Proposta recebida em Julho 2023 e aceite para publicação em Novembro 2023.

Transferência de estilo textual arbitrário em português

Arbitrary Portuguese text style transfer

Pablo Botton da Costa
Universidade de São Paulo (EACH-USP)

Ivandré Paraboni
Universidade de São Paulo (EACH-USP)

Resumo

Na Geração automática de língua natural, modelos de transferência de estilo textual arbitrário objetivam a reescrita de um texto usando qualquer novo conjunto de características estilísticas desejado. Em se tratando do idioma português, entretanto, observa-se que os recursos linguístico-computacionais necessários para o desenvolvimento de modelos deste tipo ainda são consideravelmente escassos em comparação à língua inglesa. Assim, como um primeiro passo em direção ao desenvolvimento de métodos avançados deste tipo, o presente trabalho investiga a questão da transferência de estilo textual arbitrário com o uso de paráfrases em português, combinadas ao uso de modelos neurais construídos a partir de arquiteturas do tipo *sequência-para-sequência* e por refinamento de grandes modelos de língua. Além dos modelos de reescrita textuais propriamente ditos, o estudo apresenta também recursos inéditos para a tarefa na forma de um córpus de paráfrases e de um modelo de *embeddings* validado nas tarefas de similaridade e simplificação sentencial, com resultados comparáveis ao estado da arte.

Palavras chave

geração de língua natural, transferência de estilo arbitrário, paráfrases, sequência-para-sequência, grandes modelos de língua

Abstract

In Automatic Natural Language Generation, arbitrary style transfer models aim to rewrite a text using any desired new set of stylistic features. In the case of the Portuguese language, however, we notice that the resources required for the development of models of this type are still considerably scarce compared to those dedicated to the English language. Thus, as a first step towards the development of advanced methods of this kind, the present work investigates the issue of arbitrary style transfer with the aid of paraphrases in Portuguese, combined with the use of neural models built from sequence-to-sequence architectures and by refining a number of large language models. In addition to the textual rewriting models themselves, the study also presents novel resources for the task in the form of a corpus of paraphrases and a model of embeddings validated in both sentence similarity and simplification tasks, with results comparable to the state of the art.

Keywords

natural language generation, arbitrary style transfer, paraphrases, sequence-to-sequence, large language models

1. Introdução

O uso de métodos de aprendizado neural tem se tornado prevalente em diversas áreas do Processamento de Língua Natural (PLN), incluindo a Geração de Língua Natural (GLN) (Gatt & Krahmer, 2018; Naseem et al., 2021; Dong et al., 2022). Dentre múltiplas aplicações atuais deste tipo, destacamos no presente estudo a tarefa computacional de *transferência de estilo* (Jin et al., 2022; Hu et al., 2022), que consiste em reescrever um texto sem alterar seu conteúdo semântico, mas em um estilo[1] diferente do original. Modelos de transferência de estilo têm sido aplicados em cenários específicos como a transferência de sentimentos (Hu et al., 2017; Shen et al., 2017; Li et al., 2018; Luo et al., 2019), formalidade (Xu et al., 2012; Wang et al., 2019) e outros, de modo geral fazendo uso de um córpus alinhado de pares de frases em um estilo-origem e um estilo-alvo de interesse.

Mais recentemente, a tarefa de transferência de estilo passou a ser investigada também em cenários onde os estilos-alvo da geração são *arbitrários* (Krishna et al., 2020; Reif et al., 2022; Suzgun et al., 2022), ou seja, utilizando-se de modelos que permitem a reescrita de textos com qualquer novo conjunto de características estilísticas desejado. Abordagens deste tipo dis-

[1] Considerando-se aqui a definição de estilo proposta em Jin et al. (2022), que engloba qualquer atributo que varia de um texto origem para um texto alvo, e não apenas no sentido estritamente linguístico do termo.

DOI: 10.21814/lm.15.2.410

pensam o uso de córpus previamente alinhados, que são substituídos pelo uso de algum tipo de alinhamento parcial pela geração do tipo *sequência-para-sequência* (Riley et al., 2021), pelo refinamento de grandes modelos de língua pré-treinados (LLMs), ou por meio de paráfrases (Krishna et al., 2020). Esta última opção será também o foco do presente trabalho.

De forma mais específica, o presente estudo aborda a tarefa computacional de geração textual de estilo baseada na noção de *quase-paráfrases*, definidas em (Bhagat & Hovy, 2013) como um par de frases ou sentenças que transmitem aproximadamente o mesmo significado usando palavras diferentes. As sentenças 1 e 2 a seguir, que podem ser vistas como sendo duas formas distintas (e.g., mais ou menos formal) de expressar uma mesma ideia geral, são exemplos de quase-paráfrases.

1 Foi informado pela *instituição* que seus ônibus *teriam capacidade para* 40 alunos cada.

2 A *escola falou* que seus ônibus *acomodam* 40 alunos cada.

Considerando-se que um conjunto de exemplos em um estilo fonte (como 1 acima) e outro estilo alvo (como 2) esteja disponível em forma de um córpus alinhado de paráfrases, métodos de aprendizado neural podem ser aplicados para identificar as relações de paráfrase entre seus termos, que podem então ser utilizadas na reescrita de uma sentença qualquer mesmo que esta não ocorra no conjunto de treino. Em outras palavras, modelos deste tipo, que efetivamente implementam a tarefa computacional de transferência de estilo, são capazes de reescrever um texto arbitrário qualquer no estilo fonte para o estilo alvo, como no caso da reescrita de um texto formal como o exemplo 1 acima em uma versão mais informal, como o exemplo 2.

A transferência de estilo baseada em paráfrases é de grande interesse científico, e possui uma ampla gama de aplicações possíveis (Jin et al., 2022). No entanto, no caso específico da língua portuguesa, observa-se que tanto os métodos de Geração de Língua Natural como os recursos linguístico-computacionais necessários para esse fim ainda são consideravelmente escassos em comparação com o estado da arte disponível para a língua inglesa. Assim, como um primeiro passo em direção ao desenvolvimento de métodos avançados deste tipo, o presente trabalho objetiva investigar a questão da transferência de estilo textual arbitrário por meio de paráfrases em português, apresentando recursos inéditos para a tarefa e métodos de transferência de estilo treinados e avaliados com base nestes recursos.

As principais contribuições previstas neste estudo são as seguintes:

- Modelos computacionais de transferência de estilo textual arbitrário em português construídos a partir de arquiteturas do tipo *sequência-para-sequência* e por refinamento de grandes modelos de língua.
- Resultados de avaliação dos modelos propostos com uso de métricas automáticas e avaliação humana.
- Córpus de paráfrases alinhadas em nível sentencial.
- Modelo de *embeddings* de paráfrases para transferência de estilo.
- Resultados de avaliação destes recursos nas tarefas de simplificação e similaridade sentenciais, comparáveis ao estado da arte.

O restante deste artigo é organizado da seguinte forma. A seção 2 sumariza a pesquisa recente em transferência de estilo textual arbitrário para o idioma inglês. A seção 3 descreve a construção do córpus de paráfrases a ser utilizado como base para os modelos propostos. A seção 4 descreve a construção e validação de um modelo de *embeddings* de paráfrases tomando como exemplo as tarefas de simplificação e similaridade sentencial. Com base nestes recursos, a seção 5 apresenta nossa abordagem principal de reescrita sentencial para transferência de estilo. Finalmente, a seção 6 apresenta as conclusões do presente estudo e opções de trabalhos futuros.

2. Trabalhos relacionados

A tarefa de transferência de estilo textual arbitrário ainda é relativamente pouco explorada na pesquisa em GLN. A seguir revisamos brevemente os estudos em Krishna et al. (2020); Riley et al. (2021); Reif et al. (2022) e outros trabalhos a eles relacionados, que propõem a utilização de grandes modelos de língua como forma de se obter e gerar dados sintéticos em cenários em que haja poucos recursos disponíveis para a tarefa. Um levantamento mais detalhado da área é apresentado em Jin et al. (2022).

O trabalho em Krishna et al. (2020) propõe um *framework* de geração do tipo *texto-para-texto* dito universal, ou seja, para qualquer estilo, sem a necessidade prévia de dados alinhados. O *framework* de geração é composto de

duas etapas. A primeira é responsável pela conversão do texto de entrada em uma paráfrase, desempenhada por um modelo previamente treinado em um conjunto aprimorado de pares de paráfrases do córpus *PARANMT-50M* (Wieting et al., 2017). Os pares candidatos são então selecionados a partir do córpus com base em filtros com o objetivo de maximizar sua diversidade lexical e complexidade, e usados para o ajuste de pesos do modelo GPT-2. A segunda etapa consiste de um modelo de geração do tipo *sequência-para-sequência* treinado para inverter paráfrases em onze estilos. Por exemplo, um texto de entrada como "*I'd jump in there, no doubt*", reescrito em estilo de paráfrase durante a etapa anterior, seria convertido para o estilo original como "*No lie... I would jump in*". Cada *framework* é treinado individualmente, e a transferência de estilo é realizada por meio da substituição do módulo de inversão de paráfrases pelo módulo de inversão de paráfrases do estilo alvo desejado. Por exemplo, dado um texto de entrada e o objetivo de reescrevê-lo no estilo da língua inglesa moderna, o modelo gera a paráfrase dessa entrada, e a reescreve no estilo-alvo desejado com o uso respectivo modelo de inversão de paráfrase.

O modelo proposto, chamado STRAP, é avaliado com uso de cinco métricas automáticas e humanas a partir dos córpus supervisionados *Shakespeare* (Xu et al., 2012) e *Formality* (Rao & Tetreault, 2018). As métricas de avaliação automática utilizadas foram (1) a força da transferência de estilo (aferida por um classificador de estilos treinado a partir de textos dos córpus *Shakespeare* e *Formality*), (2) a semelhança de paráfrases (calculada como a semelhança de cosseno da representação distribuída dos pares de textos a partir dos pesos apresentados em Wieting et al. (2021)), (3) a fluência do texto gerado (computada por um modelo de classificação treinado a partir do córpus de aceitabilidade gramatical COLA, como em Warstadt et al. (2019)), (4) a média geométrica entre as métricas de força da transferência, similaridade de paráfrases e fluência dos textos, e (5) o produto entre as métricas força da transferência, similaridade de paráfrases e a fluência dos textos dividido pelo tamanho da sentença alvo. Para as métricas de avaliação humana, foi proposto a um grupo de juízes medir (1) a adequação semântica, (2) a dissimilaridade lexical, e (3) a semelhança estilística dos textos gerados pelos modelos, conforme os critérios do *crowdsourcing*.

O trabalho em Riley et al. (2021) propõe a adaptação do modelo T5 (Raffel et al., 2020) como um extrator de características para a tarefa de transferência de estilo, seguindo a abordagem em Biber & Conrad (2019) de treinamento de uma rede neural do tipo *denoising auto-encoder* condicionado a um vetor de estilo. Dado que o vetor de estilo original é desconhecido, a abordagem constrói de forma conjunta um modelo neural para induzir a representação de estilo do texto. Por exemplo, dado um texto origem e o objetivo de reescrevê-lo no estilo contrário ao sentimento original da frase, o modelo gera dois novos exemplos: o primeiro tem estilo similar ao da entrada, mas com palavras diferentes, e o segundo é um texto aleatório do conjunto de treinamento que pertence ao estilo-alvo do objetivo de geração. Estas três entradas – os dois exemplos e o texto origem — são então submetidos ao modelo para que sejam combinados de tal forma a reescrevê-los no estilo-alvo desejado.

O modelo obtido, chamado *TextSETTR*, é avaliado a partir da combinação entre métricas humanas e automáticas em relação ao conjunto de testes. Para a métrica de avaliação automática, optou-se pela utilização das métricas BLEU e força da transferência de estilo aferida por meio de um classificador treinado para identificar estilos. O modelo foi avaliado em três tarefas de transferência de estilo usando os córpus *Amazon Reviews*, *Shakespeare* e o córpus de textos bíblicos em Carlson et al. (2018). Dentre outros resultados, observa-se que modelo *TextSETTR* apresenta uma melhoria discreta em relação ao trabalho em Biber & Conrad (2019).

Finalmente, o trabalho em Reif et al. (2022) propõe o uso da técnica de engenharia de *prompts* em combinação com métodos de destilação de conhecimento a partir de grandes modelos de língua da família GPT (Brown et al., 2020) para a tarefa de transferência de estilo. Por exemplo, dado um texto de entrada e o objetivo de reescrevê-lo no estilo de poesia, e.g., parnasiana, o modelo é instruído com o *prompt* '*transforme o texto a seguir em estilo de poesia parnasiana* $\rightarrow x'$, resultando no texto x' reescrito no estilo-alvo desejado. O trabalho propõe a construção de cinco modelos: os *baselines GPT-3-ada*, *GPT-3-curie* e *GPT-3-davinci*, que não utilizam *prompts*; um modelo que utiliza uma rede neural do tipo *transformers Lambda* treinada em um córpus composto por 1.95B documentos de domínio público, e um modelo chamado *lambdaFinne* que utiliza a mesma arquitetura base do anterior, porém com os pesos ajustados em um córpus com curadoria de alta qualidade para o domínio conversacional.

A avaliação destes modelos foi realizada de duas formas. Uma considerou estilos atípicos ou sem padrão definido, correspondendo aos ajustes

mais frequentes feitos por usuários de uma ferramenta inteligente de edição de textos, e a outra considerou os estilos pré-definidos de sentimento (presente no córpus Yelp em Zhang et al. (2015)) e formalidade (do córpus GYAFC em Rao & Tetreault (2018)).

Os estudos de transferência de estilo aqui discutidos são todos dedicados ao idioma inglês, e fazem uso de recursos linguístico-computacionais ainda escassos para o idioma português, como córpus alinhados, *embeddings* de paráfrases e grandes modelos de língua. Estas observações foram levadas em conta na construção de recursos básicos deste tipo para o português, e no método de reescrita sentencial baseada em paráfrases a serem abordados pelo presente estudo.

3. Córpus de paráfrases PTPARANMT

O uso de paráfrases tem se destacado como um método robusto para reescrita de texto em um estilo-alvo diferente do seu estilo original (Krishna et al., 2020, 2022). Um estudo deste tipo requer, entretanto, um córpus paralelo contemplando textos alinhados em nível de sentenças, ou seja, um conjunto de textos origem a ser parafraseado, e um segundo conjunto de textos com o mesmo significado, porém com características lexicais e sintáticas modificadas. Em virtude da dificuldade de obtenção de um recurso linguístico deste tipo em português, e com qualidade e volumes adequados para a tarefa de transferência de estilo, optou-se por criar um novo córpus de paráfrases aos moldes de Wieting et al. (2017), aqui denominado *PTPARANMT*. A construção e avaliação deste conjunto de dados é o foco desta seção.

O córpus *PTPARANMT* foi criado a partir da tradução reversa, ou seja, traduzindo-se textos de um idioma para outro, e novamente para o idioma original como forma de provocar modificações léxicas e estruturais com pouca ou nenhuma perda de significado. Por exemplo, o texto "eu te amo" poderia ser traduzido para o inglês como "i love you," e então traduzido no sentido reverso para "te amo." Técnicas de tradução reversa têm sido aplicadas com sucesso em tarefas como tradução automática (Hoang et al., 2018), transferência de estilo (Krishna et al., 2020, 2022) e sumarização automática (Beddiar et al., 2021), além da própria construção de córpus a partir de recursos já disponíveis (Gonçalo Oliveira & Alves, 2021).

Foram tomados por base três córpus existentes (*Europarl*, *ParaCrawl* e *Tapaco*) utilizando-se das porções alinhadas Português-Inglês de cada um. A escolha deste par de idiomas foi motivada pela alta incidência de paráfrases de qualidade obtidas a partir destes córpus para fins de tradução automática em Wieting et al. (2017). O córpus *Europarl*, já utilizado em Wieting & Gimpel (2018) para geração de paráfrases, é constituído de transcrições das seções públicas do parlamento europeu em 21 idiomas, e possui 1.960.407 pares de sentenças alinhadas Português-Inglês. O córpus *ParaCrawl*, já utilizado em Bañón et al. (2020) para a tarefa de tradução automática, é constituído de textos provenientes de páginas da web em 42 idiomas, e possui 84.921.510 pares de sentenças alinhadas Português-Inglês. Por se tratar de um córpus extenso, foi realizada uma amostragem de 19 milhões de pares deste córpus. Finalmente, o córpus *Tapaco*, já utilizado em Shliazhko et al. (2022) na tarefa de identificação de paráfrases, é constituído de textos da base *Tatoeba*, construída por *crowdsourcing* em 73 idiomas, e possui 110.000 pares de sentenças alinhadas Português-Inglês.

A construção do córpus *PTPARANMT* pode ser dividida em duas etapas. Na primeira, foi gerada uma base de pares de textos alinhados em nível sentencial com uso de tradução reversa a partir dos três córpus de entrada. Na segunda, o conjunto de textos foi filtrado com base em um limite de confiança estimado por meio de avaliação humana, resultando no córpus final. Estas duas etapas são discutidas individualmente a seguir.

3.1. Base de textos alinhados

O córpus *PTPARANMT* consiste de dois conjuntos de textos em português alinhados em nível sentencial, aqui denominados Origem e Alvo. O conjunto Origem foi obtido a partir da tradução do original inglês para o português, e o conjunto Alvo foi obtido pela tradução dos textos originais em português para o inglês, e posteriormente traduzidos de volta para o português.

Para as etapas de tradução e tradução reversa, foi utilizado o *framework OPUS-MT* (Tiedemann & Thottingal, 2020). A tradução do conjunto Origem usou a versão do *framework* en-ROMANCE, e a etapa de tradução reversa usou as versões ROMANCE-en e en-ROMANCE para a tradução para o inglês e para o português, respectivamente.

A tradução e tradução reversa dos conjuntos Origem e Alvo geraram 21 milhões de pares de textos alinhados em português. A Tabela 1 sumariza as estatísticas descritivas deste conjunto de dados a partir da amostra dos pares presentes em cada uma das três origens textuais consideradas. Estas estatísticas descrevem o tamanho médio das sentenças, a métrica de paráfrases *para-score* calculada pela distância média de cosseno entre as *embeddings* sentenciais de forma relativa aos pares de paráfrases, cf. Wieting et al. (2017), o grau de sobreposição média de bigramas e trigramas entre pares de sentenças, escores BLEU (Post, 2018), distância de edição (DE) e quantidade total de sentenças.

Conforme ilustrado na Tabela 1, o córpus *Europarl* tem a melhor média para a métrica de similaridade de paráfrases. O córpus *Tapaco*, por outro lado, apresenta em média as menores frases e as menos diversas.

Dado que o conjunto de dados inclui textos provenientes da Internet, foram excluídas as sentenças com confiança $\leq 0,97$ para o classificador,[2] que determina se a linguagem resultante do processo é ou não português. Além disso, foram removidas as sentenças com menos de três palavras, ou que apresentavam erros como palavras repetidas ou reticências. A Tabela 2 apresenta exemplos de textos do conjunto Origem e Alvo apresentando relação de paráfrase do par de texto e a confiança do classificador da língua portuguesa em relação ao texto Origem.

3.2. Filtragem

Após a construção do conjunto de textos inicial, foi conduzida uma avaliação humana com o objetivo de reduzi-lo ao subconjunto de paráfrases de maior confiança, ou seja, eliminando-se tanto quanto possível o ruído existente. Para este fim, oito falantes nativos da língua portuguesa foram solicitados a avaliar 63 pares de paráfrases provenientes de uma amostragem obtida a partir de diferentes faixas da métrica *para-score*. Entretanto, dado que os textos podem conter diversos tipos de imperfeição (e.g., decorrente do córpus de origem ou do processo de tradução, dentre outras possibilidades), foi solicitado aos participantes que avaliassem cada um dos pares de frases quanto aos possíveis *erros* de paráfrase (ou seja, o quanto uma frase pode ser considerada como sendo uma paráfrase da sua contrapartida) e fluência.

[2]Language Detection Library for Java, `http://code.google.com/p/language-detection/`

A avaliação de erros usou uma versão adaptada das instruções apresentadas em Agirre et al. (2012) na qual os rótulos ('alto', 'médio' e 'baixo') originalmente propostos para avaliar a força da paráfrase ou fluência do texto foram usados de forma invertida para representar o grau de erro segundo estes dois critérios.

Os três graus de erros de paráfrase considerados foram os seguintes:

- erro *baixo*: as sentenças devem ter o mesmo significado, mas alguns detalhes sem importância podem diferir.
- erro *médio*: as sentenças devem ser aproximadamente equivalentes, com algumas informações importantes faltando ou diferem um pouco.
- erro *alto*: as sentenças não são equivalentes, mesmo que compartilhando pequenos detalhes.

De forma análoga, a avaliação de fluência considerou os três graus de erro a seguir:

- erro *baixo*: o texto não deve conter erros gramaticais.
- erro *médio*: o texto deve possuir um ou dois erros gramaticais.
- erro *alto*: o texto deve possuir mais de dois erros gramaticais, ou não soa natural em português.

Foi alcançada uma concordância entre anotadores de $0,89$ de acordo com o índice kappa de Cohen (Landis & Koch, 1977). A Tabela 3 apresenta os resultados da avaliação humana agrupados em intervalos de *para-score*, considerando-se as medidas da média da sobreposição de trigrama, o número de vezes que cada texto Origem ou Alvo recebeu um escore de fluência 0, 1 ou 2 (denominados Fluência.O e Fluência.A, respectivamente), e o número de vezes que cada texto Origem ou Alvo recebeu um escore de paráfrase 0, 1 ou 2.

Conforme pode ser observado na Tabela 3, os pares ruidosos estão, em sua maioria, confinados ao primeiro intervalo $(0,24, 0,631)$. Este intervalo compreende 19,23% de todos os pares que possuem uma forte relação de paráfrase. Além disso, nos dois intervalos mais altos, 84% dos pares possuem uma relação forte de paráfrases. Nos intervalos baixos, por outro lado, uma inspeção manual dos dados originais revelou principalmente erros de alinhamento em nível de sentença, o que também foi reportado no trabalho para a língua inglesa em Wieting & Gimpel

Origem	Tamanho	para-score	sobrepos.	BLEU	DE	Sentenças
Europarl	25,08	0,80	0,44	71,65	61,26	1,9 M
ParaCrawl	14,45	0,75	0,34	54,03	47,16	19,0 M
Tapaco	5,89	0,72	0,47	63,59	10,86	0,110 M

Tabela 1: Estatísticas de 100.000 amostras de paráfrases de cada origem textual utilizada.

para-score	Confiança	Origem	Alvo
0,19	0,98	*com certeza.*	*tem toda a razão.*
0,36	0,94	*— ela é colombiana.*	*ela é irlandesa.*
0,51	1,00	*embora, como você tenha visto, o temido erro do milênio não tenha se materializado, ainda assim as pessoas de vários países sofreram uma série de desastres naturais que realmente foram terríveis.*	*como puderam constatar, o grande "bug do ano 2000" não aconteceu. em contrapartida, os cidadãos de alguns dos nossos países foram vítimas de catástrofes naturais verdadeiramente terríveis.*
0,70	1,00	*senhora presidente, gostaria de saber se haverá uma mensagem clara do parlamento esta semana sobre o nosso descontentamento sobre a decisão de hoje se recusar a renovar o embargo de armas sobre a indonésia, considerando que a grande maioria neste parlamento aprovou o embargo de armas na indonésia no passado?*	*senhora presidente, gostaria de saber se esta semana o parlamento terá oportunidade de manifestar a sua inequívoca posição de descontentamento face à decisão, hoje tomada, de não renovar o embargo de armas destinadas à indonésia, tendo em atenção que a grande maioria da assembleia apoiou o referido embargo quando este foi decretado.*
0,97	0,98	*ele reuniu diferentes exemplos.*	*reuniu diferentes exemplos.*
1,00	1,00	*é tão fácil. . .*	*é tão fácil. . .*

Tabela 2: Pares de amostras ordenados por grau de relação de paráfrase (para-score) e grau de confiança da sentença de Origem em relação à língua portuguesa.

(2018). Para a métrica de fluência, aproximadamente 82,53% dos textos oriundos do processo de tradução reversa são fluentes.

O subconjunto de pares de paráfrases dos dois níveis superiores será utilizado nos experimentos descritos nas próximas seções. Estatísticas descritivas são apresentadas na Tabela 4.

3.3. Exemplos de paráfrases obtidas

Como forma de exemplificar a qualidade dos pares de paráfrases presentes no córpus *PTPARANMT*, a Tabela 5 apresenta exemplos de textos e suas paráfrases selecionados aleatoriamente. Para facilitar a ilustração, os exemplos são agrupados informalmente em três categorias de erro (baixa, média e alta) conforme seu valor de *para-score*. De forma análoga, a Tabela 6 apresenta exemplos de textos do conjunto Origem apresentando erro de fluência baixo, médio ou alto. Em ambos os casos, observa-se que a qualidade dos textos utilizados no processo de construção do córpus (seções 3.1 e 3.2) é variável, o que se reflete naturalmente na qualidade das paráfrases obtidas.

4. Indução de embeddings sentenciais

O córpus de paráfrases *PTPARANMT* foi utilizado para indução de *embeddings* sentenciais para uso nos experimentos de transferência de estilo a serem relatados na Seção 5, privilegiando-se para este fim os pares de sentença que permitissem maximizar a variedade lexical com preservação da semântica original do texto. Assim como no estudo em Wieting & Gimpel (2018) para a língua inglesa, optamos por selecionar os pares de sentenças cujo *para-score* fosse $\geq 0,4$, e foram excluídos os pares cuja sobreposição de trigramas fosse $\leq 0,7$. A aplicação destes filtros resultou em um subconjunto de 13 milhões de pares de paráfrases a ser utilizado como conjunto de treino para as *embeddings*.

As *embeddings* foram geradas a partir de pares de córpus conforme a metodologia em Wieting et al. (2021). Essa metodologia propõe uma arquitetura neural composta por uma camada densa projetada para um vetor que combina, através da média, todas as representações de todos os trigramas possíveis para uma sentença s. O objetivo da modelagem é, a partir de

Agrupamentos por *para-score*	# pares	Sobreposição Trigramas	Fluência.O			Fluência.A			Paráfrase		
			0	1	2	0	1	2	0	1	2
$(0,24, 0,631)$	13	$0,22 \pm 0,12$	10	11	05	16	05	05	05	13	08
$(0,631, 0,752)$	12	$0,38 \pm 0,10$	12	05	07	16	04	04	16	05	03
$(0,752, 0,841)$	13	$0,47 \pm 0,12$	08	10	08	14	09	03	22	02	02
$(0,841, 0,913)$	12	$0,47 \pm 0,17$	16	07	01	12	10	02	20	04	00
$(0,913, 1,0)$	13	$0,63 \pm 0,16$	12	09	05	14	08	04	22	04	00

Tabela 3: Avaliação humana dos textos coletados.

	Origem	Alvo
Pares	21.355.451	18.650.340
Palavras	1.349.479	815.791
Tam. sentenças	6,14	6,15

Tabela 4: Estatísticas descritivas do córpus.

uma sequência de palavras (s), extrair um vetor denso de *embeddings* de tamanho k minimizando-se a função de similaridade entre pares similares e, para casos negativos, maximizando-se a função de recompensa.

Para os hiper-parâmetros, seguimos as mesmas configurações usadas em Wieting et al. (2021), com *batch* de tamanho 128, *margin* σ de 0,4, e a taxa de *annealing* em 150. Um *tokenizador* do tipo *SentencePiece* (Kudo & Richardson, 2018) foi pré-treinado usando-se o córpus com vocabulário máximo de 50.000 *tokens*. O otimizador utilizado foi o *adam* (Kingma & Ba, 2015) com uma taxa de aprendizado inicial em 0,001, e iterando-se o modelo por um total de 25 épocas. Como forma de otimizar o processo de treinamento, optou-se pela utilização do método *mega-batch* com tamanho 100, e *dropout* inicial para a camada intermediária de 0,0.

A partir de uma sentença s_i, o treinamento propriamente dito consistiu em selecionar como alvo negativo uma sentença aleatória t'_i que não fosse uma paráfrase t_i. Esta seleção é feita com base nas sentenças do conjunto Alvo do córpus, em todos os casos selecionando-se o exemplo negativo com menor similaridade de cosseno em relação à sentença. Esta estratégia objetivou obter pares positivos que fossem mais similares do que os pares de sentença negativos com a seguinte margem α:

$$\min_{(\sigma_{origem}, \sigma_{alvo})} \sum \alpha - \cos_\sigma(s_i, t_i) + \cos_\sigma(s_i, t'_i)$$

O modelo assim definido foi treinado com uso do algoritmo de *back-propagation*. Como forma de ilustrar a qualidade das *embeddings* geradas, foram conduzidos dois experimentos de avaliação intrínseca envolvendo tarefas tradicionais de PLN que guardam certa afinidade com a detecção de paráfrases. Estes experimentos, enfocando as tarefas de simplificação, e similaridade sentencial, são descritos individualmente nas seções a seguir.

4.1. Simplificação sentencial

A tarefa de simplificação sentencial considerada para avaliação das *embeddings PTPARANMT* consistiu em criar versões de menor complexidade lexical e sintática de um texto de entrada nos moldes definidos pelos córpus PorSimplesSent2 e PorSimplesSent3 descritos em Leal et al. (2018).

Foram desenvolvidos três modelos de simplificação sentencial do tipo *sequência-para-sequência* com arquitetura neural do tipo *transformer*, aqui denominados S2S+PTPARANMT, *S2S+Glove* e *S2S+Random*. No modelo *S2S+PTPARANMT*, optamos por utilizar um esquema de treinamento *sequência-para-sequência* com *transfer-learning* usando os pesos das *embeddings* induzido a partir do córpus *PTPARANMT*. No modelo *S2S+Glove* usamos *embeddings* do tipo GloVe (Pennington et al., 2014). No modelo *S2S+Random*, optamos por utilizar um esquema de inicialização de pesos aleatória.

Como hiper-parâmetros dos modelos, definimos um *batch* de tamanho 32, e usamos o otimizador *adam* com uma taxa de aprendizado inicial de 0,001 em um total de 20000 épocas com um intervalo de tolerância de 10 épocas para paradas preemptivas. Optamos por usar a métrica de avaliação automática BLEU por se tratar de um problema de geração de texto. As *embeddings* GloVe utilizadas são disponibilizadas em Hartmann et al. (2017).

Como pré-processamento das versões 2 e 3 do córpus PorSimplesSent, utilizamos um *tokenizador* do tipo *SentencePiece* (Kudo & Richardson, 2018) previamente treinado com base nos dados do córpus *PTPARANMT*. A divisão do córpus foi de 80% para treino, 10% para testes e 10% para o conjunto de treinamento para a validação.

Erro	Origem	Alvo
Baixo	*graças a esse apoio, podemos corrigir o ângulo da câmera.*	*graças a este suporte, podemos corrigir o ângulo da câmera.*
Médio	*eu vou revelar a resposta para você e, em seguida, dar-lhe uma meditação guiada para incorporar os conceitos em sua mente subconsciente para que você seja naturalmente mais feliz sozinho ou para você aprender a ser sozinho.*	*pode ser realmente feito? im irá revelar a resposta para você e, em seguida, dar-lhe uma meditação guiada para incorporar os conceitos na sua mente subconsciente, de modo que você naturalmente seja ainda mais feliz sozinho.*
Alto	*uma vez por turno: você pode desligar o material de 1 xyz desta carta, em seguida, alvo de 1 monstro que o seu adversário controla; equipa- o para esta carta.*	*uma vez por turno, você pode selecionar uma face voltada para cima do monstro synchro seu oponente controla, e equipá-la a este cartão.*

Tabela 5: Pares de amostras ordenados por grau de erro de paráfrase.

Erro	Origem
Baixo	*infundindo os clássicos europeus com toque lúdico e moderno os nossos programas experimentais que definem marca dão vida ao le méridien e cumprem nossa promessa de desbloquear destino através de experiências criativas e culturais para os hóspedes.*
Médio	*no ano seguinte huam tchao organizou insurreição em apoio vam sientchi.*
Alto	*andrea dovizioso repsol honda 2010 motogp couro terno.*

Tabela 6: Amostras do conjunto Origem ordenadas por grau de erro de fluência.

Os modelos foram treinados na porção de testes dos conjuntos de dados propostos. A Tabela 7 sumariza as configurações empregadas e seus resultados com base nas versões 2 e 3 do córpus (v2-BLEU e v3-BLEU).

Os resultados da Tabela 7 sugerem que o modelo *S2S+PTPARANMT* obteve os melhores resultados gerais. Além disso, cabe destacar a importância do uso de pesos pré-treinados, ilustrada pelo desempenho inferior do modelo *S2S+Random* em relação aos dois primeiros.

4.2. Similaridade sentencial

Como um segundo cenário de avaliação das *embeddings PTPARANMT*, consideramos a tarefa de estimativa de similaridade sentencial. Essa tarefa consistiu em decidir - em uma escala de 0 a 5 - quão próximas duas sentenças são entre si tomando-se por base o *benchmark ASSIN* 2 (Real et al., 2020). Com este propósito, foi conduzido um experimento comparando-se dois modelos refinados a partir dos pesos das *embeddings* geradas, aqui denominados *PTPARANMT* e *Siamese+PTPARANMT*, e dois sistemas de *baseline* do tipo *transformer*, aqui denominados *para-multi-MiniLM-L12-v2* e *dpr-ctx-enc-bert-base-multi*.

O modelo *PTPARANMT* utiliza a distância cosseno entre as *embeddings* dos pares de sentença, enquanto *Siamese+PTPARANMT* utiliza uma rede recorrente do tipo LSTM para a sentença Origem e outra para a Alvo, tal que ambas são otimizadas de modo a minimizar a distância cosseno entre os pares candidatos. O modelo de *baseline para-multi-MiniLM-L12-v2* usa uma arquitetura *transformer* similar ao apresentado em Reimers & Gurevych (2019), utilizando os pesos multilíngues do modelo original em um formato de *transfer-learning*. O *baseline dpr-ctx-enc-bert-base-multi*, por outro lado, adota o princípio de passagem densa de representação (do inglês *Dense Passage Retrieval*), em um formato similar ao apresentando em Karpukhin et al. (2020), também utilizando o peso multilíngue do modelo original em um formato de *transfer-learning*.

Como hiper-parâmetros dos modelos, definimos um *batch* de tamanho 32, otimização *adam* para uma taxa de aprendizado inicial de 0,001 em um total de 1000 épocas com um intervalo de tolerância de 10 épocas para paradas preemptivas.

Para pré-processamento do córpus *ASSIN* 2, utilizamos um *tokenizador* do tipo *SentencePiece* (Kudo & Richardson, 2018), que foi treinado previamente nos dados do córpus *PTPARANMT*. O conjunto de dados *ASSIN* 2 contém 10.000 pares de frases em português brasileiro, sendo 6.500 pares para treinamento, 500 para validação e 3.000 para teste. As configurações empregadas e seus resultados — representados pelas métricas *Mean*

Modelo	Tam.	Neurônios	Camadas	Pré-treino?	v2-BLEU	v3-BLEU
S2S+PTPARANMT	1024	512	12	Sim	79,8	52,0
S2S+Glove	300	512	12	Sim	70,2	43,9
S2S+Random	300	512	12	Não	60,9	30,9

Tabela 7: Modelos e resultados de simplificação sentencial.

Square Error (MSE) e correlação de Pearson — são sumarizados na Tabela 8, com os melhores resultados de cada métrica em destaque.

A Tabela 8 mostra que o modelo *Siamese +PTPARANMT* obteve os melhores resultados gerais para ambas as métricas de avaliação. Com base neste resultado, o modelo *Siamese+PTPARANMT* foi então comparado também com os sistemas participantes da iniciativa *ASSIN* 2. Os resultados desta avaliação são sumarizados na Tabela 9, novamente com os melhores resultados de cada métrica em destaque.

Os resultados da Tabela 9 indicam que o modelo *Siamese+PTPARANMT* supera as alternativas tanto no que diz respeito à métrica MSE (i.e., comparado ao sistema Stilingue em Fonseca & Alvarenga (2019)), quanto no que diz respeito à correlação de Pearson (comparado ao sistema IPR em Rodrigues et al. (2019b)).

5. Reescrita sentencial baseada em paráfrases

Os resultados do uso do córpus e *embeddings PTPARANMT* nas tarefas de simplificação (Seção 4.1) e similaridade sentencial (Seção 4.2) sugerem que estes recursos podem ser utilizados na aplicação-fim do presente estudo, ou seja, à tarefa de transferência de estilo arbitrário baseada em paráfrases. Para investigar esta possibilidade, foi conduzido um experimento deste tipo utilizando modelos neurais do tipo *sequência-para-sequência* para reescrita sentencial em um estilo-alvo de interesse. A escolha dessa arquitetura foi motivada pela sua relativa simplicidade de implementação, e pelo bom desempenho geral observado em tarefas de geração de texto (Goodfellow et al., 2016; Goldberg, 2016; Gatt & Krahmer, 2018).

O experimento realizado consistiu em desenvolver e comparar duas estratégias de reescrita sentencial — aqui denominadas *S2S* e *paraPTT5* — com três modelos de *baseline*, aqui denominados *REF*, *Cópia* e *Ingênuo*. Estas cinco configurações são sumarizadas na Tabela 10 e detalhadas a seguir.

O modelo *S2S* consiste de uma arquitetura neural do tipo *sequência-para-sequência* com atenção do tipo geral utilizando o *framework* de geração de texto *openNMT* com as mesmas configurações descritas em Bahdanau et al. (2014), e com inicialização de pesos no formato xavier (Glorot & Bengio, 2010). O modelo *paraPTT5* consiste de uma arquitetura neural de pesos ajustados a partir do modelo de língua pré-treinado *PTT5* (Carmo et al., 2020). O ajuste dos pesos originalmente fornecidos pelo modelo fez uso da biblioteca *Transformers Hugging Face* (Wolf et al., 2020). Em ambos os casos, termos desconhecidos do vocabulário do modelo foram representados por *tokens* artificiais *UNKNOWN*.

Como sistemas de *baseline* para o experimento, consideramos três modelos que imitam transformações simples sobre a entrada textual: o modelo *REF* apenas reproduz o texto alvo como saída, e representa assim o limite de desempenho máximo possível para a tarefa; *Cópia* é uma simulação de um modelo de geração que simplesmente repete a entrada como saída; e *Ingênuo* é um modelo que gera como saída uma seleção aleatória das palavras de entrada, com probabilidade $p = 0,5$, e concatena a elas um texto também aleatório do conjunto Alvo.

5.1. Conjunto de dados

A partir da versão original do córpus *PTPARANMT* (descrita na seção 3), foram aplicados filtros sistemáticos com o objetivo de produzir um conjunto de dados de modo que as características de diversidade lexical e sintática fossem maximizadas. Esse tipo de técnica tem sido aplicada com resultados positivos em tarefas para a língua inglesa como reescrita de paráfrases (Wieting et al., 2021) e transferência de estilo (Krishna et al., 2020).

Para filtrar o córpus *PTPARANMT*, utilizamos as mesmas duas métricas originalmente propostas em Krishna et al. (2020). A métrica *sobreposição* foi usada para medir a diversidade lexical dos textos a partir da contagem de sobreposição de trigramas de co-ocorrência entre os pares, e a métrica *para-score* foi usada para medir a similaridade de paráfrase entre dois textos.

Modelo	Tam.	Neurônios	Camadas	MSE	Pearson
PTPARANMT	1024	1024	1	0,31	0,781
Siamese+PTPARANMT	300	300	2	**0,24**	**0,828**
para-multi-MiniLM-L12-v2	768	3072	12	2,27	0,761
dpr-ctx-enc-bert-base-multi	768	3072	12	0,56	0,560

Tabela 8: Modelos de similaridade sentencial e resultados obtidos.

Modelo	Submissão	MSE	Pearson
Siamese+PTPARANMT	-	**0,24**	**0,828**
Stilingue (Fonseca & Alvarenga, 2019)	3	0,47	0,817
L2F	BL	0,52	0,778
IPR (Rodrigues et al., 2019b)	1	0,52	0,826
ASAPPy (Santos et al., 2019)	1-ptbr	0,58	0,730
DeepLearning Brasil (Rodrigues et al., 2019a)	Ensemble	0,59	0,785
NILC (Cabezudo et al., 2020)	2	0,64	0,729
baseline	Overlap	0,75	0,577
PUCPR (de Souza et al., 2019)	comNILC	0,85	0,678
LIACC	2	1,02	0,459
baseline	BoW-2	1,15	0,175

Tabela 9: Comparação do modelo Siamese+PTPARANMT com participantes da iniciativa *ASSIN* 2.

Modelo	Tamanho	Neurônios	Pré-treinamento?
S2S	200	200	Não
paraPTT5	768	3072	Sim
REF	–	–	Não
Cópia	–	–	Não
Ingênuo	–	–	Não

Tabela 10: Configurações dos modelos neurais do tipo *sequência-para-sequência*.

Para selecionar a melhor configuração de parâmetros de filtro do córpus, foi utilizado um modelo neural, aqui denominado *PTT5finne*, com pesos ajustados a partir do modelo público *PTT5-base* (Carmo et al., 2020). Esse modelo passou por iterações com os parâmetros de sobreposição de texto e *para-score* em um esquema de busca em *grid*, visando maximizar a métrica de similaridade semântica *bert-score*. A escolha das métricas *para-score* e sobreposição como filtros foi fundamentada nos resultados apresentados em (Krishna et al., 2020). No referido estudo, propôs-se o treinamento de um modelo de reescrita de paráfrases em inglês (STRAP) a partir de um subconjunto filtrado de pares artificiais de paráfrases, utilizando essas métricas como critério. Além disso, a adição da métrica *bert-score* foi motivada por sua significativa correlação com a avaliação humana de textos em nível sentencial (Zhang et al., 2020). A Figura 1 ilustra o mapa de calor do *bert-score* obtido por meio desse processo de busca (melhor visualizado em formato eletrônico).

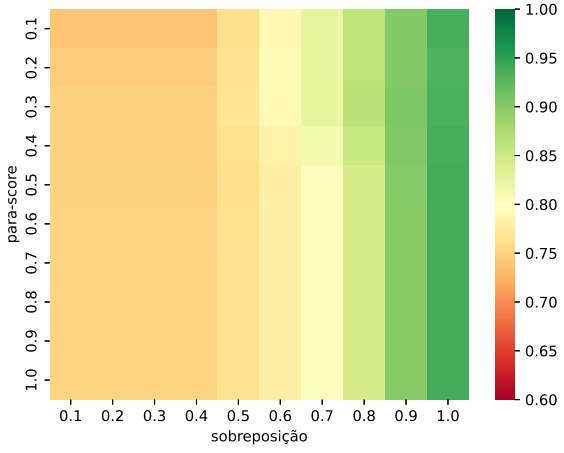

Figura 1: *bert-score* para diferentes versões do córpus *PTPARANTM* com base nas medidas de para-score e sobreposição.

O mapa de calor na Figura 1 revela que o modelo *PTT5finne* alcança 0,775 pontos de *bert-score* no intervalo de valores (0,6:0,6). Em outras palavras, nos respectivos intervalos das métricas de sobreposição e *para-score*, o modelo *PTT5finne* apresentou o melhor desempenho em termos de correlação semântica (*bert-score*) entre o texto gerado e o texto do conjunto-alvo. Estes intervalos são similares aos apresentados para as métricas de *para-score* e sobreposição em Krishna et al. (2020). A filtragem dos dados do córpus com base neste parâmetro resultou em 73.476 pares de paráfrase. Este conjunto foi particionado em uma porção de desenvolvimento de 90% (59.584 pares de sentenças), e 10% de teste (7.347 pares). Do conjunto de desenvolvimento, uma porção de 5% (3.306 pares) foi utilizada para validação e iteração dos hiper-parâmetros, e o restante foi utilizado para treino.

5.2. Avaliação

Os modelos propostos — *S2S* e *paraPTT5* — foram treinados por *back-propagation*, e sua capacidade de generalização foi avaliada em relação ao conjunto de testes. O modelo*S2S* foi otimizado pelo método *adagrad* com uma taxa inicial de 0,15 por um total de 200 mil épocas, e o modelo *paraPTT5* foi otimizado pelo método *adam* com uma taxa inicial de 0,00005 por um total de 2 épocas. Em ambos os casos, seguimos a configuração adotada em Kaplan et al. (2020) definindo um tamanho de *batch* com 20 pares de frases, e limitando as sequências de entrada e saída a sentenças de tamanho 50. Por fim, os dois modelos foram otimizados com o objetivo de minimizar o erro médio das sub-palavras geradas em relação às sub-palavras alvo, ou entropia cruzada.

Os modelos *S2S* e *paraPTT5* foram submetidos à avaliação intrínseca e humana. Além disso, no caso da avaliação intrínseca, estes modelos foram comparados também aos sistemas de *baseline REF*, *Cópia* e *Ingênuo*. As duas modalidades de avaliação são discutidas individualmente nas próximas seções.

5.2.1. Avaliação intrínseca

A avaliação intrínseca realizada foi baseada em cinco métricas de qualidade dos textos gerados pelos modelos em relação aos texto-alvo do conjunto de teste. Três destas métricas são de motivação computacional, a saber: *bert-score*, *ROUGE* e *para-score*. Para *ROUGE* e *bert-score*, utilizamos as implementações disponíveis na biblioteca *Transformers Hugging Face* (Wolf et al., 2020). Os pesos pré-treinados da métrica *bert-score* foram obtidos a partir da versão pública *bert-base-multilingual-cased* do modelo.

A estas métricas, foram acrescidas duas medidas de complexidade da escrita discutidas em Leal et al. (2023): a métrica de complexidade lexical Brunet e a métrica de complexidade sintática da distância do grafo. A escolha dessas duas métricas foi motivada pelos resultados apresentados em Leal et al. (2018) para a tarefa de identificação da complexidade textual, e pela facilidade de replicação das mesmas para o português. Em ambos os casos, optamos por reimplementar o código apresentado em Leal et al. (2023) com uso do pacote *spaCy* em Python.

A métrica Brunet relaciona a taxa dos erros tipográficos com o tamanho do texto, e apresenta valores típicos entre 10 e 20, sendo que um texto mais rico (e complexo) produz valores menores. A distância do grafo estima a complexidade sintática de uma sentença representada na forma de um grafo de dependências considerando a relação entre suas palavras e a distância do arco de dependência entre elas, apresentando valores maiores (indicativos de maior complexidade) à medida que as distâncias de dependência aumentam.

A Tabela 11 apresenta os resultados obtidos para a tarefa de reescrita sentencial com base nas métricas selecionadas. O melhor resultado obtido por um dos dois modelos propostos — seja *S2S* ou *paraPTT5* — de acordo com cada métrica de avaliação é destacado em negrito.

Os resultados da Tabela 11 sugerem que, de modo geral, os modelos propostos *S2S* e *paraPTT5* são superiores às alternativas. O modelo *paraPTT5* é superior para as métricas de complexidade sentencial (complexidade de distância do grafo e leiturabilidade Brunet), enquanto que o modelo *S2S* se destacou nas métricas *para-score*, *bert-score* e *ROUGE*. Além disso, observa-se que o *baseline REF*, que produz sempre a saída esperada (e assim obtém valores máximos para as métricas de similaridade textual como *para-score*, *bert-score*, etc.) apresenta resultados inferiores aos dos modelos propostos para todas as demais métricas por falta de diversidade linguística, ilustrando a necessidade de equilíbrio entre múltiplos critérios conflitantes para o sucesso da tarefa de reescrita.

5.2.2. Avaliação humana

Em complemento à avaliação intrínseca discutida na seção anterior, foi realizada também uma análise humana dos textos gerados pelos mode-

	Modelos				
Métrica	S2S	paraPTT5	REF	Cópia	Ingênuo
para-score	**88,7**	88,1	100,0	81,4	29,9
bert-score	**91,7**	90,6	100,0	87,6	68,3
ROUGE-1	**76,5**	70,2	100,0	54,8	23,3
ROUGE-2	**61,1**	54,4	100,0	30,8	5,8
distância do grafo	27,9	**22,9**	29,1	29,7	48,6
Brunet	5,5	**5,2**	5,5	5,5	6,5

Tabela 11: Avaliação intrínseca dos textos gerados.

los *S2S* e *paraPTT5*. Para este fim, oito falantes nativos da língua portuguesa foram solicitados a avaliar um total de 102 pares de paráfrases provenientes de uma amostragem do conjunto de teste do córpus quanto ao grau de relação de paráfrase do par, e também com relação ao grau de fluência de cada texto individualmente. Ambas avaliações seguiram as mesmas diretrizes discutidas na seção 3.2, ou seja, atribuindo-se escores de 0 a 2 a cada par de frases representado o grau da relação de paráfrase entre elas, e escores 0 a 2 a cada frase individual representando seu nível de fluência.

Foi alcançada uma concordância entre anotadores de $0,88$ de acordo com o índice kappa de Cohen (Landis & Koch, 1977). Os pares de frases foram agrupados em intervalos de 10 com base na métrica de similaridade de paráfrases *para-score*, e os resultados na Tabela 12 sumarizam o número de vezes que cada escore de paráfrase e fluência (Fluência.O e Fluência.A de textos Origem e Alvo, respectivamente) foi escolhido pelos avaliadores das 102 amostras.

Conforme observado na Tabela 12, os pares ruidosos de ambos os modelos estão majoritariamente confinados aos primeiros dois intervalos $((0.23, 0.672), (0.672, 0.766))$, o que corresponde a 30 - 46% de todos os pares que possuem uma forte relação de paráfrase. Além disso, no intervalo superior, 86,63% (para *S2S*) e 90% (para *paraPTT5*) dos pares possuem uma forte relação de paráfrase.

Em complemento a estes resultados, a Tabela 13 apresenta os valores médios para as métricas de fluência e grau de relação de paráfrase de cada par obtidos pelos dois modelos avaliados.

A Tabela 13 indica que o modelo *paraPTT5* obtém os melhores resultados tanto com base na fluência textual quanto na qualidade das paráfrases geradas, o que pode ser tomado como indício de que LLMs são efetivamente úteis para a tarefa de transferência de estilo textual por meio de paráfrases, e coloca assim a questão de qual seria o desempenho do modelo *paraPTT5* sem o refinamento dos pesos a partir do LLM. Em uma análise *post hoc* (aqui não detalhada) na qual o modelo *paraPTT5* foi substituído pelo modelo *S2S*, observamos que *paraPTT5* seria a melhor alternativa, confirmando assim a superioridade da técnica de transferência de aprendizado com LLMs.

5.2.3. Exemplos de textos gerados

Como forma de exemplificar a qualidade dos pares de paráfrases gerados pelo modelo *paraPTT5*, a Tabela 14 apresenta amostras aleatórias de textos-origem e suas paráfrases. Para facilitar a visualização, os exemplos são agrupados informalmente em três categorias do grau de erro (baixa, média e alta) conforme o grau de erro na relação de paráfrase de cada par avaliado.

Finalmente, a Tabela 15 apresenta amostra de textos-origem gerados para exemplificar a fluência dos textos gerados através do modelo paraPTT5. Os exemplos são divididos informalmente em três categorias do grau de erro (baixa, média e alta), de acordo com a pontuação de fluência do texto.

Os exemplos apresentados nas Tabelas 14 e 15 apresentam uma série de erros de natureza semântica e superficial proporcionais ao grau de erro reportado. Estes erros são, em grande parte, decorrentes do método de criação e filtragem do córpus *PTPARANMT* (seções 3 e 5.1). Em especial, observamos que o método de tradução reversa, por se tratar de um processo automático, pode ter introduzido de forma não intencional ruídos durante o processo de produção do estilo-alvo desejado, como já relatado em trabalhos similares para a língua inglesa Wieting & Gimpel (2018); Krishna et al. (2020). Além disso, observamos que no processo de filtragem dos dados foram priorizados os pares de sentenças de maior *para-score*, o que privilegia a preservação da semântica com certo detrimento à forma superficial.

Modelo	Intervalo para-score	# Pares	Fluência.O 0	1	2	Fluência.A 0	1	2	Paráfrase 0	1	2
S2S	(0.23, 0.672)	10	13	5	2	12	4	4	6	11	3
	(0.672, 0.766)	10	9	7	4	3	8	9	6	6	8
	(0.766, 0.827)	10	9	7	4	4	5	11	8	11	1
	(0.827, 0.926)	10	14	5	1	8	8	4	11	9	0
	(0.926, 1.0)	10	9	5	6	9	6	5	18	1	1
paraPTT5	(0.207, 0.638)	11	11	4	7	13	3	6	8	10	4
	(0.638, 0.747)	10	12	7	1	10	7	3	9	7	4
	(0.747, 0.828)	10	13	5	2	10	7	3	12	8	0
	(0.828, 0.896)	10	11	5	4	10	5	5	10	8	2
	(0.896, 1.0)	11	12	8	2	12	8	2	19	3	0

Tabela 12: Avaliação humana dos textos gerados.

Modelo	Fluência	Paráfrase
S2S	75,0%	87,0%
paraPTT5	83,2%	90,4%

Tabela 13: Resultados médios de fluência e paráfrase.

5.3. Considerações

Os resultados da avaliação dos modelos *S2S* e *paraPTT5* permitem uma série de observações. Em primeiro lugar, destaca-se a importância do ajuste de pesos baseado em LLMs e da seleção criteriosa de paráfrases para compor o conjunto de dados, sem os quais o texto gerado não seria minimamente aceitável do ponto de vista de um leitor humano.

Em segundo lugar, observa-se que a técnica de transferência de conhecimento permitiu ao modelo *paraPTT5* produzir textos mais coesos e semelhantes a paráfrases se comparado ao modelo *S2S*. Uma possível explicação para esse resultado é que *S2S* é um modelo neural sequencial do tipo *transformer* que não utiliza transferência de conhecimento. Assim, a correlação positiva observada nos resultados do modelo *paraPTT5* no conjunto de dados de teste parece estar relacionada à qualidade das representações internas dos LLMs.

6. Conclusões

Este trabalho apresentou um primeiro estudo em transferência de estilo textual arbitrário utilizando paráfrases em Português, tratando da construção do córpus de paráfrases *PTPARANMT* e *embeddings* de mesmo nome, e do uso destes recursos na tarefa de reescrita sentencial baseada em paráfrases.

Os recursos linguístico-computacionais construídos[3] foram inicialmente validados nas tarefas de simplificação e similaridade sentencial, obtendo resultados superiores aos das alternativas consideradas. No caso da tarefa de similaridade sentencial, os resultados obtidos foram inclusive superiores aos reportados na área tomando-se por base o *benchmark ASSIN* 2 (Real et al., 2020).

Uma vez estabelecidos estes resultados iniciais, o córpus *PTPARANMT* foi então empregado na tarefa para a qual havia sido realmente desenvolvido, ou seja, a transferência de estilo arbitrário em Português. Foram propostos para esse fim dois modelos principais, sendo um baseado na arquitetura *sequência-para-sequência* e o outro obtido pelo refino de um grande modelo de língua existente. Os modelos propostos foram avaliados de forma intrínseca e com auxílio de juízes humanos, sugerindo-se a importância do ajuste de pesos baseado em LLMs e da seleção criteriosa de paráfrases para compor o conjunto de treino, e a superioridade da técnica de transferência de aprendizado a partir de LLMs em relação à arquitetura *sequência-para-sequência* na tarefa em questão.

O presente estudo deixa diversas oportunidades de trabalho futuro. Por exemplo, uma alternativa relevante à presente abordagem, e que tem se popularizado rapidamente nas áreas do PLN e GLN, seria o uso de técnicas de engenharia de *prompts* e aprendizado *few-shoot* com uso de LLMs (Min et al., 2023). Estudos recentes de transferência de estilo, como em Troiano et al. (2023), têm demonstrado ganho significativo em relação ao uso de modelos pré-treinados em cenários onde o estilo-alvo é arbitrário ou desconhecido (Krishna et al., 2022; Reif et al., 2022). No entanto, por ser uma inovação científica re-

[3] https://github.com/pablocosta/paperLinguamaticaTSTBR

Erro	Texto original	Paráfrase
baixo	*aqui está nossa revisão do estilo de vida de wall street*	*aqui está nossa avaliação sobre walls-treet lifestyle*
médio	*assim no modelo de privatização pura estado interferiria menos no seb exceto no que considerava privatização*	*assim no modelo puro de privatização estado interferia menos no seb com exceção do que*
alto	*ao adicionar 75 por cento em cima dos seus gastos e comprar árvores por essa quantidade tudo que você gasta é devol-vido você*	*ao adicionar 75 ao exceder os gastos e comprar árvores por essa quantia todos os gastos são*

Tabela 14: Amostras aleatórias de textos produzidos pelo modelo paraPTT5 com diferentes graus de erro de paráfrase.

Erro de fluência	Texto gerado
Baixo	*acesse fórum de viagens do tripadvi-sor sobre portland e faça perguntas ao*
Médio	*alimentação no mercado central mercado central 27 km*
Alto	*inscreverse para atualizações seja notificado quando atualizarmos in-formações sobre vgoc*

Tabela 15: Amostras aleatórias de textos gerados pelo modelo paraPTT5 com diferentes graus de erro de fluência.

cente, métodos deste tipo ainda apresentam desempenho inferior ao das abordagens que fazem refinamento de LLMs pré-treinados nos casos em que um conjunto de dados de proporção significativa esteja disponível (Scao & Rush, 2021; Puri et al., 2023; Liu et al., 2023), havendo portanto oportunidade para mais pesquisas.

Além disso, consideramos também a construção de um córpus de paráfrases composto de estilos reais (i.e., produzidos por diferentes autores humanos) no lugar de um conjunto de dados sintético como o presente córpus *PTPARANMT*. Uma iniciativa deste tipo proporcionaria não apenas um maior grau de realismo à tarefa computacional, mas poderia também auxiliar na redução do ruído proveniente do método de tradução automática aqui empregado.

Agradecimentos

O segundo autor conta com apoio do processo nro. 2021/08213-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

Agirre, Eneko, Daniel Cer, Mona Diab & Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. Em *1st Joint Conference on Lexical and Computational Semantics*, 385–393.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. ArXiv [cs.CL]. 10.48550/arXiv.1409.0473.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins & Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. Em *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4555–4567. 10.18653/v1/2020.acl-main.417.

Beddiar, Djamila Romaissa, Md Saroar Jahan & Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media* 24. 100153. 10.1016/j.osnem.2021.100153.

Bhagat, Rahul & Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics* 39(3). 463–472. 10.1162/COLI_a_00166.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press. 10.1017/CBO9780511814358.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. Em *34th International Conference on Neural Information Processing Systems*, 1877–1901.

Cabezudo, Marco Antonio Sobrevilla, Marcio Lima Inácio, Ana Carolina Rodrigues, Edresson Casanova & Rogério Figueredo de Souza. 2020. Nilc at assin 2: exploring multilingual approaches. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 49–58.

Carlson, Keith, Allen Riddell & Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science* 5(10). 171920. 10.1098/rsos.171920.

Carmo, Diedre, Marcos Piau, Israel Campiotti, Rodrigo Nogueira & Roberto Lotufo. 2020. PTT5: Pretraining and validating the PT5 model on Brazilian Portuguese data. ArXiv [cs.CL]. 10.48550/arXiv.2008.09144.

Dong, Chenhe, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen & Min Yang. 2022. A survey of natural language generation. *ACM Computing Surveys* 55(8). 173. 10.1145/3554727.

Fonseca, Evandro & João Paulo Reis Alvarenga. 2019. Wide and deep transformers applied to semantic relatedness and textual entailment. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 68–77.

Gatt, Albert & Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61(1). 65–170.

Glorot, Xavier & Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. Em *13th International Conference on Artificial Intelligence and Statistics (AIStat)*, 249–256.

Goldberg, Yoav. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57. 345–420.

Gonçalo Oliveira, Hugo & Ana Alves. 2021. AIA-BDE: um corpo de perguntas, variações e outras anotações. *Linguamática* 13(2). 19–35. 10.21814/lm.13.2.350.

Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016. *Deep learning*. MIT Press.

Hartmann, Nathan S., Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues & Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 122–131.

Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari & Trevor Cohn. 2018. Iterative back-translation for neural machine translation. Em *2nd Workshop on Neural Machine Translation and Generation*, 18–24. 10.18653/v1/W18-2703.

Hu, Zhiqiang, Roy Ka-Wei Lee, Charu C. Aggarwal & Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter* 24(1). 14–45. 10.1145/3544903.3544906.

Hu, Zhiting, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov & Eric P Xing. 2017. Toward controlled generation of text. Em *International Conference on Machine Learning*, 1587–1596.

Jin, Di, Zhijing Jin, Zhiting Hu, Olga Vechtomova & Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics* 48(1). 155–205. 10.1162/coli_a_00426.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu & Dario Amodei. 2020. Scaling laws for neural language models. ArXiv [cs.LG]. 10.48550/arXiv.2001.08361.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen & Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. 10.18653/v1/2020.emnlp-main.550.

Kingma, Diederik P. & Jimmy Ba. 2015. Adam: A method for stochastic optimization. Em *3rd International Conference on Learning Representations (ICLR)*, 10.48550/arXiv.1412.6980.

Krishna, Kalpesh, Deepak Nathani, Xavier Garcia, Bidisha Samanta & Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. Em *60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7439–7468. 10.18653/v1/2022.acl-long.514.

Krishna, Kalpesh, John Wieting & Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. Em *Empirical Methods in Natural Language Processing (EMNLP)*, 737–762. 10.18653/v1/2020.emnlp-main.55.

Kudo, Taku & John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 66–71. 10.18653/v1/D18-2012.

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33. 159–174. 10.2307/2529310.

Leal, Sidney Evaldo, Magali Sanches Duran & Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. Em *27th International Conference on Computational Linguistics (COLING)*, 401–413.

Leal, Sidney Evaldo, Magali Sanchez Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2023. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation* 10.1007/s10579-023-09693-w.

Li, Juncen, Robin Jia, He He & Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. ArXiv [cs.CL]. 10.48550/arXiv.1804.06437.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi & Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55(9). 10.1145/3560815.

Luo, Fuli, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui & Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. ArXiv [cs.CL]. 10.48550/arXiv.1905.10060.

Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz & Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys* 56(2). 1–40. 10.1145/3605943.

Naseem, Usman, Imran Razzak, Shah Khalid Khan & Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing* 20(5). 1–35. 10.1145/3434237.

Pennington, J., R. Socher & C. D. Manning. 2014. GloVe: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. Em *3rd Conference on Machine Translation*, 186–191. 10.18653/v1/W18-6319.

Puri, Ravsehaj Singh, Swaroop Mishra, Mihir Parmar & Chitta Baral. 2023. How many data samples is an additional instruction worth? Em *Findings of the Association for Computational Linguistics: EACL*, 1042–1057. 10.18653/v1/2023.findings-eacl.77.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li & Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21(140). 1–67.

Rao, Sudha & Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 129–140. 10.18653/v1/N18-1012.

Real, Livy, Erick Fonseca & Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: a quick overview. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 406–412. 10.1007/978-3-030-41505-1_39.

Reif, Emily, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch & Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. Em *60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 837–848. 10.18653/v1/2022.acl-short.94.

Reimers, Nils & Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. 10.18653/v1/D19-1410.

Riley, Parker, Noah Constant, Mandy Guo, Girish Kumar, David Uthus & Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. Em *59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3786–3800. 10.18653/v1/2021.acl-long.293.

Rodrigues, Ruan Chaves, Jéssica Rodrigues da Silva, Pedro Vitor Quinta de Castro, Nádia Silva & Anderson da Silva Soares. 2019a. Multilingual transformer ensembles for Portuguese natural language tasks. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 27–38.

Rodrigues, Rui, Paula Couto & Irene Rodrigues. 2019b. IPR: The semantic textual similarity and recognizing textual entailment systems. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 39–48.

Santos, José, Ana Alves & Hugo Gonçalo Oliveira. 2019. ASAPPpy: a Python framework for Portuguese STS. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 14–26.

Scao, Teven Le & Alexander M. Rush. 2021. How many data points is a prompt worth? Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2627–2636. 10.18653/v1/2021.naacl-main.208.

Shen, Tianxiao, Tao Lei, Regina Barzilay & Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems* 30.

Shliazhko, Oleh, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova & Tatiana Shavrina. 2022. mGPT: Few-shot learners go multilingual. ArXiv [cs.CL]. 10.48550/arXiv.2204.07580.

de Souza, João Vitor Andrioli, Lucas E. S. Oliveira, Yohan Bonescki Gumiel, Deborah Ribeiro de Carvalho & Cláudia Maria Cabral Moro. 2019. Incorporating multiple feature groups to a siamese neural network for semantic textual similarity task in Portuguese texts. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 59–68.

Suzgun, Mirac, Luke Melas-Kyriazi & Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2195–2222. 10.18653/v1/2022.emnlp-main.141.

Tiedemann, Jörg & Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. Em *22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, 479–480.

Troiano, Enrica, Aswathy Velutharambath & Roman Klinger. 2023. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering* 29(4). 849–908. 10.1017/S1351324922000407.

Wang, Yunli, Yu Wu, Lili Mou, Zhoujun Li & Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3573–3578. 10.18653/v1/D19-1365.

Warstadt, Alex, Amanpreet Singh & Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7. 625–641. 10.1162/tacl_a_00290.

Wieting, John & Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. Em *56th Annual Meeting of the Association for Computational Linguistics*, 451–462. 10.18653/v1/P18-1042.

Wieting, John, Kevin Gimpel, Graham Neubig & Taylor Berg-Kirkpatrick. 2021. Paraphrastic representations at scale. ArXiv [cs.CL]. 10.48550/arXiv.2104.15114.

Wieting, John, Jonathan Mallinson & Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. Em

Conference on Empirical Methods in Natural Language Processing (EMNLP), 274–285. doi 10.18653/v1/D17-1026.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. doi 10.18653/v1/2020.emnlp-demos.6.

Xu, Wei, Alan Ritter, Bill Dolan, Ralph Grishman & Colin Cherry. 2012. Paraphrasing for style. Em *International Conference on Computational Linguistics (COLING)*, 2899–2914.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger & Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. Em *International Conference on Learning Representations*, on–line.

Zhang, Xiang, Junbo Zhao & Yann LeCun. 2015. Character-level convolutional networks for text classification. Em *28th International Conference on Neural Information Processing Systems*, 649–657.