na opção de escolha. Ao final do formulário também foram feitas algumas perguntas como grau de escolaridade, a língua nativa, e se o participante era formado em alguma área de especialidade na linguagem tratada, tal como curso de letras. Apesar destas perguntas, o formulário se manteve anônimo, uma vez que não era possível identificar os participantes por meio dessas informações.

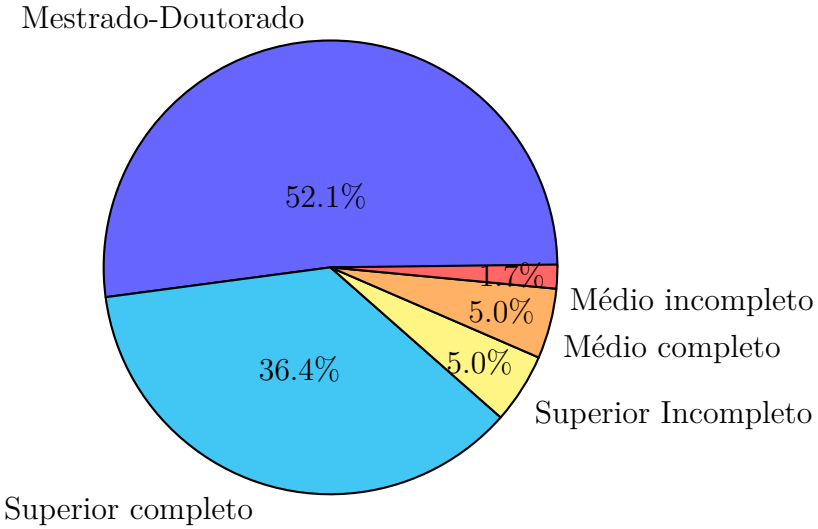

Figura 11 – Distribuição de grau de escolaridade pelos 120 participantes do formulário anônimo.

O gráfico de pizza da figura 11 mostra que mais de 50% das pessoas que participaram tinham mestrado, doutorado ou alguma pós graduação, porém podemos ver que tivemos pessoas de até ensino médio incompleto participando.

O gráfico da Figura 12 abaixo mostra que 10.7% de pessoas que participaram possuíam algum curso que tinha como objeto de estudo linguagem tratada, como Letras , Direito e Jornalismo. Esta pergunta foi feita para analisar se um maior conhecimento específico de linguagem traria alguma vantagem para os participantes, ou se a métrica continuaria julgando apenas por frequência de uso da palavra e tamanho da sentença. Se pessoas com maior grau de conhecimento da estrutura da linguagem tiverem maiores

acertos, seria uma evidência de que características específicas da linguagem, como por exemplo a sintaxe, deveriam ser consideradas na métrica.

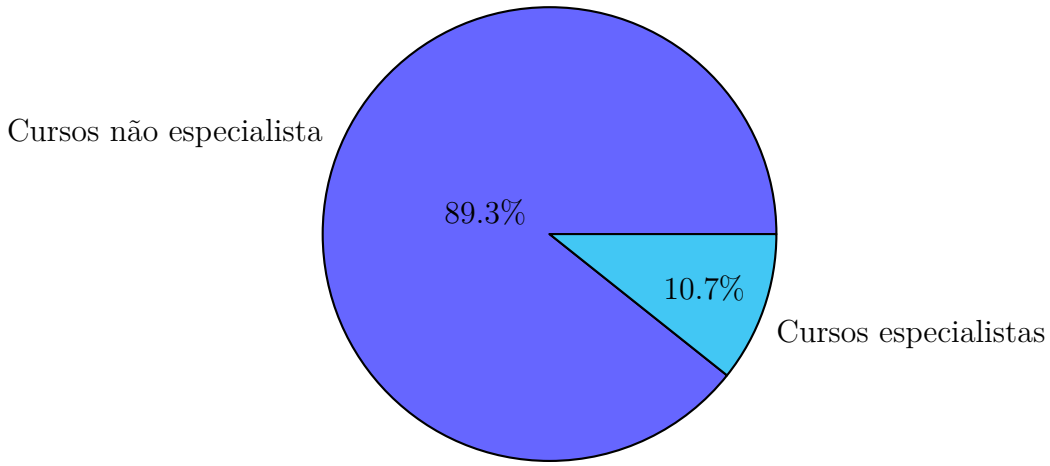

Figura 12 – Quantidade de pessoas com alto grau de especialidade na língua portuguesa.

	Todos	**Apenas Pós Graduação**	**Curso Especializado**
min acertos	19	23	27
max acertos	34	34	31
mediana	30	31	30
%mediana	75.00%	77.50%	75.00%
média	29.33	30.06	29.33
% média	73.33%	75.15%	73.33%
media excluindo extr.	29.44	30.11	29.33
% média interna	73.60%	75.28%	73.33%

Tabela 13 – Dados extraídos do formulário, onde vemos o resultado geral e o resultado isolado para pessoas com pós graduação.

A tabela 13 mostra alguns dados extraídos do formulário. O dado mais importante é a mediana de acertos geral, que chega a 75% das 40 perguntas listadas, independente do grau de escolaridade. Como este teste também é dependente da qualidade do *corpus* de paráfrase utilizado, alguns resultados negativos podem ser frutos da má qualidade das paráfrases, escolhidas aleatoriamente no embaralhamento do *corpus*. Podemos notar também que, apesar das pessoas com pós graduação terem acertado em média 1 questão a mais do que o geral, as que possuem alto grau de conhecimento na língua portuguesa acabaram tendo o mesmo resultado geral, mostrando que aspectos técnicos da linguagem não foram relevantes para medir a simplicidade de cada sentença.

O gráfico da figura 13 mostra que 3 questões tiveram taxa de acerto de menos de 10%, enquanto 19 das 40 questões tiveram taxa de acerto acima de 90%. Esta taxa de apenas 3 questões sendo amplamente em desacordo entre os usuários e a métrica mostra o baixo grau de erro da ISiM.

Analisando esses resultados, podemos ver que valores emitidos pela métrica ISiM coincide as escolhas da maioria dos participantes do formulário. Apoiado por esses resul-

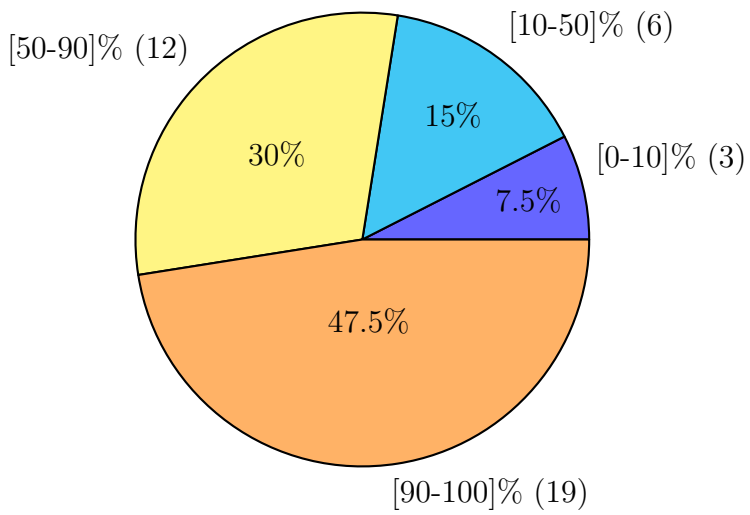

Figura 13 – Quantidade de questões em cada intervalo de % de acerto. Entre parênteses está o número de questões pra cada porcentagem

tados, foi criado um modelo onde dado um par de paráfrases, o modelo gera um arquivo csv contendo a ordem correta de cada sentença do par, mostrando qual é a mais complexa e qual é mais simples, obtendo de forma rápida a versão complexo/simples do *corpus* escolhido. Com o uso da métrica ISiM, a execução do modelo é rápida, o que facilita a tarefa de transformação. Vale a pena ressaltar que quanto melhor for o *corpus* de paráfrase utilizado, melhor será o *corpus* de par complexo/simples gerado pela métrica.

5.2.2 Refinando *corpus* Paralelo

Como visto anteriormente, a métrica ISiM pode ser usada para transformar um *corpus* de paráfrases em um *corpus* paralelo de sentenças complexo/simples para ajudar no treinamento de modelos voltados a tarefa de simplificação. A segunda proposta argumenta que a métrica também pode refinar *corpus* existentes de pares complexo/simples, para torná-los mais apropriados, descartando pares de sentenças que a métrica julgar "incorretos".

Em (ŠTAJNER; BÉCHARA; SAGGION, 2015) vemos, por exemplo, que no *corpus* paralelo de simplificação da Wikipedia há uma alta incidência de simplificações que não são significativamente mais simples ou que são imprecisas, não estando alinhadas ou estando apenas parcialmente alinhadas com a simplificação da sentença. Essa questão é uma das principais dificuldades enfrentadas pelos modelos de simplificação existentes, que muitas vezes não conseguem gerar sentenças simplificadas e acabam deixando as frases de entrada sem alterações. Portanto, nota-se a importância do refinamento de *corpora* existentes, o que pode contribuir para a melhoria do treinamento dos modelos.

Para esta análise foram utilizados dois conhecidos *corpus* paralelos, o Turkcorpus (XU; CALLISON-BURCH; NAPOLES, 2015; XU et al., 2016) e o ASSET (ALVA-

MANCHEGO et al., 2020). Ambos são amplamente utilizados por modelos e tem sua eficiência medida pelas métricas BLUE e SARI (comentadas no início do trabalho). Estes *corpus* foram utilizados em modelos recentes (2020-2022) que representam o estado da arte para as métricas SARI e BLEU, como o modelo MUSS (BART+ACCESS Supervised) (MARTIN et al., 2020), o Control Prefixes (CLIVE; CAO; REI, 2022), o TST (OMELIANCHUK; RAHEJA; SKURZHANSKYI, 2021) e o MUSS (BART+ACCESS Unsupervised) (MARTIN et al., 2020).

Os corpora utilizam a *English Wikipedia* para obter as sentenças originais (é possível notar que nos dois as bases de dados originais são do mesmo tamanho), apenas as sentenças de referências que são diferente entre os *corpus*. O objetivo aqui é analisar se a pontuação SARI destes *corpus* aumentaria se removêssemos pares de sentenças das quais a métrica julgasse estarem trocados, ou seja, a sentença complexa seria na verdade mais simples do que seu par. Os resultado dos dois *corpus* são analisados separadamente a seguir.

5.2.2.1 Turkcorpus

O Tukcorpus pode ser obtido online, sendo dividido em teste e *tuning*. Ambos são formados por 3 tipos de sentenças: a sentença normal, a sentença simplificada e 8 sentenças de referências para que o SARI possa calcular sua pontuação. O *corpus* de teste possui 359 linhas (10 sentenças cada) e o de *tuning* 2 mil linhas.

Para realizar a filtragem desse *corpus*, iremos utilizar a métrica ISiM comparando a sentença considerada "normal" (complexa) e a sentença considerada "simples" do *corpus*. Caso a métrica aponte que a sentença "normal" seja mais simples do que a sentença "simples", iremos remover a linha (com todas as sentenças de referência) onde esse par de sentença se encontra. Também consideramos duas possibilidades para aplicar a métrica: levando em conta as *stop words* e removendo as *stop words*.

Implementamos um código que lê o *corpus*, compara as sentenças "normal" e "simples", e as remove caso sejam consideradas trocadas. Isso é feito quase que instantaneamente pela métrica ISiM, pois são realizados apenas contas matemáticas. A seguir, temos a Tabela 14 com os resultados do *corpus* de teste.

	Teste Original	**Teste com stop words**	**Teste sem stop words**
Número de pares	359	304	282
Pares Removidos	0	55	77
Pontuação SARI	0.35732	0.35870	0.35750
Aumento Pontuação	0.0%	**0.3862%**	**0.0503%**

Tabela 14 – Pontuação SARI do *corpus* Turkcorpus de teste após utilização da métrica ISiM para refinamento.

Nota-se que, apesar de pequena no caso sem *stop words*, há uma melhoria na pon-

tuação em relação à pontuação do SARI em ambos os casos. Levando-se em conta as *stop words*, obtivemos uma melhoria de 0.3862% da pontuação, que em casos de treinamento de um modelo pode fazer diferença nos resultados. A seguir temos as pontuações para o *corpus* de *tuning* na tabela 15.

	Teste Original	**Teste com stop words**	**Teste sem stop words**
Número de pares	2000	1631	1563
Pares Removidos	0	368	436
Pontuação SARI	0.35477	0.35565	0.35494
Aumento Pontuação	0.0%	**0.2480%**	**0.0479%**

Tabela 15 – Pontuação SARI do *corpus* Turkcorpus de tuning após utilização da métrica ISiM para refinamento.

Mais uma vez podemos notar que em ambos os casos a métrica ISiM melhorou a pontuação do SARI para o *corpus* de *tuning*, mostrando que as remoções de sentenças consideradas incorretamente posicionadas foram corretas. Além disso, é importante mostrar o seguinte teste que prova que a melhoria na pontuação do SARI não foi ao acaso. Na tabela 16 temos os resultados de 5 remoções aleatórias de 368 pares de sentenças, a mesma quantidade que a métrica ISiM removeu ao analisar o coprus. Note que a média de melhoria da pontuação SARI foi de 0.016912%, sendo que em alguns casos a pontuação SARI piorou para o novo *corpus*. Note, também, que a pontuação de 0.2480% obtida pelo *corpus* após a métrica ISiM remover 368 sentenças é 14.66 vezes maior do que a média da pontuação SARI na remoção aleatória de sentenças.

Pontuação SARI padrão	**Pontuação pós remoção aleatória**	%
0.35477	0.35470	**-0.019731**
0.35477	0.35496	**0.053556**
0.35477	0.354758	**-0.003382**
0.35477	0.35475	**-0.005356**
0.35477	0.35498	**0.059475**

Tabela 16 – Resultado da pontuação SARI do *corpus* TURKCORPUS quando removemos aleatoriamente 368 pares de sentenças do *corpus*.

Realizando o teste de significância estatística, é possível afirmar que a o aumento percentual da pontuação SARI devido à intervenção da métrica ISiM foi positiva. Utilizamos o método t-student (STUDENT, 1908) para realizar o teste. A utilização da distribuição t de Student é baseada em uma série de considerações metodológicas. Primeiramente, a natureza dos dados analisados justifica essa escolha. Estamos lidando com uma amostra relativamente pequena de diferenças nas pontuações SARI. A distribuição t-Student é conhecida por ser adequada para amostras pequenas e situações onde a variância populacional não é previamente estabelecida.

Além disso, as suposições necessárias para a aplicação do teste t são atendidas em nossos dados. As análises preliminares indicaram que os dados das pontuações SARI são

aproximadamente normalmente distribuídos, uma condição essencial para a aplicação da distribuição t. Além disso, as observações das pontuações SARI são independentes entre si, cumprindo outra suposição fundamental do teste t.

O objetivo desta análise é comparar a média das diferenças nas pontuações SARI antes e depois da remoção de pares de sentenças. O teste t pareado é particularmente apropriado para esta situação, pois permite verificar se a média das diferenças é estatisticamente significativa. Esta abordagem é ideal para detectar se houve uma mudança significativa nas pontuações SARI após a intervenção aplicada.

O primeiro passo para iniciar o t-Student é calcular a média ($\bar{x}$) e o desvio padrão (s) das pontuações SARI após as remoções aleatórias.

Os valores aleatórios sâo: $[0.35470, 0.35496, 0.354758, 0.35475, 0.35498]$

Precisamos calcular as seguintes variáveis:

a média $\bar{x}$:

$$\bar{x} = \frac{\sum \text{Pontuações Aleatórias}}{n}$$
$$\bar{x} = 0,3486828$$

o desvio padrão (s):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$
$$s = 0.0000521$$

onde x_i é a i-ésima pontuação resultante da remoção aleatória e n o número total de pontuações dessas remoções.

Como pontuação SARI observada após a remoção dos pares identificados pela métrica é $x_0 = 0.35870$, o cálculo do valor T segue da seguinte maneira:

$$t = \frac{x_0 - \bar{x}}{s/\sqrt{n}}$$

$$t = -14.081842940839437$$

Para calcular o valor P associado ao valor T, usamos a distribuição t com n - 1 graus de liberdade (df = 5-1 = 4). O valor P para um teste bicaudal é:

$$p = 2xP(T > t)$$

Usando o código da biblioteca estatística scipy.stats do Python para calcular o valor P, temos:

$$p \approx 0.0001475$$

Como o valor p de 0.0001475 é menor que 0.05, rejeitamos a hipótese nula e concluímos que a diferença observada na pontuação SARI após a remoção dos pares identificados pela métrica ISiM é estatisticamente significativa.

5.2.2.2 ASSET

O *corpus* ASSET pode ser obtido online e é dividido em um subcorpus de teste e um de validação, onde o primeiro possui também 359 linhas de sentenças e o segundo 2000 linhas. No ASSET a divisão entre as sentença é diferente. Enquanto o TurkCorpus possui uma sentença original e uma sentença simplificada e mais 8 sentenças de referência, no ASSET temos apenas a sentença original seguida de 10 sentenças de referência. Para fazer o refinamento nesse caso, foi feita a comparação da sentença original com todas as sentenças de referência. Caso a maioria das sentenças de referência fossem consideradas mais complexas do que a original, a linha seria removida. Foi construído um modelo utilizando a métrica ISiM para gerar o *corpus* final, e na tabela 18 temos o resultado para o *corpus* ASSET de teste.

Mas antes de verificar os resultados da tabela, para o *corpus* ASSET também fizemos o teste randômico de eliminação de sentenças, para ver se o comportamento aqui também não melhoraria a pontuação SARI do *corpus*. Foram eliminadas 125 sentenças das 2000 presentes no *corpus*, assim como no teste real. A tabela 17 mostra que, em 5 tentativas aleatórias, a pontuação do SARI alterou em média -0.000348%, enquanto utilizando a métrica ISiM obtivemos melhoria de 0.1073%.

Pontuação SARI padrão	**Pontuação pós remoção aleatória**	%
0.34457	0.34449	**-0.023217**
0.34457	0.34460	**0.008707**
0.34457	0.344594	**0.006965**
0.34457	0.344592	**0.006385**
0.34457	0.344568	**-0.000580**

Tabela 17 – Resultado da pontuação SARI sobre o *corpus* ASSET após a remoção aleatoria de 125 pares de sentenças do *corpus*.

Note que para este *corpus*, as melhorias foram de menor valor se comparadas ao corpus TurkCorpus, mas novamente em ambos os casos houve um aumento da pontuação do SARI, indicando que as remoções foram bem pontuadas por esta métrica. Outro fator importante é o tempo de análise. Enquanto a métrica ISiM percorre todo o *corpus* em

	Teste Original	**Teste com stop words**	**Teste sem stop words**
Número de pares	359	282	339
Pares Removidos	0	19	20
Pontuação SARI	0.34523	0.34542	0.34537
Aumento Pontuação	0.0%	**0.0550%**	**0.0405%**

Tabela 18 – Resultado da pontuação SARI sobre o *corpus* ASSET de teste após utilização da métrica ISiM para refinamento

pouco mais de 1 segundo, a métrica SARI demorou 18 minutos e 20 segundos para analisar o mesmo *corpus* com 2000 linhas de sentenças.

A tabela 19 mostra o resultado para o *corpus* de validação.

	Teste Original	**Teste com stop words**	**Teste sem stop words**
Número de pares	2000	1875	1755
Pares Removidos	0	125	245
Pontuação SARI	0.34457	0.34494	0.34486
Aumento Pontuação	0.0%	**0.1073%**	**0.0841%**

Tabela 19 – Resultado da pontuação SARI sobre o *corpus* ASSET de validação após utilização da métrica ISiM para refinamento

O número de remoções de sentenças foi bem menor neste caso do que no Turkcorpus. Isto pode ser atribuído à metodologia utilizada para remoção da sentença onde 50% ou mais das sentenças de referências deveriam ser consideradas mais complexas do que a original para que a métrica ISiM as removesse do *corpus*. Ainda assim vimos uma melhoria na pontuação SARI. Independentemente do método usado para eliminar sentenças, ou da presença ou ausência de *stop words*, a métrica em nenhum teste piorou a pontuação do *corpus*.

Mais uma vez, calculamos o valor da relevância estatística desses resultados, e tivemos o seguinte:

Calculando a média ($\bar{x}$) e o desvio padrão (s) das pontuações SARI após as remoções aleatórias.

Os valores aleatórios sâo: $[0.34449, 0.34460, 0.344594, 0.344592, 0.344568]$

média $\bar{x}$:

$$\bar{x} = \frac{\sum \text{Pontuações Aleatórias}}{n}$$

$$\bar{x} = 0, 0.3445688$$

desvio padrão (s):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$s = 0.00001828$$

A pontuação SARI observada após a remoção dos pares identificados pela métrica é $x_0 = 0.34494$, o cálculo do valor T segue da seguinte maneira:

$$t = \frac{x_0 - \bar{x}}{s/\sqrt{n}}$$

$$t = -18.159461$$

Para calcular o valor P associado ao valor T, temos:

$$p = 2xP(T > t)$$

$$p \approx 0.0000540$$

Como o valor p de 0.0000540 é menor que 0.05, rejeitamos a hipótese nula e concluímos que a diferença observada na pontuação SARI após a remoção dos pares identificados pela métrica ISiM é estatisticamente significativa.

De acordo com os resultados exibidos nas tabelas acimas, vimos que nossa métrica foi capaz de criar e refinar *corpus* de maneira eficiente e rápida. Apesar das porcentagens de ganho serem na casa de 0.38% a 0.05%, se analizarmos o ranking de pontuação [3] SARI de alguns modelos, vemos que a variação de pontuação de uma posição para outra também está dentro dessa magnitude, o que mostra a significância dos resultados obtidos.

Visto que encontrar *corpus* de pares de frase complexo/simples é muito difícil devido a sua raridade, esta utilidade da métrica ISiM se mostrou de extrema importância para a área de simplificação de textos, podendo ampliar a quantidade de corpora disponível tanto quanto melhorar a qualidade dos existentes, provendo uma melhoria no treinamento de modelos para simplificação de textos.

5.3 Utilizando o ChatGPT

Nesta seção utilizaremos o chatGPT (versão 3) como modelo gerador de sentenças simples. O lançamento do ChatGPT (OPENAI, 2022) pela OpenAI em novembro de 2022 inaugurou uma nova era na inteligência artificial, permitindo que indivíduos fora do ambiente acadêmico tivessem contato direto com um poderoso modelo de linguagem fácil de utilizar. A popularização da IA ganhou impulso com essa ferramenta, que utiliza o transformer GPT 3.5 como modelo de geração de texto, pre-treinado em uma imensa

[3] https://paperswithcode.com/sota/text-simplification-on-turkcorpus

base de dados multilíngue.

Atualmente, há disponível o ChatGPT que utiliza o transformer do GPT 4, porém, para os testes realizados abaixo, não foi utilizado devido a ser uma ferramenta paga. Vale ressaltar que não estamos utilizando o GPT como ferramenta, mas sim como colaborador, a fim de testar a métrica ISiM em 200 sentenças originais do *corpus* TurkCorpus. Para simplificar as sentenças, utilizamos o seguinte comando do ChatGPT.

> "Simplify the following text. Do not summarize it. Try to replace more complex words with others more commonly used in the English language: "

Após as frases simplificadas terem sido geradas pelo ChatGPT, no primeiro teste a métrica ISiM apresentou os resultados da tabela 20, onde vemos uma percentual de concordância entre a métrica e o ChatGPT já de 88% das 200 sentenças simplificadas.

Stop Words	**Parâmetro**	**Corretas**	**Erradas**	**Empates**	**%**
Sim	1.19	175	25	0	87.5
Não	1.19	176	24	0	**88.0**

Tabela 20 – Resultados da métrica ISiM usando o parâmetro padrão de penalização do tamanho da sentença (1.19).

Ao analisar as sentenças que não estavam sendo classificadas como mais simples pela ISiM, notamos que algumas delas foram transformadas em 2 sentenças. Fizemos uma modificação na forma com que pedimos ao ChatGPT que simplificasse a sentença, pedindo que não as transformassem em duas ou mais sentenças, ficando da seguinte forma:

> "Simplify the following text. Try not to summarize it. Change as much as possible the uncommon words for more popular ones. Do not split the sentence in two or more. "

Os novos resultados podem ser vistos na tabela 21, onde conseguimos uma taxa de acerto de 93.5% das 200 sentenças simplificadas pela ferramenta do ChatGPT.

Stop Words	**Parâmetro**	**Corretas**	**Erradas**	**Empates**	**%**
Sim	1.19	186	14	0	93.0
Não	1.19	187	13	0	**93.5**

Tabela 21 – Resultados da métrica ISiM usando o parâmetro padrão de penalização do tamanho da sentença (1.19).

Foram feitos mais alguns testes, uma vez que algumas sentenças ainda continuaram sendo classificadas como mais complexas do que as originais. Uma das primeiras suspeitas

Original	**Simplificada**
We have already spent millions on fire safety in our council homes and improvement works have been carried out over the last 5 years.	We've spent a lot of money to make our council homes safer from fires and we've been improving them for the last 5 years.
Sheffield Wednesday Football Club, also known as' The Owls' (so-called because the team played at Owlerton from 1899) formed in 1867 in Sheffield.	Sheffield Wednesday FC, nicknamed 'The Owls' due to their former stadium at Owlerton, was established in Sheffield in 1867.

Tabela 22 – Sentenças geradas pelo ChatGPT que foram classificadas como mais complexas do que a original pela métrica ISiM.

foi a abreviação de palavras e a contração de palavras na língua inglesa (Veja a Tabela 22).

Note que trechos como "We have" foram contraídos para "We've", que possui uma pontuação mais baixa na frequência de utilização. Também vemos a abreviação de "Football Club" para "FC", que também possui uma pontuação menor, logo influenciou diretamente numa pontuação ruim para a sentença gerada. Outra situação encontrada foi a escolha do ChatGPT por palavras que não eram realmente mais frequentes do que as da sentença original, como os exemplos da tabela 23.

Original	**Simplificada**
The Framework enables us to assess all family members so that we can identify the right support for families and work together with other professionals to support them.	The Framework allows us to evaluate all family members so that we can determine the appropriate support for families and collaborate with other professionals to assist them.
By controlling how cities are constructed and making sure that Part IIA of the Environmental Protection Act 1990 is followed.	Through the town planning development control process and the enforcement of Part IIA of the Environmental Protection Act 1990.

Tabela 23 – Frases geradas pelo ChatGPT que foram classificadas como mais complexas do que a original pela métrica ISiM

Na primeira sentença, a escolha de trocar "right support" para "appropriate support" e a palavra "support" para "assist" influenciaram negativamente na pontuação da sentença. Já no segundo exemplo, a palavra "town" é mais frequente do que a palavra "cities", e a utilização de "town planning development control process" substituindo "controlling how cities are constructed", acabou introduzindo palavras mais complexas para tentar construir a sentença de uma forma mais explicativa.

Outra abordagem que fizemos utilizando o chatGPT foi perguntar a ele qual das duas sentenças desse *corpus* que ele mesmo gerou era mais simples. O resultado foi melhor do que o anterior, tendo a ISiM concordado com a simplificação de 97% das sentenças simplificadas [4] pelo ChatGPT, porém temos duas questões para serem analisadas aqui. A

[4] https://github.com/Mucida/frases-chatgpt

primeira é que, uma vez que foi o próprio modelo que gerou as sentenças, então ele pode estar sendo tendencioso. A segunda questão é saber se ele seria consistente nessas respostas se perguntarmos isso mais de uma vez. Para sanar a primeira dúvida, perguntamos novamente ao ChatGPT qual das sentenças era mais simples, mas dessa vez passando o *corpus* original e não os com as sentenças simples que ele mesmo gerou. Novamente ele teve uma alta taxa de acerto, cerca de 90%. Porém, ao perguntarmos novamente qual era a sentença mais simples ele trocou o resultado de algumas sentenças. Selecionamos algumas sentenças aleatórias do *corpus* para fazermos outros testes, e o modelo realmente alterava a sua resposta se perguntássemos mais de uma vez. Tentamos também acrescentar nessa pergunta, o motivo da escolha da sentença mais simples. O ChatGPT escolhia a sentença, dava sua explicação, mas após 2 ou 3 repetições da pergunta, ele alterava tanto a escolha da sentença quanto a própria explicação. Logo, por falta de consistência nos resultados, não podemos utilizar o ChatGPT como métrica de comparação entre sentenças para indicar a simplicidade de cada uma delas.

5.4 O Modelo Mucimples

Nesta seção, apresentamos nosso modelo de simplificação de sentenças baseado no modelo T5 (RAFFEL et al., 2020), chamado "Mucimples". O T5 é um modelo *encoder-decoder* pré-treinado em uma combinação de tarefas supervisionadas e não supervisionadas, convertidas em problemas de geração de texto. Com o objetivo de aperfeiçoar o modelo, realizamos o *fine tuning* utilizando as técnicas de *Transfer Learning*, que consiste em aproveitar o conhecimento adquirido em tarefas anteriores para solucionar um novo problema relacionado. No nosso caso, o problema era a geração de sentenças simplificadas a partir de sentenças complexas.

Para melhorar ainda mais o desempenho da tarefa de geração de texto simplificado, empregamos o *corpus* Turkcorpus, aprimorado pela métrica ISiM, conforme descrito na seção 5.2.2.1. Como já mencionado, este *corpus* consiste em pares de sentenças complexas e simplificadas e foi utilizado para aperfeiçoar o modelo T5 através da etapa de *fine tuning*. Adicionalmente, foram experimentadas diferentes combinações de parâmetros de treinamento, e a escolha final foi determinada empiricamente, com base na avaliação da qualidade dos textos gerados, lembrando que consideramos que a métrica ISiM e o modelo já exigem que todas as sentenças utilizadas sejam gramaticalmente corretas por padrão.

Em nosso estudo, realizamos uma etapa crucial para o treinamento do modelo T5: a inclusão da tag "simplify:"antes de cada sentença complexa presente no *corpus* utilizado.

Com essa modificação, quando quisermos simplificar um dado texto utilizando o modelo, precisamos colocar antes da sentença a tag "Simplify: "para que o modelo entenda qual tarefa está sendo exigida.

No código de refinamento do modelo T5, escolhemos de forma empírica após vários testes, os parâmetros mais adequados para gerar textos simplificados, onde *input_ids* são as sequências de entrada; *attention_mask* garante que os tokens de preenchimento das entradas sejam ignorados; *max_length* é o comprimento máximo da sequência a ser gerada, tendo valores entre 0 e o infinito, sendo que no nosso modelo definimos 128; *num_beams* é o número de feixes da beam search, que deve estar entre 1 e infinito onde 1 significa nenhhum bean (sem bean search), que foi o nosso caso; *repetition_penalty* é o parâmetro para penalidade de repetição de palavras, estando entre 1.0 e infinito onde 1.0 significa nenhuma penalidade, deixamos esse valor bem alto, em 1.000.000, para que o modelo tentasse sempre encontrar novas palavras (de preferência mais simples) para substituir as atuais; *length_penalty* é a penalidade exponencial ao comprimento da sentença sendo gerada, aqui o valor em nosso modelo é 10; *early_stopping* interrompe a pesquisa quando tem pelo menos pelo menos *num_beams* sentenças terminadas por lote, aqui deixamos como falso.

Note que o fator de penalização por repetição foi fortemente aumentado, visando evitar a utilização das mesmas palavras da sentença original e fazer com que o modelo opte por termos mais simples. A penalização por comprimento da sentença também foi mais alta, que vai de encontro ao que é proposto pela métrica ISiM, onde é sugerido que sentenças mais curtas tendem a ser menos complexas. Ainda temos os parâmetros que finalmente treinam o modelo T5. Estes também foram atribuídos por avaliação empírica e de acordo com os limites de recursos computacionais disponíveis.

```
model_params={
    "MODEL":"t5-base",
    "TRAIN_BATCH_SIZE":16,
    "VALID_BATCH_SIZE":16,
    "TRAIN_EPOCHS":12,
    "VAL_EPOCHS":5,
    "LEARNING_RATE":3e-5,
    "MAX_SOURCE_TEXT_LENGTH":80,
    "MAX_TARGET_TEXT_LENGTH":80,
    "SEED": 42
}
```

O treinamento desse modelo foi rápido, com apenas 12 épocas de treinamento e 5 de avaliação, com batches de tamanho 16. O treinamento durava cerca de 2 horas num computador com as seguintes configurações: Processador AMD Ryzen 7 3800X 8-Core 3.89 GHz; 32 gb de RAM 3200mhz; GPU RTX 4070 ti; HD SSD 1tb.

Para comparação de resultados, foram utilizados 6 modelos de simplificação de sentenças encontrados no *Hugging Face*. São eles: BigSalmon/SimplifyText, twigs/bart-

text2text-simplifier, JorgeSarry/est5base-simplify, mrm8488/t5-small-finetuned text-simplification, mrm8488/t5-base-finetuned-turk-text-simplification, philippelaban/keep_it_simple. Fizemos testes com 10 sentenças diferentes para cada um dos modelos acima e com o modelo Mucimples. A Tabela 24 abaixo mostra as sentenças que foram utilizadas como teste, e no capítulo de apêndices (7) são exibidos os resultados gerados por todos os modelos para cada sentença de teste. O principal motivo da utilização dos modelos do *Hugging Face* foi por serem modelos já funcionais disponibilizados pela própria plataforma. Outros modelos parecidos encontrados em artigos científicos específicos não disponibilizavam o código para teste, ou o código não funcionava mais (muitas vezes devido a incompatibilidade de versão Python).

sentença
He will This is only a test that I am doing to check how far we can go with text simplification using Transformers to achieve that goal.
However the first 15 minutes of each meeting are available for members of the public to voice opinions or raise questions.
Terms such as "undies"for underwear and "movie"for "moving picture"are oft-heard terms in English.
I am very fatigued today.
I was already despondent about what happened last fortnight.
He seems gleeful with this news today, so we are going to throw him a gathering to celebreate this great day.
The single most powerful asset we all have is our mind.
Fives is a British sport believed to derive from the same origins as many racquet sports.
Receiving the highest accolade in the contest was incredible.
He will grab every chance to show his ostentatious lifestyle.

Tabela 24 – sentenças utilizadas nos testes.

A tabela 25 mostra os resultados dos testes de cada uma das 10 sentenças para cada modelo.

Modelo	**Simplificou**	**Complexificou**	**Repetiu a sentença**	**Alterou o sentido**	**sentença incompleta**	**Não gerou**
mucimples	**6**	1	2	1	0	0
BigSalmon	1	0	0	7	0	2
twigs	1	0	4	1	4	0
JorgeSarry	2	0	3	0	5	0
mrm8488/t5-small	0	0	4	2	4	0
mrm8488/t5-base	2	1	1	0	5	1
philippelaban	0	0	0	8	0	2

Tabela 25 – Resultados apresentados por cada modelo ao rodas as 10 sentenças de teste.

Note que o Mucimples teve o melhor resultado no quesito simplificação de texto, tendo apenas 1 sentença gerada com pontuação pior do que a original, 2 sentenças foram repetidas pelo modelo sem alterações e apenas 1 teve o sentido alterado. Podemos considerar que 60% das vezes ele melhorou a sentença e que em 90% das vezes a sentença não sofreu nenhum prejuízo de sentido. Comparando com o resultado de outros modelos onde no máximo 2 sentenças foram consideradas realmente mais simples do que a original, podemos perceber uma vantagem do nosso modelo quando é exigido apenas uma simplificação da sentença, sem uma tentativa de sumarização ou alteração de muitas palavras, que podem ocasionar perda ou alteração da semântica, observada em outros modelos. Podemos observar também que alguns modelos possuem um número máximo de caracteres muito baixo, o que fazia com que sentenças maiores acabassem não sendo geradas de forma completa.

6 Conclusão

Esta pesquisa focou na simplificação de texto em linguagem natural. Dentro deste escopo foi apresentada uma métrica (denominada de ISim) e um modelo de linguagem, apoiado pela métrica criada. A métrica ISiM, proposta nesta pesquisa, se apresenta como uma contribuição no campo da simplificação textual. Este estudo demonstrou que a ISiM é uma ferramenta útil para avaliar e comparar a complexidade textual em diversos contextos. Sua aplicação inicial, apresentada na análise de simplificação de textos, revelou bons resultados, primeiramente no formulário anônimo onde obteve 75% de acerto envolvendo frases do *corpus* TaPaCo, e especialmente nos *corpora* SimPA e TurkCorpus, onde alcançou respectivamente 80,27% e 96,94% de acerto na classificação de complexidade em um par de frases, destacando-se pelo seu desempenho independente de idioma e independente de intervenção humana.

Os resultados deste estudo também demonstraram a versatilidade da métrica ISiM, estendendo-se desde a transformação de *corpora* de paráfrase até o desenvolvimento de um modelo generativo de simplificação textual, o que valida e comprova a hipótese de que a frequência das palavras e o tamanho da sentença é capaz de medir simplificação de texto de forma a auxiliar no aprimoramento de *corpus* paralelos voltados para a tarefa de simplificação de texto. A métrica se mostrou útil na otimização de *corpora* existentes, onde obteve melhorias de 0,3862% e 0.2480% na pontuação SARI do *corpus* TurkCorpus de teste e *tuning*, e de 0,0550% e 0,1073% no *corpus* ASSET de teste e *tuning*. Também validamos a a criação de novos *corpora*, obtendo os 75% de acerto na classificação de complexidade de par de sentenças comentados anteriormente. Tudo isso contribui para a expansão do número de bases disponíveis e da qualidade dessas bases textuais, permitindo melhorar o treinamento de modelos de simplificação de texto. Isso mostra que a métrica ISiM permitiu criar e/ou melhorar *corpora* que já eram amplamente utilizados na área de simplificação textual, utilizando a confiabilidade da métrica SARI para demonstrar tal melhoria.

Por último, apresentamos a aplicação da métrica na criação de um modelo gerador de textos, simplificando frases de maneira mais eficaz do que os concorrentes diretos encontrados na literatura. Nosso modelo se destacou não só na qualidade da simplificação dos textos, como também no tamanho, uma vez que aceita frases maiores, sem repetição da frase modelo e sem ,consequente, alteração no sentido da frase gerada em 90% das tentativas.

A métrica ISiM pode ser aprimorada com a incorporação de critérios adicionais provenientes de índices de análise de complexidade amplamente reconhecidos, como o *Gunning Fog*, o *Coleman-Liau* e o *SMOG*, abrindo caminho para uma abordagem mais

abrangente da avaliação da complexidade textual.

Este estudo se concentrou na aplicação da métrica nas línguas portuguesa e inglesa, mas postulamos que ela pode ser aplicada em línguas que usam o alfabeto latino, sistema de escrita alfabético e estrutura de frase similar. Futuras pesquisas devem explorar a eficácia da ISiM em outros grupos de idiomas, como os de origem germânica, oriental e semítica, como o mandarim, japonês e árabe, ampliando assim seu potencial alcance e impacto.

Outra limitação da métrica é não garantir a equivalência semântica entre as sentenças analisadas. Diante disso, reconhecemos a necessidade de aprimoramentos futuros para abordar esse aspecto crítico, considerando a inclusão de medidas de similaridade semântica baseadas na vetorização densa das sentenças.

Por fim, a continuação deste trabalho poderia direcionar-se para contextos mais específicos e adaptados, refinando *corpora* de acordo com necessidades particulares de aplicação em modelos de geração de texto simplificado. Este avanço não apenas aprimoraria a precisão da simplificação textual, mas também enriqueceria a área de Processamento de Linguagem Natural, fortalecendo a base de conhecimento e a qualidade das soluções nesse domínio em constante evolução.

7 Material Suplementar

Frase 1	**Frase 2**
Você gostou da festa?	Que tal foi a festa?
Ele ficou manco depois da queda.	Ele ficou coxo depois da queda.
A taxa de mortalidade está mais baixa.	A taxa de mortalidade diminuiu.
Vês a coroa?	Você vê uma coroa?
Vocês gostariam de ajudar, não gostariam?	Você gostaria de ajudar, não é?
O sorriso que enviais vos é devolvido.	O sorriso que você envia retorna a você.
Este é um encontro muito importante.	Esta é uma reunião muito importante.
Eu não o tenho visto há muito tempo.	Faz um tempão que não a vejo.
Ele perdeu as esperanças e se matou com veneno.	Ele perdeu esperança e suicidou-se com veneno.
Eu devo proferir um discurso?	Eu tenho que dar um discurso?
Eu tenho algo a lhe dizer, Tom.	Tom, tenho algo a dizer para você.
Não conseguimos convencê-lo.	Não conseguimos persuadi-lo.
Confia, desconfiando.	A desconfiança causa segurança.
Japoneses são asiáticos.	O Japão está na Ásia.
Eu sou do meu amado, seu desejo o traz a mim.	Eu pertenço ao meu amado, e ele me deseja.
Eu adoraria saber se minha bagagem chegará logo.	Eu quero saber se a minha bagagem vai chegar.
Uma xícara de chá, por favor.	Pode me trazer um copo de chá?
Você está fazendo mau uso da sua autoridade.	Você está abusando de sua autoridade.
Nunca aperte esse botão.	Nunca pressione este botão.
Nunca vou esquecê-la.	Nunca te esquecerei.
Caminhar é o melhor exercício.	Caminhar é um excelente exercício.
É por isso que eu estou atrasado.	Eis o porquê de eu ter chegado atrasado.
O escritor está trabalhando em um novo livro.	Esse escritor está escrevendo agora um novo livro.
Você está nervoso.	Elas ficam nervosas.
Tom arrumou a sala de estar.	Tom pôs a sala de estar em ordem.
Tu sabes em que sala se realizará a conferência?	Você sabe em qual sala a reunião será realizada?
A mulher é muito bonita.	Esta mulher tem uma aparência muito boa.
Elas nos fizeram trabalhar por toda a noite.	Eles nos fizeram trabalhar a noite inteira.
Eu quero te beijar.	Gostaria de beijá-lo.
Tu eras o meu herói.	Você era meu herói.
Divertirei as crianças.	Eu divertirei as crianças.
Seja qual for a vida que levardes, eu vos apoiarei.	Seja qual for a vida que leves, eu vou apoiar-te.
As pessoas sempre reclamam do tempo.	As pessoas sempre se queixam do tempo.
Eu espero que você venha para a minha festa de aniversário.	As senhoras me darão prazer se vierem à festa de meu aniversário.
Você fez isto por si mesmo?	Fizeste-o só?
Para começar, tens que deixar de fumar.	Primeiro, você deve parar de fumar.
Ela não gostou de tudo.	Nem tudo foi do agrado dela.
Eu estudo espanhol.	Aprendo espanhol.
Meu pai vai se aposentar aos 60 anos.	Meu pai aposentar-se-á aos 60 anos.
Terminaste com o livro?	Você já leu esse livro?

Tabela 26 – Frases utilizadas no formulário anônimo para validação da métrica ISiM

Modelos	**Frase**	**Resultado**
original	This is only a test that I am doing to check how far we can go with text simplification using Transformers to achieve that goal.	3.24409
mucimples	a test that I am doing to check how far we can go with text simplification using Transformers for this goal.	3.31944
BigSalmon/SimplifyText	This is only a test, not a comprehensive plan. This is only a test that I am doing right now.	alterou sentido
twigs/bart-text2text-simplifier	This is only a test that I am doing to see how far we can go with	incompelto
JorgeSarry/est5base-simplify	This is only a test that I am doing to check how far we can go with	incompelto
mrm8488/t5-small-finetuned-text-simplification	This is only a test that I am doing to check how far we can go with text	incompelto
mrm8488/t5-base-finetuned-turk-text-simplification	to see how far we can go with text simplification using Transformers to achieve that goal	incompelto
philippelaban/keep_it_simple	This is only a test, not a final solution. This is only a test, not a final solution.	alterou sentido

Tabela 27 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	However the first 15 minutes of each meeting are available for members of the public to voice opinions or raise questions.	3.38380
mucimples	Nevertheless the first 15 minutes of each meeting are available for members to voice opinions or raise questions.	3.35839
BigSalmon/SimplifyText	However the 15 minutes are not publicly available to members of the public.	alterou sentido
twigs/bart-text2text-simplifier	However , the first 15 minutes of each meeting are available for members of the public to	incompelto
JorgeSarry/est5base-simplify	However the first 15 minutes of each meeting are available for members of the public to voice	incompelto
mrm8488/t5-small-finetuned-text-simplification	The first 15 minutes of each meeting are available for members of the public to voice opinions or raise	incompelto
mrm8488/t5-base-finetuned-turk-text-simplification	the first 15 minutes of each meeting are available for members of the public to voice opinions or raise	incompelto
philippelaban/keep_it_simple	However, the 15 minutes do not include any discussion of policy or candidates.	alterou sentido

Tabela 28 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	Terms such as "undies"for underwear and "movie"for "moving picture"are oft-heard terms in English.	3.27910
mucimples	undies for underwear and the term movie is often heard in English.	3.61729
BigSalmon/SimplifyText	.....	alterou sentido
twigs/bart-text2text-simplifier	The term underwear is often used in English to refer to underwear .	alterou sentido
JorgeSarry/est5base-simplify	Terms such as "undies"for underwear and "movie"for "moving	incompelto
mrm8488/t5-small-finetuned-text-simplification	Terms such as "undies "for underwear and "movie "for "mov	incompelto
mrm8488/t5-base-finetuned-turk-text-simplification	"undies"for underwear and "movie"for "moving picture"	incompelto
philippelaban/keep_it_simple	.....	alterou sentido

Tabela 29 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	I am very fatigued today.	4.03623
mucimples	I am very tired today.	4.29108
BigSalmon/SimplifyText	I have lost my sense of timing. I want to hit the ground running but am tired of trying. I am not sure yet how I will perform.	continuou o texto
twigs/bart-text2text-simplifier	I am very fatigued today .	repetiu
JorgeSarry/est5base-simplify	I am very fatigued.	4.19759
mrm8488/t5-small-finetuned-text-simplification	I am very fatigued today.	repetiu
mrm8488/t5-base-finetuned-turk-text-simplification	I am very tired today.	4.29108
philippelaban/keep_it_simple	It's been a long time since I sat at my desk and looked at the screen. Now I'm in a room full of people, and I am fatigued.	continuou o texto

Tabela 30 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	I was already despondent about what happened last fortnight.	3.61485
mucimples	I was already despondent about what happened last fortnight.	repetiu
BigSalmon/SimplifyText	I had already expected the result of last fortnight to be a lot more complicated than it actually was. I was already despondent about what would happen if I saw this result.	alterou sentido
twigs/bart-text2text-simplifier	I was already despondent about what happened last week .	3.77221
JorgeSarry/est5base-simplify	I was despondent about what happened last fortnight.	3.69142
mrm8488/t5-small-finetuned-text-simplification	I was already despondent about what happened last fortnight .	repetiu
mrm8488/t5-base-finetuned-turk-text-simplification	. . . . .	erro
philippelaban/keep_it_simple	I knew that I had to get better at what I do. I had to think twice about what I did and why. I was already feeling out of sorts.	alterou sentido

Tabela 31 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	He seems gleeful with this news today, so we are going to throw him a gathering to celebreate this great day.	3.25295
mucimples	he seems happy with this news today, so we are going to throw him an event for his celebration of great day.	3.40570
BigSalmon/SimplifyText	.....	alterou sentido
twigs/bart-text2text-simplifier	He seems gleeful with this news today , so we are going to throw him a	incompelto
JorgeSarry/est5base-simplify	He seems gleeful con this news today, so we are going to throw	incompelto
mrm8488/t5-small-finetuned-text-simplification	He seems gleeful with this news today, so we are going to throw him	incompelto
mrm8488/t5-base-finetuned-turk-text-simplification	he seems happy with this news today, so we are going to throw him a gathering	incompelto
philippelaban/keep_it_simple	.....	alterou sentido

Tabela 32 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	The single most powerful asset we all have is our mind.	3.83957
mucimples	The single most powerful asset we all have is our mind.	repetiu
BigSalmon/SimplifyText	We know what we want and what we need, and we have a clear plan. The single most powerful asset we have is our mind. We know what we need and what we want,	continuou o texto
twigs/bart-text2text-simplifier	The single most powerful asset we all have is our mind .	repetiu
JorgeSarry/est5base-simplify	The single most powerful asset we all have is our mind.	repetiu
mrm8488/t5-small-finetuned-text-simplification	The single most powerful asset we all have is our mind .	repetiu
mrm8488/t5-base-finetuned-turk-text-simplification	the single most powerful asset we all have is our mind.	repetiu
philippelaban/keep_it_simple	We can see how much we can improve, and how much progress we can make. The single most powerful asset we have is our mind.	continuou o texto

Tabela 33 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	Fives is a British sport believed to derive from the same origins as many racquet sports.	3.23518
mucimples	fives is a British sport believed to derive from the same origin as many other sports like tennis and ballroom dancing.	alterou sentido
BigSalmon/SimplifyText	Fives is considered to be part of a broader tradition of racquet sports, such as football or baseball.	3.28396
twigs/bart-text2text-simplifier	Fives is a British sport believed to come from the same origins as many racquet	incompelto
JorgeSarry/est5base-simplify	Fives is a British sport believed to derive from the same origins	incompelto
mrm8488/t5-small-finetuned-text-simplification	Fives is a sport in the British Isles .	alterou sentido
mrm8488/t5-base-finetuned-turk-text-simplification	fives is a British sport believed to have the same origins as many racque	incompelto
philippelaban/keep_it_simple	Fives is very fast and fast at the net, but not at full speed. Fives is not an easy target for a driver, but it can	alterou sentido

Tabela 34 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	Receiving the highest accolade in the contest was incredible.	3.73122
mucimples	Getting the highest accolade in this contest was incredible.	3.73634
BigSalmon/SimplifyText	Receiving the highest honour was undoubtedly impressive. He was recognised as one of the game's best players.	alterou sentido
twigs/bart-text2text-simplifier	Receiving the highest accolade in the contest was incredible .	repetiu
JorgeSarry/est5base-simplify	Receiving the highest accolade in the contest was incredible.	repetiu
mrm8488/t5-small-finetuned-text-simplification	The highest accolade in the contest was incredible .	alterou sentido
mrm8488/t5-base-finetuned-turk-text-simplification	the highest award in the contest was incredible.	4.07959
philippelaban/keep_it_simple	Receiving the highest prize was undoubtedly important. The players were so effective, they even managed to score	alterou sentido

Tabela 35 – Resultados do teste de geração de frase simplificada.

Modelos	**Frase**	**Resultado**
original	grab every chance to show his ostentatious lifestyle.	3.57046
mucimples	He will grab every chance to show his extravagant lifestyle.	3.61695
BigSalmon/SimplifyText	He will walk to the gym and then chat away with friends. He will not be bothered by the fact that he will not be able to eat much.	alterou sentido
twigs/bart-text2text-simplifier	He will grab every chance to show his ostentatious lifestyle .	repetiu
JorgeSarry/est5base-simplify	He will grab every chance to show his ostentatious lifestyle.	repetiu
mrm8488/t5-small-finetuned-text-simplification	He will grab every chance to show his ostentatious lifestyle .	repetiu
mrm8488/t5-base-finetuned-turk-text-simplification	he will grab every chance to show off his ostentatious lifestyle.	3.52944
philippelaban/keep_it_simple	He will also look to cash in on his celebrity status. He will also try to sell his personality. He will not be afraid of appearing ostentatious.	alterou sentido

Tabela 36 – Resultados do teste de geração de frase simplificada.

Referências

AFONSO, S. A floresta sintá (c) tica como recurso. 2004.

ALKALDI, W.; INKPEN, D. Text simplification to specific readability levels. **Mathematics**, MDPI, v. 11, n. 9, p. 2063, 2023.

ALVA-MANCHEGO, F. et al. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2020. p. 4668–4679.

ALVA-MANCHEGO, F.; SCARTON, C.; SPECIA, L. The (un) suitability of automatic evaluation metrics for text simplification. **Computational Linguistics**, MIT Press, v. 47, n. 4, p. 861–889, 2021.

ANDROUTSOPOULOS, I.; MALAKASIOTIS, P. A survey of paraphrasing and textual entailment methods. **Journal of Artificial Intelligence Research**, v. 38, p. 135–187, 2010.

ANDRYCHOWICZ, M. et al. Learning to learn by gradient descent by gradient descent. **Advances in neural information processing systems**, v. 29, 2016.

BABU, A. et al. Non-autoregressive semantic parsing for compositional task-oriented dialog. **arXiv preprint arXiv:2104.04923**, 2021.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014.

BANU, A.; FATIMA, S. S.; KHAN, K. U. R. Information content based semantic similarity measure for concepts subsumed by multiple concepts. **IJWA**, v. 7, n. 3, p. 85–94, 2015.

BLINOVA, S. et al. Simsum: Document-level text simplification via simultaneous summarization. In: **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2023. p. 9927–9944.

BORMUTH, J. R. Readability: A new approach. **Reading research quarterly**, JSTOR, p. 79–132, 1966.

BOWMAN, S. R. et al. A large annotated corpus for learning natural language inference. **arXiv preprint arXiv:1508.05326**, 2015.

BROWN, T. et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, 2020.

CARROLL, J. B.; DAVIES, P. Barry rich man. **Word Frequency Book. Boston: Hough ton Mifflin**, 1971.

CELIKYILMAZ, A.; CLARK, E.; GAO, J. Evaluation of text generation: A survey. **arXiv preprint arXiv:2006.14799**, 2020.

CHOROWSKI, J. K. et al. Attention-based models for speech recognition. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2015. p. 577–585.

CLIVE, J.; CAO, K.; REI, M. Control prefixes for parameter-efficient text generation. In: **Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)**. [S.l.: s.n.], 2022. p. 363–382.

COLEMAN, M.; LIAU, T. L. A computer readability formula designed for machine scoring. **Journal of Applied Psychology**, American Psychological Association, v. 60, n. 2, p. 283, 1975.

CROSSLEY, S. A.; ALLEN, D. B.; MCNAMARA, D. S. Text readability and intuitive simplification: A comparison of readability formulas. **Reading in a foreign language**, ERIC, v. 23, n. 1, p. 84–101, 2011.

CROSSLEY, S. A. et al. A linguistic analysis of simplified and authentic texts. **The Modern Language Journal**, Wiley Online Library, v. 91, n. 1, p. 15–30, 2007.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. v. 1, p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>.

DEVLIN, S.; UNTHANK, G. Helping aphasic people process online information. In: **Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility**. [S.l.: s.n.], 2006. p. 225–226.

DONG, Y. et al. Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing. **arXiv preprint arXiv:1906.08104**, 2019.

EVANS, R.; ORASAN, C.; DORNESCU, I. An evaluation of syntactic simplification rules for people with autism. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)**. [S.l.], 2014.

FLESCH, R. A new readability yardstick. **Journal of applied psychology**, American Psychological Association, v. 32, n. 3, p. 221, 1948.

FONSECA, E. R. et al. Visão geral da avaliação de similaridade semântica e inferência textual. **Linguamática**, v. 8, n. 2, p. 3–13, 2016.

GANIN, Y.; LEMPITSKY, V. Unsupervised domain adaptation by backpropagation. In: PMLR. **International conference on machine learning**. [S.l.], 2015. p. 1180–1189.

GASPERIN, C. et al. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. **The proceedings of SEMISH-XXXVI seminário integrado de software e hardware**, Citeseer, p. 387–401, 2009.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.

GOODMAN, N. Problems and projects. 1972.

GRABAR, N.; SAGGION, H. Evaluation of automatic text simplification: Where are we now, where should we go from here. In: ATALA. **Traitement Automatique des Langues Naturelles**. [S.l.], 2022. p. 453–463.

GUNNING, R. The fog index after twenty years. **Journal of Business Communication**, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 2, p. 3–13, 1969.

HARISPE, S. et al. Semantic similarity from natural language and ontology analysis. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers, v. 8, n. 1, p. 1–254, 2015.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. **arXiv preprint arXiv:1801.06146**, 2018.

HU, D. An introductory survey on attention mechanisms in nlp problems. In: SPRINGER. **Proceedings of SAI Intelligent Systems Conference**. [S.l.], 2019. p. 432–448.

IBGE. **Taxa de analfabetismo no brasil**. 2019. Disponível em: <https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18317-educacao.html>.

INAF. Indicador de analfabetismo funcional. **Instituto Paulo Montenegro, Ação Educativa. São Paulo**, Instituto Paulo Montenegro, 2018. Disponível em: <https://acaoeducativa.org.br/wp-content/uploads/2018/08/Inaf2018_Relat%C3%B3rio-Resultados-Preliminares_v08Ago2018.pdf>.

KAUCHAK, D. Improving text simplification language modeling using unsimplified text data. In: **Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)**. [S.l.: s.n.], 2013. p. 1537–1546.

KIM, J.; STOREY, V. C. Construction of domain ontologies: Sourcing the world wide web. **International Journal of Intelligent Information Technologies (IJIIT)**, IGI Global, v. 7, n. 2, p. 1–24, 2011.

KINCAID, J. P. et al. **Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel**. [S.l.], 1975.

KLARE, G. R. et al. **Measurement of readability**. [S.l.]: Iowa State University Press, 1963.

KRASHEN, S. The lexile framework: The controversy continues. **California School Library Journal**, v. 25, n. 2, p. 29–31, 2002.

KRIZ, R.; APIDIANAKI, M.; CALLISON-BURCH, C. Simple-qe: Better automatic quality estimation for text simplification. **arXiv preprint arXiv:2012.12382**, 2020.

LENNON, C.; BURDICK, H. The lexile framework as an approach for reading measurement and success. **electronic publication on www. lexile. com**, 2004.

LEVY, O.; GOLDBERG, Y. Neural word embedding as implicit matrix factorization. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2014. p. 2177–2185.

LI, Z.; SHARDLOW, M. How do control tokens affect natural language generation tasks like text simplification. **Natural Language Engineering**, Cambridge University Press, p. 1–28, 2024.

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

MADDELA, M. et al. Lens: A learnable evaluation metric for text simplification. **arXiv preprint arXiv:2212.09739**, 2022.

MARTIN, L. et al. Muss: multilingual unsupervised sentence simplification by mining paraphrases. **arXiv preprint arXiv:2005.00352**, 2020.

MARTIN, L. et al. Controllable sentence simplification. **arXiv preprint arXiv:1910.02677**, 2019.

MAX, A. Writing for language-impaired readers. In: SPRINGER. **International Conference on Intelligent Text Processing and Computational Linguistics**. [S.l.], 2006. p. 567–570.

MCENERY, T. **Corpus linguistics**. [S.l.]: Oxford University Press Inc, 2012. v. 978019.

MILLER, G. A.; GILDEA, P. M. How children learn words. **Scientific American**, JSTOR, v. 257, n. 3, p. 94–99, 1987.

MUCIDA, L.; OLIVEIRA, A.; POSSI, M. A language-independent metric for measuring text simplification that does not require a parallel corpus. In: **The International FLAIRS Conference Proceedings**. [S.l.: s.n.], 2022. v. 35.

NOY, N. F.; MCGUINNESS, D. L. et al. **Ontology development 101: A guide to creating your first ontology**. [S.l.]: Stanford knowledge systems laboratory technical report KSL-01-05 and . . . , 2001.

OMELIANCHUK, K.; RAHEJA, V.; SKURZHANSKYI, O. Text simplification by tagging. **arXiv preprint arXiv:2103.05070**, 2021.

OPENAI. **Introducing ChatGPT**. 2022. Acessado em: 20 de Maio de 2023. Disponível em: <https://openai.com/blog/chatgpt>.

PAETZOLD, G.; SPECIA, L. Understanding the lexical simplification needs of non-native speakers of english. In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. [S.l.: s.n.], 2016. p. 717–727.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2002. p. 311–318.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543.

PETERS, M. E. et al. Deep contextualized word representations. corr abs/1802.05365 (2018). **arXiv preprint arXiv:1802.05365**, 2018.

RADFORD, A. et al. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019.

RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **arXiv preprint arXiv:1910.10683**, 2019.

RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **The Journal of Machine Learning Research**, JMLRORG, v. 21, n. 1, p. 5485–5551, 2020.

RELLO, L. et al. Simplify or help? text simplification strategies for people with dyslexia. In: **Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility**. [S.l.: s.n.], 2013. p. 1–10.

ROCHA, G.; CARDOSO, H. L. Recognizing textual entailment: challenges in the portuguese language. **Information**, Multidisciplinary Digital Publishing Institute, v. 9, n. 4, p. 76, 2018.

RONG, X. word2vec parameter learning explained. **arXiv preprint arXiv:1411.2738**, 2014.

SCARTON, C.; PAETZOLD, G.; SPECIA, L. Simpa: A sentence-level simplification corpus for the public administration domain. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. [S.l.: s.n.], 2018.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010.

SCHERRER, Y. Tapaco: A corpus of sentential paraphrases for 73 languages. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION (ELRA). **Proceedings of The 12th Language Resources and Evaluation Conference**. [S.l.], 2020.

SHARDLOW, M. A survey of automated text simplification. **International Journal of Advanced Computer Science and Applications**, v. 4, n. 1, p. 58–70, 2014.

SIDDHARTHAN, A. A survey of research on text simplification. **ITL-International Journal of Applied Linguistics**, John Benjamins, v. 165, n. 2, p. 259–298, 2014.

SMITH, D. R. et al. **The Lexile Scale in Theory and Practice. Final Report.** [S.l.]: MetaMetrics, Inc., 1989.

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 and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**. [S.l.: s.n.], 2015. p. 823–828.

STENNER, A. J. Measuring reading comprehension with the lexile framework. In: **Proceedings of the North American Conference on AdolescentAdult Literacy**. [S.l.: s.n.], 1996.

STENNER, A. J.; III, M. S.; BURDICK, D. S. Toward a theory of construct definition. **Journal of educational measurement**, JSTOR, p. 305–316, 1983.

STUDENT. The probable error of a mean. **Biometrika**, JSTOR, p. 1–25, 1908.

SUN, H.; ZHOU, M. Joint learning of a dual smt system for paraphrase generation. In: **Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. [S.l.: s.n.], 2012. p. 38–42.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2014. p. 3104–3112.

TALBURT, J. The flesch index: An easily programmable readability analysis algorithm. In: **Proceedings of the 4th annual international conference on Systems documentation**. [S.l.: s.n.], 1986. p. 114–122.

VADLAMANNATI, S.; ŞAHIN, G. G. Metric-based in-context learning: A case study in text simplification. **arXiv preprint arXiv:2307.14632**, 2023.

VASWANI, A. et al. Attention is all you need. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2017. p. 5998–6008.

XENOULEAS, S. et al. SUM-QE: a BERT-based summary quality estimation model. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 6005–6011. Disponível em: <https://aclanthology.org/D19-1618>.

XU, W.; CALLISON-BURCH, C.; NAPOLES, C. Problems in current text simplification research: New data can help. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 3, p. 283–297, 2015.

XU, W. et al. Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 4, p. 401–415, 2016.

YANG, Z. et al. Xlnet: Generalized autoregressive pretraining for language understanding. **Advances in neural information processing systems**, v. 32, 2019.

ZHANG, X.; LAPATA, M. Sentence simplification with deep reinforcement learning. **arXiv preprint arXiv:1703.10931**, 2017.

ZHAO, S. et al. Integrating transformer and paraphrase rules for sentence simplification. **arXiv preprint arXiv:1810.11193**, 2018.