UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

PAULO BERLANGA NETO

Simplificação Automática de Sentenças com Ênfase no Método Split-and-Rephrase

Ribeirão Preto–SP

2021

PAULO BERLANGA NETO

Simplificação Automática de Sentenças com Ênfase no Método Split-and-Rephrase

Versão Corrigida

Versão original encontra-se na FFCLRP/USP.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador: Prof. Dr. Evandro Eduardo Seron Ruiz

Ribeirão Preto–SP

2021

PAULO BERLANGA NETO

Sentence Simplification with Emphasis on Split-and-Rephrase Method

Corrected Version

The original version is found at FFCLRP/USP.

Dissertation presented to Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) from the Universidade de São Paulo (USP), as part of the requirements to hold the Master of Science degree.

Field of Study: Applied Computing.

Supervisor: Prof. Dr. Evandro Eduardo Seron Ruiz

Ribeirão Preto–SP

2021

Paulo Berlanga Neto
Simplificação Automática de Sentenças com Ênfase no Método Split-and-Rephrase.
Ribeirão Preto–SP, 2021.
120p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras
de Ribeirão Preto da USP, como parte das exigências para
a obtenção do título de Mestre em Ciências,
Área: Computação Aplicada.

Orientador: Prof. Dr. Evandro Eduardo Seron Ruiz

1. Processamento de Linguagem Natural. 2. Simplificação de Sentenças. 3. Split-and-Rephrase.

Paulo Berlanga Neto

Simplificação Automática de Sentenças com Ênfase no Método Split-and-Rephrase

Modelo canônico de trabalho monográfico acadêmico em conformidade com as normas ABNT.

Orientador:
Prof. Dr. Evandro Eduardo Seron Ruiz

Professor
Convidado 1

Professor
Convidado 2

Ribeirão Preto–SP
2021

Este trabalho é dedicado ao meu avô José Toniollo (in memoriam),
meu maior exemplo de conduta para a vida.

Agradecimentos

Agradeço primeiramente ao meu orientador e amigo, Prof. Dr. Evandro Eduardo Seron Ruiz, pela confiança depositada desde o início, por todos os conselhos e pela ajuda incessante, mesmo em tempos complicados de pandemia e distanciamento.

Agradeço à todos os demais professores, colegas de classe e de laboratório, com quem também tive a chance de aprender muito nesta jornada, e à todos os queridos funcionários da USP pela receptividade e suporte de sempre, em especial, à Lucia Akemi Tatsuno Rodrigues.

Agradeço aos meus pais e à minha (em breve) futura esposa, pelo incentivo constante neste percurso, pelo contínuo compartilhamento de ideias sobre este projeto, por cada uma das pequenas realizações comemoradas, e pela nossa vida de sempre.

Por fim, agradeço ao meu saudoso avô José Toniollo (*in memoriam*), por todos os ensinamentos e pelo exímio convívio na infância, pela educação e o gosto por esportes, pelo exemplo de vida, de humildade e sabedoria.

"Creativity is intelligence having fun!"
(Albert Einstein)

Resumo

Simplificação de texto (ST) é um processo de transformação da linguagem natural para redução de sua complexidade e aumento de sua compreensão. No cerne deste problema, está a necessidade de uma preservação semântica adequada em conjunto com a melhoria da inteligibilidade. No campo de processamento de linguagem natural, abordagens recentes para a tarefa de simplificação automática de textos têm visto este processo de maneira holística ou abrangente. Muitas das ideias aplicadas são emprestadas pela tarefa de tradução automática, considerando que a simplificação pode ser vista como uma ação de tradução monolíngue entre um texto complexo e simples. Ao ponderar ainda que textos considerados complexos podem conter uma parcela de sentenças simples em sua composição, estudos recentes têm endereçado aspectos específicos da linguagem no âmbito de sentenças. Nesta pesquisa, visitamos a tarefa de simplificação automática de sentenças, apresentando características de abordagens recentes e propondo a construção de um *pipeline* computacional próprio para o aprendizado artificial do método *split-and-rephrase*. Este método busca particionar uma sentença singular de entrada em duas ou mais sentenças reescritas de saída que juntas mantêm o significado equivalente, com a concepção de que sentenças mais curtas beneficiam a compreensão na leitura humana e aprimoram o desempenho de tarefas relacionadas em processamento de linguagem natural.

Palavras-chave: Processamento de linguagem natural. Simplificação de sentenças. Particionamento e reformulação.

Abstract

Text Simplification (TS) is transforming natural language to reduce its complexity and improve its comprehension. At the heart of this problem is the need for adequate semantic preservation together with improved readability. In natural language processing, recent approaches to the automatic text simplification task have seen this process holistically. Many of the ideas applied are borrowed from the machine translation (MT) task since simplification can be considered a monolingual translation between complex and simple texts. Given that texts considered complex may contain a portion of simple sentences in their composition, recent studies have considered specific aspects of the language at the sentence level. In this research, we visited the sentence simplification task, presenting characteristics of recent approaches and proposing the construction of a computational pipeline to address the *split-and-rephrase* method. This method seeks to split one single input sentence into two or more output sentences that retain equivalent meaning, conceptualizing the notion that shorter sentences benefit human reading comprehension and improve the performance of natural language processing-related tasks.

Keywords: Natural language processing. Sentence simplification. *Split-and-rephrase*.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
EW	English Wikipedia
FKGL	Flesch-Kincaid Grade Level
FRE	Flesch Reading Ease
GloVe	Global Vectors for Word Representation
GRU	Gated Recurrent Unit
INAF	Indicador de Alfabetismo Funcional
LSTM	Long Short-Term Memory
MLM	Masked Language Modeling
NLTK	Natural Language Toolkit
NMT	Neural Machine Translation
NTS	Neural Text Simplification
PLN	Processamento de Linguagem Natural
POS	Part-of-Speech
RNN	Recurrent Neural Network
SARI	System output Against References and against the Input sentence
SEW	Simple English Wikipedia
Seq2Seq	Sequence-to-Sequence
TS	Text Simplification
UCCA	Universal Cognitive Conceptual Annotation

Lista de símbolos

$\in$	Pertence
$\sum$	Letra grega maiúscula sigma (somatório)
σ	Letra grega minúscula sigma
ζ	Letra grega minúscula zeta

Sumário

Introdução

Simplificação de texto (ST) é um processo de transformação da linguagem natural para redução de sua complexidade e aumento de sua compreensão (SHARDLOW, 2014). A transformação do vocabulário ou da estrutura de um texto para uma abordagem simplificada pode beneficiar indivíduos com habilidades textuais limitadas, tais como populações com baixos níveis de educação, crianças, não nativos, pessoas que possuem transtornos de aprendizagem (como autismo, dislexia e afasia), entre outros públicos (ŠTAJNER; CALIXTO; SAGGION, 2015; GUO; PASUNURU; BANSAL, 2018).

Embora a simplicidade de um texto pareça algo intuitivamente óbvio, em termos técnicos esta simplificação não possui uma definição precisa. Métricas de avaliação clássicas consideram fatores como o comprimento das sentenças, a contagem de sílabas e outras análises superficiais do texto para caracterização da complexidade (SHARDLOW, 2014). No entanto, em alguns casos, um texto pode ser mais longo mas também mais explicativo e fácil de entender, motivando a busca de uma análise mais profunda sobre coesão e coerência (GRAESSER et al., 2004).

Entre os fatores relativos a compreensão de leitura de um texto, temos a legibilidade (apresentação gráfica do texto) e a inteligibilidade (uso de palavras mais frequentes e estruturas sintáticas menos complexas) (SCARTON; ALUÍSIO, 2010). A inteligibilidade de textos é um tópico de pesquisa relevante, com forte impacto pedagógico, especialmente ligado ao desenvolvimento de materiais para auxílio ao aprendizado. Os estudos nesta área, geralmente buscam criar uma escala de dificuldade para a avaliação do nível de complexidade dos textos (CURTO; MAMEDE; BAPTISTA, 2014).

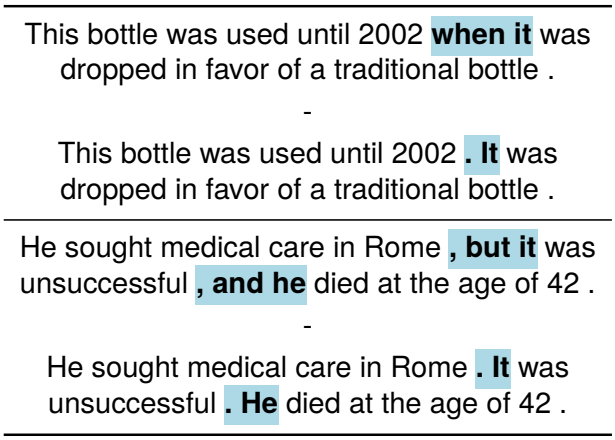
O processamento de linguagem natural (PLN) consiste em um campo de estudo das capacidades e limitações das máquinas para compreensão, interpretação e geração das línguas humanas, da maneira como são escritas ou faladas. De modo geral, promove interação entre os ramos da ciência da computação, inteligência artificial e linguística (RESHAMWALA; MISHRA; PAWAR, 2013). Em meio às tarefas desenvolvidas e aprimoradas por este campo, encontram-se, por exemplo, a tradução automática de textos, a sumarização automática, a correção ortográfica, a análise sintática, a análise de sentimentos, entre outras, bem como a tarefa de simplificação automática de textos, objeto de estudo fundamental desta pesquisa.

Além da vantagem implicitamente obtida para a leitura humana, processos de simplificação de textos também viabilizam ganhos de desempenho para outras atividades em processamento de linguagem natural, como nas tarefas de extração de informação e tradução automática (NIKLAUS et al., 2019b; ŠTAJNER; POPOVIĆ, 2019). Ou seja, é válido considerar que dentro da linguística computacional, as tarefas por vezes recorrem umas às outras.

Frequentemente, textos considerados complexos podem conter uma parcela de sentenças simples em sua composição. Para que a tarefa de simplificação seja mais eficaz e precisa, estudos geralmente direcionam esforços para identificação de complexidades analisando aspectos no âmbito de sentenças (LEAL; DURAN; ALUÍSIO, 2018), e buscam a simplificação de sentenças (ALVA-MANCHEGO; SCARTON; SPECIA, 2020). Embora tenha havido um progresso significativo nos últimos anos, modelos e *pipelines* atuais ainda não são capazes de executar a tarefa de simplificação de sentenças de forma totalmente automática, e até ao presente não fornecem desempenhos diretamente úteis para usuários finais (ALVA-MANCHEGO; SCARTON; SPECIA, 2020).

Split-and-rephrase, método proposto por Narayan et al. (2017), é uma abordagem recente dentro do estudo de simplificação de sentenças que tem atraído interesse no campo de processamento de linguagem natural. Seu objetivo é particionar uma dada sentença complexa de entrada em duas ou mais sentenças simplificadas de saída, as quais, em conjunto, manterão um significado equivalente (veja exemplos na Figura 1). O desafio neste processo é promover adequadamente as transformações sintáticas exigidas pela ação de particionamento, sem que haja perda de informação. Ou seja, preservando a gramaticalidade e o sentido original da sentença de entrada.

Figura 1 – Exemplos típicos de atuação do método *split-and-rephrase* extraídos do *corpus* WikiSplit (BOTHA et al., 2018), destacando as transformações demandadas pela ação de particionamento das sentenças. Fonte: O autor.



Motivação e Objetivos

Embora existam variados trabalhos acadêmicos sobre simplificação automática de textos no âmbito de sentenças, a literatura ainda carece de estudos aplicados ou relacionados ao português brasileiro. Em consultas realizadas em bases de dados *online* distintas, constatamos uma larga prevalência de aplicações na língua inglesa, certamente impulsionadas pelo maior número de pesquisas e recursos disponíveis para esta língua. Especificamente a respeito do método *split-and-rephrase*, não encontramos nenhuma aplicação do mesmo para a língua portuguesa.

Considerando a relevância do tema de compreensão de leitura para a ciência da computação e também para o propósito de inclusão social na aprendizagem no contexto pedagógico brasileiro, identificamos este nicho de pesquisa para a construção de um *pipeline* computacional próprio para simplificação automática de sentenças com o método *split-and-rephrase*.

Sendo assim, o objetivo do presente trabalho é investigar o tema da simplificação automática de textos no âmbito de sentenças, bem como explorar possibilidades provenientes de estudos sobre modelos computacionais, técnicas, algoritmos e métricas de avaliação do campo de PLN para a implementação de um *pipeline* próprio para atendimento do método *split-and-rephrase*, com a realização de experimentos nas línguas inglesa e portuguesa. Com base no aprendizado artificial, este *pipeline* deve automaticamente efetuar intervenções na estrutura de uma sentença complexa de entrada para realizar o particionamento e a reescrita em sentenças simplificadas de saída, idealmente mantendo a gramaticalidade e o significado original. Consideramos que esta pesquisa possa gerar uma colaboração acadêmica e científica tanto para a área de processamento de linguagem natural, como para outras áreas de interesse em textos simplificados.

Objetivo Geral

O objetivo geral deste trabalho é elaborar uma pesquisa sobre recentes abordagens de simplificação automática de sentenças, e através deste levantamento implementar um *pipeline* computacional capaz de aplicar o método *split-and-rephrase*, relacionando esta aplicação com a língua inglesa e com o português brasileiro.

Objetivos Específicos

- Pesquisar, em publicações acadêmicas dos campos de processamento de linguagem natural e da linguística, características de inteligibilidade pertinentes à tarefa de

simplificação textual para intervenção no âmbito de sentenças;

- Integrar ferramentas, modelos, algoritmos, manuais, *corpora*, métricas automáticas e quaisquer outros recursos do campo de processamento de linguagem natural, capazes de auxiliar nos processos de implementação, treinamento e avaliação do *pipeline* computacional proposto;
- Implementar e especificar o *pipeline* proposto, realizar experimentos com diferentes treinamentos, analisar, relatar e discutir os resultados obtidos através de métricas de avaliação automática.

Método de Pesquisa

Neste projeto realizamos uma pesquisa aplicada, onde, através de princípios e conceitos formulados na literatura, compilamos uma fundamentação teórica para implementação, experimentação e avaliação dos resultados do *pipeline* proposto.

Organização do Trabalho

A organização deste documento está como segue: o Capítulo 1 aborda os fundamentos teóricos para a tarefa de simplificação automática de sentenças; o Capítulo 2 apresenta alguns dos principais trabalhos acadêmicos recentes para a tarefa, que auxiliaram no embasamento para implementação e avaliação do *pipeline* proposto; o Capítulo 3 especifica os detalhes da metodologia, os dados e as métricas adotadas para implementação e avaliação do *pipeline* proposto; o Capítulo 4 apresenta os diferentes experimentos realizados, bem como os resultados e a discussão sobre eles; e o Capítulo 5 traz a conclusão com as considerações finais e trabalhos futuros.

1

Fundamentos Teóricos

Simplificação automática de texto é uma tarefa frequentemente tratada no âmbito de sentenças (MARTIN et al., 2018). Uma sentença é uma unidade importante que traz, na maioria das vezes, informações suficientes para inferência e análise de sua complexidade (LEAL et al., 2019). Neste capítulo, reunimos fundamentos teóricos para uma breve contextualização sobre a tarefa de simplificação de sentenças e a sua relação com os estudos de inteligibilidade de textos.

1.1 Simplificação de Texto no Âmbito de Sentenças

Aplicações para simplificação de sentenças objetivam simplificar o vocabulário e/ou a estrutura de sentenças complexas, tornando-as mais fáceis de serem interpretadas por leitores humanos ou por outras aplicações de processamento de linguagem natural (VU et al., 2018). Os dois tipos de simplificação dominantes na literatura, são: a simplificação lexical e a simplificação sintática (SCARTON; SPECIA, 2018). A simplificação lexical tem a função de identificar e substituir palavras ou expressões para sinônimos mais comuns ou palavras mais frequentes do vocabulário (HARTMANN; PAETZOLD; ALUÍSIO, 2018; SHARDLOW, 2014). Já a simplificação sintática, deve atuar identificando complexidades gramaticais em sentenças, reescrevendo-as em estruturas mais simples (ŠTAJNER; GLAVA, 2017; SCARTON; PAETZOLD; SPECIA, 2019).

Em atendimento à estes dois tipos de simplificação, podemos reconhecer como operações padrão: o particionamento (*split*), a exclusão, a reordenação, e a substituição dos elementos das sentenças complexas (ZHU; BERNHARD; GUREVYCH, 2010). A operação de particionamento visa quebrar uma sentença longa em várias sentenças mais curtas. A operação de exclusão remove partes menos relevantes de uma sentença para torná-la mais concisa. A operação de reordenação visa alternar a ordem das sentenças particionadas ou

das palavras de uma mesma sentença. Finalmente, a operação de substituição troca palavras ou expressões difíceis por sinônimos ou termos mais simples. Tais transformações lexicais e sintáticas nestas sentenças, aliadas à necessidade da preservação da gramaticalidade e de uma semântica original adequada, é um problema desafiador em processamento de linguagem natural, ainda longe de ser resolvido (GARBACEA et al., 2020; ALVA-MANCHEGO; SCARTON; SPECIA, 2020).

Em particular sobre a simplificação lexical, muitas das abordagens tradicionais se baseiam no uso de dicionários, *thesaurus* ou léxico-semânticos como o WordNet (MILLER, 1995). Com o advento dos vetores de representação de palavras (*word embeddings*) (MIKOLOV et al., 2013; PENNINGTON; SOCHER; MANNING, 2014; BOJANOWSKI et al., 2016), novos ganhos foram observados para a tarefa, pela capacidade de associação matemática entre palavras em espaços vetoriais de baixa dimensão, com a hipótese de que palavras que aparecem em contextos similares devem ter semânticas similares e, portanto, representações próximas (TISSIER; GRAVIER; HABRARD, 2017).

Já sobre a simplificação sintática, algumas das abordagens mais clássicas são baseadas em regras artesanais fixas, geralmente construídas para atender questões de mudança na sintaxe, por exemplo para separar sentenças coordenadas e/ou subordinadas em várias orações, ou modelar transformações de voz passiva para voz ativa (NARAYAN; GARDENT, 2016). Entretanto, embora tenham contribuído historicamente para a evolução da tarefa, a maior parte destas abordagens não se mostra flexível o bastante para executar padrões de reescrita mais refinados e específicos, dada a grande complexidade linguística existente (ALVA-MANCHEGO; SCARTON; SPECIA, 2020).

Considerando a simplificação de sentenças como um processo de 'tradução' entre sentenças complexas e simples, esta tarefa tem sido tipicamente tratada como uma variante monolíngue da tradução automática (MARTIN et al., 2020). Neste contexto, visto que o processo pode demandar múltiplas operações simultâneas em uma mesma sentença, modelos orientados por dados (*data-driven*) tem buscado aprendizado automático a partir do treinamento sobre *corpora* compostos por numerosos exemplos de alinhamentos entre sentenças complexas e simples (ZHU; BERNHARD; GUREVYCH, 2010; SCARTON; PAETZOLD; SPECIA, 2018).

Entre as diversas abordagens *data-driven* existentes, temos mais recentemente os modelos apelidados como *sequence-to-sequence* (SUTSKEVER; VINYALS; LE, 2014). Uma prática comum para a composição destes modelos é a combinação de redes neurais recorrentes com variantes projetadas para memorizar sequências a longo prazo no processo de treinamento (HOCHREITER; SCHMIDHUBER, 1997; CHO et al., 2014). Especialmente pela capacidade de atuar com predição estruturada, este tipo de modelo é estratégico para tarefas de geração de linguagem natural, como é o caso da simplificação de sentenças (WANG et al., 2016; NISIOI et al., 2017).

1.2 Avaliação de Inteligibilidade de Textos

Uma área de pesquisa multidisciplinar, normalmente relacionada com a tarefa de simplificação de sentenças, é a de avaliação de inteligibilidade de textos. O nível de inteligibilidade de um texto pode estar relacionado com a facilidade de leitura proporcionada pela escolha de um dado conteúdo, estilo, organização, estrutura sintática ou vocabulário simples que atenda às habilidades e a motivação dos leitores (LEAL et al., 2020). Esta é uma área que tem longo histórico de influência nas áreas da educação e da psicolinguística (ŠTAJNER; NISIOI; HULPUș, 2020), sendo tradicionalmente aplicada para medir a complexidade de livros escolares, manuais técnicos, documentos públicos, legendas de televisão para deficientes auditivos, entre diversos outros fins (MARTINS et al., 1996).

Segundo DuBay (2004), ainda na década de 1980, calculava-se a existência de cerca de 200 fórmulas propostas para medir a inteligibilidade de textos na língua inglesa, entre as quais se destacam as de Senter e Smith (1967), Coleman e Liau (1975) e Kincaid et al. (1975). Porém, uma limitação conhecida por parte destas fórmulas é o fato de calcularem índices sobre características superficiais do texto, não conseguindo capturar a coesão e nem avaliar mais profundamente as razões e correlações de fatores que tornam um dado texto mais difícil de ser compreendido (SCARTON; ALUÍSIO, 2010).

No sentido de capturar a coesão e a coerência dos textos, estratégias mais avançadas buscam calcular índices de acordo com vários níveis de análise linguística, como o léxico, o sintático, o discursivo e o conceitual (GRAESSER et al., 2004; SCARTON; ALUÍSIO, 2010). Estudos mais recentes, buscam ainda correlacionar a complexidade do texto com métricas baseadas no rastreamento ocular de leitores humanos, aliando-as a outros atributos linguísticos e psicolinguísticos nestes experimentos (LEAL et al., 2020).

Do ponto de vista do modelo de complexidade conceitual proposto por Kintsch e Dijk (1978), para que se obtenha a coerência e a plena compreensão do texto, é necessário que haja, por parte do leitor, o conhecimento prévio tanto de suas proposições e conceitos individuais quanto as suas interligações. Sendo assim, a dificuldade de um texto pode ser vista como a quantidade de lacunas na coerência do texto e o esforço exigido pelo leitor para repará-las por meio de inferências (ŠTAJNER; NISIOI; HULPUș, 2020).

Embora estudos evidenciem que os conectores lógicos, denominados conjunções ou marcadores discursivos, desempenham papel estratégico no estabelecimento da coerência textual, estes geralmente ocasionam aumento no tempo de leitura quando presentes em sentenças longas, demandando mais esforço para o entendimento apropriado do texto (REBELLO et al., 2019). Já por outro lado, sentenças mais curtas favorecem a manutenção dos sentidos na memória, sendo este um dos parâmetros sintáticos capazes de afetar a inteligibilidade de um texto (REBELLO et al., 2019).

Por fim, é preciso considerar ainda que as características de complexidade de um texto podem ter pesos diferentes de acordo com um dado público-alvo (SIDDHARTHAN, 2014). Por exemplo, alguns pacientes afásicos apresentam dificuldades para ler sentenças com alta carga cognitiva, como sentenças longas com estruturas sintáticas complexas, enquanto estudantes de uma segunda língua podem não compreender um texto apenas pela imposição de um vocabulário raro ou específico (MARTIN et al., 2020). Sendo assim, em um olhar mais amplo, diferentes objetivos podem demandar diferentes estratégias de simplificação textual (REBELLO et al., 2019).

2

Trabalhos Relacionados

Neste capítulo reunimos alguns dos principais trabalhos acadêmicos recentes relacionados à tarefa de simplificação de sentenças em processamento de linguagem natural.

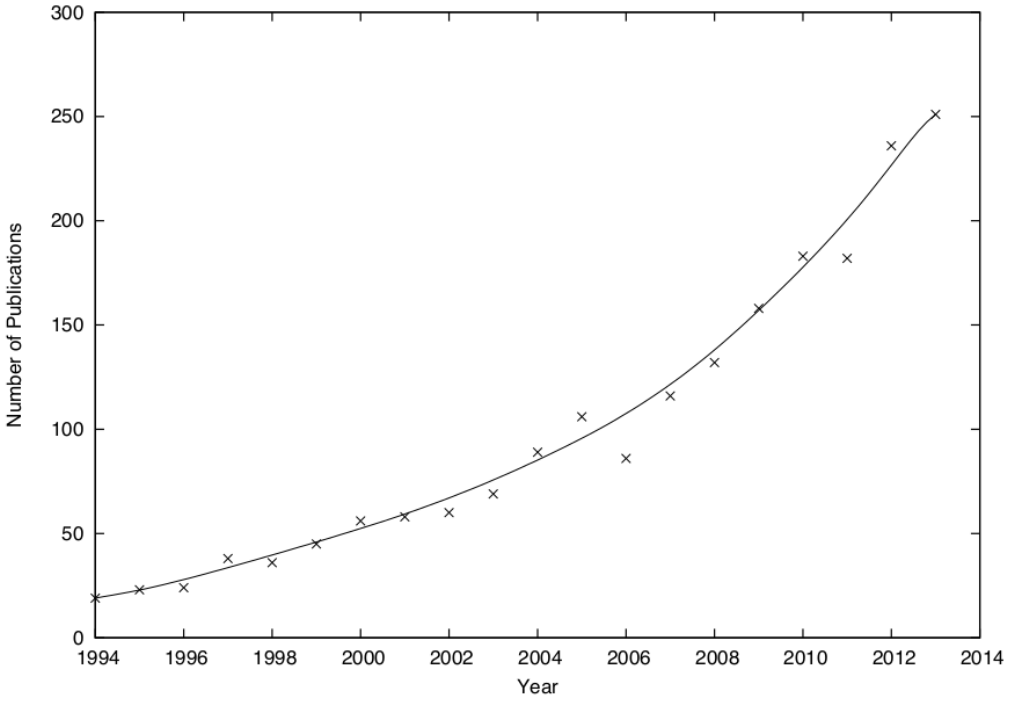
Efetuamos buscas em bases de dados *online* distintas e, conforme esperado, constatamos uma prevalência considerável de aplicações de simplificação na língua inglesa. Ainda assim, é possível encontrar trabalhos diversificados aplicados em línguas advindas do *latim* como, espanhol, francês, italiano e português. É necessário mencionar ainda, a existência de trabalhos direcionados para outras línguas, tais como o japonês, o chinês, o árabe, entre outras. Este último grupo, no entanto, foi desconsiderado nesta revisão por questões de inconformidade, dado que estas línguas possuem alfabetos e estruturas sintáticas completamente diferentes daquilo que se objetiva aplicar neste estudo.

Os trabalhos foram divididos entre: (1) *surveys* sobre a tarefa; (2) algumas das principais aplicações recentes na língua inglesa, contemplando uma seção específica sobre *split-and-rephrase*; e (3) contribuições relacionadas ao português brasileiro.

2.1 *Surveys*

No *survey* de Shardlow (2014), as aplicações de simplificação automática de textos são divididas entre os seguintes tópicos: simplificação lexical; simplificação sintática; simplificação por geração de explicações (*explanation generation*); simplificação baseada em modelos de tradução automática estatística (*statistical machine translation*); e aplicações em línguas além do inglês (*non-English approaches*). Nesta pesquisa, o autor descreve os desafios enfrentados pela tarefa de simplificação de texto, trazendo um pouco de sua história e evolução. Ele ressalta que esta é uma atividade não-trivial, e que já na época de sua publicação (2014) vinha registrando crescimento contínuo no número de publicações, conforme a representação da Figura 2. A pesquisa sintetiza recursos, sistemas e técnicas inerentes à tarefa e trabalhos relacionados em várias línguas até o ano de sua publicação.

Figura 2 – Número de publicações retornadas pelo Google Scholar, na busca pela *String*: "Text Simplification" OR "Lexical Simplification" OR "Syntactic Simplification". Fonte: Shardlow (2014, p. 1).



No *survey* de Siddharthan (2014), o autor também compila um amplo conjunto de trabalhos sobre simplificação de textos até o ano de sua publicação, 2014. Exemplos de sentenças simplificadas e características linguísticas são analisadas, bem como modelos computacionais de simplificação, evidenciando lacunas e sugerindo novas direções para o avanço da tarefa. São destacados ainda, estudos comportamentais que avaliam a necessidade de adaptação desta simplificação à diferentes contextos e categorias de leitores. Alguns exemplos de linguagens citadas, são: as voltadas ao maternalês (infantilizada para bebês e crianças), as direcionadas às populações não-nativas, aos indivíduos com deficiência auditiva ou transtornos como afasia e dislexia, entre outros públicos. O autor também pontua que, em alguns casos, a simplificação pode envolver redundância para enfatização de pontos-chave do texto com mais clareza e/ou sumarização para exclusão de informações periféricas e desnecessárias.

No *survey* mais recente de Alva-Manchego, Scarton e Specia (2020), são relatados trabalhos antigos e atuais sobre simplificação automática de sentenças com foco em abordagens *data-driven*, que buscam aprendizado por meio de *corpora* alinhados compostos por exemplos de sentenças complexas e simplificadas. Este método para o aprendizado artificial, é o paradigma dominante nos dias atuais (MARTIN et al., 2020). Também são especificados os modelos *baselines*, os principais *corpora* e as métricas de avaliação atuais para simplificação de sentenças. Os autores reforçam que, embora tenha havido

um recente avanço considerável na tarefa, as abordagens atuais ainda não são capazes de executar a simplificação de forma totalmente automática para proveito direto dos usuários finais. Portanto, a pesquisa é construída no sentido de revisar trabalhos recentes e destacar pontos fortes e fracos de cada um, com uma análise crítica e empírica para identificar onde a área pode ser melhorada.

2.2 Aplicações na Língua Inglesa

São diversos os trabalhos relacionados à tarefa de simplificação de sentenças na língua inglesa. Dada esta grande quantidade, aliada ao fato de que dois dos *surveys* mencionados datam do ano de 2014, nos restringimos a apresentar nesta seção apenas algumas das principais contribuições, levando em conta os trabalhos do ano de 2014 em diante.

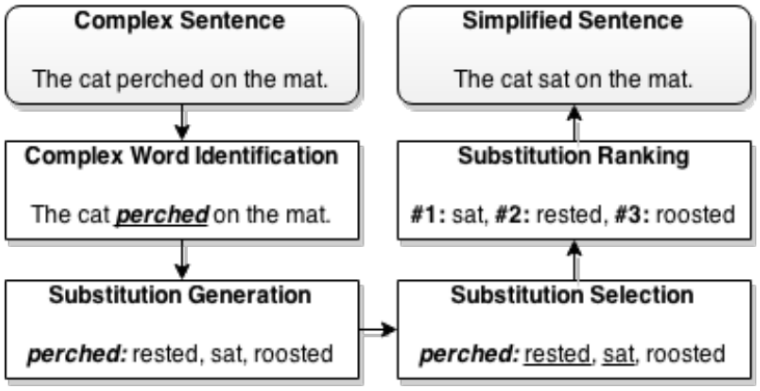
No trabalho proposto por Kauchak et al. (2014), os autores examinam sistematicamente o nível de inteligibilidade de textos médicos por meio de um classificador binário (rótulos 'complexo' e 'simples'), utilizando seis diferentes algoritmos de aprendizado de máquina: árvores de decisão, regressão linear, *random forests*, *Naïve Bayes*, *K–Nearest Neighbors* (KNN) e *Support Vector Machines* (SVM). O experimento utilizou um conjunto de dados de 118.000 sentenças alinhadas e coletadas a partir da enciclopédia Wikipedia no inglês convencional[1] e no inglês simplificado[2] (EW-SEW) (COSTER; KAUCHAK, 2011). Os atributos considerados para esta classificação binária, foram as seguintes características de cada sentença: número de caracteres e palavras; análise do vocabulário pela frequência de palavras/*unigrams* no Google Web Corpus; análises de *part-of-speech* (número de substantivos, adjetivos, verbos, advérbios e outras classes); e conceitos de densidade, especificidade e ambiguidade na linguagem natural.

Já no trabalho de Glavaš e Štajner (2015), os autores apresentam uma abordagem de aprendizado não-supervisionado para simplificação lexical (LightLS), que não se baseia em *corpora* alinhados e nem em léxico-semânticos típicos como o WordNet (MILLER, 1995). Em vez disso, são adotadas representações de vetores de palavras, com o emprego do GloVe (*Global Vectors for Word Representation*) (PENNINGTON; SOCHER; MANNING, 2014) para explorar a semelhança semântica das palavras e assim selecionar alternativas simples para simplificação das palavras complexas. Os autores reportam resultados competitivos aos de sistemas que usam *corpora* simplificados.

Também com foco em simplificação lexical, o artigo de Paetzold e Specia (2015) apresenta a ferramenta LEXenstein como o primeiro *framework* de código aberto para desenvolvimento e *benchmarking* de modelos de simplificação lexical. Similarmente a outros

[1] <https://en.wikipedia.org/wiki/English_Wikipedia>
[2] <https://simple.wikipedia.org/wiki/Main_Page>

Figura 3 – *Pipeline* de simplificação lexical: na sentença de exemplo, a palavra 'perched' é substituída por 'sat'. Fonte: Paetzold e Specia (2015, p. 1).



autores (SHARDLOW, 2014; FERRÉS et al., 2016) e conforme a ilustração da Figura 3, o trabalho conceitua que um *pipeline* ideal para simplificação lexical deve contemplar as seguintes etapas de análise: (1) identificação de palavras complexas nas sentenças; (2) geração de palavras candidatas à substituição; (3) seleção de palavras adequadas pela desambiguação e; (4) escolha da palavra mais simples com base em um *ranking*.

No artigo de Barbu et al. (2015), os autores apresentam a ferramenta Open Book, como uma plataforma para simplificação de textos nos níveis lexical, sintático e semântico. O estudo aborda três módulos direcionados para simplificação especificamente à população autista: simplificação de conceitos complexos mediante a associação de imagens gráficas; detecção de língua para identificação de expressões idiomáticas ou figuras de linguagem presentes em um texto (por exemplo, *'abandonar o barco' = 'desistir'*); e simplificação de texto por meio de sumarização. Neste estudo, são sintetizados alguns dos obstáculos encontrados por autistas no que diz respeito a compreensão de texto, como a dificuldade em utilizar o contexto semântico para resolver ambiguidade de palavras e a dificuldade de utilização de pronomes de referência nas sentenças.

Já o trabalho de Xu et al. (2016), apresenta como principal contribuição o SARI, *System output Against References and against the Input sentence*, primeira métrica automática criada especificamente para avaliação de sistemas de simplificação. Esta métrica compara a sentença de predição contra a sentença de entrada e suas referências de simplificação possíveis, calculando a média aritmética do *F-score* de *n-grams* (onde $n = 1$, 2, 3 e 4) de três diferentes operações: adição (*add*), retenção (*keep*) e exclusão (*del*). Especificamente, ela recompensa operações de adição em que uma palavra gerada na sentença simplificada não consta na sentença original, mas está presente nas sentenças de referência. E também recompensa por palavras mantidas ou excluídas na sentença simplificada de acordo com as referências (ZHAO et al., 2018; MARTIN et al., 2020). A pontuação final é uma soma equilibrada destes parâmetros. A representação deste método

pode ser dada pela Equação 2.1:

$$
\begin{aligned}
&\text{ope} \in [\text{add}, \text{keep}, \text{del}] \\
f_{\text{ope}}(n) &= \frac{2 \times p_{\text{ope}}(n) \times r_{\text{ope}}(n)}{p_{\text{ope}}(n) + r_{\text{ope}}(n)} \\
F_{\text{ope}} &= \frac{1}{k} \sum_{n=[1,\ldots,k]} f_{\text{ope}}(n) \\
SARI &= \frac{F_{\text{add}} + F_{\text{keep}} + F_{\text{del}}}{3}
\end{aligned}
\tag{2.1}
$$

De acordo com os experimentos observados em Xu et al. (2016), o SARI pode se correlacionar com julgamentos humanos sobre a simplicidade das sentenças, e consegue capturar noções de gramaticalidade e preservação de significado. No entanto, pelo foco nos aspectos lexicais, outros trabalhos apontam que a métrica é mais efetiva apenas para avaliação de simplificação lexical, já que o método não endereça questões diversas de transformação de estrutura (ALVA-MANCHEGO et al., 2020), por exemplo, o particionamento de sentenças demandado pelo método *split-and-rephrase*.

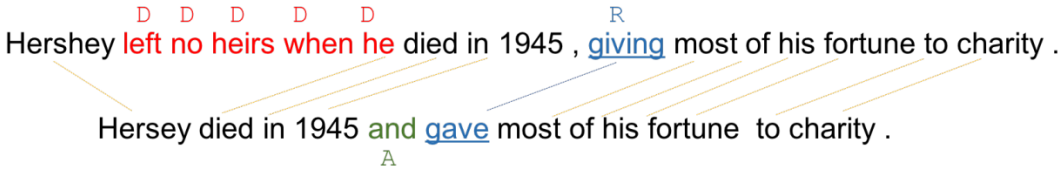
No trabalho experimental de Wang et al. (2016), os autores propõem um estudo para simplificação de sentenças por meio de modelos *sequence-to-sequence* formados por redes neurais recorrentes com arquitetura de memória a longo prazo (LSTM's) (HOCHREITER; SCHMIDHUBER, 1997). Neste trabalho, foram realizados experimentos para validação do aprendizado automático de regras de operação pelos modelos, a partir do treinamento sobre sequências simbólicas alinhadas de exemplo. O objetivo era apurar se o modelo treinado poderia captar conhecimento sobre como inverter, reordenar e substituir os elementos destas sequências de entrada para sequências de saída. Os resultados comprovam a validade do método para a simplificação de sentenças, uma vez que, de maneira análoga, ele é capaz de alterar a estrutura e o conteúdo de sentenças, invertendo, reordenando e substituindo palavras com este aprendizado automático.

De modo similar, o trabalho de referência de Nisioi et al. (2017) apresenta um modelo *sequence-to-sequence* baseado em redes LSTM's (HOCHREITER; SCHMIDHUBER, 1997) reforçado com mecanismos de atenção (LUONG; PHAM; MANNING, 2015) para simplificação de sentenças, conduzindo experimentos com textos genuínos. O modelo, denominado *Neural Text Simplification* (NTS), foi inspirado por avanços observados na tarefa de tradução automática (*Neural Machine Translation*) (NMT), com a adaptação da arquitetura original para a tarefa de simplificação. A abordagem ainda combina um modelo secundário Word2Vec (MIKOLOV et al., 2013) para obtenção de representações de palavras com baixa frequência. Os autores destacam a capacidade do modelo em realizar simultaneamente simplificação lexical e redução de conteúdo, combinando gramaticalidade e preservação do sentido das sentenças.

Em uma abordagem no âmbito de discurso, o artigo de Štajner e Glava (2017) apresenta uma proposta aplicando simplificação lexical, sintática e também simplificação semântica. O projeto é dividido em dois componentes. O primeiro, denominado EBS (*Event-Based Simplification*), elimina informações irrelevantes nas sentenças e mantém apenas as partes factuais, particionando-as em sentenças curtas e sintaticamente simples. O segundo componente, consiste em um módulo de simplificação lexical baseado em métodos de aprendizado não-supervisionado, que faz uso de *word embeddings* no GloVe (PENNINGTON; SOCHER; MANNING, 2014) para substituir termos complexos e pouco frequentes. O *corpus* de teste foi obtido a partir de 200 artigos *online*, sendo a metade deles referente a notícias em geral e a outra metade extraída da enciclopédia Wikipedia.

No trabalho de Alva-Manchego et al. (2017), os autores apresentam uma abordagem de simplificação de sentenças baseada em dados de treinamento rotulados no *corpus* de simplificação Newsela (XU; CALLISON-BURCH; NAPOLES, 2015). As anotações consistem no seguinte conjunto de operações: DELETE (D), REPLACE (R) e MOVE (M) em sentenças do lado complexo; ADD (A) em sentenças do lado simplificado e REWRITE (RW) em ambos os lados. A Figura 4 exibe um exemplo de alinhamento com algumas destas atribuições. Para predizer as operações de simplificação, o estudo treinou uma arquitetura composta por redes neurais LSTM's (HOCHREITER; SCHMIDHUBER, 1997). Os autores pontuam como diferencial deste trabalho a interpretabilidade mais fácil dos tipos de simplificação em virtude deste esquema de rotulação.

Figura 4 – Exemplo de alinhamento rotulado: as palavras suprimidas da sentença original recebem o rótulo (D) DELETE; a palavra que se alinha para uma palavra diferente recebe rótulo (R) REPLACE; e a palavra adicionada na sentença simplificada recebe rótulo (A) ADD. Fonte: Alva-Manchego et al. (2017, p. 6).



Já o trabalho recente de Sulem, Abend e Rappoport (2018b), propõe o SAMSA, *Simplification Automatic evaluation Measure through Semantic Annotation*, como métrica para avaliação de simplicidade estrutural. Nesta métrica, o *score* é maximizado quando a simplificação é um conjunto de sentenças que representam cada uma um evento semântico da sentença original de entrada. Para avaliar a simplificação, a métrica faz uso de um alinhador de palavras e do analisador semântico UCCA, *Universal Cognitive Conceptual Annotation* (ABEND; RAPPOPORT, 2013), decompondo as sentenças de entrada com base na estrutura semântica e as comparando com as saídas (MARTIN et al., 2018). Embora se trate de uma métrica conveniente, alguns trabalhos apontam a margem de aplicação limitada da mesma, visto que recursos linguísticos como analisadores semânticos

em boa qualidade estão disponíveis para apenas algumas línguas (MARTIN et al., 2018). Outros trabalhos ainda pontuam que a implementação original apresenta documentação insuficiente e requer a execução de variados *scripts* (ALVA-MANCHEGO et al., 2019), dificultando a eventual adaptação para outras línguas, como o português.

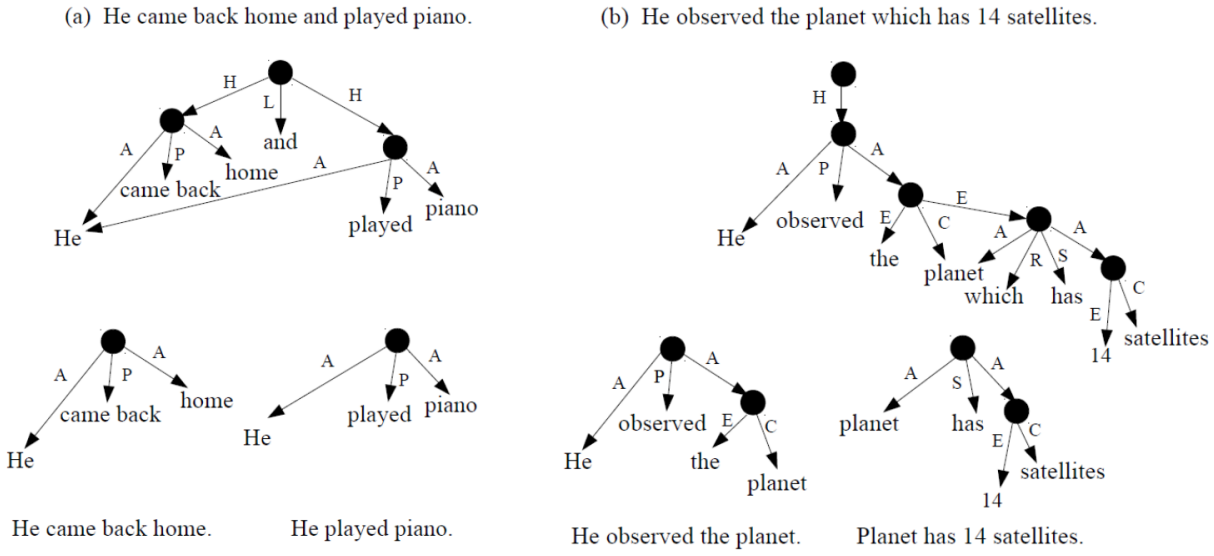
Tratando de simplificação na área médica, o trabalho de Kloehn et al. (2018) apresenta uma abordagem de simplificação léxico-semântica para melhoria no entendimento de textos médicos. Nesta proposta, apelidada de SubSimplify, termos técnicos/complexos são identificados e substituídos por sinônimos ou por uma definição gerada em língua natural. Neste segundo método, dicionários de afixos são consultados para identificação das unidades morfológicas do termo, viabilizando uma geração de explicação automaticamente (por exemplo, *hyperglycemic [hyper-glycemic] = 'extreme' or 'beyond normal' sugar-em-pertaining to*). Além do inglês, a pesquisa exibe também experimentos em espanhol. Para cada língua, foram extraídos de artigos médicos e em geral, 400 termos difíceis balanceados por frequência. Para os termos em inglês, as definições geradas pelo SubSimplify foram comparadas às dos vocabulários *Consumer Health Vocabulary* (CHV), *WordNet Synonyms and Summaries* e *Word Embedding Vector* (WEV), e para os termos em espanhol, foram comparadas às do *WordNet Summaries* e WEV para espanhol. A avaliação dos resultados foi realizada por humanos especialistas nas duas línguas, mediante aplicação de questionários na escala *Likert* (1–4).

No artigo de Sulem, Abend e Rappoport (2018c), os autores propõem um método de particionamento de sentenças, através do algoritmo *Direct Semantic Splitting* (DSS). Esta abordagem, baseada no analisador semântico UCCA, *Universal Cognitive Conceptual Annotation* (ABEND; RAPPOPORT, 2013), suporta a decomposição de sentenças em seus constituintes semânticos. A Figura 5, página 46, ilustra dois exemplos desta decomposição. O estudo busca basicamente aliar o particionamento semântico ao método de simplificação do sistema *Neural Text Simplification* (NISIOI et al., 2017). Além das métricas de avaliação automática BLEU (PAPINENI et al., 2002) (mais detalhes na Seção 3.6.1, página 71) e SARI (XU et al., 2016), avaliações humanas são realizadas por três anotadores nativos em inglês, que pontuam parâmetros de gramaticalidade, preservação de significado, simplicidade e simplicidade estrutural. Ao contrário do método *split-and-rephrase* proposto por Narayan et al. (2017), este trabalho também endereça a simplificação lexical.

2.2.1 Método *Split-and-Rephrase*

Especificamente sobre *split-and-rephrase*, o trabalho de referência de Narayan et al. (2017) traz duas principais contribuições para a concepção do método: (1) um *benchmark* com a disponibilização do *corpus* WebSplit v0.1, formado por alinhamentos entre sentenças únicas complexas e suas respectivas reescritas simplificadas que em conjunto retêm significado

Figura 5 – Exemplos de particionamento por constituintes semânticos. As seguintes categorias são adotadas pelo: Cena Paralela (H), Linker (L), Participante (A), Processo/Estado (P/S), Centro (C), Elaborador (E), Relator (R). Fonte: Sulem, Abend e Rappoport (2018c, p. 4).



equivalente (veja mais detalhes na Seção 3.5, página 63); e (2) o fornecimento de cinco diferentes modelos *baselines* para trazer *insights* sobre como endereçar o método. Entre os cinco modelos, um refere-se a um modelo probabilístico híbrido baseado em tradução automática estatística, e os demais são todos baseados em *sequence-to-sequence* compostos por redes LSTM's (HOCHREITER; SCHMIDHUBER, 1997), com diferentes configurações ou integrações com auxiliares semânticos.

Ao contrário da tarefa convencional de simplificação de sentenças, no método *split-and-rephrase*, a simplificação lexical e a exclusão de informações não são pretendidas. Idealmente, o método visa a preservação do significado original da sentença de entrada apesar da ação de particionamento em sentenças mais curtas. Sendo assim, este pode ser considerado um método particular dentro da tarefa de simplificação de sentenças (ALVA-MANCHEGO; SCARTON; SPECIA, 2020). Além disso, o trabalho de Narayan et al. (2017) esclarece diferenças para outras tarefas de reescrita de sentenças, como:

- Compressão de sentenças (*sentence compression*), que se concentra na exclusão de estruturas e palavras das sentenças, e é mais voltada para o desenvolvimento de sumarização, já que resumos reformulam textos para deixá-los mais curtos;
- Fusão de sentenças (*sentence fusion*), que consiste na combinação de duas ou mais sentenças com informações sobrepostas, preservando informações em comum e excluindo detalhes irrelevantes; e
- Paráfrase de sentenças (*sentence paraphrasing*), onde os sistemas tem o objetivo de

extrair padrões sintáticos de grandes *corpora* e puramente reescrever sentenças com significados semelhantes, reordenando e modificando sintaxes e/ou léxicos.

A Tabela 1 sintetiza as semelhanças e diferenças entre as tarefas e métodos de reescrita citados.

Tabela 1 – Semelhanças entre tarefas e métodos de reescrita de sentenças em relação as suas operações: particionamento (Split); exclusão (Del.); reformulação (Rephr.); e preservação de significado (MPre.), onde Y e N correspondem respectivamente a 'Sim' e 'Não', e o símbolo ?Y denota 'Deveria, mas a maioria não o faz'. Fonte: Narayan et al. (2017, p. 2).

Tarefa/método	**Split**	**Del.**	**Rephr.**	**MPre.**
Compression	N	Y	?Y	N
Fusion	N	Y	Y	?Y
Paraphrasing	N	N	Y	Y
Simplification	Y	Y	Y	N
Split-and-Rephrase	Y	N	Y	Y

No trabalho de Aharoni e Goldberg (2018), os autores propõem modelos *sequence-to-sequence* mais robustos para *split-and-rephrase*, com a integração de mecanismos de cópia (*copy-mechanisms)* (GU et al., 2016), que influenciam os modelos com um viés para copiar *tokens* das sentenças de entrada para as sentenças de saída, considerando que muitas das palavras originais devem ser repetidas na predição. Os autores também pontuam que o *corpus* WebSplit v0.1, proveniente do trabalho de Narayan et al. (2017), pode afetar a medição da capacidade de generalização dos modelos, devido a forma com que os dados estão divididos nos conjuntos de treinamento, validação e teste. Sendo assim, realizam uma nova divisão destes dados e disponibilizam uma versão atualizada do *corpus* para eliminar sobreposições nos dados. No presente trabalho, nos referimos a esta versão como 'WebSplit AG18' (veja mais detalhes na Seção 3.5, página 63).

No trabalho de Botha et al. (2018), os autores apresentam uma nova contribuição com o lançamento público do *corpus* WikiSplit, contendo mais de um milhão de alinhamentos de exemplo entre sentenças complexas e suas reescritas particionadas. A compilação de tal *corpus* é realizada por uma heurística de seleção de sentenças obtidas a partir do histórico de edições da enciclopédia Wikipedia. Para cada sentença complexa C e a sua respectiva reescrita particionada $S = (S_1, S_2)$, são exigidos que C e S_1 tenham um mesmo *trigram* de prefixo, C e S_2 tenham um mesmo *trigram* de sufixo e S_1 e S_2 tenham diferentes *trigrams* de sufixo. Ao contrário das versões anteriores do WebSplit, os autores ressaltam que o WikiSplit oferece sentenças naturalmente expressas com vocabulário diverso, embora muitos exemplos apresentem ruídos (veja mais detalhes na Seção 3.5, página 63). O trabalho traz um impacto positivo para o método *split-and-rephrase*, elevando resultados com modelos treinados neste *corpus*.

No trabalho de Niklaus et al. (2019b), os autores estudam fenômenos sintáticos e apresentam o DisSIM, uma abordagem de particionamento de sentenças recursiva que aplica um conjunto de 35 regras fixas escritas para decomposição de sentenças, orientadas para gerar saídas estruturalmente regulares e oferecer suporte a outras tarefas de PLN, como a extração de informação. Cada sentença curta de saída gerada busca representar uma unidade semântica mínima. O trabalho de Niklaus et al. (2019a), de mesma autoria, apresenta uma linha similar, mas com a integração da língua alemã.

Entre as aplicações atuais para realizar as transformações exigidas pelo particionamento de sentenças, Niklaus et al. (2019b) classifica as abordagens em três tipos:

- Baseadas em regras sintáticas fixas (*syntax-driven rule-based approaches*), que usam de regras fixas manualmente definidas para detectar pontos onde as sentenças devem ser particionadas (SIDDHARTHAN; MANDYA, 2014; FERRÉS et al., 2016);
- Baseadas em análise semântica (*semantic parsing approaches*), que visam decompor sentenças em unidades semânticas mínimas individualmente nas sentenças de saída (NARAYAN; GARDENT, 2014; SULEM; ABEND; RAPPOPORT, 2018c); e
- Baseadas em dados (*data-driven approaches*), onde o ponto de particionamento e as transformações necessárias são aprendidos automaticamente a partir do treinamento em *corpora* alinhados com exemplos de sentenças complexas e simples (NARAYAN et al., 2017; AHARONI; GOLDBERG, 2018).

Por fim, mais recentemente, o trabalho de Niklaus, Freitas e Handschuh (2019) compila o *corpus* MinWikiSplit, com a execução do *framework* DisSIM (NIKLAUS et al., 2019b) sobre os dados WikiSplit (BOTHA et al., 2018). A principal contribuição do *corpus* MinWikiSplit é viabilizar que modelos treinados nele aprendam o particionamento em sentenças mais curtas, potencialmente contrariando abordagens de simplificação mais conservadoras que tendem a realizar um único particionamento ou replicar a sentença de entrada sem nenhuma transformação (veja mais detalhes na Seção 3.5, página 63).

2.3 Aplicações na Língua Portuguesa

Os trabalhos sobre simplificação de sentenças em língua portuguesa são comparativamente escassos. Apresentamos nesta seção as principais aplicações para o português brasileiro em ordem cronológica anual.

O artigo de Aluísio et al. (2008) apresenta o projeto PorSimples como um dos primeiros trabalhos para construção de sistemas de simplificação de texto em português brasileiro, com o intuito de facilitar o acesso à informação para pessoas com baixo nível de

alfabetização. O termo 'letramento' (alfabetização) é referido para designar a capacidade da população brasileira em usar efetivamente as habilidades de leitura e escrita. Entre várias contribuições, este estudo rendeu o primeiro manual de simplificação sintática para o português brasileiro (SPECIA; ALUíSIO; PARDO, 2008). Este manual recomenda como determinados fenômenos sintáticos devem ser simplificados e se baseia em um estudo gramatical sobre seis *corpora* de gêneros de textos diferentes.

No artigo de Watanabe et al. (2009), os autores apresentam o Facilita como parte integrante do projeto PorSimples (ALUÍSIO et al., 2008). Esta ferramenta oferece uma base de auxílio à leitura para facilitar a compreensão de textos no português brasileiro em páginas *web* e aplicativos. Nesta publicação, são tratados os recursos da ferramenta, os aspectos de desenvolvimento e o seu *design*. A tecnologia assistiva viabiliza usuários com baixo nível de alfabetização a compreenderem conteúdos na *web* por meio de um *plug-in* para navegadores, aplicando principalmente tarefas de sumarização e simplificação sintática de textos. O estudo também aborda sobre o INAF (Indicador de Alfabetismo Funcional) e as respectivas classificações para a população brasileira através das quatro seguintes categorias:

1. Analfabeto: para indivíduos que não realizam leituras de palavras e frases;
2. Rudimentar: para indivíduos que somente compreendem informações explícitas em textos curtos, como anúncios ou uma carta breve;
3. Básico: para indivíduos que podem ler e compreender textos de média extensão, encontrando informações mesmo que por inferências;
4. Avançado: para indivíduos que podem compreender textos mais longos, capazes de relacionar suas partes, comparar e interpretar informações, distinguir fatos de opiniões, fazer inferências e sintetizar informações.

Ainda como parte do projeto PorSimples (ALUÍSIO et al., 2008), no trabalho de Scarton et al. (2010), os autores apresentam a ferramenta Simplifica. Esta tecnologia incentiva redatores, em geral, a elaborarem textos simplificados no português brasileiro. A ferramenta apresenta três módulos, sendo um para simplificação lexical, outro para simplificação sintática, e outro para avaliação do nível de complexidade dos textos de entrada, estabelecendo mapeamentos para um de três níveis de alfabetização definidos pelo INAF (rudimentar, básico ou avançado). Este projeto contempla três dicionários: um primeiro contendo vocabulário do cotidiano de jovens; um segundo composto por palavras frequentes extraídas de textos de notícias e jornais; e um terceiro formado por palavras concretas. Além deles, o dicionário Unitex-PB (MUNIZ, 2004) é utilizado para a busca do lema das palavras do texto.

Em uma contribuição dedicada à avaliação de inteligibilidade de textos, o trabalho de Scarton e Aluísio (2010) descreve a adaptação das métricas da ferramenta Coh-Metrix (GRAESSER et al., 2004) para o português brasileiro, apresentando o Coh-Metrix-Port. São demonstrados os recursos adotados no projeto, bem como duas aplicações com esta ferramenta: 1) a avaliação de textos jornalísticos e sua versão para crianças, mostrando as diferenças entre os textos complexos e simples; e 2) a criação de classificadores binários, analisando a influência de diferentes gêneros (jornalístico e de divulgação científica) em seus desempenhos. A precisão do melhor classificador treinado foi de 97%, alcançada com uma implementação em *Support Vector Machines* (SMO) no WEKA (WITTEN et al., 1999). As 41 métricas definidas pelo projeto, estão divididas em:

1. Contagens Básicas: número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças, sílabas por palavras, número de verbos, número de substantivos, número de advérbios, número de adjetivos, número de pronomes, incidência de palavras de conteúdo (substantivos, adjetivos, advérbios e verbos) e incidência de palavras funcionais (artigos, preposições, pronomes, conjunções e interjeições);
2. Constituintes: incidência de sintagmas nominais, modificadores por sintagmas nominais e palavras antes de verbos principais;
3. Frequências: frequência de palavras de conteúdo e mínimo das frequências de palavras de conteúdo;
4. Conectivos: incidência de todos os conectivos, incidência de conectivos aditivos positivos, incidência de conectivos temporais positivos, incidência de conectivos causais positivos, incidência de conectivos lógicos positivos, incidência de conectivos aditivos negativos, incidência de conectivos causais negativos, incidência de conectivos temporais negativos e incidência de conectivos lógicos negativos;
5. Operadores Lógicos: incidência de operadores lógicos, número de 'e', número de 'ou', número de 'se' e número de negações;
6. Pronomes, Tipos e *Tokens*: incidência de pronomes pessoais, pronomes por sintagmas e relação tipo/*token*;
7. Hiperônimos: hiperônimos de verbos;
8. Ambiguidades: ambiguidade de verbos, de substantivos, de adjetivos e de advérbios;
9. *Readability formula*: Índice Flesch Reading Ease (PT-BR) (MARTINS et al., 1996).

No artigo de Woloszyn et al. (2016), os autores apresentam a pesquisa e o desenvolvimento de uma biblioteca de código aberto, para análise de linguagem natural de

textos em português brasileiro, sugerindo métricas para estudos de inteligibilidade de textos. Além da biblioteca proposta, denominada Pylinguistics[3], o trabalho fornece uma análise empírica sobre aspectos de inteligibilidade do jornalismo científico brasileiro e uma comparação com o jornalismo público em geral, abordando características textuais que poderiam tornar o texto científico mais acessível ao público geral. Uma das referências adotadas é a ferramenta Coh-Metrix-Port (SCARTON; ALUÍSIO, 2010).

Já no trabalho de Hartmann, Paetzold e Aluísio (2018), os autores tratam a simplificação de texto tendo como público-alvo as crianças, ressaltando a importância do tema para melhoria do nível de compreensão de leitura e aprendizagem dos alunos na educação escolar. O trabalho compila um *corpus* para o português brasileiro próprio para simplificação lexical, denominado SIMPLEX-PB, e apresenta um *benchmarking* para avaliação de abordagens de simplificação lexical.

No artigo de Leal, Duran e Aluísio (2018), os autores apresentam o *corpus* PorSimplesSent, construindo três cenários diferentes de *corpora* para a tarefa de avaliação de inteligibilidade de sentenças na língua portuguesa. Todos foram compilados e disponibilizados publicamente[4]. De acordo com os autores, o melhor cenário deste trabalho, foi um *corpus* composto por uma compilação de 4.888 sentenças alinhadas. Um modelo de classificação também foi apresentado e considerou 17 características lexicais, sintáticas e psicolinguísticas para identificar o nível de inteligibilidade das sentenças, alcançando uma precisão de 74,2%.

Especificamente sobre *split-and-rephrase* na língua portuguesa, com base na literatura pesquisada, não encontramos nenhuma aplicação ou *corpus* direcionado para o atendimento do método. Desta forma, com a disponibilização do conjunto PorSimplesSent pelo trabalho de Leal, Duran e Aluísio (2018), incorporamos alternativamente no presente trabalho, uma adaptação deste *corpus* para o teste do *pipeline* de simplificação de sentenças no português brasileiro (veja mais detalhes na Seção 3.5, página 63).

[3] <https://github.com/vwoloszyn/pylinguistics>
[4] <http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources>

3

Metodologia

Ao revisarmos os capítulos anteriores sobre fundamentos teóricos e trabalhos relacionados, observamos o emprego de variadas abordagens e estratégias para viabilização da tarefa de simplificação de sentenças. Algumas destas referências inspiraram a metodologia deste trabalho. Neste capítulo descrevemos os métodos, dados e métricas automáticas envolvidos para implementar e avaliar o *pipeline* de simplificação de sentenças proposto.

3.1 Definição para *Split-and-Rephrase*

Conforme já mencionado, o *pipeline* proposto neste trabalho endereça a simplificação de sentenças pelo método *split-and-rephrase*. Uma definição válida para este método é a que segue: dada uma sentença complexa de entrada C, o objetivo do método *split-and-rephrase* é produzir um texto de saída simplificado S consistindo em uma sequência de sentenças $S_1, S_2, \ldots, S_n$, $n \geq 2$, de forma que S preserve o significado de C (AHARONI; GOLDBERG, 2018).

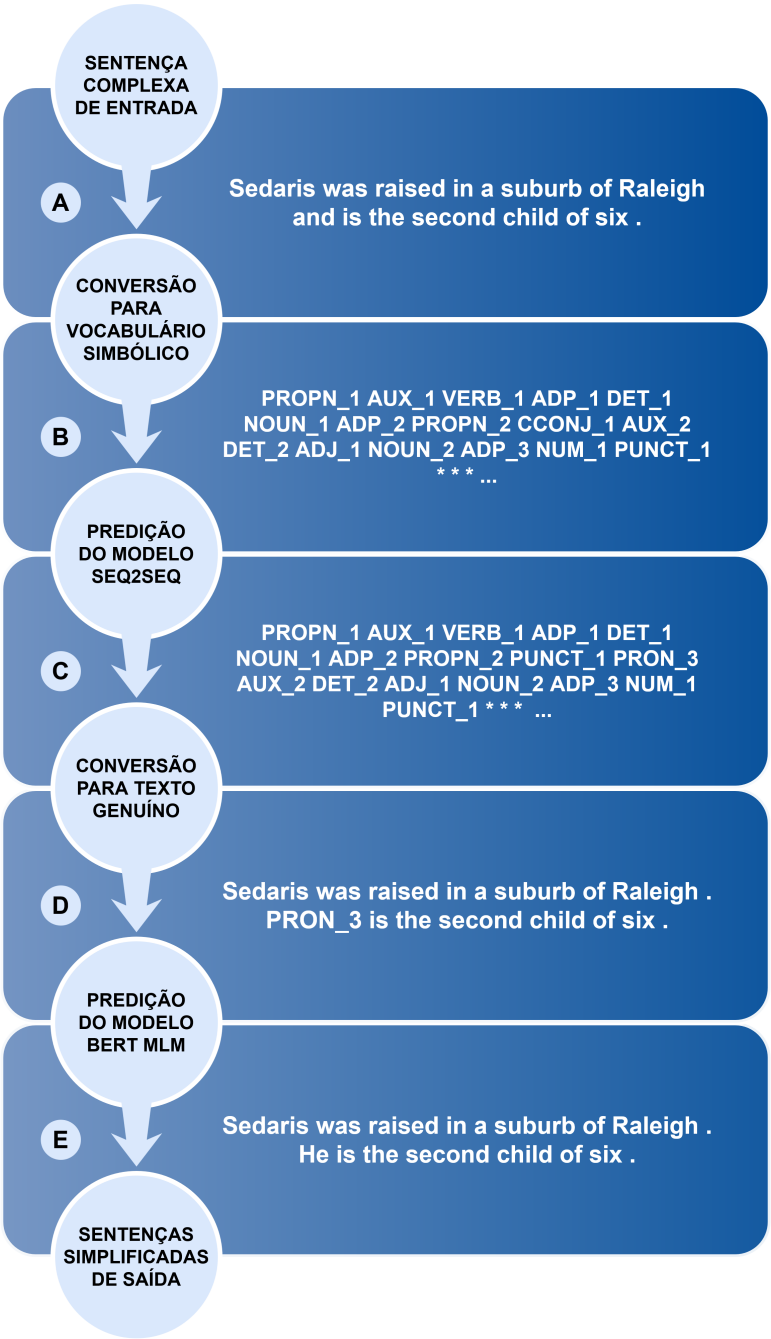
3.2 Apresentação do *Pipeline*

Propomos um *pipeline* composto por dois elementos principais: (1) um modelo de redes neurais *sequence-to-sequence* equipado com mecanismos de atenção, treinado a partir de um vocabulário simbólico customizado; e (2) um modelo de linguagem mascarada (*masked language modeling*), pré-treinado com o BERT MLM (DEVLIN et al., 2019).

Uma visão geral ilustrada do *pipeline* e a representação da integração dos elementos pode ser conferida na Figura 6, página 54. Na configuração proposta, para realizar uma predição no *pipeline*, a sentença complexa de entrada (A) passa por uma etapa de pré-processamento onde a sequência de *tokens* do texto é convertida em uma sequência de

itens do vocabulário simbólico (explicado adiante), que é fornecida ao modelo de redes neurais *sequence-to-sequence* (B). Este modelo, treinado com o auxílio de mecanismos de atenção, prediz uma saída com base no conhecimento aprendido sobre como transformar a sequência de itens recebida (C). Esta sequência de saída é então reconvertida em uma sequência de sentenças em texto genuíno (D), que em seguida é enviada para o modelo de linguagem mascarada pré-treinado BERT MLM, responsável por substituir itens do vocabulário não reconvertidos em texto, produzindo a predição final com as sentenças reescritas/simplificadas (E).

Figura 6 – Ilustração do *pipeline* para *split-and-rephrase*. Fonte: O autor.



3.3 Vocabulário Simbólico

No contexto deste trabalho e em processamento de linguagem natural em geral, um vocabulário pode ser entendido como um conjunto válido de *tokens* exclusivos para o treinamento de um modelo. A seguir detalhamos as técnicas para construção dos vocabulários simbólicos adotados para a posterior alimentação dos modelos. Tal abordagem foi viabilizada por um processo de *part-of-speech tagging*, e inspirada pelo estudo experimental de Wang et al. (2016), que treinou modelos sequenciais a partir de sequências simbólicas alinhadas para validação do aprendizado automático de operações para simplificação de sentenças.

3.3.1 *Part-of-Speech Tagging*

Part-of-speech tagging, ou rotulação morfo-sintática, é o processo de atribuição de rótulos aos elementos de um texto, sendo este um facilitador para diversas tarefas em PLN. Tais rótulos são normalmente correspondentes às propriedades gramaticais semelhantes das palavras, como classes gramaticais, características morfológicas ou dependências sintáticas. Dado que uma única palavra pode ter múltiplos significados, este processo pode, por exemplo, auxiliar na desambiguação de sentidos de uma palavra de acordo com o contexto da sentença, sendo bastante útil para a escolha de um sinônimo adequado no âmbito da simplificação lexical (FERRÉS et al., 2016).

Similarmente ao trabalho experimental de Wang et al. (2016), em vez de treinarmos modelos com extensos vocabulários formados por palavras genuínas, aproveitamos o recurso de um *part-of-speech tagger* para compilarmos vocabulários simbólicos, generalizando itens exclusivamente pela concatenação dos rótulos (*part-of-speech tags*) e suas respectivas recorrências (índices) observadas nas sentenças alinhadas dos conjuntos de treinamento. A Figura 7, página 56, lista o grupo de rótulos universais advindos do Spacy POS Tagger[1] adotados neste trabalho, com base na proposta de anotação sintática *Universal Dependencies*[2].

A Figura 8, página 56, ilustra de maneira mais específica a extração dos itens para compilação dos vocabulários simbólicos. Na implementação customizada adotada, cada *token* do lado complexo do alinhamento (A) é convertido em um símbolo formado por seu respectivo rótulo (*part-of-speech tag*) e um índice equivalente à sua ordem na sentença, constituindo-se assim um vetor sequencial simbólico. Os mesmos símbolos gerados são então atribuídos ao lado simplificado do alinhamento (B) considerando as novas posições de cada *token*. Neste segundo momento, são também permitidas repetições

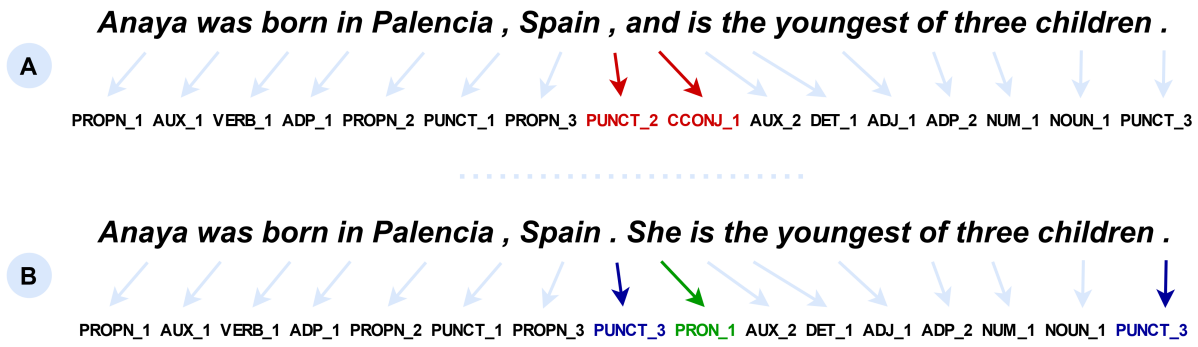
[1] <https://spacy.io/api/tagger/>
[2] <https://universaldependencies.org/docs/u/pos/>

Figura 7 – Rótulos universais em destaque. Fonte: O autor.

ADJ = adjective , **ADP** = adposition , **ADV** = adverb , **AUX** = auxiliary verb , **CCONJ** = coordinating conjunction , **DET** = determiner , **INTJ** = interjection , **NOUN** = noun , **NUM** = numeral , **PART** = particle , **PRON** = pronoun , **PROPN** = proper noun , **PUNCT** = punctuation , **SCONJ** = subordinating conjunction , **SYM** = symbol , **VERB** = verb , **X** = other

(em azul), omissões (em vermelho) e criações de novos símbolos para *tokens* não vistos (em verde), constituindo-se assim um segundo vetor simbólico. Após a realização deste mesmo processo contra todas as sentenças alinhadas do conjunto de treinamento, reunimos um exemplar único de cada símbolo gerado para compilação do vocabulário final. Já os alinhamentos formados pelos vetores simbólicos são posteriormente utilizados para alimentação e treinamento do modelo (veja mais detalhes na Seção 3.4, página 57).

Figura 8 – Exemplos de itens gerados para compilação dos vocabulários simbólicos. Fonte: O autor.



Com este contexto detalhado, a estratégia simbólica adotada é o fator chave para habilitar os modelos treinados a atuarem paralelamente tanto na língua inglesa quanto na língua portuguesa. Em vez de lidarem diretamente com textos/palavras genuínas, eles são capazes de aprender a interpretar itens de um vocabulário customizado comum, já que tais símbolos representam informações gramaticais similares e atuam como atributos compartilhados entre as línguas (STODDEN; KALLMEYER, 2020).

Devido à recorrência de padrões baseados na sintaxe de particionamento em ambas as línguas, o conhecimento específico sobre como transformar as sequências simbólicas pode ser capturado de acordo, graças à natureza de observação sequencial dos modelos de redes neurais *sequence-to-sequence* equipados com mecanismo de atenção (BAHDANAU; CHO; BENGIO, 2015). Desta forma, embora tais modelos tenham sido treinados única e indiretamente a partir de *corpora* na língua inglesa, o conhecimento absorvido estabelece um potencial de aplicação tanto em sentenças do inglês quanto do português brasileiro.

Abordaremos estes pontos com mais detalhes na Seção 4, página 73.

3.4 Modelagem de Sequências

Em processamento de linguagem natural, modelos sequenciais são escolhas canônicas para modelagem de textos. Desde a inspiração nos processos estocásticos por modelos ocultos de Markov (BAUM; PETRIE, 1966), até as abordagens de redes neurais recorrentes (JORDAN, 1986; ELMAN, 1990) e suas variantes arquiteturais de memória a longo prazo, como a *Long Short-Term Memory*, (LSTM) (HOCHREITER; SCHMIDHUBER, 1997) e a *Gated Recurrent Unit*, (GRU) (CHO et al., 2014). Estudos mais recentes, têm endereçado a combinação entre conjuntos destes tipos de redes neurais para viabilizar a construção de modelos apelidados como *sequence-to-sequence* (*seq2seq*), também conhecidos como *encoder-decoder*. Nos tópicos a seguir, contextualizamos a evolução destas redes neurais até a especificação adotada neste trabalho.

3.4.1 Redes Neurais Recorrentes

Nos últimos anos, as redes neurais ressurgiram como poderosos modelos de aprendizado de máquina aliados à crescente capacidade de processamento computacional, produzindo resultados do estado da arte em áreas de estudo distintas (GOLDBERG, 2016). Especificamente no campo de processamento de linguagem natural, onde os dados são geralmente estruturados em sequências de tamanhos variados, como palavras (sequências de letras) e sentenças (sequências de palavras), destaca-se em particular a classe de redes neurais recorrentes (JORDAN, 1986; ELMAN, 1990).

Uma rede neural recorrente (RNN) é uma generalização natural de redes neurais *feedforward* adequada para tratar sequências, na qual as conexões entre as unidades internas formam um ciclo para verificar o histórico de entradas anteriores. Para uma sequência de dados $(x_1, \ldots, x_T)$, onde cada $t \in \{1, \ldots, T\}$, o estado oculto h_t de uma rede neural recorrente pode ser atualizado por $h_t = f(h_{t-1}, x_t)$, onde f representa uma função de ativação (WANG et al., 2016). Devido a sua essência de persistência, redes deste tipo podem facilmente mapear sequências de entrada para sequências de saída em forma de predições estruturadas (SUTSKEVER; VINYALS; LE, 2014). Este tipo de rede neural é bastante útil para tarefas *text-to-text*, como a tradução automática, a análise sintática (*parsing*), a sumarização de textos, entre outras aplicações para previsões de séries temporais.

No entanto, uma RNN básica, popularmente conhecida como *vanilla*, tem uma limitação importante, que é o fato do seu treinamento poder ser prejudicado pelo problema

de *vanishing gradient* (WANG et al., 2016). Basicamente, durante o processamento de longas sequências, ela tende a perder uma de suas principais virtudes de aprendizado, que é a capacidade de persistência.

3.4.2 *Long Short-Term Memory*

A LSTM, *Long Short-term Memory*, concebida pelo trabalho de Hochreiter e Schmidhuber (1997), é uma variante arquitetural de redes neurais recorrentes, com capacidade de aprendizado de dependências a longo prazo. Isto é, por conta de grupos de células internas de memória especializadas, esta arquitetura tende a não sofrer com o problema de *vanishing gradient* (SUTSKEVER; VINYALS; LE, 2014).

De maneira similar a uma RNN básica, a arquitetura LSTM é capaz de atualizar seu estado oculto sequencialmente. No entanto, estas atualizações dependem das células de memória, que são reguladas por três tipos diferentes de *gates*: um *forget gate*, responsável por controlar as informações memorizadas que devem ser esquecidas; um *input/update gate*, que decide como atualizar as informações memorizadas; e um *output gate*, responsável por controlar as informações de saída (WANG et al., 2016).

Conforme especificado por Chung et al. (2014), diferentemente de uma unidade recorrente básica, que calcula a soma ponderada de um sinal de entrada e aplica uma função não-linear, cada enésima unidade j de uma LSTM mantém uma dada memória c_t^j no tempo t. A saída h_t^j, ou a ativação de uma unidade LSTM, é então equivalente a:

$$h_t^j = o_t^j \tanh\left(c_t^j\right),$$

onde o_t^j é um *output gate* que modula a exposição do conteúdo da memória. Este *output gate* é calculado por:

$$o_t^j = \sigma\left(W_o \mathrm{x}_t + U_o \mathrm{h}_{t-1} + V_o \mathrm{c}_t\right)^j,$$

onde σ é uma função sigmóide logística, W_o corresponde aos pesos atribuídos para o vetor de entrada x_t, U_o corresponde aos pesos atribuídos para o vetor do estado anterior h_{t-1}, e V_o corresponde a uma matriz diagonal. A célula de memória c_t^j é atualizada esquecendo parcialmente a memória existente e adicionando um novo conteúdo na memória $\tilde{c}_t^j$:

$$c_t^j = f_t^j c_{t-1}^j + i_t^j \tilde{c}_t^j,$$

onde o novo conteúdo da memória é:

$$\tilde{c}_t^j = \tanh\ \left(W_c \mathrm{x}_t + U_c \mathrm{h}_{t-1}\right)^j.$$

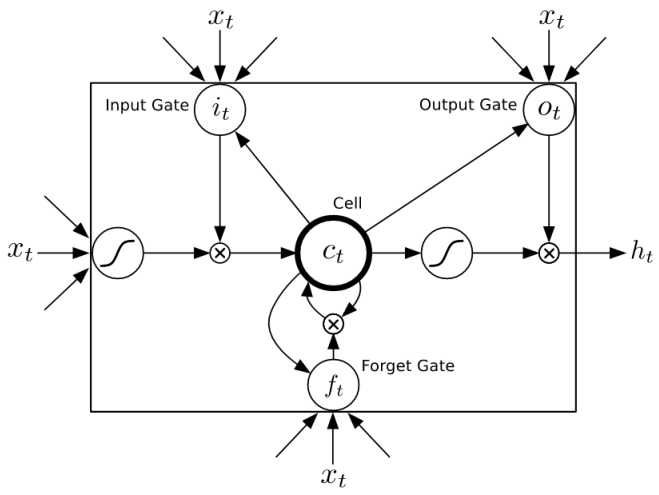
A ação para regular até que ponto uma memória existente deve ser esquecida, é modulada pelo *forget gate* f_t^j. Já o grau em que um novo conteúdo deve ser adicionado à célula de memória é modulado por um *input gate* i_t^j. Estes *gates* são computados por:

$$f_t^j = \sigma \left(W_f \mathrm{x}_t + U_f \mathrm{h}_{t-1} + V_f \mathrm{c}_{t-1} \right)^j ,$$

$$i_t^j = \sigma \left(W_i \mathrm{x}_t + U_i \mathrm{h}_{t-1} + V_i \mathrm{c}_{t-1} \right)^j ,$$

onde V_f e V_i são matrizes diagonais. A Figura 9 ilustra esta formulação.

Figura 9 – Célula de memória da variante arquitetural LSTM. Fonte: Graves (2013, p. 5).



Com o fluxo de informações cuidadosamente regulado pelos *gates*, esta variante arquitetural apresenta boa capacidade de persistência para análise de longas sequências, e promoveu avanço considerável para o aprendizado automático em aplicações, especialmente as que demandam predições estruturadas (BOSCO; PILATO; SCHICCHI, 2018).

3.4.3 *Gated Recurrent Unit*

A GRU, *Gated Recurrent Unit*, consolidada pelo trabalho de Cho et al. (2014), pode ser vista como uma simplificação da variante LSTM, e também foi proposta para eliminar o problema de *vanishing gradient*. Similarmente a uma unidade LSTM, a GRU possui *gates* que modulam o fluxo de informações dentro da unidade, porém, não possui células de memória separadas (CHUNG et al., 2014).

A ativação h_t^j da GRU no tempo t, é uma interpolação linear entre a ativação anterior h_{t-1}^j e a ativação candidata $\tilde{h}_t^j$:

$$h_t^j = (1 - z_t^j) h_{t-1}^j + z_t^j \tilde{h}_t^j,$$

onde um *update gate* z_t^j decide o quanto a unidade atualiza sua ativação ou conteúdo. Este *update gate* é computado por:

$$z_t^j = \sigma \left(W_z \mathrm{x}_t + U_z \mathrm{h}_{t-1} \right)^j .$$

O procedimento para obtenção de uma soma linear entre o estado existente e o estado recém-calculado é semelhante ao de uma unidade LSTM. Entretanto, a GRU não possui mecanismos para controlar o grau a que seu estado é exposto, e sim expõe o seu estado todo a cada tempo. A ativação candidata $\tilde{h}_t^j$ é computada de maneira similar a uma unidade recorrente tradicional:

$$\tilde{h}_t^j = \tanh \left(W \mathrm{x}_t + U(\mathrm{r}_t \odot \mathrm{h}_{t-1}) \right)^j ,$$

onde r_t é um conjunto de *reset gates* e $\odot$ é uma multiplicação elemento a elemento. Quando desligado (r_t^j próximo a 0), o *reset gate* efetivamente faz a unidade agir como se estivesse lendo o primeiro símbolo de uma sequência de entrada, permitindo que ela esqueça o estado computado anteriormente. Este *reset gate* r_t^j é computado similarmente ao *update gate*:

$$r_t^j = \sigma \left(W_r \mathrm{x}_t + U_r \mathrm{h}_{t-1} \right)^j .$$

Experimentos realizados pelo trabalho de Chung et al. (2014), apontam que as composições GRU superam as composições LSTM em termos de tempo de convergência, bem como em termos de atualização e generalização de parâmetros no treinamento.

3.4.4 *Sequence-to-Sequence*

Os modelos de redes neurais *sequence-to-sequence* (*seq2seq*) (SUTSKEVER; VINYALS; LE, 2014), também denominados *encoder-decoder*, são abordagens largamente aplicadas para processamento de sequências em modo *end-to-end*. Nestas abordagens, um codificador (*encoder*) comprime uma sequência original de elementos de entrada em vetores de estado oculto, que então são analisados pelo decodificador (*decoder*) para emissão de uma determinada sequência de saída. Deste modo, cada elemento gerado nesta sequência de saída pode ser condicionado pelos elementos anteriormente processados (MA; SUN, 2017; BEROV; STANDVOSS, 2018). Neste contexto, estão também envolvidos os mecanismos de atenção (*attention-mechanisms*). Dado que, ao longo do treinamento, nem todos os elementos das sequências de entrada devem contribuir igualitariamente para as sequências de saída, tais mecanismos permitem o estabelecimento de referências em pontos específicos nas sequências originais e a transmissão dessas instâncias para as saídas do decodificador (BAHDANAU; CHO; BENGIO, 2015).

Uma prática de arquitetura comum para a composição das camadas dos codificadores e decodificadores, é a combinação de um conjunto de redes neurais recorrentes,

sendo geralmente adotadas as variantes de memória a longo prazo citadas (LSTM's ou GRU's) (HOCHREITER; SCHMIDHUBER, 1997; CHO et al., 2014).

3.4.4.1 Especificação dos Modelos *Sequence-to-Sequence*

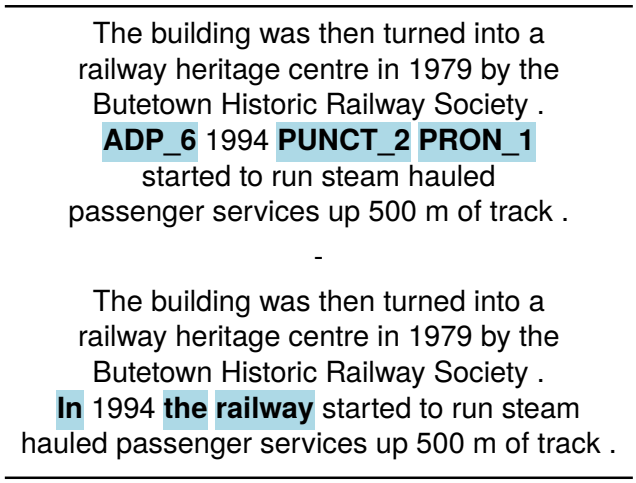
Neste trabalho, adotamos a implementação de modelos de redes neurais recorrentes com a arquitetura *sequence-to-sequence*, composta por *Gated Recurrent Units*, (GRU's), e equipados com mecanismos de atenção (CHO et al., 2014; BAHDANAU; CHO; BENGIO, 2015). Na configuração adotada, a camada de atenção (*attention layer*) foi conectada às camadas (GRU's) do codificador e do decodificador, sendo ambas compostas por 100 unidades. Entre os hiperparâmetros de treinamento, empregamos um *batch size* de tamanho 200 e o número de épocas foi variado entre os experimentos realizados (veja mais detalhes no Capítulo 4, página 73). Utilizamos uma função de perda de entropia cruzada categórica e aplicamos o algoritmo de otimização de Adam (KINGMA; BA, 2015) para atualização dos pesos nas redes iterativamente.

Ao desconsiderar, por um instante, a simbologia imposta pelo vocabulário simbólico, uma concepção para o aprendizado automático da simplificação de sentenças pelo método *split-and-rephrase* nesta abordagem, é a que segue: dado o alinhamento de uma sentença complexa $C = (c_1, c_2, \ldots, c_l)$ com as respectivas sentenças particionadas/simplificadas $S = (s_1, s_2, \ldots, s_{l'})$, onde c_i e s_i são *tokens* do mesmo vocabulário, e l e l' são os comprimentos de cada sequência, a implementação do modelo *sequence-to-sequence* é proposta para modelar a probabilidade condicional $p(S|C)$, sendo o treinamento realizado para a maximização desta probabilidade.

Visto ainda que tal abordagem não exerce o aprendizado diretamente com dados categóricos, utilizamos representações *one-hot encoding* obtidos a partir dos itens do vocabulário simbólico. Neste processo, para cada item w dos vocabulários simbólicos foi atribuído um identificador numérico único w_{ID} que está entre 1 e $|V|$, onde V é o tamanho do conjunto de itens do vocabulário. Cada item w é então representado por um vetor binário V-dimensional preenchido pelo número 0 (zero), exceto na posição w_{ID}, onde o valor equivale a 1 (um) (VAJJALA et al., 2020).

Conforme já mencionado, o aprendizado artificial com a representação dos vocabulários simbólicos torna esta abordagem apta para produzir predições de sequências simbólicas. Consequentemente, uma etapa de reconversão é demandada para o obtenção da representação em texto/palavras genuínas. Nesta etapa, parte dos itens da predição podem não constar na sentença original de entrada para serem reconvertidos, deixando lacunas no texto, conforme ilustração na parte superior da Figura 10, página 62. Para o preenchimento destas lacunas, empregamos modelos de linguagem mascarada, BERT *Masked Language Modeling* (DEVLIN et al., 2019).

Figura 10 – Exemplo de atuação do BERT MLM. Fonte: O autor.



3.4.5 *BERT Masked Language Modeling*

Modelos de linguagem são importantes na linguística computacional graças à capacidade de auxílio em variadas tarefas (KANÉ et al., 2019). Neste contexto, temos o *n-gram*, um tipo de modelo de linguagem referente a uma subsequência de n elementos em meio a uma dada sequência de *tokens*. O *n-gram* envolve o conceito de predição ao considerar que num dado texto, caracteres e palavras apresentam determinadas cadeias que podem ser observadas dentro de um modelo de linguagem. Esta tarefa de predição pode ser entendida como uma tentativa de estimar a função P de probabilidade condicional: $P(w_n|w_1, \ldots, w_{n-1})$ (MANNING; MANNING; SCHÜTZE, 1999), ou seja, uma estimação da probabilidade de uma palavra, w_n, dada a ocorrência das anteriores, $w_1, \ldots, w_{n-1}$.

O BERT, acrônimo para *Bidirectional Encoder Representations from Transformers*, é proposto para treinar representações de linguagem bidirecionais profundas com base na arquitetura Transformer (VASWANI et al., 2017). Em vez de prever a próxima palavra de acordo com a sequência de palavras anteriores, o modelo de linguagem BERT *Masked Language Modeling* (MLM), realiza a predição considerando o contexto formado pelas palavras anteriores e posteriores à uma palavra mascarada (QIANG et al., 2020).

A Figura 10 apresenta a atuação do BERT MLM no *pipeline* deste trabalho. Na parte superior, os itens destacados nas sentenças não foram reconvertidos em texto/palavras genuínas, prejudicando a gramaticalidade da saída. Na estratégia adotada, substituímos esses itens, um por vez, pelo símbolo indicador `<mask>`, e executamos as predições em modelos pré-treinados BERT MLM para o preenchimento das lacunas. Na parte inferior da imagem, vemos a predição final após a execução completa deste processo.

Para a seleção de palavras mais adequadas ao preenchimento das máscaras, adotamos mais especificamente a seguinte estratégia: dada a sequência de *tokens* das sentenças

simplificadas/particionadas de saída $S = (s_1, s_2, \ldots, s_l)$ e a sequência de *tokens* da sentença de entrada complexa $C = (c_1, c_2, \ldots, c_{l'})$, onde l e l' são os comprimentos de cada sequência, para cada item s_i não reconvertido a partir de c_i, mascaramos s_i em S usando o símbolo `<mask>` e alimentamos S no modelo pré-treinado BERT MLM. Este modelo então considera o contexto de S para gerar um *ranking* de classificação com cinco *tokens* candidatos $t_1, t_2, \ldots, t_n$, $n = 5$. Seguindo esta classificação (de 1 para 5), verificamos a existência de cada candidato t_i em C, aceitando o primeiro *token* encontrado como o escolhido para preenchimento da máscara. Se nenhum dos candidatos estiver presente em C, o primeiro candidato t_1 é o escolhido para o preenchimento da máscara. O processo é finalizado quando todos os itens em S estiverem reconvertidos.

Para uso do BERT MLM na língua inglesa, adotamos o modelo pré-treinado na Wikipedia em inglês e no Book Corpus. Para o português brasileiro, empregamos o modelo BERTimbau (*large*), disponibilizado pelo trabalho de Souza, Nogueira e Lotufo (2020).

3.5 Caracterização dos Dados

Tradicionalmente, parte dos *corpora* para aprendizado automático e avaliação da tarefa de simplificação de sentenças na língua inglesa, são construídos a partir de alinhamentos de textos complexos e simples extraídos de artigos da Wikipedia, que possui uma versão no inglês convencional[3] (EW) e outra no inglês simplificado[4] (SEW) (ALVA-MANCHEGO et al., 2020). Na versão simplificada, os colaboradores são incentivados a escreverem textos em linguagem simples, por exemplo, encurtando sentenças ou usando palavras simples do inglês (OGDEN, 1930). Em consideração a estas observações, os trabalhos de Zhu, Bernhard e Gurevych (2010) e de Coster e Kauchak (2011), construíram respectivamente o PWKP e o EW-SEW, dois dos *corpora* alinhados mais conhecidos para a tarefa convencional de simplificação de sentenças.

No entanto, outros trabalhos ressaltam que tais alinhamentos automáticos podem produzir exemplos inconsistentes (XU; CALLISON-BURCH; NAPOLES, 2015). Em alguns casos, os lados simplificados podem aparecer tão complexos quanto os lados originais, sendo poucas as palavras substituídas ou eliminadas, e com muitas sentenças deixadas inalteradas. Por este motivo, no trabalho de Xu, Callison-Burch e Napoles (2015), os autores compilam o Newsela, um *corpus* de licença restrita que contém simplificações diversas produzidas por profissionais que aplicam transformações de reescrita criando 5 níveis distintos de simplificação (COOPER; SHARDLOW, 2020).

Todavia, uma análise documentada no trabalho de Narayan et al. (2017), aponta

[3] <https://en.wikipedia.org/wiki/English_Wikipedia>
[4] <https://simple.wikipedia.org/wiki/Main_Page>

que tanto o último *corpus* quanto os anteriores apresentam baixas incidências de exemplos com sentenças particionadas, sendo inadequados para treinar o aprendizado automático do particionamento em sentenças curtas e sintaticamente simplificadas, conforme a premissa do método *split-and-rephrase* (NIKLAUS; FREITAS; HANDSCHUH, 2019). A seguir, especificamos quatro *corpora* recentes e próprios para o atendimento do método *split-and-rephrase* na língua inglesa e também um *corpus* auxiliar adaptado para o português brasileiro, todos aplicados neste trabalho.

3.5.1 WebSplit v0.1

O WebSplit v0.1 é um *corpus* inaugural para treinamento, validação e teste do método *split-and-rephrase*, criado pelo trabalho de Narayan et al. (2017). É composto por 1.100.166 pares de sentenças escritas a partir de tuplas (*subject* / *property* / *object*) do *corpus* para geração de linguagem natural, WebNLG (GARDENT et al., 2017).

Na composição deste *corpus*, visto que uma única sentença complexa pode ser mapeada para um conjunto de referências estruturalmente simplificadas S_n, o número real de sentenças complexas distintas, $|C|$, está na ordem de apenas 4.438. A quantidade média de sentenças no lado simplificado do alinhamento é de 4,99, com os extremos variando entre 2 e 7 sentenças. O tamanho do vocabulário é de apenas 3.311 *tokens*, dos quais alguns se referem a entidades nomeadas (NARAYAN et al., 2017).

Neste *corpus*, a maior parte dos exemplos traz estruturas uniformes, predominantemente compostas por sequências de orações coordenadas, aumentadas ocasionalmente com uma cláusula relativa ou adverbial (NIKLAUS; FREITAS; HANDSCHUH, 2019). A Figura 11 apresenta três alinhamentos de exemplo deste *corpus*, mapeados de uma única sentença complexa para diferentes referências estruturalmente simplificadas.

Figura 11 – Exemplos do *corpus* WebSplit v0.1. Fonte: O autor.

Max Huiberts owns AZ Alkmaar which has 17023 members .
-
Max Huiberts owns AZ Alkmaar . **<SPLIT>** AZ Alkmaar has 17023 members .
The owner of AZ Alkmaar is Max Huiberts . **<SPLIT>** AZ Alkmaar has 17023 members .
Max Huiberts is the owner of AZ Alkmaar . **<SPLIT>** AZ Alkmaar has 17023 members .

3.5.2 WebSplit AG18

No trabalho de Aharoni e Goldberg (2018), os autores propõem uma nova divisão de dados para o WebSplit v0.1, após verificarem que parte das sentenças dos conjuntos de validação e teste também apareciam no conjunto de treinamento, o que pode afetar a medição da capacidade de generalização dos modelos. Conforme mencionado anteriormente, no presente trabalho, nos referimos a esta versão atualizada do *corpus* pelo nome de WebSplit AG18. Igualmente a versão anterior, uma única entrada complexa pode estar mapeada para diversas referências, com a possibilidade de múltiplos particionamentos no lado simplificado do alinhamento. A Figura 12 apresenta doze alinhamentos deste *corpus*, mapeados a partir de uma única sentença complexa para diferentes referências estruturalmente simplificadas, algumas com mais de um particionamento.

Figura 12 – Exemplos do *corpus* WebSplit AG18. Fonte: O autor.

108 St. Georges Terrace boasts 50 floors and is located in Perth , Australia .
-
108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is in Australia . **<SPLIT>** 108 St Georges Terrace has a floor count of 50 .
The 108 St. Georges Terrace is located in Perth , Australia . **<SPLIT>** 108 St Georges Terrace has a floor count of 50 .
108 St. Georges Terrace is located in Perth , Australia . **<SPLIT>** There are 50 floors at 108 St Georges Terrace .
The 108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is in Australia . **<SPLIT>** 108 St Georges Terrace has a floor count of 50 .
108 St. Georges Terrace is located in Perth , Australia . **<SPLIT>** 108 St Georges Terrace has a floor count of 50 .
108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is located in Australia . **<SPLIT>** 108 St Georges Terrace has a floor count of 50 .
The 108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is located in Australia . **<SPLIT>** 108 St Georges Terrace has a floor count of 50 .
108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is located in Australia . **<SPLIT>** There are 50 floors at 108 St Georges Terrace .
The 108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is in Australia . **<SPLIT>** There are 50 floors at 108 St Georges Terrace .
The 108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is located in Australia . **<SPLIT>** There are 50 floors at 108 St Georges Terrace .
The 108 St. Georges Terrace is located in Perth , Australia . **<SPLIT>** There are 50 floors at 108 St Georges Terrace .
108 St. Georges Terrace is located in Perth . **<SPLIT>** Perth is in Australia . **<SPLIT>** There are 50 floors at 108 St Georges Terrace .

3.5.3 WikiSplit

O WikiSplit é apresentado no trabalho de Botha et al. (2018). Este *corpus* contém mais de um milhão de alinhamentos, divididos em conjuntos de treinamento, *tune*, validação e teste, onde cada sentença complexa é mapeada para uma única referência estruturalmente simplificada, com um único particionamento (*split*). Os conjuntos apresentam um vocabulário rico e variado com sentenças mais naturalmente expressas em relação às versões do WebSplit, que devido ao vocabulário pequeno apresentam exemplos repetitivos. No WikiSplit, o vocabulário total é composto por mais de 633 mil diferentes itens cobrindo mais de 33 milhões de *tokens*.

Conforme admitido pelos autores, o *corpus* apresenta muitos ruídos (BOTHA et al., 2018). Uma amostra de 100 sentenças coletadas no próprio trabalho, indicou que 68% dos exemplos se mostraram corretos, enquanto os 32% restantes continham algum ruído, seja por fatos não suportados (*unsupported facts*) ou informações ausentes (*missing statements*). Em uma inspeção manual, verificamos também diversas sentenças compostas por caracteres especiais, por exemplo, com caracteres dos alfabetos chinês e árabe. Ainda assim, os autores mostram que modelos treinados neste *corpus* produzem resultados notavelmente melhores em relação ao WebSplit. A Figura 13, página 67, traz exemplos de alinhamentos extraídos deste *corpus*.

3.5.4 MinWikiSplit

O *corpus* MinWikiSplit, disponibilizado pelo trabalho de Niklaus, Freitas e Handschuh (2019), é composto por mais de 203 mil sentenças cujas referências simplificadas são formadas por sentenças mais curtas. Cada sentença complexa é mapeada para uma única referência simplificada particionada em duas ou mais sentenças. De acordo com os autores, estas são orações com unidades semânticas mínimas que não podem ser mais decompostas em proposições significativas. Este conjunto foi criado a partir do *corpus* WikiSplit através da execução do DisSIM (NIKLAUS et al., 2019b), um *framework* que aplica 35 tipos de regras fixas para decompor estruturas de sentenças coordenadas e subordinadas.

Considerando que os modelos treinados no WikiSplit podem absorver o aprendizado de particionar sentenças de entrada em exatamente duas sentenças de saída, os autores propõem este *corpus* com a contribuição de habilitar modelos treinados a realizarem mais de um único particionamento por sentença de entrada. Em uma inspeção manual, verificamos alinhamentos com reescritas curtas, porém, com uma parcela excessivamente particionada e iniciada pelas palavras 'This is' em cada sentença. Tais particionamentos também levam parte das referências a ficarem incorretas. A Figura 14, página 68, traz

Figura 13 – Exemplos do *corpus* WikiSplit: exemplo correto (A); exemplo com fatos não suportados (*unsupported facts*) (B); exemplo com informações ausentes (*missing statements*) (C); e exemplo com caracteres especiais (D). Fonte: O autor.

(A) Exemplo correto
Street Rod is the first in a series of two games released for the PC and Commodore 64 in 1989. - Street Rod is the first in a series of two games. **<SPLIT>** It was released for the PC and Commodore 64 in 1989.
(B) Exemplo com fatos não suportados
When the police see Torco's injuries, they send Ace to a clinic to be euthanized, but he escapes and the clinic worker covers up his incompetence. - When the police see Torco's injuries **to his neck, they believe it is a result of Ace biting him**. **<SPLIT>** They send Ace to a clinic to be euthanized, but he escapes and the clinic worker covers up his incompetence.
(C) Exemplo com informações ausentes
Catalina Ponor - A Romanian Olympic gold medal gymnast , **she competed in the 6th competition , where** she failed the " Flying Pillar " in the First Stage . - Catalina Ponor - A Romanian Olympic gold medal gymnast . **<SPLIT>** She failed the " Flying Pillar " in the First Stage .
(D) Exemplo com caracteres especiais
Zeinab Elobeid Yousif (1952 – 19 March 2016) (Arabic : زينب العبيد يوسف) , was the first female Sudanese Aircraft Engineer to be licensed by the Civil Aviation Authority (United Kingdom) . - Zeinab Elobeid Yousif (1952 – 19 March 2016) (Arabic : زينب العبيد يوسف) , was A Sudanese aircraft engineer . **<SPLIT>** She was the first Sudanese female to be licensed by the Civil Aviation Authority (United Kingdom) .

exemplos para ilustração. Ao contrário dos *corpora* anteriores, este *corpus* não apresenta originalmente nenhuma divisão entre conjuntos de treinamento, validação e teste.

3.5.5 PorSimplesSent

O PorSimplesSent é um *corpus* compilado e detalhado no trabalho de Leal, Duran e Aluísio (2018), sendo originalmente proposto para avaliação de inteligibilidade de textos no português brasileiro. Este conjunto foi construído a partir do *corpus* de simplificação de texto

Figura 14 – Exemplos do *corpus* MinWikiSplit: exemplo correto (A); e exemplos incorretos com gramaticalidade e significado prejudicados (B e C). Fonte: O autor.

(A) Exemplo correto
He later fell ill on a trip to Santiago and died on May 13 , 1841 , and was buried in the church at Santiago . - He later fell ill on a trip to Santiago . **<SPLIT>** He later died on May 13 , 1841 . **<SPLIT>** He later was buried in the church at Santiago .
(B) Exemplo incorreto
The trip time to Montevideo by car is approximately 20 minutes , while by bus it is 1 hour and 15 minutes . - The trip time to Montevideo by car is approximately 20 minutes . **<SPLIT>** **It is 1 hour** . **<SPLIT>** **This is by bus** . **<SPLIT>** **It is 15 minutes** . **<SPLIT>** **This is by bus** .
(C) Exemplo incorreto
This requires no understanding of the material being indexed therefore leads to more uniform indexing but this is at the expense of the true meaning being interpreted . - **This requires** . **<SPLIT>** No understanding of the material leads to more uniform indexing . **<SPLIT>** This material is being indexed therefore . **<SPLIT>** This is at the expense of the true meaning . **<SPLIT>** This true meaning is being interpreted .

PorSimples (CASELI et al., 2009) e disponibilizado em três formatos (PorSimplesSent1, PorSimplesSent2, e PorSimplesSent3), cada um apresentando três níveis diferentes de simplificação: *de Original para Natural*; *de Original para Forte*; e *de Natural para Forte*.

Neste trabalho, empregamos o formato do PorSimplesSent1, que repete as sentenças originais do lado complexo do alinhamento, para formação de exemplos alinhados com uma única referência, mas que podem apresentar múltiplos particionamentos. Entre os níveis de simplificação, escolhemos o *de Natural para Forte*, que reflete melhor um *corpus* próprio para o método *split-and-rephrase*, já que os demais níveis também aplicam simplificação lexical e redução de conteúdo. Com este conjunto, selecionamos apenas pares com particionamentos no lado simplificado do alinhamento, extraindo no total 719 pares de sentenças alinhadas para teste no português brasileiro. A Figura 15, página 69, ilustra um exemplo de alinhamento obtido, onde o lado simplificado apresenta dois particionamentos.

3.5.6 Síntese

A Tabela 2, página 69, sintetiza as divisões originais dos conjuntos de treinamento, *tune*, validação e teste de cada *corpus*, considerando o número de alinhamentos pelas sentenças

Figura 15 – Exemplo do *corpus* PorSimplesSent. Fonte: O autor.

Na base militar, Bush se reuniu com o comandante americano das tropas no país árabe, general David Petraeus, e o embaixador dos EUA no Iraque, Ryan Crocker.

-

Na base militar, Bush se reuniu com o comandante americano das tropas no país árabe e o embaixador dos EUA no Iraque. **<SPLIT>** O comandante americano das tropas é o general David Petraeus. **<SPLIT>** O embaixador dos EUA no Iraque é Ryan Crocker.

complexas distintas. Nesta tabela, são denotadas também a quantidade de referências para cada sentença complexa distinta (QR) e a quantidade de particionamentos (*splits*) presentes do lado simplificado do alinhamento (QS), onde $n >= 1$. O MinWikiSplit (*) não apresenta divisões em conjuntos, sendo tratado na íntegra por este trabalho como conjunto de treinamento. Já o PorSimplesSent (**), proposto originalmente para avaliação de inteligibilidade de textos, foi adaptado neste trabalho para um conjunto auxiliar de teste.

Tabela 2 – Número de alinhamentos por sentença complexa distinta. Fonte: O autor.

Corpus	**Treinamento**	**Tune**	**Validação**	**Teste**	**QR**	**QS**
WebSplit v0.1	4.438	-	554	554	n	n
WebSplit AG18	4.506	-	535	503	n	n
WikiSplit	989.944	5.000	5.000	5.000	1	1
MinWikiSplit	203.309 (*)	-	-	-	1	n
PorSimplesSent	-	-	-	719 (**)	1	n

3.6 Métricas de Avaliação

De acordo com Saggion et al. (2015), avaliações de simplificação de texto para leitura humana devem levar em conta as respectivas necessidades de cada população-alvo. Entretanto, como o envolvimento de tais populações nem sempre é possível, uma prática comum é a avaliação mais genérica baseada em pontuações humanas para julgar as sentenças de saída. Já os trabalhos que têm a possibilidade de comparar as sentenças de saída com referências padrão-ouro (*gold standard*), geralmente fazem uso de métricas de avaliação automáticas.

Na literatura de modo geral, métricas automáticas para avaliação de simplificação de sentenças visam se correlacionar com os seguintes parâmetros (MARTIN et al., 2018):

- Gramaticalidade (ou fluência): quão gramaticalmente correta foi a predição?
- Preservação de significado (ou adequação): quão bem o significado da sentença original foi preservado na predição?
- Simplicidade: quão simples foi a predição?

O trabalho de Martin et al. (2018), pontua a existência de um *trade-off* entre os parâmetros de gramaticalidade, preservação de significado e simplicidade, uma vez que a otimização de uma dessas dimensões geralmente leva as outras a resultados mais baixos. Por exemplo, a melhor maneira de garantir a gramaticalidade e a preservação do significado seria deixar a sentença original inalterada, resultando em nenhuma simplificação. Nesta lógica, o desafio reside na combinação adequada destas dimensões.

Outra linha de avaliação pode se dar por premissas de inteligibilidade propostas por índices de *readability formulas*. Entre os exemplos clássicos deste tipo de fórmula, estão o Flesch Reading Ease (FRE) (FLESCH, 1948) e o Flesch–Kincaid Grade Level (FKGL) (KINCAID et al., 1975), que são calculadas pelas combinações lineares do número médio de palavras por sentença e do número médio de sílabas por palavra. Porém, tais abordagens não consideram a gramaticalidade e a preservação do significado nas saídas, e devem ser interpretadas com cautela (ALVA-MANCHEGO et al., 2019).

Uma vez que a tarefa de simplificação de sentenças pode ser considerada uma tarefa de tradução monolíngue, métricas advindas da tradução automática baseadas em *n-grams*, como o BLEU (PAPINENI et al., 2002) e outras variantes, também são largamente empregadas (NISIOI et al., 2017; NARAYAN et al., 2017). O estudo de Martin et al. (2018), aponta que tais métricas se correlacionam com os julgamentos humanos de gramaticalidade e preservação de significado, enquanto a correlação com o parâmetro de simplicidade pode estar mais relacionada com cálculos básicos inspirados nas fórmulas de inteligibilidade citadas, como médias de palavras e sílabas.

Dado este contexto e seguindo os trabalhos de referência sobre *split-and-rephrase* de Narayan et al. (2017), Aharoni e Goldberg (2018) e Botha et al. (2018), adotamos o BLEU (PAPINENI et al., 2002) para avaliação automática das sentenças de saída, bem como os seguintes cálculos para estimativa de qualidade sobre elas: número médio de sentenças simplificadas de saída por sentenças complexas de entrada (#S/C); e número médio de *tokens* por sentenças simplificadas de saída (#T/S). Também, seguindo Niklaus, Freitas e Handschuh (2019), relatamos a porcentagem de saídas totalmente copiadas das entradas (%SAME). Para o cálculo das métricas utilizamos o pacote padrão de funções provenientes do EASSE (ALVA-MANCHEGO et al., 2019). Detalhamos a seguir cada uma delas.

3.6.1 BLEU

BLEU (PAPINENI et al., 2002), acrônimo para *Bilingual Evaluation Understudy*, é uma métrica automática originalmente proposta para avaliação de tradução automática, que mede sobreposições de sentenças de saída contra múltiplas sentenças de referência, pelo cálculo de *n-grams* correspondentes entre as partes (SAGGION et al., 2015; SURYA et al., 2019). Primeiramente, é calculada a média de precisões dos *n-grams* modificados, p_n, usando *n-grams* até o comprimento N (tipicamente $N = 4$) e o somatório de pesos positivos w_n. Em seguida, dado c o comprimento de uma saída candidata e r o comprimento médio das referências, a pontuação é obtida com a multiplicação por um fator de penalidade de brevidade BP, que penaliza saídas menores que as referências, conforme as Equações 3.1 e 3.2 seguintes:

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{3.1}$$

$$\mathrm{BLEU} = \mathrm{BP} \cdot exp\left(\sum_{n=1}^{N} w_n \, log \, p_n\right) \tag{3.2}$$

Os valores obtidos podem ser apresentados com variações entre 0 e 1, ou alternativamente entre 0 e 100, onde quanto mais alta a pontuação, melhor o resultado observado (ALVA-MANCHEGO; SCARTON; SPECIA, 2020).

Seguindo a linha convencional dos trabalhos de referência para o método *split-and-rephrase*, adotamos o BLEU no âmbito de sentenças e desconsideramos o uso de quaisquer versões variantes adaptadas por métodos de suavização (*smoothing*).

3.6.2 Estimativas de Qualidade

De maneira similar às fórmulas de inteligibilidade mencionadas, os cálculos para estimativa de qualidade empregados, usam como base as características superficiais do texto (NARAYAN et al., 2017; NIKLAUS; FREITAS; HANDSCHUH, 2019). Uma vantagem deste tipo de abordagem é a desnecessidade de referências de simplificação, possibilitando assim a aplicação para diferentes línguas, como neste trabalho a inglesa e a portuguesa.

O atributo #S/C é proposto para computar o número médio de sentenças simplificadas de saída, onde x representa o número total de sentenças de saída e y o número total de sentenças de entrada, visando medir a capacidade do *pipeline* em particionar sentenças complexas em várias reescritas simplificadas: $\#S/C = x \div y$.

Já o atributo #T/S, é proposto para computar o número médio de *tokens* de saída, onde z representa o número total de *tokens* e x o número total de sentenças de saída, visando medir a capacidade do *pipeline* em produzir reescritas mais curtas: $\#T/S = z \div x$.

Por fim, o atributo %SAME é proposto para especificar o percentual de sentenças de saída que foram totalmente copiadas das sentenças de entrada, com o intuito de analisar o nível de conservadorismo do *pipeline*.

4

Resultados

Neste capítulo, detalhamos os diferentes experimentos realizados com base na metodologia apresentada. Em seguida, relatamos e discutimos os resultados obtidos com o *pipeline* proposto, através de uma inspeção manual para analisar aspectos das simplificações nas línguas inglesa e portuguesa, bem como as limitações encontradas.

4.1 Experimentos

Realizamos três diferentes experimentos envolvendo três treinamentos de modelos de redes neurais *sequence-to-sequence*. Cada treinamento foi realizado a partir de seleções de dados distintas dos conjuntos WikiSplit e MinWikiSplit. Conforme já citado, estes são os dois *corpora* que apresentam vocabulário e sintaxe mais ricos e um grande número de alinhamentos, sendo ideais para os propósitos de treinamento. O *Experimento 1* (E1) diz respeito a uma validação preliminar da proposta do *pipeline*, publicada no evento KDMILE - *VIII Symposium on Knowledge Discovery, Mining and Learning*, com o artigo *Experimenting Sentence Split-and-Rephrase Using Part-of-Speech Labels* (Anexo A). Já os demais experimentos, *Experimento 2* (E2) e *Experimento 3* (E3), apresentam configurações mais avançadas, onde o *pipeline* integral é adotado contra conjuntos completos de teste dos *corpora* do estado da arte WebSplit v0.1, WebSplit AG18, WikiSplit, e também contra a adaptação a partir do PorSimplesSent. Estes dois últimos experimentos foram publicados no evento RANLP 2021 - *Recent Advances in Natural Language Processing*, com o artigo *Split-and-Rephrase in a Cross-Lingual Manner: a Complete Pipeline* (Anexo B).

4.1.1 Experimento 1

Nesta versão preliminar proposta, utilizamos o conjunto de treinamento do *corpus* WikiSplit, onde empiricamente realizamos dois recortes para seleção dos dados de treinamento do

modelo *sequence-to-sequence*, reservando uma parcela para sua posterior validação. O primeiro recorte foi realizado para selecionar alinhamentos com comprimento de no mínimo 15 e no máximo 30 *tokens*, considerando cada lado do alinhamento. Selecionamos sentenças com tokenizações equivalentes entre os tokenizadores do NLTK[1] e do Spacy[2] formadas unicamente por caracteres alfanuméricos, vírgulas, pontos e espaços em branco na tentativa de minimizar ruídos nos dados. Este primeiro recorte selecionou 171.133 alinhamentos. O segundo recorte foi realizado para selecionar especificamente alinhamentos onde o lado simplificado possuía um pronome como primeiro elemento da segunda sentença (por exemplo: *'Seth Stewart is from Stow, Ohio. He grew up with his mother and father.'*). Neste experimento, evidenciamos a importância desta classe gramatical para a coesão e encadeamento lógico das sentenças simplificadas. Este segundo recorte extraiu 63.623 alinhamentos, consolidando a seleção final para participação no experimento.

A partir desta seleção final, recuperamos aleatoriamente 63.000 pares de sentenças para extração do vocabulário simbólico. Foram coletados 139 itens diferentes para formação deste vocabulário e o treinamento do modelo *sequence-to-sequence* foi realizado com os respectivos alinhamentos em até 1.000 épocas. Os 623 pares restantes foram utilizados para a validação básica do comportamento do método *split-and-rephrase*. Por se tratar de uma versão preliminar, o *pipeline* completo com o emprego do BERT MLM não consta na publicação, porém aqui relatamos os resultados também com a integração deste modelo. A Tabela 3 apresenta as pontuações obtidas com base na avaliação automática pelo BLEU juntamente aos cálculos de estimativa de qualidade equivalentes.

Tabela 3 – Resultados obtidos com o *Experimento 1*. Fonte: O autor.

Conjunto	**BLEU Seq2Seq + BERT MLM**	**BLEU Seq2Seq - BERT MLM**	**#S/C**	**#T/S**	**%SAME**
623 sentenças	**78.50**	74.72	2.00	11.95	0

4.1.2 Experimento 2

Neste experimento, adotamos o conjunto de treinamento do *corpus* WikiSplit (BOTHA et al., 2018). Novamente selecionamos alinhamentos com sentenças formadas apenas por caracteres alfanuméricos, vírgulas, pontos e espaços em branco na tentativa de eliminar caracteres estrangeiros e especiais para minimizar ruídos nos dados. O recorte extraiu 485.120 alinhamentos, consolidando o conjunto de treinamento para este experimento. Com a execução da implementação customizada para extração do vocabulário simbólico, obtivemos 247 itens diferentes para treinamento do modelo *sequence-to-sequence*. Este modelo foi projetado para receber até 100 *tokens* de entrada e treinado em 10 épocas. Em

[1] <https://www.nltk.org/>
[2] <https://spacy.io/>

seguida, o integramos ao modelo de linguagem mascarada BERT MLM, pré-treinado nas línguas inglesa e portuguesa. A Tabela 4 exibe os resultados obtidos contra os conjuntos de teste completos dos *corpora* WebSplit v0.1, WebSplit AG18, WikiSplit e no PorSimplesSent.

Tabela 4 – Resultados obtidos com o *Experimento 2*. Fonte: O autor.

Conjunto	**BLEU Seq2Seq + BERT MLM**	**BLEU Seq2Seq - BERT MLM**	**#S/C**	**#T/S**	**%SAME**
WebSplit v0.1	**58.34**	53.47	2.17	12.52	0.014
WebSplit AG18	**60.01**	55.08	2.21	10.70	0.019
WikiSplit	**68.92**	65.59	2.05	20.45	0.071
PorSimplesSent	**65.00**	61.11	2.06	14.95	0

4.1.3 Experimento 3

Neste experimento, adotamos para treinamento o *corpus* MinWikiSplit (NIKLAUS; FREITAS; HANDSCHUH, 2019). Primeiramente estabelecemos um limite para seleção de alinhamentos com comprimento máximo de 100 *tokens* considerando cada lado do alinhamento, já que este *corpus* contém uma porção de sentenças longas que elevariam excessivamente o consumo dos recursos computacionais no processo de treinamento. Este primeiro recorte extraiu 197.496 alinhamentos. Em seguida, repetimos a abordagem anterior selecionando pares de sentenças alinhadas formadas apenas por caracteres alfanuméricos, vírgulas, pontos e espaços em branco, finalmente consolidando um conjunto de treinamento de 122.104 alinhamentos. O vocabulário simbólico obtido pela implementação customizada foi composto por 230 itens diferentes e o treinamento do modelo *sequence-to-sequence* realizado em 40 épocas. Por fim, integramos o modelo treinado ao modelo de linguagem mascarada BERT MLM, pré-treinado nas línguas inglesa e portuguesa. A Tabela 5 exibe os resultados obtidos contra os mesmos conjuntos de teste mencionados.

Tabela 5 – Resultados obtidos com o *Experimento 3*. Fonte: O autor.

Conjunto	**BLEU Seq2Seq + BERT MLM**	**BLEU Seq2Seq - BERT MLM**	**#S/C**	**#T/S**	**%SAME**
WebSplit v0.1	**57.86**	54.17	3.12	10.34	0.043
WebSplit AG18	**58.45**	54.76	3.17	9.03	0.033
WikiSplit	**44.65**	39.48	5.81	11.78	0.011
PorSimplesSent	**49.52**	42.72	4.61	10.07	0

4.2 Avaliação Automática

Ao analisarmos os resultados das Tabelas 3, 4 e 5, verificamos que a melhor pontuação BLEU obtida foi a de 78.50 no *Experimento 1*. Esta pontuação superou os 74.72 obtidos

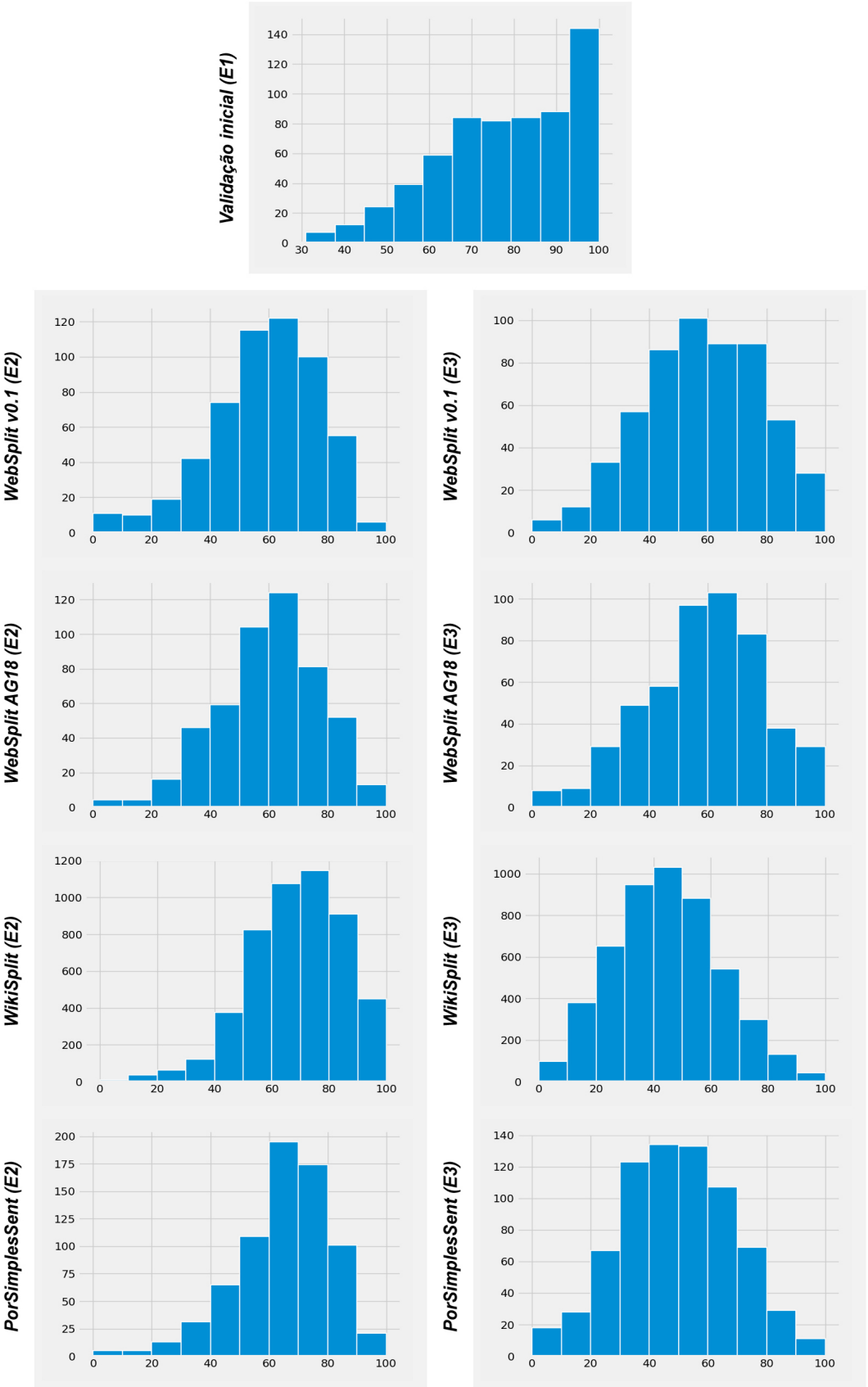
anteriormente sem o emprego do modelo de linguagem mascarada, demonstrando a boa capacidade do BERT MLM em preencher máscaras com os respectivos *tokens* esperados pelas sentenças de referência.

Já analisando os resultados obtidos com o BLEU nos *Experimentos 2 e 3*, verificamos uma pontuação inferior esperada devido ao aumento da complexidade do problema: nestes experimentos, os dois modelos *sequence-to-sequence* foram treinados a partir de alinhamentos com grandes variações de reescrita para fenômenos sintáticos diversificados, ao contrário do *Experimento 1*, que trabalhou apenas com os exemplos baseados em pronomes para aprendizado do *split-and-rephrase*. Além disso, nos *Experimentos 2 e 3*, estabelecemos uma extensão de 100 *tokens* para treinamento, habilitando os modelos a receberem sentenças de entrada com esta longa dimensão, ao contrário do *Experimento 1*, onde limitamos a dimensão para somente 30. Tal elevação no número de *tokens* de entrada, possibilitou que os *Experimentos 2 e 3* fossem realizados contra os conjuntos de teste completos dos *corpora* do estado da arte para o método *split-and-rephrase*.

Entre os *Experimentos 2 e 3*, notamos uma diferença sensível nas estatísticas de estimativa de qualidade, conforme expresso pelas Tabelas 4 e 5. Na Tabela 5, tanto os números mais altos da coluna #S/C, quanto os números mais baixos da coluna #T/S, demonstram que o *Experimento 3* realizou mais particionamentos que o *Experimento 2*, portanto gerando mais sentenças de saída e ao mesmo tempo mais curtas. Isso confirma a hipótese de que o treinamento com os dados do MinWikiSplit potencializa o número de particionamentos em relação aos dados do WikiSplit. Já os números da coluna %SAME, por sua vez, ilustram o baixo conservadorismo do *pipeline* em todos os cenários, mostrando a tendência de interceptar praticamente todas as sentenças de entrada para realizarem as transformações do *split-and-rephrase*, deixando poucas entradas inalteradas.

Podemos analisar pelos histogramas projetados com as pontuações obtidas pelo BLEU, na Figura 16, página 77, que muitas das predições alcançaram correspondências perfeitas contra as referências esperadas, especialmente no *Experimento 1* e no *Experimento 2* contra o conjunto de teste do *corpus* WikiSplit. Visto que as referências deste último conjunto esperavam um único particionamento, as pontuações BLEU do *Experimento 3* foram naturalmente prejudicadas, já que ele demonstrou a tendência de promover vários particionamentos para cada sentença de entrada.

Figura 16 – Histogramas com as pontuações BLEU (horizontal) em relação a quantidade de predições (vertical) para os *Experimentos 1, 2 e 3*. Fonte: O autor.



4.3 Discussão

Para a realização de uma análise mais detalhada, inspecionamos manualmente algumas das predições geradas pelos respectivos experimentos, trazendo exemplos que ajudam a explicar as pontuações BLEU obtidas e também as estatísticas de estimativa de qualidade. As análises estão divididas para a língua inglesa e portuguesa, seguidas das limitações encontradas e da síntese.

4.3.1 Resultados na Língua Inglesa

Experimento 1

Iniciamos a análise com algumas predições na Tabela 6, página 79, todas extraídas do *Experimento 1*.

No *Exemplo A*, observamos um exemplo de predição correta, onde a sentença de entrada foi interceptada pelo particionamento em orações coordenadas com a supressão da vírgula ‘,’ e da conjunção ‘and’, e a colocação do ponto final ‘.’ seguido da palavra ‘She’. Dado que esta predição correspondeu perfeitamente com a referência esperada, foi atribuída a ela a pontuação BLEU máxima (100.00 pontos).

No *Exemplo B*, vemos também um caso de pontuação máxima do BLEU, onde, além do particionamento para uma oração subordinada, o pronome ‘He’ foi devidamente posicionado, embora o mesmo não tenha aparecido na sentença de entrada. Este é um tipo de resultado que mantém a ligação semântica entre os componentes individuais e a relação de coerência entre as sentenças de saída (ZHU; BERNHARD; GUREVYCH, 2010; NIKLAUS et al., 2019b).

Já no *Exemplo C*, vemos outra reescrita correta realizada pelo *pipeline*, mas que teve a pontuação BLEU prejudicada pela falta de mais referências. Embora tenha preservado a gramaticalidade e o significado original da entrada com sucesso, a pontuação atribuída foi de apenas 63.23 pontos, justificando um dos motivos recorrentes de baixas pontuações indevidas, tanto neste quanto nos demais experimentos. Esta atribuição indevida, evidencia que métricas automáticas requerem dados confiáveis para produzirem resultados mais precisos, idealmente com múltiplas referências (MARTIN et al., 2018).

No *Exemplo D*, temos mais um caso de predição correta, mas que foi prejudicado pelo ruído de *unsupported facts*, presente no *corpus* WikiSplit. Pelo fato da referência apresentar as palavras ‘secrete primarily mucus’, esta se torna uma referência incorreta para comparação com a predição, ilustrando outro motivo recorrente da diminuição indevida da pontuação BLEU, que nesta predição alcançou 68.25 pontos.

Tabela 6 – Exemplos de predições em inglês a partir do *Experimento 1*. Fonte: O autor.

Exemplo A	
Entrada	She laid down on the top of the mountain directly above the Crystal , and looked directly up into the beam of light from the three suns .
Referência	She laid down on the top of the mountain directly above the Crystal . She looked directly up into the beam of light from the three suns .
Predição E1 (100.00)	She laid down on the top of the mountain directly above the Crystal . She looked directly up into the beam of light from the three suns .
Exemplo B	
Entrada	Dr. Francia was a creole with an advanced law degree who used only three men in his leading of the country .
Referência	Dr. Francia was a Creole with an advanced law degree . He used only three men in his leading of the country .
Predição E1 (100.00)	Dr. Francia was a Creole with an advanced law degree . He used only three men in his leading of the country .
Exemplo C	
Entrada	Alexander Frey is an American symphony orchestra conductor , virtuoso organist , pianist and harpsichordist .
Referência	Alexander Frey is an American symphony orchestra conductor . He is also known as a virtuoso organist , pianist and harpsichordist .
Predição E1 (63.23)	Alexander Frey is an American symphony orchestra conductor . Frey is virtuoso organist, pianist and harpsichordist .
Exemplo D	
Entrada	The cardiac glands of the stomach are few in number and occur close to the cardiac orifice where the esophagus joins the stomach .
Referência	The cardiac glands of the stomach secrete primarily mucus . They are few in number and occur close to the cardiac orifice where the esophagus joins the stomach .
Predição E1 (68.25)	The cardiac glands of the stomach are few in number . They occur close to the cardiac orifice where the esophagus joins the stomach .

Experimentos 2 e 3

Na Tabela 7, página 80, exemplificamos algumas das predições obtidas nos *Experimentos 2 e 3*.

No *Exemplo A*, ilustramos duas predições equivalentes geradas pelos diferentes experimentos com o particionamento adequado para a sentença com subordinação. Ambas alcançaram a pontuação BLEU máxima, já que corresponderam exatamente a uma das duas referências esperadas do *corpus*.

No *Exemplo B*, ambos os experimentos geraram predições corretas em termos de gramaticalidade e preservação de significado ao particionar a subordinação. No entanto, apenas a *Predição E3* alcançou a pontuação BLEU máxima, pois foi a única que correspondeu exatamente a uma das referências, sendo a *Predição E2* punida de modo indevido. Constatamos que o aprendizado absorvido pelo *Experimento 3* tende a repetir elementos (no exemplo o substantivo 'Christian Panucci'), enquanto o aprendizado do *Experimento 2*

Tabela 7 – Exemplos de predições em inglês a partir dos *Experimentos 2 e 3*. Fonte: O autor.

Exemplo A (extraído do *corpus* WebSplit AG18)	
Entrada	Asser Levy Public Baths is located in Manhattan , which is a part of New York .
Referência 1	Asser Levy Public Baths is located in Manhattan . Manhattan is a part of New York .
Referência 2	Asser Levy Public Baths is located in Manhattan . Manhattan is part of New York .
Predição E2 (100.00)	Asser Levy Public Baths is located in Manhattan . Manhattan is a part of New York .
Predição E3 (100.00)	Asser Levy Public Baths is located in Manhattan . Manhattan is a part of New York .
Exemplo B (extraído do *corpus* WebSplit v0.1)	
Entrada	A.S. Livorno Calcio are managed by Christian Panucci who is attached to the club Genoa CFC .
Referência 1	A.S. Livorno Calcio are managed by Christian Panucci . Christian Panucci is attached to the club Genoa CFC .
Referência 2	A.S. Livorno Calcio is managed by Christian Panucci . Christian Panucci played football for Genoa C.F.C.
Referência 3	A.S. Livorno Calcio are managed by Christian Panucci . Christian Panucci played football for Genoa C.F.C.
Referência 4	A.S. Livorno Calcio is managed by Christian Panucci . Christian Panucci is attached to the club Genoa CFC .
Predição E2 (82.53)	A.S. Livorno Calcio are managed by Christian Panucci . He is attached to the club Genoa CFC .
Predição E3 (100.00)	A.S. Livorno Calcio are managed by Christian Panucci . Christian Panucci is attached to the club Genoa CFC .
Exemplo C (extraído do *corpus* WikiSplit)	
Entrada	Gavin confessed to the murder of George Pollard and was held in the Round House until he was hung on the 6th April 1844 , his body was buried south of the Round House .
Referência	Gavin confessed to the murder of George Pollard and was held in the Round House until he was hanged on 6 April 1844 . His body was buried south of the Round House .
Predição E2 (84.02)	Gavin confessed to the murder of George Pollard and was held in the Round House until he was hung on the 6th April 1844 . His body was buried south of the Round House .
Predição E3 (73.90)	Gavin confessed to the murder of George Pollard . Gavin was held in the Round House until he was hung on the 6th April 1844 . His body was buried south of the Round House .
Exemplo D (extraído do *corpus* WikiSplit)	
Entrada	Eric Bell is a musician born in 1947 and was the lead guitarist for Thin Lizzy .
Referência	Eric Bell is a musician born in Belfast in 1947 . He was the lead guitarist for Thin Lizzy .
Predição E2 (70.02)	Eric Bell is a musician born in 1947 . Bell was the lead guitarist for Thin Lizzy .
Predição E3 (52.91)	Eric Bell is a musician . The musician is born in 1947 . The musician was the lead guitarist for Thin Lizzy .

tende a promover mais pronomes (no exemplo, o pronome pessoal ‘He’).

No *Exemplo C*, observamos novamente que uma mesma sentença de entrada complexa foi transformada em diferentes saídas: a *Predição E2* trouxe um único particionamento, enquanto a *Predição E3* apresentou dois particionamentos, novamente reforçando que o aprendizado sobre o *corpus* MinWikiSplit captura a tendência de múltiplos particionamentos. Embora tais resultados apresentem diversificadas formas de simplificação de uma mesma entrada, novamente as pontuações foram indevidamente diminuídas pela falta de mais referências para comparação.

No *Exemplo D*, vemos que a pontuação BLEU foi novamente prejudicada por ruídos do *corpus* WikiSplit, dado que a informação ‘in Belfast’ da referência não consta na sentença de entrada. Tais fatos não suportados (*unsupported facts*) tornam esta referência inconsistente. No entanto, embora diferentes, ambas as predições atenderam perfeitamente o método *split-and-rephrase*: a *Predição E2* copiou o sobrenome ‘Bell’ para estabelecer ligação com o substantivo ‘Eric Bell’, enquanto a *Predição E3* o referencia nas demais sentenças como ‘The musician’.

4.3.2 Resultados na Língua Portuguesa

Experimentos 2 e 3

Na Tabela 8, página 82, apresentamos predições dos *Experimentos 2 e 3* com exemplos extraídos do conjunto PorSimplesSent, empregado para o teste do *pipeline* na língua portuguesa.

Nos *Exemplos A e B*, ambas as *Predições E2* particionaram orações coordenadas com a supressão do ponto final ‘.’ e da conjunção ‘e’, e adotaram o pronome ‘Ele’ na segunda sentença para referenciarem componentes da primeira sentença. Já as *Predições E3*, repetiram respectivamente os elementos, ‘O projeto Gemini’ e ‘O Fórum Social Mundial’, correspondendo com as referências. Ou seja, os exemplos ilustram o mesmo comportamento observado nas sentenças em inglês, visto que o *Experimento 2* tende a promover pronomes e o *Experimento 3* tende a replicar tais constituintes repetidamente.

No *Exemplo C*, temos novamente saídas distintas entre as predições, com a promoção do marcador discursivo ‘Mas’ na *Predição E2*, e a repetição das palavras ‘A’ e ‘doença’ na *Predição E3*. Embora ambas tenham produzido gramaticalidade perfeita e sentidos corretos, consideramos que a *Predição E2* desempenha um papel mais adequado em termos semânticos, mantendo uma ideia de contradição.

No *Exemplo D*, as predições também apresentam particionamentos distintos. A *Predição E3* assimilou que a vírgula e a conjunção ‘e’ da sentença original deveriam

ser suprimidas para inserção de pontos finais ‘.’ seguidos das palavras ‘A’ e ‘técnica’, promovendo dois particionamentos. Já na *Predição E2*, apenas a conjunção foi eliminada, e a ação realizada pelo *pipeline* foi a colocação do pronome ‘Ela’ para estabelecer referência com a sentença anterior, produzindo apenas um particionamento.

Tabela 8 – Exemplos de predições em português a partir dos *Experimentos 2 e 3*. Fonte: O autor.

Exemplo A	
Entrada	O projeto Gemini é resultado de uma associação de sete países e envolve a construção de dois telescópios com um espelho de oito metros de diâmetro.
Referência	O projeto Gemini é resultado de uma associação de sete países. O projeto Gemini envolve a construção de dois telescópios com um espelho de oito metros de diâmetro.
Predição E2 (84.02)	O projeto Gemini é resultado de uma associação de sete países . Ele envolve a construção de dois telescópios com um espelho de oito metros de diâmetro .
Predição E3 (100.00)	O projeto Gemini é resultado de uma associação de sete países . O projeto Gemini envolve a construção de dois telescópios com um espelho de oito metros de diâmetro .
Exemplo B	
Entrada	O Fórum Social Mundial começa no sábado e vai até o dia 25 no país africano.
Referência	O Fórum Social Mundial começa no sábado. O Fórum Social Mundial vai até o dia 25 no país africano.
Predição E2 (71.05)	O Fórum Social Mundial começa no sábado . Ele vai até o dia 25 no país africano .
Predição E3 (100.00)	O Fórum Social Mundial começa no sábado . O Fórum Social Mundial vai até o dia 25 no país africano .
Exemplo C	
Entrada	A doença é estudada de forma intensa no mundo todo, mas tem um mecanismo de ação ainda pouco compreendido.
Referência	A doença é estudada de forma intensa no mundo todo. Mas tem um mecanismo de ação ainda pouco compreendido.
Predição E2 (100.00)	A doença é estudada de forma intensa no mundo todo . Mas tem um mecanismo de ação ainda pouco compreendido .
Predição E3 (82.32)	A doença é estudada de forma intensa no mundo todo . A doença tem um mecanismo de ação ainda pouco compreendido .
Exemplo D	
Entrada	A técnica existe, pode ser utilizada e já foi testada.
Referência	A técnica existe. A técnica pode ser utilizada. A técnica já foi testada.
Predição E2 (28.07)	A técnica existe , pode ser utilizada . Ela foi testada .
Predição E3 (100.00)	A técnica existe . A técnica pode ser utilizada . A técnica já foi testada .

Conforme expresso pelos exemplos nas Tabelas 6, 7 e 8, nota-se que o mesmo comportamento aprendido para *split-and-rephrase* na língua inglesa, pôde ser refletido e aplicado para sentenças no português brasileiro, com as predições apresentando transformações similares baseadas em padrões sintáticos compartilhados que podem atuar como atributos em comum entre as línguas (STODDEN; KALLMEYER, 2020).

4.3.3 Limitações

Na Tabela 9, compilamos algumas limitações e comportamentos indesejados produzidos pelo *pipeline* nos experimentos realizados.

O *Exemplo A* apresenta uma predição rigorosamente igual a sentença de entrada. Embora o *pipeline* tenha se mostrado pouco conservador e o número de sentenças sem interceptação tenha ficado próximo a zero em todos os experimentos, observamos que algumas poucas sentenças não receberam simplificação quando deveriam.

O *Exemplo B* mostra um exemplo típico de informações de saída que não deveriam estar na predição. A segunda sentença 'Ele é bom' não consta na sentença de entrada. No contexto de simplificação de sentenças, este tipo de geração de linguagem natural incorreta é conhecida como *hallucination*. Observamos que este fenômeno foi causado na maioria das vezes por predições indevidas do modelo de linguagem mascarada, BERT MLM.

No *Exemplo C*, a predição particionou um aposto promovendo a reordenação dos elementos da sentença de entrada, porém a palavra 'was' apareceu repetida. Tal efeito incorreto é ocasionalmente produzido pelos modelos *sequence-to-sequence*. Em particular, o *Experimento 3* apresentou mais vezes este comportamento de repetição de palavras, em função das excessivas repetições nos dados de treinamento do MinWikiSplit.

Tabela 9 – Exemplos de predições indesejadas. Fonte: O autor.

Exemplo A	
Entrada	Guy Mathez is a Swiss football manager from Canton of Jura .
Referência	Guy Mathez is a Swiss football manager . He comes from Canton of Jura .
Predição	Guy Mathez is a Swiss football manager from Canton of Jura .
Exemplo B	
Entrada	Além disso, o resultado da pesquisa ajuda países pobres, onde o regime curto é o único acessível.
Referência	Além disso, o resultado da pesquisa ajuda países pobres. Nos países pobres, o regime curto é o único acessível.
Predição	O resultado da pesquisa ajuda países pobres , onde o regime curto é o único acessível . Ele é bom .
Exemplo C	
Entrada	Born in Huzhou , Zhejiang , Qian was trained in traditional Chinese philology , and was a student of Zhang Binglin .
Referência	Born in Huzhou , Zhejiang , Qian was trained in traditional Chinese philology . He was a student of Zhang Binglin .
Predição	Qian was born in Huzhou , Zhejiang . Qian was trained in traditional Chinese philology . Qian was was a student of Zhang Binglin .

4.3.4 Síntese

De um modo geral, as estatísticas de estimativa de qualidade e os exemplos de predições apresentados, comprovam que a simplificação sintática decorrente do método *split-and-rephrase* gera aumento no número de sentenças de saída, pela desejável quebra das sentenças de entrada. Ao mesmo tempo, observamos que tais transformações diminuíram a incidência de palavras funcionais nas sentenças de saída, como conjunções, preposições e pronomes relativos, conforme pontuado pelo trabalho de Rebello et al. (2019).

Entre os fenômenos sintáticos, observamos que o *pipeline* proposto pôde executar com sucesso particionamentos para diferentes tipos de orações coordenadas, subordinadas e apostos, tanto na língua inglesa quanto no português brasileiro. Estas estão entre as principais transformações equivalentes entre as línguas, conforme pontuado pelo trabalho de Scarton et al. (2017), que aplicou simultaneamente simplificação sintática no inglês e em línguas relacionadas ao português, como a espanhola e a italiana.

Especificamente sobre o emprego do modelo de linguagem mascarada, BERT MLM, este se mostrou mais preciso para a língua inglesa, conseguindo preencher lacunas com palavras mais adequadas para o contexto das sentenças. Embora, em alguns casos, palavras inadequadas tenham levado ao comportamento indesejado de *hallucination*. Especialmente pela predição de pronomes pessoais, notamos que as reescritas simplificadas mantiveram a ligação semântica entre as sentenças de saída.

Com relação a análise pelo BLEU, parte das predições refletiram em pontuações baixas indevidamente, seja pelos ruídos apresentados nos conjuntos ou pela falta de referências adequadas. Dado que as formas de simplificação podem variar, o cenário ideal para uma avaliação automática é a existência de múltiplas referências corretas para uma estimação mais precisa (NARAYAN; GARDENT, 2014). A falta de dados paralelos apropriados é um fator limitante conhecido na tarefa de simplificação de sentenças (NARAYAN; GARDENT, 2016; GARBACEA et al., 2020), pois além de escassos, geralmente são alinhados automaticamente e apresentam tais inconsistências (MARTIN et al., 2018). Especificamente neste trabalho, focado no método *split-and-rephrase*, *unsupported facts* e *missing facts* observados nos dados de treinamento e teste, foram os fatores que mais trouxeram obstáculos.

Por fim, estudos pontuam ainda que a tarefa de simplificação de sentenças precisa desenvolver métricas mais adequadas para endereçar múltiplas transformações de reescrita, especialmente decorrentes de alterações sintáticas (ALVA-MANCHEGO et al., 2020). Alguns trabalhos afirmam que o BLEU não é idealmente adequado, por não se correlacionar diretamente com parâmetros de simplicidade (SULEM; ABEND; RAPPOPORT, 2018a). Já outros estudos consideram que a avaliação humana protocolar é a forma mais apropriada (SAGGION et al., 2015). No entanto, esta última geralmente é

realizada com um baixo número de predições, e por consequência acaba considerando apenas uma pequena parte dos resultados (SULEM; ABEND; RAPPOPORT, 2018b). Em suma, conforme pontuado por Garbacea et al. (2020), não existe na literatura atual um consenso estabelecido para a medição exata da simplificação, e tais avaliações são percebidas em parte como insuficientes, prejudicando até mesmo as comparações entre trabalhos relacionados.

5

Conclusão

Simplificação automática de sentenças é uma tarefa que tem potencial para auxiliar pessoas menos proficientes em leitura a compreenderem textos, e também otimizar resultados de outras tarefas relacionadas em processamento de linguagem natural. Esta é uma tarefa desafiadora, que segue em constante evolução para alcançar níveis mais satisfatórios e diretamente úteis para usuários finais.

Neste trabalho, contextualizamos a tarefa de simplificação automática de sentenças com foco no método *split-and-rephrase*. Este método busca particionar uma sentença única de entrada em duas ou mais sentenças reescritas de saída que juntas mantêm o significado original, conceitualizando que sentenças curtas são geralmente melhores para a compreensão humana e para o processamento de outras aplicações automáticas.

5.1 Contribuições

Implementamos um *pipeline* para simplificação de sentenças através do método *split-and-rephrase*. A proposta foi composta por um modelo de redes neurais *sequence-to-sequence* equipado com mecanismos de atenção, integrado a um modelo de linguagem mascarada pré-treinado, BERT MLM. O primeiro visa o aprendizado automático sobre a recorrência de sequências de *part-of-speech tags* (classes de palavras) observadas nos dados de treinamento para absorver o comportamento de particionamento, e o segundo fornece palavras para complementar as reescritas de saída de acordo com seus respectivos contextos.

Realizamos três experimentos distintos, onde, para treinamento dos três modelos *sequence-to-sequence*, empregamos uma estratégia de vocabulário simbólico, para que o conhecimento sobre padrões sintáticos absorvido a partir de dados no inglês, pudesse ser aplicado também no português brasileiro. Observamos que essa estratégia reduziu drasticamente o tamanho do vocabulário para apenas alguns itens, adicionalmente otimizando tempos de convergência no treinamento.

Para validar e testar o método proposto, envolvemos os quatro *corpora* do estado da arte para o método *split-and-rephrase*, e também um *corpus* adaptado para o português brasileiro. Em uma inspeção manual detalhada, confirmamos que o *pipeline* idealizado pôde particionar sentenças complexas de entrada em sentenças de saída mais curtas, frequentemente preservando a gramaticalidade e o significado equivalente com êxito. Por outro lado, algumas das predições revelaram erros comuns de geração de linguagem natural, como repetição ou omissão de *tokens* e 'alucinação' de informações indesejadas.

5.2 Trabalhos Futuros

Em trabalhos futuros, pretendemos explorar o *pipeline* em línguas próximas do português, como o espanhol. Além disso, seria interessante promover uma avaliação dos benefícios do método *split-and-rephrase* para aplicações de PLN relacionadas, como na tarefa de extração de informação, por exemplo.

Em situações em que a simplificação precisa ser direcionada para determinados alvos ou públicos específicos, o encaixe de um modelo classificador no *pipeline* para distinguir se a sentença de entrada deve ou não ser simplificada, também seria uma abordagem interessante em trabalhos futuros, no sentido de regular quebras excessivas. Consideramos que sentenças demasiadamente curtas podem não ser ideais para leitura humana, mas podem eventualmente beneficiar interpretações por outras tarefas de PLN.

Por fim, entendemos que a construção de um *corpus* paralelo próprio para aprendizado e teste do método *split-and-rephrase* para o português brasileiro, composto por fenômenos sintáticos diversificados entre os alinhamentos, poderia trazer novos ganhos, sendo este também um desejável trabalho futuro.

Referências

ABEND, O.; RAPPOPORT, A. Universal Conceptual Cognitive Annotation (UCCA). In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2013. p. 228–238.

AHARONI, R.; GOLDBERG, Y. Split and rephrase: Better evaluation and stronger baselines. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. [S.l.: s.n.], 2018. p. 719–724.

ALUÍSIO, S. M. et al. Towards brazilian portuguese automatic text simplification systems. In: ACM. *Proceedings of the eighth ACM symposium on Document engineering*. [S.l.], 2008. p. 240–248.

ALVA-MANCHEGO, F. et al. Learning how to simplify from explicit labeling of complex-simplified text pairs. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. [S.l.: s.n.], 2017. p. 295–305.

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. et al. Easse: Easier automatic sentence simplification evaluation. 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): System Demonstrations*. [S.l.: s.n.], 2019. p. 49–54.

ALVA-MANCHEGO, F.; SCARTON, C.; SPECIA, L. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, MIT Press, v. 46, n. 1, p. 135–187, 2020.

BAHDANAU, D.; CHO, K. H.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: *3rd International Conference on Learning Representations, ICLR 2015*. [S.l.: s.n.], 2015.

BARBU, E. et al. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, v. 42, n. 12, p. 5076 – 5086, 2015. ISSN 09574174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2015.02.044>.

BAUM, L. E.; PETRIE, T. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, JSTOR, v. 37, n. 6, p. 1554–1563, 1966.

Referências

BEROV, L.; STANDVOSS, K. Discourse embellishment using a deep encoder-decoder network. In: *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)*. [S.l.: s.n.], 2018. p. 11–16.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, v. 5, 07 2016.

BOSCO, G. L.; PILATO, G.; SCHICCHI, D. A Neural Network model for the Evaluation of Text Complexity in Italian Language: A Representation Point of View. In: . Prague, Czech republic: [s.n.], 2018. v. 145, p. 464 – 470. ISSN 18770509. Disponível em: <http://dx.doi.org/10.1016/j.procs.2018.11.108>.

BOTHA, J. A. et al. Learning to split and rephrase from wikipedia edit history. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2018. p. 732–737.

CASELI, H. M. et al. Building a brazilian portuguese parallel corpus of original and simplified texts. In: *Advances in Computational Linguistics, Research in Computer Science, (CICLing-2009), volume 41*. [S.l.: s.n.], 2009. p. 59—-70.

CHO, K. et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1724–1734.

CHUNG, J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: *NIPS 2014 Workshop on Deep Learning, December 2014*. [S.l.: s.n.], 2014.

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.

COOPER, M.; SHARDLOW, M. "CombiNMT: An exploration into neural text simplification models". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 5588–5594. ISBN 979-10-95546-34-4. Disponível em: <https://www.aclweb.org/anthology/2020.lrec-1.686>.

COSTER, W.; KAUCHAK, D. Simple english wikipedia: a new text simplification task. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2011. p. 665–669.

CURTO, P.; MAMEDE, N.; BAPTISTA, J. Automatic readability classifier for European Portuguese. *System*, v. 5, p. 6, 2014.

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, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2019. p. 4171–4186.

DUBAY, W. H. *The Principles of Readability. A brief introduction to readability research*. 2004.

ELMAN, J. L. Finding structure in time. *Cognitive science*, Wiley Online Library, v. 14, n. 2, p. 179–211, 1990.

Referências

FERRÉS, D. et al. YATS: yet another text simplifier. In: SPRINGER. *International Conference on Applications of Natural Language to Information Systems*. [S.l.], 2016. p. 335–342.

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

GARBACEA, C. et al. An empirical study on explainable prediction of text complexity: Preliminaries for text simplification. *CoRR*, abs/2007.15823, 2020. Disponível em: <https://arxiv.org/abs/2007.15823>.

GARDENT, C. et al. Creating training corpora for nlg micro-planning. In: *55th annual meeting of the Association for Computational Linguistics (ACL)*. [S.l.: s.n.], 2017.

GLAVAŠ, G.; ŠTAJNER, S. Simplifying lexical simplification: Do we need simplified corpora? 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. 63–68.

GOLDBERG, Y. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, v. 57, p. 345–420, 2016.

GRAESSER, A. C. et al. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, Springer, v. 36, n. 2, p. 193–202, 2004.

GRAVES, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, Citeseer, 2013.

GU, J. et al. Incorporating copying mechanism in sequence-to-sequence learning. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2016. p. 1631–1640.

GUO, H.; PASUNURU, R.; BANSAL, M. Dynamic multi-level multi-task learning for sentence simplification. In: *Proceedings of the 27th International Conference on Computational Linguistics*. [S.l.: s.n.], 2018. p. 462–476.

HARTMANN, N. S.; PAETZOLD, G. H.; ALUÍSIO, S. M. SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese. In: VILLAVICENCIO, A. et al. (Ed.). *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2018. p. 272–283. ISBN 978-3-319-99722-3.

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

JORDAN, M. I. Serial order: a parallel distributed processing approach. technical report, june 1985-march 1986. 5 1986. Disponível em: <https://www.osti.gov/biblio/6910294>.

KANÉ, H. et al. Towards neural similarity evaluator. In: *Workshop on Document Intelligence at NeurIPS 2019*. [S.l.: s.n.], 2019.

KAUCHAK, D. et al. Text simplification tools: Using machine learning to discover features that identify difficult text. In: . Waikoloa, HI, United States: [s.n.], 2014. p. 2616 – 2625. ISSN 15301605. Disponível em: <http://dx.doi.org/10.1109/HICSS.2014.330>.

KINCAID, J. P. et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training, University of Central Florida, 1975.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: *ICLR (Poster)*. [S.l.: s.n.], 2015.

KINTSCH, W.; DIJK, T. A. V. Toward a model of text comprehension and production. *Psychological review*, American Psychological Association, v. 85, n. 5, p. 363, 1978.

KLOEHN, N. et al. Improving Consumer Understanding of Medical Text: Development and Validation of a New SubSimplify Algorithm to Automatically Generate Term Explanations in English and Spanish. *J. Med. Internet Res.*, v. 20, n. 8, p. e10779, 08 2018.

LEAL, S. E.; DURAN, M. S.; ALUÍSIO, S. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In: *Proceedings of the 27th International Conference on Computational Linguistics*. [S.l.: s.n.], 2018. p. 401–413.

LEAL, S. E. et al. Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural. In: IN: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY AND COLLOCATES. *Embrapa Gado de Leite–Artigo em anais de congresso (ALICE)*. [S.l.], 2019.

LEAL, S. E. et al. Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches. In: *Proceedings of the 28th International Conference on Computational Linguistics*. [S.l.: s.n.], 2020. p. 5821–5831.

LUONG, M.-T.; PHAM, H.; MANNING, C. D. Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2015. p. 1412–1421.

MA, S.; SUN, X. A semantic relevance based neural network for text summarization and text simplification. *Computational Linguistics*, v. 1, n. 1, 2017.

MANNING, C. D.; MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT press, 1999.

MARTIN, L. et al. Controllable sentence simplification. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 4689–4698. ISBN 979-10-95546-34-4. Disponível em: <https://www.aclweb.org/anthology/2020.lrec-1.577>.

MARTIN, L. et al. Reference-less quality estimation of text simplification systems. In: *1st Workshop on Automatic Text Adaptation (ATA)*. [S.l.: s.n.], 2018.

MARTINS, T. B. et al. *Readability formulas applied to textbooks in brazilian portuguese*. 1996. Notas do ICMSC–USP, 28.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119.

Referências

MILLER, G. A. Wordnet: A lexical database for english. *Commun. ACM*, ACM, New York, NY, USA, v. 38, n. 11, p. 39–41, nov. 1995. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/219717.219748>.

MUNIZ, M. C. M. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB*. Tese (Doutorado) — Universidade de São Paulo, 2004.

NARAYAN, S.; GARDENT, C. Hybrid simplification using deep semantics and machine translation. In: *The 52nd annual meeting of the association for computational linguistics*. [S.l.: s.n.], 2014. p. 435–445.

NARAYAN, S.; GARDENT, C. Unsupervised sentence simplification using deep semantics. In: *Proceedings of the 9th International Natural Language Generation conference*. [S.l.: s.n.], 2016. p. 111–120.

NARAYAN, S. et al. Split and rephrase. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2017. p. 606–616.

NIKLAUS, C. et al. Dissim: A discourse-aware syntactic text simplification framework for english and german. In: *Proceedings of the 12th International Conference on Natural Language Generation*. [S.l.: s.n.], 2019. p. 504–507.

NIKLAUS, C. et al. Transforming complex sentences into a semantic hierarchy. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2019. p. 3415–3427.

NIKLAUS, C.; FREITAS, A.; HANDSCHUH, S. Minwikisplit: A sentence splitting corpus with minimal propositions. In: *Proceedings of the 12th International Conference on Natural Language Generation*. [S.l.: s.n.], 2019. p. 118–123.

NISIOI, S. et al. Exploring neural text simplification models. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short papers)*. [S.l.: s.n.], 2017. v. 2, p. 85–91.

OGDEN, C. K. Basic english: A general introduction with rules and grammar. Paul Treber, 1930.

PAETZOLD, G.; SPECIA, L. Lexenstein: A framework for lexical simplification. In: *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. [S.l.: s.n.], 2015. p. 85–90.

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*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318. Disponível em: <https://www.aclweb.org/anthology/P02-1040>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://www.aclweb.org/anthology/D14-1162>.

QIANG, J. et al. LSBert: A Simple Framework for Lexical Simplification. *arXiv preprint arXiv:2006.14939*, 2020.

REBELLO, B. M. et al. Efeito da simplificação sintática sobre a compreensão de leitura de crianças do ensino fundamental. *Audiology-Communication Research*, SciELO Brasil, v. 24, 2019.

RESHAMWALA, A.; MISHRA, D.; PAWAR, P. Review on natural language processing. *IRACST – Engineering Science and Technology: An International Journal (ESTIJ)*, v. 3, p. 113–116, 02 2013.

SAGGION, H. et al. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Trans. Access. Comput.*, ACM, New York, NY, USA, v. 6, n. 4, p. 14:1–14:36, maio 2015. ISSN 1936-7228. Disponível em: <http://doi.acm.org/10.1145/2738046>.

SCARTON, C. et al. MUSST: a multilingual syntactic simplification tool. In: *Proceedings of the IJCNLP 2017, System Demonstrations*. [S.l.: s.n.], 2017. p. 25–28.

SCARTON, C. et al. SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In: *Proceedings of the NAACL HLT 2010 Demonstration Session*. [S.l.: s.n.], 2010. p. 41–44.

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.; PAETZOLD, G. H.; SPECIA, L. Text simplification from professionally produced corpora. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: Elsevier Inc., 2019. p. 3504–3510.

SCARTON, C.; SPECIA, L. Learning simplifications for specific target audiences. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. [S.l.: s.n.], 2018. p. 712–718.

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.

SENTER, R.; SMITH, E. A. *Automated readability index*. [S.l.], 1967.

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.

SIDDHARTHAN, A.; MANDYA, A. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2014. p. 722–731.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2020. p. 403–417.

SPECIA, L.; ALUíSIO, S.; PARDO, T. A. S. *Manual de Simplificação Sintática para o Português*. 2008. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional (NILC-TR-08-06).

ŠTAJNER, S.; CALIXTO, I.; SAGGION, H. Automatic text simplification for spanish: Comparative evaluation of various simplification strategies. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing*. [S.l.: s.n.], 2015. p. 618–626.

ŠTAJNER, S.; GLAVA, G. Leveraging event-based semantics for automated text simplification. *Expert Systems with Applications*, v. 82, p. 383 – 395, 2017. ISSN 09574174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2017.04.005>.

ŠTAJNER, S.; NISIOI, S.; HULPUș, I. CoCo: A tool for automatically assessing conceptual complexity of texts. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. [S.l.: s.n.], 2020. p. 7179–7186.

ŠTAJNER, S.; POPOVIĆ, M. Automated Text Simplification as a Preprocessing Step for Machine Translation into an Under-resourced Language. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. [S.l.: s.n.], 2019. p. 1141–1150.

STODDEN, R.; KALLMEYER, L. A multi-lingual and cross-domain analysis of features for text simplification. In: *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*. Marseille, France: European Language Resources Association, 2020. p. 77–84. ISBN 979-10-95546-45-0. Disponível em: <https://www.aclweb.org/anthology/2020.readi-1.12>.

SULEM, E.; ABEND, O.; RAPPOPORT, A. Bleu is not suitable for the evaluation of text simplification. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2018. p. 738–744.

SULEM, E.; ABEND, O.; RAPPOPORT, A. Semantic structural evaluation for text simplification. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 685–696. Disponível em: <https://aclanthology.org/N18-1063>.

SULEM, E.; ABEND, O.; RAPPOPORT, A. Simple and effective text simplification using semantic and neural methods. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 162–173. Disponível em: <https://aclanthology.org/P18-1016>.

SURYA, S. et al. Unsupervised neural text simplification. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 2058–2068. Disponível em: <https://aclanthology.org/P19-1198>.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 3104–3112.

TISSIER, J.; GRAVIER, C.; HABRARD, A. Dict2vec : Learning word embeddings using lexical dictionaries. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 254–263. Disponível em: <https://www.aclweb.org/anthology/D17-1024>.

VAJJALA, S. et al. *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. [S.l.]: O'Reilly Media, 2020.

VASWANI, A. et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

VU, T. et al. Sentence simplification with memory-augmented neural networks. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 79–85. Disponível em: <https://aclanthology.org/N18-2013>.

WANG, T. et al. An experimental study of LSTM encoder-decoder model for text simplification. *CoRR*, abs/1609.03663, 2016. Disponível em: <http://arxiv.org/abs/1609.03663>.

WATANABE, W. M. et al. Facilita: Reading assistance for low-literacy readers. In: *Proceedings of the 27th ACM International Conference on Design of Communication*. New York, NY, USA: ACM, 2009. (SIGDOC '09), p. 29–36. ISBN 978-1-60558-559-8. Disponível em: <http://doi.acm.org/10.1145/1621995.1622002>.

WITTEN, I. H. et al. Weka: Practical machine learning tools and techniques with java implementations. 1999.

WOLOSZYN, V. et al. Pylinguistics: an open source library for readability assessment of texts written in portuguese. *Revista de Sistemas de Informação da FSMA*, v. 18, 2016. ISSN 1983-5604.

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*, v. 4, p. 401–415, 2016. Disponível em: <https://www.aclweb.org/anthology/Q16-1029>.

ZHAO, S. et al. Integrating transformer and paraphrase rules for sentence simplification. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 3164–3173. Disponível em: <https://aclanthology.org/D18-1355>.

ZHU, Z.; BERNHARD, D.; GUREVYCH, I. A monolingual tree-based translation model for sentence simplification. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, 2010. p. 1353–1361. Disponível em: <https://www.aclweb.org/anthology/C10-1152>.

Anexos

A

Experimenting Sentence Split-and-Rephrase Using Part-of-Speech Labels

Experimenting Sentence Split-and-Rephrase Using Part-of-Speech Labels

P. Berlanga Neto and E. Y. Okano and E. E. S. Ruiz

Departamento de Computação e Matemática, FFCLRP
Universidade de São Paulo (USP).
Av. Bandeirantes, 3900, Monte Alegre. 14040-901, Ribeirão Preto, SP – Brazil
[pauloberlanga, okano700, evandro]@usp.br

Abstract. Text simplification (TS) is a natural language transformation process that reduces linguistic complexity while preserving semantics and retaining its original meaning. This work aims to present a research proposal for automatic simplification of texts, precisely a split-and-rephrase approach based on an encoder-decoder neural network model. The proposed method was trained against the WikiSplit English corpus with the help of a part-of-speech tagger and obtained a BLEU score validation of 74.72%. We also experimented with this trained model to split-and-rephrase sentences written in Portuguese with relative success, showing the method's potential.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

Keywords: natural language processing, neural networks, sentence simplification

1. INTRODUCTION

Text Simplification (TS) is the process of modifying natural language to reduce complexity and improve both readability and understandability [Shardlow 2014]. A simplified vocabulary or a simplified text structure can benefit different publics, such as people with low education levels, children, non-native individuals, people who have learning disorders (such as autism, dyslexia, and aphasia), among others [Štajner et al. 2015]. Although the simplicity of a text seems intuitively obvious, it does not have a precise definition in technical terms. Traditional assessment metrics consider factors such as sentence length, syllable count, and other linguistic features of the text to identify it as elaborate or not [Shardlow 2014].

In the last decades, several models and systems have been developed to improve the task of automatic text simplification. They are primarily based on two main approaches: lexical simplification (LS) and/or syntactic simplification (SS) [Shardlow 2014]. Lexical simplification has the goal to identify and replace words or expressions for synonyms that can be understood by a larger audience [Hartmann et al. 2018; Shardlow 2014]. On the other hand, syntactic simplification must identify grammatical complexities presented in the sentences and rewrite them in simpler structures [Tajner and Glava 2017; Scarton et al. 2019].

Since complex texts often contain a portion of simple sentences in their structure, some late approaches have focused on the analysis of specific aspects at the sentence level, expanding a study branch known as sentence simplification (SentS). Split-and-rephrase is a SentS method proposed by Narayan and colleagues [Narayan et al. 2017] that aims to split a complex sentence into shorter sentences while preserving the meaning of the original sentence. This method conceptualizes the notion that shorter sentences are generally better understood by the majority of the individuals. Neverthe-

less, it may also be easier to be processed by NLP (Natural Language Processing) systems, facilitating tasks such as parsers, semantic role labelers, and machine translation systems.

In this paper, we make the main contribution in exploring split-and-rephrase from word labels gained from a part-of-speech tagger. We focus on coordinate clauses, a relative dependence clause to the main sentence clause. Also, we add a minor improvement to this experiment by performing a simple transfer learning. We train split-and-rephrase systems in English sentences and apply the learned knowledge to Portuguese sentences. Based on the literature surveyed, this is the first reference to an automatic split-and-rephrase method successfully used to the Portuguese language. In the next section, we describe some previous work on text simplification. In Section 3 we present the WikiSplit corpus [Botha et al. 2018] and our suggestion of simplification process. In Section 4 we present the results obtained by our proposed model. In Section 5, we briefly discuss these results, while in Section 6 we have the conclusion and the expected challenges to the future.

2. RELATED WORK

Advanced neural computational methods have transformed the task of text simplification. For a survey of the academic work before 2014, see the excellent survey from Advaith Siddharthan [Siddharthan 2014]. A quick search on Google Scholar for the string query *"Text Simplification" OR "Lexical Simplification" OR "Syntactic Simplification"* for documents published up to 2014 resulted in 3,200 hits.

Wang and colleagues [Wang et al. 2016] affirm that, up to 2016, some of the TS techniques were limited to either lexical-level applications or manually defining a large number of rules. In this same paper, the authors propose to use a Long Short-Term Memory (LSTM) encoder-decoder neural model for sentence-level TS. By applying this model, they examined operation rules such as reversing, sorting and replacing constituents from sequence pairs, which has the potential of sentence simplification.

Later, Nisioi and collaborators [Nisioi et al. 2017] present another attempt at using encoder-decoder neural networks to model TS, named Neural Text Simplification. They apply this model to simultaneously perform lexical simplification and content reduction, inspired by the success observed in the Neural Machine Translation approach [Bahdanau et al. 2014].

Vu and colleagues [Vu et al. 2018] also worked to simplify the content and structure of complex sentences. For this reason, they adopt an architecture with augmented memory capacities called Neural Semantic Encoders [Munkhdalai and Yu 2017] for sentence simplification. They have demonstrated the effectiveness of their approach by automatic evaluation measures and human judgments.

As for the simplification of texts written in Portuguese, we highlight the work of Hartman et al. [Hartmann et al. 2018] that targets lexical simplification compiling the SIMPLEX-PB, the first available corpus of lexical simplification for Brazilian Portuguese. In the article by Scarton et al. [Scarton et al. 2010], the authors present the 'Simplifica' tool, also as an integral part of the PorSimples project [Aluísio et al. 2008]. This technology encourages writers to write simplified texts in Brazilian Portuguese. The tool has two modules, one for simplifying lexical terms and the other for assessing the complexity of the input texts.

Concerning datasets for research in the split-and-rephrase task, until very recently, the WebSplit corpus introduced by Narayan et al. [Narayan et al. 2017] was the main corpus used as a benchmark for the split-and-rephrase job, nowadays WikiSplit [Botha et al. 2018] is the latest reference corpus, and it is the dataset used in this paper.

3. DATA AND METHODS

WikiSplit, by Botha and his AI team at Google [Botha et al. 2018], is a corpus for the split-and-rephrase task. It is composed of one million naturally occurring sentence rewrites obtained from mining English Wikipedia's edit history. WikiSplit is a public corpus, freely available[1], and licensed under CC BY-SA 4.0 [2].

The dataset is released as text files formatted as tab-separated values. It contains 1,004,994 English sentences, each split into two sentences that together preserve the original meaning. This corpus is divided into four datasets, the training dataset with 989,944 sentences, and the other three datasets, tune, validation, and test, containing 5,000 sentences each. The sentences below are an example of one may see in this corpus:

> Street Rod is the first in a series of two games released for the PC and Commodore 64 in 1989 .
>
> Street Rod is the first in a series of two games . <:::::> It was released for the PC and Commodore 64 in 1989 .

The string <:::::> marks the start of the split-up sentences.

We define the split-and-rephrase task as follows. Given a complex sentence C, the goal is to produce a simplified text T consisting of a sequence of sentences $T_1, T_2, \ldots, T_n$, $n \geq 2$, in such a way that T preserves the meaning of C.

3.1 Sentence selection

Given the vast amount of aligned sentence pairs in the WikiSplit corpus, two specific cuts were made in the original training dataset to train the proposed model. The first cut was selecting the alignments with a length of, at least 15, and a maximum of 30 time steps per sentence. We considered only sentences that had equivalent counts between NLTK[3] and Spacy[4] tokenizers with no special characters. This first cut selected 171,133 alignments.

The second cut aimed to individually select compound sentences formed by a main clause and a coordinate clause. See the highlighted example below extracted from the WikiSplit corpus. This second cut extracted 63,623 alignments, thus consolidating the final selection.

Original sentence
Jes was recruited to be on the Rock Of Love show while she was bartending in downtown Chicago at a bar called Rizzo's.
Original split-and-rephrased sentences
Jes was recruited to be on the Rock Of Love show. <:::::> **She** was bartending in down town Chicago at a bar called Rizzo's.

Regarding the part-of-speech classification, we adopted the Spacy POS tagger[5]. Contrary to some approaches that seek training the model with an extensive vocabulary, we generalize this learning solely by grammatical classes and by their respective recurrences. This way, we obtained a small set of attributes capable of optimizing training times. It took approximately two hours for the training

[1] https://github.com/google-research-datasets/wiki-split
[2] https://creativecommons.org/licenses/by-sa/4.0/
[3] https://www.nltk.org/
[4] https://spacy.io/
[5] https://spacy.io/api/tagger/

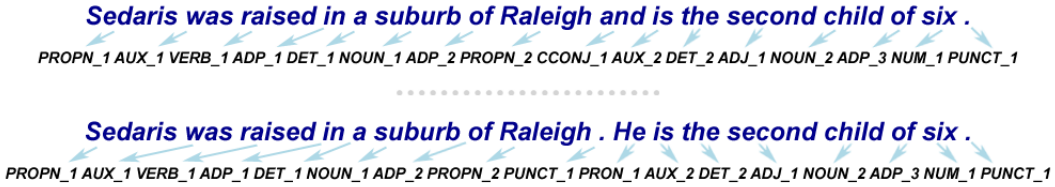


Fig. 1. Sample of the assigned parts-of-speech tags used to train the encoder-decoder neural model.

process in a multi-user computer environment. By using POS tags, the learned model may also split sentences in other languages than the training dataset itself, as the Brazilian Portuguese.

3.2 Model specification

We follow the proposal of Bahdanau and colleagues [Bahdanau et al. 2014], implementing an encoder-decoder neural network model based on the sequence-to-sequence (seq2seq) architecture, composed of recurrent neural networks with GRU (Gated Recurrent Unit) gating mechanisms [Cho et al. 2014]. This seq2seq composition is an appropriate architecture for training an aligned corpus comprising original sentences and their simplified, split versions.

Unlike traditional sequence-to-sequence approaches, which promote the compression of original sequences into a fixed context vector, the method proposed by Bahdanau an co-authors [Bahdanau et al. 2014] presents the attention mechanism that suggests associations between specific context vectors at each of the output time steps. This mechanism makes it possible to establish references at particular points in the original sequences, and enable the transmission of these instances to the decoder outputs.

In this experiment, the proposed architecture configuration had an attention layer connected to encoder-decoder GRU layers, both composed of 50 units. We used a batch size of 200 and trained the model in 1,000 epochs. One hundred thirty-nine (139) different labels generated by the part-of-speech tagger represented the vocabulary size. They were composed of grammatical classes and numbers (see Figure 1 example), together with the wildcard character '*' for padding. We applied Adam optimization algorithm [Kingma and Ba 2015] to update the weights of the networks iteratively and a categorical-cross entropy loss function. The Keras library[6] was used.

4. RESULTS

We trained the model using 63,000 random sentence pairs from the selected data (around 99% of the alignments). We recall that these selected sentences are composed of a main clause and a coordinate clause. We validated the method in the remaining 623 sentence pairs following a similar approach adopted by Botha and colleagues [Botha et al. 2018]. Even though this validation set seems small, there were enough predictions to analyze the expected split behavior.

Although some studies consider human judgments on grammaticality, meaning preservation and simplicity the most reliable method for evaluating the sentence simplification task, it is a common practice to use automatic metrics [Alva-Manchego et al. 2019]. Following the WikiSplit work, we adopted the BLEU (Bilingual Evaluation Understudy) method [Papineni et al. 2002] to validate the results. This metric was originally created to analyze results obtained by translation algorithms, but nowadays is also used by the literature [Vu et al. 2018; Nisioi et al. 2017; Štajner et al. 2015] to evaluate text simplification.

[6] `https://keras.io/`

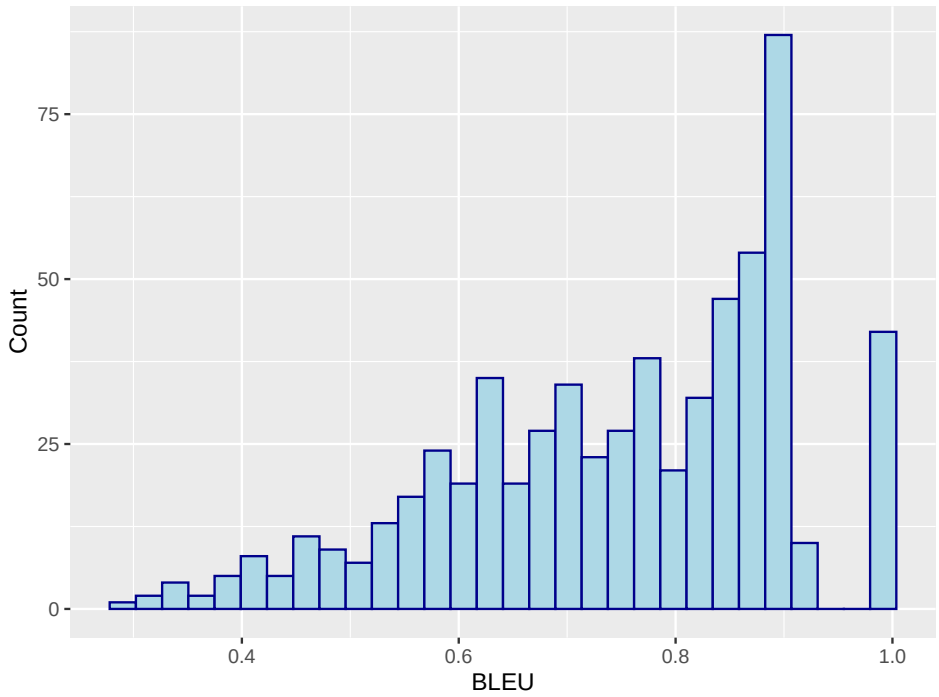


Fig. 2. Histogram of the BLEU scores for the 623 experiment sentences.

Using the proposed model against the validation set, we obtained a BLEU score of 74.72 ± 15.17. In Table I, we present some example results. The 'Input' corresponds to the original sentence from WikiSplit. The 'Reference' corresponds to the split version from WikiSplit, and the 'Predicted' is the result of the proposed approach. Figure 2 show a histogram of the BLEU scores for all the 623 validation sentences in the experiment. One may notice that there are many sentences with a perfect BLEU score. The complete log containing all the 623 predictions is publicly available[7].

The extracted examples in Table I show some behaviors that kept a high BLEU score and others that decreased its count. In Example 1, the predicted sentence was strictly the same as the expected reference sentence, thus keeping the BLEU score at its maximum. In Example 2, the word 'also' does not appear in the input sentence, then the proposed model could not predict this word according to the expected reference sentence. Despite that, it produced a perfect meaning sentence, close to the generated sentence. In Example 3, we have a similar problem since the neural model predicted a pronoun that could not be found in the input sentence. Thus it was not able to resolve which pronoun would best fit the sentence. In Example 4, we can observe a mix of both situations described in Example 2 and Example 3, with the prediction skipping a pronoun and also skipping some information present in the reference sentence like 'capital Quito'.

4.1 Experiments in Brazilian Portuguese

We also examined the ability of the proposed method to split-and-rephrase sentences written in Brazilian Portuguese, reusing the neural model trained with the part-of-speech labels gained from the English sentences of the WikiSplit corpus. We loaded the compiled model and made some predictions using the Spacy POS tagger [Honnibal and Montani 2017] for Brazilian Portuguese.

[7] `https://github.com/pauloberlanga/split-and-rephrase/`

Example 1	
Input	He was first elected in 2005 and represents the British Columbia New Democratic Party .
Reference	He was first elected in 2005 . He represents the British Columbia New Democratic Party .
Predicted (100.0%)	He was first elected in 2005. He represents the British Columbia New Democratic Party.
Example 2	
Input	They enrolled at New York Central College , an interracial institution in Cortland , New York , and worked as cleaning servants to support themselves .
Reference	They enrolled at New York Central College , an interracial institution in Cortland , New York . They also worked as cleaning servants to support themselves .
Predicted (90.0%)	They enrolled at New York Central College, an interracial institution in Cortland, New York. They worked as cleaning servants to support themselves.
Example 3	
Input	GTS Technologies is a paint finishing and mechanical handling systems company , registered in Wolverhampton , England .
Reference	GTS Technologies is a paint finishing and mechanical handling systems company . It is registered in Wolverhampton , England .
Predicted (85.7%)	GTS Technologies is a paint finishing and mechanical handling systems company. PRON_1 is registered in Wolverhampton, England.
Example 4	
Input	Pichincha is an active volcano in the country of Ecuador and gives its name to the entire province .
Reference	Pichincha is an active volcano in the country of Ecuador directly beneath its capital Quito . It gives its name to the entire province .
Predicted (63.4%)	Pichincha is an active volcano in the country of Ecuador. PRON_1 gives its name to the entire province.

Table I. Examples predicted by the model using WikiSplit sentences. See the BLEU score under the 'Predicted' tag.

Table II shows some rephrased sentences in Portuguese. We did not apply the BLEU metric here since we have no correct references for these sentences. In Example 5, we see that the word 'e' from the input sentence was successfully replaced by the '.' (period) character in the prediction sentence. The period was also followed by an expected pronoun. We consider it an exciting result although the skipped pronoun, since the model intercepted the correct terms just like for the English predictions. In Example 6, we see the skipped pronoun with repetitive words, sampling one of our wrong predictions.

Example 5	
Input	João era o melhor aluno de matemática e tirava nota 10 em todas as provas .
Predicted	João era o melhor aluno de matemática. PRON_1 tirava nota 10 em todas as provas.
Example 6	
Input	A seleção brasileira de futebol representa uma das equipes mais gloriosas da história futebolística, já tendo vencido 5 vezes a Copa do Mundo Fifa e diversos outros torneios.
Predicted	A seleção brasileira de futebol representa uma das equipes mais gloriosas da história . . PRON_1 tendo vencido vezes a copa do mundo diversos outros outros torneios.

Table II. Examples of some Brazilian Portuguese sentences predicted by the model.

5. DISCUSSIONS

The results confirm that the proposed encoder-decoder neural model could split a complex sentence into shorter sentences, most of the time preserving the meaning of the original sentence successfully. More than that, it also showed the potential to simplify sentences written in Brazilian Portuguese. On the other hand, some of the predictions skipped familiar words by the reference sentence and brought the grammatical classes instead. This is justified by the fact that our model does not treat lexical issues. We consider this as a minor problem but an essential question to solve.

As the model generates each word of the prediction sentence considering a soft-search on a set of references and all the previously created words, it showed the capability to align and learn this split-and-rephrase behavior simultaneously.

Regarding the BLEU score, the skipped words and some mistaken repetitive words reflected a low score for some predictions. Additionally, a certain number of the reference sentences in the WikiSplit corpus contain new information that does not effectively simplify the text, as observed in the reference sentence from the Example 4, Table I. We view that that extra information also contributed to harm the BLEU score.

6. CONCLUSIONS

We proposed a novel encoder-decoder neural model for sentence simplification through the use of the split-and-rephrase method. The model relies on the recurrence of the part-of-speech tags to automatically learn this split behavior. We also showed that the model trained against an English dataset, can split and rephrase sentences in Brazilian Portuguese with exciting results, performing a simple form of transfer learning.

One of the challenges left for the future is constructing a lexical approach to properly place skipped words in the predicted sentences, like the skipped pronouns according to their referenced constituents. Another exciting experience would be to promote the model evaluation under a proper Brazilian Portuguese corpus and also by human judgment assessments.

Acknowledgements

This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), under process number 2018/03129-8.

REFERENCES

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G., Caseli, H. M., and Fortes, R. P. A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*. ACM, New York, NY, United States, pp. 15–22, 2008.

Alva-Manchego, F., Martin, L., Scarton, C., and Specia, L. EASSE: Easier automatic sentence simplification evaluation. 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): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, pp. 49–54, 2019.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2014.

Botha, J. A., Faruqui, M., Alex, J., Baldridge, J., and Das, D. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pp. 732–737, 2018.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014.

Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese. In *Computational Processing of the Portuguese Language*, A. Villavicencio, V. Moreira, A. Abad, H. Caseli, P. Gamallo, C. Ramisch, H. Gonçalo Oliveira, and G. H. Paetzold (Eds.). Springer International Publishing, Cham, pp. 272–283, 2018.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *In Proceedings of the International Conference on Learning Representations (ICLR)*. Curran Associates, Inc., San Diego, CA, USA., 2015.

Munkhdalai, T. and Yu, H. Neural semantic encoders. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Vol. 1. Association for Computational Linguistics, Valencia, Spain, pp. 397, 2017.

Narayan, S., Gardent, C., Cohen, S. B., and Shimorina, A. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pp. 606–616, 2017.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pp. 85–91, 2017.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318, 2002.

Scarton, C., Oliveira, M., Candido Jr., A., Gasperin, C., and Aluísio, S. SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In *Proceedings of the NAACL HLT 2010 Demonstration Session*. Association for Computational Linguistics, Los Angeles, California, pp. 41–44, 2010.

Scarton, C., Paetzold, G. H., and Specia, L. Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Elsevier Inc., Miyazaki, Japan, pp. 3504–3510, 2019.

Shardlow, M. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications* 4 (1): 58–70, 2014.

Siddharthan, A. A survey of research on text simplification. *ITL–International Journal of Applied Linguistics* 165 (2): 259–298, 2014.

Štajner, S., Calixto, I., and Saggion, H. Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. "INCOMA Ltd. Shoumen, BULGARIA", Hissar, Bulgaria, pp. 618–626, 2015.

Tajner, S. and Glava, G. Leveraging event-based semantics for automated text simplification. *Expert Systems with Applications* vol. 82, pp. 383 – 395, 2017.

Vu, T., Hu, B., Munkhdalai, T., and Yu, H. Sentence Simplification with Memory-Augmented Neural Networks, 2018.

Wang, T., Chen, P., Amaral, K., and Qiang, J. An experimental study of LSTM encoder-decoder model for text simplification, 2016.

B

Split-and-Rephrase in a Cross-Lingual Manner: a Complete Pipeline

Split-and-Rephrase in a Cross-Lingual Manner: a Complete Pipeline

Paulo Berlanga Neto and **Evandro Eduardo Seron Ruiz**
Department of Computer Science and Mathematics
Faculty of Philosophy, Sciences and Letters at Ribeirão Preto
University of São Paulo (USP)
{pauloberlanga,evandro}@usp.br

Abstract

Split-and-rephrase is a challenging task that promotes the transformation of a given complex input sentence into multiple shorter sentences retaining equivalent meaning. This rewriting approach conceptualizes that shorter sentences benefit human readers and improve NLP downstream tasks attending as a preprocessing step. This work presents a complete pipeline capable of performing the split-and-rephrase method in a cross-lingual manner. We trained sequence-to-sequence neural models as from English corpora and applied them to predict the transformations in English and Brazilian Portuguese sentences jointly with BERT's masked language modeling. Contrary to traditional approaches that seek training models with extensive vocabularies, we present a non-trivial way to construct symbolic ones generalized solely by grammatical classes (POS tags) and their respective recurrences, reducing the amount of necessary training data. This pipeline contribution showed competitive results encouraging the expansion of the method to languages other than English.

1 Introduction

Text Simplification (TS) is the process of modifying natural language to reduce complexity and improve both readability and understandability (Shardlow, 2014). A simplified vocabulary or a simplified text structure can benefit people with limited language skills, such as those with low education levels, children, non-native speakers, and individuals with learning impairments (e.g., autism, dyslexia, or aphasia) (Štajner et al., 2015; Guo et al., 2018). Furthermore, when applied as a preprocessing step, TS may also improve the performance of several natural language processing (NLP) tasks, such as parsing, machine translation, semantic role labeling, text summarization, information extraction, among others (Niklaus et al., 2019a; Štajner and Popović, 2019).

This bottle was used until 2002 **when it** was dropped in favor of a traditional bottle .

This bottle was used until 2002 **. It** was dropped in favor of a traditional bottle .

He sought medical care in Rome **, but it** was unsuccessful **, and he** died at the age of 42 .

He sought medical care in Rome **. It** was unsuccessful **. He** died at the age of 42 .

Figure 1: Basic split-and-rephrase examples highlighting the transformations promoted by the split action.

Most work on TS has concentrated on analyzing specific characteristics at the sentence level, fashioning the task of sentence simplification (SS) (Alva-Manchego et al., 2020). SS applications aim to identify and solve two main aspects: lexical complexity, which refers to difficult words or expressions in the text (e.g., non-frequent words, specific terminologies, foreign words, etc.) (Štajner et al., 2020; Narayan and Gardent, 2014); and syntactic complexity, which refers to the length of the sentences and their grammatical complexities (e.g., number of subordinate or coordinate clauses, unusual sentence structures, depth of the syntactic tree, among others) (Štajner et al., 2020; Rebello et al., 2019).

Split-and-rephrase, proposed by Narayan and co-authors (Narayan et al., 2017), is a novel sentence simplification task that has attracted much research interest in the NLP field. Its goal is to split and rephrase a complex input sentence into shorter sentences that retain equivalent meaning (see examples in Figure 1). Neither deletion nor lexical/phrasal simplification is intended. The core of this process is to properly make the syntactic transformations required by the split action (e.g., turn a relative clause into a main clause).

This work innovates from previous split-and-rephrase methods. We present a complete pipeline capable of performing the split-and-rephrase challenge by combining trained sequence-to-sequence neural models that rely on symbolic vocabularies accompanied by BERT's masked language modeling. The main contribution is to construct a cross-lingual solution that deals both with English and Portuguese sentences. In addition, we enhanced a preliminary work (Berlanga et al., 2020) promoting analysis against complete reference test sets and comparing results to similar models/pipelines. To the best of our knowledge, this is the first complete pipeline to address split-and-rephrase in a cross-lingual manner, encouraging the expansion of the method to languages other than English.

2 Related Work

As discussed by Narayan and colleagues (Narayan et al., 2017), split-and-rephrase method must be distinguished from other sentence rewriting tasks, such as sentence compression, sentence fusion, and sentence paraphrasing. Furthermore, in contrast to the conventional sentence simplification task, split-and-rephrase does not entail loss of information, thus targeting the meaning preservation despite the split behavior (Alva-Manchego et al., 2020).

In an observational study, Gasperin and colleagues (Gasperin et al., 2009) stated that sentence splitting was the most frequent syntactic simplification operation used by an annotator when creating simplified texts. Among the techniques to perform the transformations required by sentence splitting, Niklaus et al. (Niklaus et al., 2019a) segregates them into three classes: (a) *Syntax-driven rule-based approaches* that use a set of hand-written rules to detect points where sentences may be split (Siddharthan and Mandya, 2014; Ferrés et al., 2016); (b) *Semantic parsing based approaches* that aim to decompose sentences into minimal semantic units that may be split into individual output sentences (Narayan and Gardent, 2014; Sulem et al., 2018); and (c) *Data-driven approaches* where the splitting point and transformations are learned automatically from training in aligned corpora of complex-simple sentences (Narayan et al., 2017; Aharoni and Goldberg, 2018).

Concerning split-and-rephrase previous works, Narayan et al. (Narayan et al., 2017) recently presented data-driven baseline models to help with some insights about the task, together with the WebSplit benchmark corpus. After that, Aharoni and Goldberg (Aharoni and Goldberg, 2018) established more robust baselines augmenting sequence-to-sequence neural models with copy-mechanism (Gu et al., 2016), and also released an updated version of WebSplit to reduce overlap in the data splits. Given the small vocabulary and the unnatural linguistic expressions present in WebSplit corpora, Botha et al. (Botha et al., 2018) compiled the WikiSplit corpus reuniting more than one million naturally occurring sentence rewrites obtained from mining English Wikipedia's edit history. Later, Niklaus et al. (Niklaus et al., 2019b) constructed the MinWikiSplit corpus running DisSim framework (Niklaus et al., 2019a) over the WikiSplit data and applied a set of 35 hand-written transformation rules to decompose source sentences in more split simplified counterparts.

As for the Portuguese language's split-and-rephrase task, based on the literature surveyed, we found no specific corpus built for this purpose. However, Leal et al. (Leal et al., 2018) made available the PorSimplesSent data set, a Brazilian Portuguese corpus to study sentence readability assessment, which we incorporated into this work to further test our pipeline.

3 Methodology and Data

We define the split-and-rephrase task as follows. Given a complex sentence C, the goal is to produce a simplified text T consisting of a sequence of sentences $T_1, T_2, \ldots, T_n$, $n \geq 2$, in such a way that T preserves the meaning of C.

In this section we specify the details about the implementation of our proposed pipeline and all the above mentioned split-and-rephrase corpora employed in this work.

3.1 Pipeline Specification

Our complete pipeline is composed of two main elements: (1) one trained sequence-to-sequence neural model that relies on a given custom symbolic vocabulary explained ahead; and (2) the BERT's masked language modeling. The overview of the pipeline is illustrated in Figure 2. Below we present these elements and how they are integrated.

Sequence-to-sequence neural models Our constructed models were based on the conventional encoder-decoder architecture composed of Gated Recurrent Unit (GRU) neural networks with attention mechanism (Bahdanau et al., 2014; Cho et al.,

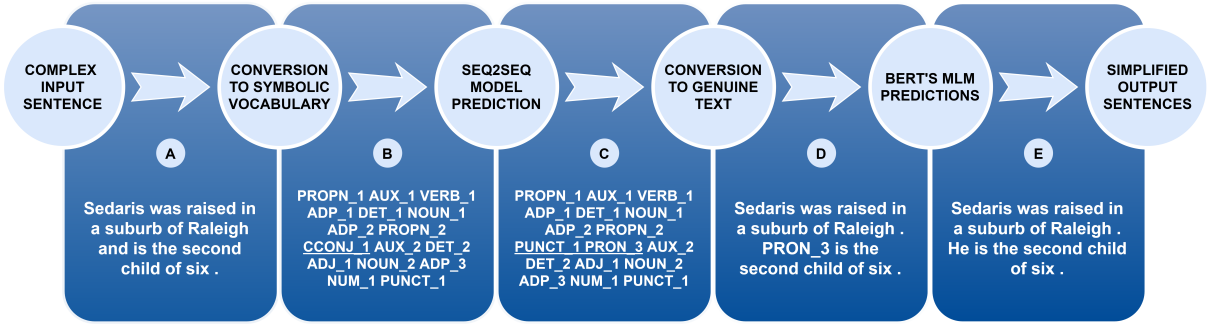


Figure 2: Illustration of our complete pipeline. To perform a prediction in the pipeline, the complex input sentence (A) passes through a preprocessing step to convert the text into symbolic vocabulary (B). This converted symbolic sequence is given to a sequence-to-sequence neural model that produces an output based on the learned knowledge on how to split such items (C). The model's output symbolic sequence is then reconverted to genuine text (D) and fed into BERT's masked language modeling to generate the simplified output sentences filling eventual gaps (E).

2014). Such mechanism makes it possible to establish references at particular points in original sequences, and enable the transmission of these instances to the decoder outputs. This approach is known to be an appropriate strategy for training models in aligned corpora and has shown excellent results for text-to-text NLP tasks (Raffel et al., 2019). The attention layer is connected to the encoder-decoder GRU layers, both composed of 100 units. We employed a batch size of 200 and the training process lasts 10 epochs in *Training Setup 1* and 40 epochs in *Training Setup 2*. These two distinct setups are further discussed in Section 4. We used categorical-cross entropy loss function and applied Adam optimization algorithm (Kingma and Ba, 2014) to update the networks' weight iteratively.

Symbolic vocabulary Contrary to traditional approaches that seek training models with extensive genuine vocabularies, we feed our sequence-to-sequence neural models with custom symbolic ones generalized uniquely by the concatenation of grammatical classes (POS tags) and their respective recurrences (indexes) observed in the aligned sentence pairs from the training data sets. The wildcard character '$\star$' is used for padding. The custom implementation to build such vocabulary is illustrated in detail in Figure 3. We found this strategy drastically reduced the vocabulary size to only a few items optimizing training process times.

This symbolic vocabulary approach is the key factor that enables our models to work in a cross-lingual manner: instead of dealing with genuine texts, they are capable to understand items in common gained from sentences of different languages, namely English and Portuguese, given that features such as grammatical classes are standard across these languages (Stodden and Kallmeyer, 2020). In addition, due to the existence of syntax-based patterns behind splitting in both languages, the specific knowledge on how to split the sentences may be captured accordingly thanks to the sequential observation nature of sequence-to-sequence neural models with attention mechanism.

BERT's masked language modeling Since our models are trained with alignments of symbolic sequences (such the example in Figure 3), they may predict symbolic sequences of items that need to be reconverted to genuine texts. But such predicted items are not always present in the complex input sentence to be converted back (see example in Figure 4). To fill this gap, we employed BERT's masked language modeling (MLM) (Devlin et al., 2018). BERT is proposed to train deep bidirectional language representations based on the Transformer architecture (Vaswani et al., 2017). Instead of predicting the next word in a sequence given the history, MLM predicts missing tokens in a sequence given its left and right context (Qiang et al., 2020). For the English language, we adopted the pre-trained model on English Wikipedia and Book Corpus. For the Brazilian Portuguese language, we employed the large trained model from BERTimbau work (Souza et al., 2020).

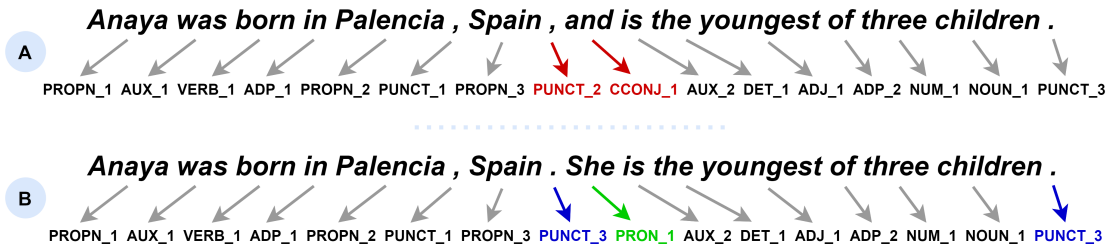


Figure 3: Examples of items collected to compose our symbolic vocabularies. Each token from the complex side of the alignment (A) is converted to a symbol formed by its respective POS tag and an index equivalent to its order of appearance in the sentence. These same symbols are then assigned to the simplified side of the alignment (B) considering the new positions of each token now allowing repetitions (in blue) and omissions (in red), together with new symbols likewise created for possible new tokens (in green). We made use of the Spacy POS tagger[1].

The building was then turned into a railway heritage centre in 1979 by the Butetown Historic Railway Society .
ADP_6 1994 **PUNCT_2** **PRON_1** started to run steam hauled passenger services up 500 m of track .

The building was then turned into a railway heritage centre in 1979 by the Butetown Historic Railway Society .
In 1994 **the** **railway** started to run steam hauled passenger services up 500 m of track .

Figure 4: BERT's MLM application example. In the first block, the highlighted items could not be reconverted to genuine text, harming the meaning of the output. To use BERT's MLM, we replace such items one at a time with the <mask> symbol while executing the predictions to fill the gaps. The second block illustrates the final output after the complete execution.

More specifically, given an output sequence of simplified sentences S, for each item i_n not reconverted from the complex input sentence C, we mask i_n on S using the <mask> symbol and feed S into MLM. MLM then considers the context of S to generate a ranking of five candidate tokens $c_1, c_2, \ldots, c_n$, $n = 5$. Following the ranking order (1 to 5), we check each c_n candidate's existence in C, accepting the first one encountered as the chosen token to fill the mask. If none of the candidates are present in C, the first candidate token c_1 is chosen from the ranking to fill the mask.

3.2 Data

The five different corpora involved in this work are composed of aligned complex-simple counterparts (non-split and split sentences), therefore ideal for training sequence-to-sequence neural models. We present them as follows.

WebSplit v0.1 Narayan et al. (Narayan et al., 2017) launched this corpus as the first data set to address the split-and-rephrase task. It is composed of 1,100,166 sentences written from RDF tuples. Due to the fact that one single complex sentence may map to a set of S_n structurally simplified references, the actual number of distinct complex sentences, $|C|$, is in the order of 4,5K;

WebSplit AG18 Aharoni & Goldberg (Aharoni and Goldberg, 2018) arguing they could achieve more robust results from their split-and-rephrase models, proposed a new train-development-test data split corpus. They randomly divided the distinct complex sentences from the original WebSplit corpus across the TDT sets to ensure that every possible RDF relation is represented in the training set, and every RDF triplet is conferred in only one of the splits;

WikiSplit Botha and colleagues (Botha et al., 2018) introduced this corpus presenting a language-agnostic method for extracting split-and-rephrase rewrites from Wikipedia edit histories. Each single complex sentence maps to a single simplified reference containing only one split. Compared to WebSplit versions, this data set has a more rich and varied vocabulary over naturally expressed sentences, despite being slightly noisy. The authors showed that models trained on this data set produced dramatically better results;

[1] https://spacy.io/api/tagger/

Corpus	Training set	Dev. set	Test set
WebSplit v0.1 (Narayan et al., 2017)	-	554	554
WebSplit AG18 (Aharoni and Goldberg, 2018)	-	535	503
WikiSplit (Botha et al., 2018)	989,944	5,000	5,000
MinWikiSplit (Niklaus et al., 2019b)	203,309	-	-
PorSimplesSent (Leal et al., 2018)	-	-	719

Table 1: Number of involved alignments in this work considering distinct complex sentences.

MinWikiSplit This corpus is composed of 203K sentences whose referred simplified references are composed of shorter, syntactically simplified counterparts. As they specify, these are clauses with a 'minimal semantic unit that cannot be further decomposed into meaningful propositions' (Niklaus et al., 2019b). For this reason, the main contribution of this corpus is to possibly enable models to learn to perform more than one single split per complex input sentence. The authors did not state any division in train-development-test sets;

PorSimplesSent This is a corpus for sentence-based readability assessment in Portuguese. It is constructed from the PorSimples text simplification corpus (Caseli et al., 2009) and combines three levels of simplifications: *from Original to Natural*; *from Natural to Strong*; and *from Original to Strong* pairs (Leal et al., 2018). In this work, we employed the specific version of *from Natural to Strong* pairs that, in our view, better reflects a split-and-rephrase corpus. We selected only pairs with splits in the simplified side of the alignments, extracting 719 sentence pairs to test the pipeline in Brazilian Portuguese language.

For training purposes, we used both WikiSplit and MinWikiSplit training sets as they contain more rich and varied vocabulary with diverse syntax (see Section 4). The validations throughout implementation were performed using WebSplit v0.1, WebSplit AG18, and WikiSplit development sets. At last the results were obtained from WebSplit v0.1, WebSplit AG18, WikiSplit test sets, and PorSimplesSent (see Section 4.1). Table 1 summarizes the number of involved alignments from each corpus/set considering distinct complex sentences.

4 Experiments

We assembled two different training setups concerning the sequence-to-sequence neural models attending different corpora as follows.

Training Setup 1 From WikiSplit (Botha et al., 2018) training corpus, we selected aligned sentence pairs formed only by alphanumerical characters, commas, periods and whitespaces, eliminating any foreign/special characters as this corpus is slightly noisy as admitted by the authors[2]. This cut extracted 485,120 alignments, consolidating the training set for this first setup. We then executed our aforementioned custom implementation to construct the symbolic vocabulary and obtained 247 different items to train the first model;

Training Setup 2 From MinWikiSplit (Niklaus et al., 2019b) corpus, we first established a limit to select aligned sentences with a maximum length of 100 tokens, due to the fact that this corpus has few long sentences that would lead to long padding. This first cut extracted 197,496 alignments. We then repeated the prior setup selecting aligned sentence pairs formed only by alphanumerical characters, commas, periods and whitespaces, finally consolidating a training set of 122,104 alignments. The symbolic vocabulary obtained by our custom implementation was composed of 230 different items to train the second model.

4.1 Results

Following Narayan et al. (Narayan et al., 2017), Aharoni and Goldberg (Aharoni and Goldberg, 2018) and Botha et al. (Botha et al., 2018) reference works, we report the results in sentence-level through BLEU (Papineni et al., 2002), BiLingual Evaluation Understudy, which is a primarily known metric borrowed from machine translation. It calculates modified n-gram precision as follows: (i)

[2]Despite this training selection, the final predicted sentences by the pipeline can normally still have special characters assigned by the reconversion process.

count the maximum number of times that an n-gram occurs in any of the references; (ii) clip the total count of each candidate n-gram by its maximum reference count; and (iii) add these clipped counts up, and divide by the total (unclipped) number of candidate words (Alva-Manchego et al., 2020). Also following reference works, we report the average number of simplified output sentences per complex input sentence (#S/C); and the average number of tokens per simplified output sentence (#T/S). Lastly, following Niklaus et al. (Niklaus et al., 2019b) we report the percentage of simplified output sentences that were totally copied from the complex input sentence without any modification (%SAME).[3]

Table 2 reports the obtained results against the full test sets when performing the complete pipeline with both trained models, considering the aforementioned different setups. Our best BLEU score was obtained with the model built by the *Training setup 1* in the WikiSplit test set, closely followed by the score in the PorSimplesSent data. We also highlight the #S/C and #T/S features obtained with model from *Training setup 2*, pointing that this fashion attempts to split complex input sentences into shorter ones than the model from *Training setup 1*. The %SAME column values in turn illustrate our proposal's low conservatism, tending to virtually intercept all the input sentences to perform the split-and-rephrase rewriting transformations (see detailed discussion in Section 5).

As expressed by the scores in Table 3 alongside other approaches scores (Copy512 and DisSim) in the WikiSplit test set, we established our pipeline as a competitive method. Copy512 is the strongest baseline reported by Aharoni and Goldberg (Aharoni and Goldberg, 2018) work. It is a sequence-to-sequence neural model augmented with a copy-mechanism (Gu et al., 2016) that bias the model towards copying tokens from the complex input sentences, taking into account that many of them should appear in the simplified output sentences. DisSim framework, by Niklaus et al. (Niklaus et al., 2019a), is a recursive sentence splitting approach, that applies a set of 35 hand-written rules to decompose a wide range of linguistic constructs, more oriented to generate simple and regular structures to support downstream semantic applications and faster generalization in machine learning tasks.

[3]The metrics/quality estimation features were achieved with EASSE package (Alva-Manchego et al., 2019).

To encourage further research analysis, our complete logs containing all the predictions from *Training setup 1* in the WikiSplit test set are publicly available[4]. One may notice many sentences achieved meaning preservation and perfect matches against the expected references.

5 Discussion

To achieve a detailed analysis, we manually inspected some of the predictions from the pipeline with the two built models bringing some examples to help explain the scores illustrated in Tables 2 and 3. These extracted examples are in Table 4 and show general patterns with some of the exciting behaviors produced by our method.

In *Example 1*, the same complex input sentence is transformed into different simplified outputs according to their training setups: *Output 1* performed a single split whereas *Output 2* performed two splits. This different behavior explains the higher numbers in the #S/C column and the lower numbers in the #T/S column from *Training setup 2*. These two measures confirmed the hypothesis that models trained in MinWikiSplit might capture the tendency to split source sentences into multiple output ones. Such multiple sentences may not be good for humans readers, but may benefit NLP downstream tasks.

In *Example 2*, even though both setups generated perfect outputs in terms of meaning preservation, only the *Output 2* achieved maximum BLEU score since it is the unique that matched perfectly against one of the references. This brings the evidence that BLEU requires high-quality data to produce more precise outcomes, ideally with multiple correct references (Martin et al., 2019). Another limitation from BLEU is the low correlation with simplicity when sentence splitting is performed, but it still holds the high correlation with human assessments of grammaticality and meaning preservation (Alva-Manchego et al., 2020).

In *Example 3*, we note interesting contrasts produced by the models from the distinct training setups. While *Output 1* retained the same structure from the complex input sentence, *Output 2* promoted the reordering of the words preserving equivalent meaning and showing low conservatism. The only little mistake is observed by the repetition of word "was" in the *Output 2*.

[4]https://github.com/pauloberlanga/split-and-rephrase-pipeline/

Training setup 1	BLEU	#S/C	#T/S	%SAME
WebSplit v0.1 (Test set)	58.34	2.17	12.52	0.014
WebSplit AG18 (Test set)	60.01	2.21	10.70	0.019
WikiSplit (Test set)	**68.92**	2.05	20.45	0.071
PorSimplesSent	65.00	2.06	14.95	**0**
Training setup 2	**BLEU**	**#S/C**	**#T/S**	**%SAME**
WebSplit v0.1 (Test set)	57.86	3.12	10.34	0.043
WebSplit AG18 (Test set)	58.45	3.17	**9.03**	0.033
WikiSplit (Test set)	44.65	**5.81**	11.78	0.011
PorSimplesSent	49.52	4.61	10.07	**0**

Table 2: Results obtained by the pipeline when applying both models built from the training setups[5].

Models/pipelines	BLEU	#S/C	#T/S	%SAME
Training setup 1	68.92	2.05	20.45	0.07
Training setup 2	44.65	**5.81**	**11.78**	**0.01**
Copy512 (Aharoni and Goldberg, 2018)	**76.42**	2.08	16.55	13.30
DisSim (Niklaus et al., 2019a)	51.96	4.09	11.91	0.76

Table 3: Scores alongside other approaches in the WikiSplit test set.

Lastly, *Example 4* illustrates the pipeline working in a cross-lingual manner. *Output 1* produced a pronoun "Ele" (He) instead of repeating "O projeto Gemini" (The Gemini project), as observed in *Output 2*. It is exactly the same behavior seen in the English *Example 2* reflected for Brazilian Portuguese sentences. Recent studies that analyze eye movements of human readers interestingly reveal that they quickly retrieve information upon finding pronouns when referred to a close syntactic antecedent (Rebello et al., 2019).

Our detailed inspection together with the prediction logs confirmed that the pipeline could split complex input sentences into shorter simplified ones, often preserving equivalent meaning successfully. More than that, it showed ability to perform equivalent syntax transformations for different languages (English and Portuguese). On the other hand, some of the predictions reveal common mistakes from sequence-to-sequence models, such as repetition or omission of tokens and "hallucination" of new unwanted information. Another limiting factor is the noise from *unsupported* or *missing statements* observed in the referred test data sets. The low quality references eventually harmed the BLEU scores in those cases.

6 Conclusion

Split-and-rephrase task conceptualizes that shorter sentences are generally better processed by humans and by NLP downstream applications. We presented a complete pipeline for the split-and-rephrase method that attends in a cross-lingual manner English and Portuguese languages, by integrating sequence-to-sequence neural models and BERT's masked language modeling. In contrast to conventional approaches, we train models making use of symbolic vocabularies defined by a custom implementation. This approach speeds up the training process and enables the models to acquire specific knowledge on how to split symbolic sequences, then demanding only a little step to convert them back to genuine texts in respective languages. Furthermore, the pipeline is capable to foster new words to rewrite the complex input sentence, thanks to BERT's MLM predictions. Unlike most previous works on split-and-rephrase, we employed the four state-of-the-art corpora for the task and also a Brazilian Portuguese corpus, showing competitive results to equivalent approaches.

As future work, we plan to exploit our pipeline in more languages. We should also inspect the effectiveness of the Transformer architecture in replacement of the sequence-to-sequence models. Moreover, we intend to promote an extrinsic evaluation of the benefits of the split-and-rephrase method in NLP downstream applications.

[5]We refrain from report SARI and SAMSA scores. The first metric is more reliable to evaluate lexical (not structural) simplicity, and the second heavily relies on linguistic resources making the application in Portuguese language unfeasible.

Example 1 (from WikiSplit data set)	
Input	Gavin confessed to the murder of George Pollard and was held in the Round House until he was hung on the 6th April 1844 , his body was buried south of the Round House .
Ref.	Gavin confessed to the murder of George Pollard and was held in the Round House until he was hanged on 6 April 1844 . His body was buried south of the Round House .
Output 1	Gavin confessed to the murder of George Pollard and was held in the Round House until he was hung on the 6th April 1844 . • His body was buried south of the Round House .
Output 2	Gavin confessed to the murder of George Pollard . • Gavin was held in the Round House until he was hung on the 6th April 1844 . • His body was buried south of the Round House .
Example 2 (from WebSplit v0.1 data set)	
Input	A.S. Livorno Calcio are managed by Christian Panucci who is attached to the club Genoa CFC .
Ref. 1	A.S. Livorno Calcio are managed by Christian Panucci . Christian Panucci is attached to the club Genoa CFC .
Ref. 2	A.S. Livorno Calcio is managed by Christian Panucci . Christian Panucci played football for Genoa C.F.C.
Ref. 3	A.S. Livorno Calcio are managed by Christian Panucci . Christian Panucci played football for Genoa C.F.C.
Ref. 4	A.S. Livorno Calcio is managed by Christian Panucci . Christian Panucci is attached to the club Genoa CFC .
Output 1	A.S. Livorno Calcio are managed by Christian Panucci . • He is attached to the club Genoa CFC .
Output 2	A.S. Livorno Calcio are managed by Christian Panucci . • Christian Panucci is attached to the club Genoa CFC .
Example 3 (from WikiSplit data set)	
Input	Born in Huzhou , Zhejiang , Qian was trained in traditional Chinese philology , and was a student of Zhang Binglin .
Ref.	Born in Huzhou , Zhejiang , Qian was trained in traditional Chinese philology . He was a student of Zhang Binglin .
Output 1	Born in Huzhou , Zhejiang , Qian was trained in traditional Chinese philology . • Qian was a student of Zhang Binglin .
Output 2	Qian was born in Huzhou , Zhejiang . • Qian was trained in traditional Chinese philology . • Qian was was a student of Zhang Binglin .
Example 4 (from PorSimplesSent data set)	
Input	O projeto Gemini é resultado de uma associação de sete países e envolve a construção de dois telescópios com um espelho de oito metros de diâmetro.
Ref.	O projeto Gemini é resultado de uma associação de sete países. O projeto Gemini envolve a construção de dois telescópios com um espelho de oito metros de diâmetro.
Output 1	O projeto Gemini é resultado de uma associação de sete países . • Ele envolve a construção de dois telescópios com um espelho de oito metros de diâmetro .
Output 2	O projeto Gemini é resultado de uma associação de sete países . • O projeto Gemini envolve a construção de dois telescópios com um espelho de oito metros de diâmetro .

Table 4: Examples predicted by the pipeline with highlighted splitting points.

References

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and a stronger baseline. *arXiv preprint arXiv:1805.01035*.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. *arXiv preprint arXiv:1908.04567*.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

P. N. Berlanga, E. Y. Okano, and E. E. S. Ruiz. 2020. Experimenting Sentence Split-and-Rephrase Using Part-of-Speech Labels. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 169–176, Porto Alegre, RS, Brasil. SBC.

Jan A Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. *arXiv preprint arXiv:1808.09468*.

Helena M. Caseli, Tiago F. Pereira, Lúcia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Sandra M. Aluísio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. In *Advances in Computational Linguistics, Research in Computer Science, (CICLing-2009), volume 41*, pages 59—-70.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Ferrés, Montserrat Marimon, Horacio Saggion, et al. 2016. YATS: yet another text simplifier. In *International Conference on Applications of Natural Language to Information Systems*, pages 335–342. Springer.

Caroline Gasperin, Lucia Specia, Tiago Pereira, and Sandra Aluísio. 2009. Learning when to simplify sentences for natural text simplification. *Proceedings of ENIA*, pages 809–818.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. *arXiv preprint arXiv:1806.07304*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Antoine Bordes, Éric Villemonte de La Clergerie, and Benoît Sagot. 2019. Reference-less quality estimation of text simplification systems. *arXiv preprint arXiv:1901.10746*.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *The 52nd annual meeting of the association for computational linguistics*, pages 435–445.

Shashi Narayan, Claire Gardent, Shay B Cohen, and Anastasia Shimorina. 2017. Split and rephrase. *arXiv preprint arXiv:1707.06971*.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019a. Transforming complex sentences into a semantic hierarchy. *arXiv preprint arXiv:1906.01038*.

Christina Niklaus, André Freitas, and Siegfried Handschuh. 2019b. MinWikiSplit: A Sentence Splitting Corpus with Minimal Propositions. *arXiv preprint arXiv:1909.12131*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. LSBert: A Simple Framework for Lexical Simplification. *arXiv preprint arXiv:2006.14939*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Beatriz Meira Rebello, Giovanna Lima dos Santos, Clara Regina Brandão de Ávila, and Adriana de Souza Batista Kida. 2019. Efeito da simplificação sintática sobre a compreensão de leitura de crianças do ensino fundamental. *Audiology-Communication Research*, 24.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Advaith Siddharthan and Angrosh Mandya. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 722–731.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417. Springer.

Sanja Štajner, Iacer Calixto, and Horacio Saggion. 2015. Automatic text simplification for spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 618–626.

Sanja Štajner, Sergiu Nisioi, and Ioana Hulpuș. 2020. CoCo: A tool for automatically assessing conceptual complexity of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7179–7186.

Sanja Štajner and Maja Popović. 2019. Automated Text Simplification as a Preprocessing Step for Machine Translation into an Under-resourced Language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1141–1150.

Regina Stodden and Laura Kallmeyer. 2020. A multilingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 77–84, Marseille, France. European Language Resources Association.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and effective text simplification using semantic and neural methods. *arXiv preprint arXiv:1810.05104*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.