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).

Proceedings of Recent Advances in Natural Language Processing, pages 155–164
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_019

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.,

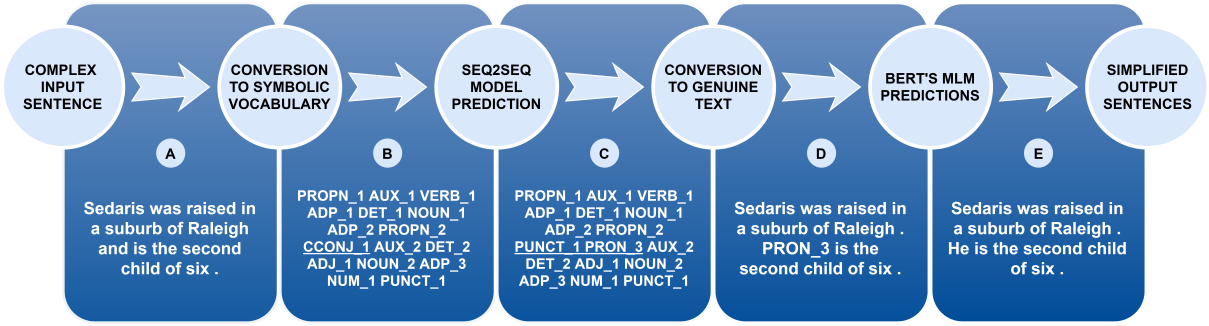

Figure 2: Illustration of our complete pipeline. To perform a prediction in the pipeline, the complex input sentence (A) passes through a preprocessing step to convert the text into symbolic vocabulary (B). This converted symbolic sequence is given to a sequence-to-sequence neural model that produces 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).

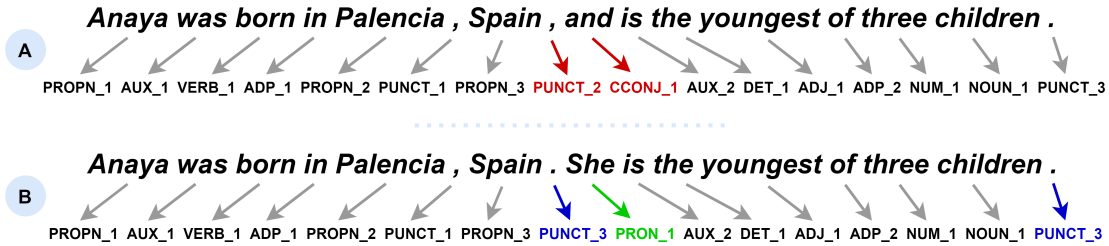

Figure 3: Examples of items collected to compose our symbolic vocabularies. Each token from the complex side of the alignment (A) is converted to a symbol formed by its respective POS tag and an index equivalent to its order of appearance in the sentence. These same symbols are then assigned to the simplified 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.

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*.

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

[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*.