GERC: Multilingual Grammatical Error Correction for the Informal Writer

João Maria Janeiro
joao.maria.j@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2022

Abstract

Grammatical Error Correction (GEC) has mostly been developed for English and for the domain of second language learners. In this thesis, we create a multilingual grammatical error correction system for the customer service domain, working for Portuguese and German. To the best of our knowledge this is the first public GEC research work for the Portuguese language based on neural networks. To adapt to this new domain, a new dataset provided by Unbabel is used. Different training regimes and data augmentation techniques, namely pre-training with public and in-domain data, creating synthetic data, and fine-tuning on the new Unbabel dataset, are explored. We also extended the ERRANT tool (Bryant et al., 2017) to support more languages, used for automatically generating annotated edits, and allowing us to score hypothesis. In addition to the explicit evaluation of the quality of these models, we also performed an implicit evaluation in which we measured the impact of our proposed GEC models as a pre-processing step for MT, measuring the quality of the German-English and Portuguese-English MT systems.

As a second contribution, we also develop a novel quality estimation (QE) approach, leveraging the more fine-grained word-level model of TransQuest (Ranasinghe et al., 2021), coupled with the T5's (Raffel et al., 2020) likelihood and COMET (Rei et al., 2021) sentence-level scores. The word-level model was used to generate a sentence-level score. It is, to the best of our knowledge, the first word level QE model developed for GEC to generate a sentence-level score, also used as a re-ranker. It is also the first time a multilingual re-ranker is introduced for GEC, as well as the first re-ranker working for the more informal customer service domain, and not just for the second learners formal domain. With this approach, we are able to improve the results of the T5-small, and T5-base models proposed by Rothe et al. (2021), by up to +2.1 and + 1.54 $F_{0.5}$ points on the CoNLL-14 test set, respectively.

Keywords: Grammatical Error Correction, Multilingual, Deep Learning, Machine Translation, Data Augmentation

1. Introduction

Humans make many mistakes when writing online. Customer service clients in need of assistance, even more. This can make it hard for people, or Natural Language Processing (NLP) systems, to to process and understand what was written. This is even more noticeable when customer service companies use automated pipelines and tools (eg. Machine Translation systems, Parsers, Name Entity Recognisers, etc) to process these generated text. These tools are commonly sensitive to noise in their input, and can give an output that has lost its original meaning.

The problems above could be solved by leveraging a Grammatical Error Correction (GEC) model that processes the given text before being sent to any other NLP tool, e.g. a machine translation (MT) system. Given an ungrammatical text containing a varying amount of errors such as morphological, lexical, syntactic, or semantic, the task of a GEC system is to fix the errors and to produce a grammatically correct sentence. An example of a GEC model working is shown in Figure 1.

I like to swimming
GEC model
I like to swim.

Figure 1: Example of a GEC model making a correction.

Current publicly available GEC systems are mostly developed for second language learners and perform poorly in the informal register, usually written by the native speakers, used for customer service (CS). The publicly available GEC systems are also mostly available for English only, and there is no previous work reporting on multilingual GEC for the CS domain.

2. Background

Most modern approaches view **GEC as a sequence to sequence task**, similar to MT, employing encoder-decoder neural architectures (Zhao et al., 2019; Rothe et al., 2021; Náplava and Straka, 2019). The issue is that these methods usually require large amounts of data to work properly, but **data is scarce for the GEC task**, even more for languages other than English.

This leads to the necessity of performing data augmentation. The two most prominent approaches to data augmentation are ***direct noise*** (Kiyono et al., 2019; Jayanthi et al., 2020; Lichtarge et al., 2019) and ***back translation*** (Sennrich et al., 2016; Xie et al., 2018; Stahlberg and Kumar, 2021). In *direct noise*, noise is injected directly into the text, by replacing or deleting characters (or spans of characters) or words (or spans of words), probabilistically. In *back translation*, a model is trained to produce an erroneous text, given clean text. The data produced by these data augmentation techniques is usually used as pre-training data (Zhao et al., 2019; Choe et al., 2019; Junczys-Dowmunt et al., 2018; Stahlberg and Kumar, 2021; Rothe et al., 2021) to pre-train a GEC model that is then commonly fine-tuned on manually created training data for the task.

The current state-of-the-art results (Rothe et al., 2021) were obtained for English with a **T5 model**, and for German with an **mT5 model**. Pre-train data was generated using *direct-noise* approaches, and fine-tune data was gold GEC data for the supported languages. An mT5-XXL was trained on the reversed data, i.e. parallel *correct/incorrect* data, and used to better corrections for the Lang-8 corpus (Mizumoto et al., 2011), **generating the cLang-8 corpus**. Several T5 and mT5 models were then trained on the proposed cLang-8 corpus.

Quality Estimation (QE) is a task to score a model's prediction, based only on the source and the predicted correction. In GEC, QE is most commonly used to to measure the quality of the model hypotheses and re-rank them accordingly. The first QE model (Chollampatt and Ng, 2018b) was based on the predictor-estimator architecture (Kim et al., 2017), based on the RNN (Elman, 1990) and CNN (LeCun et al., 1998) architectures. These models were trained to predict either the $F_{0.5}$ scores or HTER. These models, combined with other simple metrics (e.g. number of words in the hypothesis), were then used to re-rank model's hypotheses. Recently, Liu et al. (2021) proposed the current state-of-the-art approach for QE and re-ranking for GEC, named VERNet. Unfortunately, important details to reproduce their results were missing in the paper, and our re-implementation of the approach did not yield the scores reported in the paper.

Our contributions with this paper are threefold.

- We apply GEC to a new domain, customer service, usually more informal, which differs from prior work that was focused on second language learners domain.
- We extend previous work in multilingual grammatical error correction by covering Portuguese and German, having the first publicly released GEC research work for Portuguese, and also extending the GEC evaluation tool ERRANT (Bryant et al., 2017) to several languages.
- We develop a new state-of-the-art GEC reranking technique, which allows to improve already existing models with little effort, i.e. keeping the GEC model frozen and only needing to process its hypotheses with our developed re-rankers, which we will release as an easy to use framework.. It is also, to the best of our knowledge, the first multilingual and word-level re-ranking system developed for GEC, as well as the first re-ranker and QE model that were developed for a domain other than second language learners. When applied on the original German and Portuguese input segments (before they are fed to the MT systems), our GEC models and re-rankers help to improve the quality of the translations.

3. Multi-task Learning and Reranking for Grammatical Error Correction

Our experiments in this section build on top of the T5-small trained on the cLang-8 dataset, as proposed by Rothe et al. (2021). Our goal is to improve upon the results from this model. Since this model was not publicly released, our first steps were to replicate this model, which we then publicly released on HuggingFace[1].

The T5 is a large pre-trained transformer-based model that was trained on several tasks, making it very adaptable to new tasks, even with small amounts of data. The T5 model achieves state of the art results in several tasks. The T5-small has 60M parameters, and the T5-base has 220M parameters.

3.1. Model Architectural Changes

In this thesis, we first experimented with multitask learning, leveraging the state of the art T5, in an attempt to improve its performance. First, we tried to do generation and classification together, and second we tried adding the copy mechanism on top of the T5 (for the first time, to the best of our knowledge).

[1] http://huggingface.co/Unbabel/gec-t5_small

3.1.1 Predicting Tags Along With Generation

GECToR (Omelianchuk et al., 2020) tackled GEC as a classification problem, and achieved state-of-the-art results, at the time of release. GECToR leverages *KEEP, DELETE, REPLACE, APPEND* tags, as well as custom-developed transformation tags. We leverage their labels' generation method, but convert the tags to binary *KEEP/REPLACE*.

The generation loss is the cross entropy loss between the generated correction and the reference correction, and the classification loss is the cross entropy loss between the predicted binary tags and the reference binary tags. The classification is performed by a Linear layer predicting over the last hidden state of the encoder. When using the T5 model alone, only the generation loss is considered. We now consider both the T5's generation loss, as before, but now we also consider the classification loss. The loss used during training would be given by (1).

$$loss = \alpha \cdot loss_{generation} + \beta \cdot loss_{classification} \quad (1)$$

This could help the model to overcorrect less, helping the model only to change the actual errors.

We tried both by assigning α and β values manually, as well as training these parameters. Neither variant was able to imrpove upon the baseline results.

3.1.2 T5 with Copy Mechanism

Since trying to perform classification along with generation did not improve our results, we decided to tackle another option to make our model overcorrect less, which is of great importance to have a robust model, that is able to understand when to correct, and when to copy from the source. Having a model that does not overcorrect is also very important for the customer support domain. We coupled the copy mechanism (Gu et al., 2016) with the T5, following what Zhao et al. (2019) did. The combination of the copy mechanism with the generated tokens is given by (2).

$$p_{token} = \alpha \cdot p_{gen} + (1 - \alpha) \cdot p_{copy} \quad (2)$$

Before this change, i.e. in standard generation, p_{token} was equal to p_{gen} only. p_{gen} is the probability assigned by the model to each token during generation. p_{copy} is the probability assigned to each of the source tokens by the copy mechanism.

The results obtained are in Table 1, where the scores for the CoNLL-14 (Ng et al., 2014) and BEA-19 (Bryant et al., 2019) are both reported in terms of $F_{0.5}$ score. Baseline is the T5 by Rothe et al. (2021), T5+copy is the T5 and copy mechanism both trained, while f-T5 is the baseline T5 frozen, only training the copy mechanism. In p-copy, the value of α, as defined in (2), is set to 0.5 with a probability of 0.5 (i.e., half of the times the value of α is fixed to the value 0.5), in an approach similar to dropout. With this approach, we slightly outperformed our baseline in the CoNLL-14 test set, but we did not manage to improve the results on the BEA-19 test set.

Table 1: Copy mechanism results

model	CoNLL-14	BEA-19
baseline	60.68	**66.54**
T5+copy	47.61	-
f-T5 + copy	60.66	-
f-T5 + p-copy	**60.74**	66.31

3.2. Beam Search Hypotheses Re-ranking

Next, we keep the T5 model intact, and only process its predictions. The potential gains with reranking, when generation leverages beam search, are present in Table 2.

3.2.1 Re-ranking Based on Language Model's Scores

We leverage BERT (Devlin et al., 2019) to score the model hypothesis, and then re-rank based on these scores.

Several metrics were experimented: (1) Average of probabilities per sentence; (2) Sum of log probabilities per sentence; (3) Probability of the end of sentence (EOS) token; (4) Perplexity score (PPL); (5) Pseudo-likelihood (PPL); (6) Fine-tuned BERT.

The best scoring method was the Pseudo log-likelihood (Salazar et al., 2020), which achieved an $F_{0.5}$ score of 52.72 on the CoNLL-14 test set. See Table 4 for a full comparison of all re-ranking methods.

The main drawbacks, from our point of view, with these language model (LM) approaches is that they do not take into consideration the source, and the perplexity (and other metrics) based on the correction alone does not correlate well with the $F_{0.5}$ metric. LMs only take into account the fluency and the likelihood of the sentence, meanwhile in GEC we want the smallest amount of edits to correct a given sentence, which does not necessarily have to be the most fluent. To overcome this issue, we will cover methods that leverage the source in the following Sections.

3.2.2 Minimum Bayes Risk Decoding

Instead of *maximum a posteriori* (MAP) decoding, Minimum Bayes Risk Decoding (MBR) finds a candidate that minimizes the cost/risk of choosing a candidate hypothesis by comparing to a refer-

Table 2: Beam Search potential gains. Beam size refers to the number of beams/hypotheses used during beam search. *Best hyp* is the score of the highest scoring hypothesis in the beam, in terms of $F_{0.5}$ metric. *Top-1* is the score assigned to the first hypothesis (the hypothesis assigned the highest probability by the T5 model). *Avg-position* is the average position of the best scoring hypothesis in the beam. The *% reference found* is the percentage of times the reference is present in the beam. The *avg ref position* is the average position of the reference in the beam.

beam size	top-1 hyp	best hyp	avg position	% reference found	avg ref position
10	60.62	79.11	1.61	54.88	1.15
100	60.74	86.3	11.4	64.94	8.27

ence hypothesis. References in MBR are approximated by other hypotheses (pseudo-references), meaning each hypothesis will be considered a gold reference once and compared with the remaining hypotheses. The hypothesis that maximizes the scores compared to all pseudo-references is selected.

MBR has been extensively used recently with neural machine translation evaluation metrics, namely COMET (Rei et al., 2020), with some of the most prominent works being Freitag et al. (2021), Fernandes et al. (2022) and Amrhein and Sennrich (2022).

For this study, we adapted the implementation available in COMET[2], using as evaluation metric both ERRANT (Bryant et al., 2017) and M^2 (Dahlmeier and Ng, 2012).

The best method was using as metric ERRANT, which achieved an $F_{0.5}$ score of 54.49 on the CoNLL-14 test set. Table 4 shows a comparison with the remaining re-ranking methods.

This is, to the best of our knowledge, the first work developed with MBR for GEC, but the results did not match the gains obtained with MBR for MT. This is understandable since MBR can be seen as selecting the most consensual hypothesis, which works well for MT because it has some degree of freedom in the translations, where there are several possible correct translations for a given input. The same is not valid for GEC, where there are usually one or two possible minimal corrections to fix a sentence. This lack of freedom in the GEC task does not allow MBR to improve the model scores.

3.2.3 Re-ranking with COMET

COMET-QE (Rei et al., 2021) is a state of the art MT neural quality estimation metric and framework, with a dual encoder architecture, based on XLM-RoBERTa (Liu et al., 2019). This model takes into account both the source and hypothesis to assign a score. For our goals, we trained it with the objective being producing $F_{0.5}$ scores computed by ERRANT for the given source and hypothesis.

The training, development, and testing datasets were created by generating hypotheses with the T5-small for each source. These hypotheses were then scored with ERRANT. These ERRANT scores are used as our golden scores. The training set was the combination of FCE (Yannakoudakis et al., 2011), W&I (Bryant et al., 2019) and NUCLE (Dahlmeier et al., 2013). The dev set the CoNLL-13 test set (Ng et al., 2013).

The best COMET model achieved an $F_{0.5}$ score of 56.05. Table 4 shows a full comparison.

In our manaul analysis we observed that COMET's selections are somewhat conservative. This model tends to prefer corrections with the least amount of edits, even if sometimes more edits are needed.

3.2.4 Re-ranking with TransQuest

Since COMET employs a dual encoder architecture, it excels at capturing whether the *source* and *hypothesis* share the same semantics, but it does not pay much attention to individual words, and the relationship shared between words in the *hypothesis* and the *source*. Our hypothesis for why COMET did not give better results is that errors in sentences are sparse, unlike MT where for each word in the source, there is an operation in the hypothesis and target. So we thought we needed a model that could leverage self-attention in order to understand better what words in the source influence an operation in the hypothesis and create some form of alignment this way. To address this we leveraged TransQuest's framework (Ranasinghe et al., 2020). We trained the TransQuest architecture the same way we trained COMET, to predict the ERRANT scores of given source/hypothesis pairs.

The training set was the cLang-8, which performed better for the TransQuest than the training data used for COMET. The dev set used was the same as COMET.

The best scoring model achieved an $F_{0.5}$ score of 57.43 on the CoNLL-14 test set.

[2] https://github.com/Unbabel/COMET/blob/master/comet/cli/mbr.py

3.2.5 Re-ranking with Word-level TransQuest

COMET leveraged the dual encoder architecture, and TransQuest leveraged self-attention in the input, but despite the improvement of the latter, it is still a sentence-level scorer, it considers the entire sentence simultaneously to predict a score, without paying the necessary attention to the individual words. So, even if a single word in the sentence is wrong it doesn't impact the scores significantly. As mentioned previously, errors in sentences are sparse, and many sentences are error-free. This property makes it very hard for a sentence-level scorer model to understand how to score a sentence correctly. The sparsity makes it hard to predict a single score for all the corrections in a given source/hypothesis pair. We hypothesize that working more locally, at the word level, instead of the sentence level, makes it easier to address these shortcomings of previous models.

We present the first word-level QE model for GEC used to score hypotheses, leveraging TransQuest word level (Ranasinghe et al., 2021). The word level tags were obtained by using DeepSPIN's *QE corpus builder*[3]. We generate the score sentence from the binary word-level tags, with the expression in (3).

$$score(sent_k) = \frac{|OK_{hyp_k}|}{|OK_{hyp_k}| + |BAD_{hyp_k}|} \quad (3)$$

This formula calculates the ratio between the number of OK tags in the hypothesis over the sum of OK and BAD tags in the hypothesis (equivalent to the sum of the number of words and spaces in the hypothesis). From the few expressions we tried, including explicit information from source tags did not improve performance. Despite not leveraging the source information explicitly, source information was already leveraged to generate the target tags implicitly.

The best scoring model obtained an $F_{0.5}$ score of 59.86 on the CoNLL-14 test set, see Table 4 for full comparison.

3.2.6 Multi-metric Re-ranking

Thus far, we have only seen re-ranking based on a single metric, namely the QE score from one of our models.

In this section, following the work of Fernandes et al. (2022) and Chollampatt and Ng (2018a), we extend our single-metric re-ranking solution so that it considers multiple metrics.

What we did to combine the scores is a weighted sum, assigning a weight to the QE scores and the T5-likelihood. The sentence score was then given by (4).

$$score = \alpha \times T5 + (1 - \alpha) \times QE \quad (4)$$

The best scoring model was leveraging T5, COMET and Word-level TransQuest, which achieved an $F_{0.5}$ score of 62.78 on the CoNLL-14 test set, which is an improvement of over 2 $F_{0.5}$ points compared to the baseline. These are new state-of-the-art results.

Table 3: T5-base re-ranking results

Model	$F_{0.5}$
T5-base	64.0
T5 + (W-TransQuest + COMET)	**65.54**

Table 4: Results from re-ranking experiments

Experiment	$F_{0.5}$
T5-small	60.68
Language Model	52.72
MBR	54.49
COMET	56.05
TransQuest	57.43
W-TransQuest	59.86
T5 + COMET	61.6
T5 + TransQuest	62.01
T5 + (TransQuest + COMET)	62.34
T5 + W-TransQuest	62.65
T5 + (W-TransQuest + COMET)	**62.78**

3.2.7 Re-ranking T5-base

After observing the potential of our re-ranking approaches on the small T5 pre-trained model, i.e T5-small, we decided to evaluate its impact on the larger version of this model, i.e. T5-base, that has over 3 times more parameters and produces results of over 3 $F_{0.5}$ scores better than its smaller counterpart. Taking the best scoring model for the T5-small, we used it to re-rank the T5-base. The results are in Table 3. The improvement is smaller compared to the gains obtained for the T5-small, but it is still significant. The slight decrease in results is expected, since the QE models were trained on hypotheses generated with the T5-small.

4. Domain Adaptation Multilingual GEC

In this Section, we describe the development of GEC models for a new domain, the customer service (CS) domain, by leveraging the data provided

[3] https://github.com/deep-spin/qe-corpus-builder

by Unbabel. We conduct experiments in German and Portuguese. Since the T5 only supports English, for the experiments conducted in this section, the mT5 model, i.e. the multilingual version of the T5 model, is used. The mT5-small model has 300M parameters, meanwhile the T5-small has 60M parameters. We propose the data processing, and the models developed to adapt GEC models to a new domain, the Customer Service domain.

4.1. Data

ERRANT (Bryant et al., 2017) can take parallel ungrammatical/grammatical data and automatically annotate it with error types, but it was developed for English only. Boyd (2018) extended the ERRANT toolkit for German. Following this approach, we extended it to Spanish, French, and Portuguese as well (while correcting some errors in the original code for German). ERRANT is necessary to generate the reference M2 annotated files, and to compute the scores we will later use for re-ranking.

Since the dataset provided by Unbabel (gold data) does not have enough data to train a GEC model (see Table 5 for the statistics), we will perform data augmentation.

Table 5: Unbabel data sizes.

Language	Train	Dev	Test
de	2.4k	996	1.5k
pt	1k	980	1.4k

Firstly, we use Unbabel's in-domain data, inject noise through *direct-noise* approaches and use this data as pre-training data for our GEC model, which we then fine-tune with the gold Unbabel data.

Secondly, we take that same data injected with noise, reverse it, so now the source is error-free and the target is erroneous, and feed it to a model, i.e. our *backtranslation* model, which we then fine-tune on the reversed Unbabel gold data. This *backtranslation* model is then used to generate a dataset, which is used to train a GEC model.

The *direct-noise* operations we make use of are the following, i) drop spans of tokens, ii) swap tokens, iii) drop spans of characters, iv) swap characters, v) insert characters, vi) lower-case a word, vii) upper-case the first character of a word, all with a 10% probability. We also kept 5% of sentences from both datasets completely unchanged, so models also learn that sentences can be grammatical (as mentioned by Rothe et al. (2021)). Our script to inject these errors is available in GitHub[4].

4.2. Models

In this Section, different training regimes are experimented to maximize the performance in the new domain. We also experiment with some of our proposed re-ranking methods.

4.2.1 Pre-training

As a first experiment, we generated pre-training data like discussed in Section 4.1, using some Unbabel sentences and *direct-noise*. This data was first used as the actual training data for our correction model to establish a baseline if no gold data was available. See Table 6 for a comparison of all methods.

4.2.2 Pre-training and Fine-tuning

As a second experiment, we use the data from Section 4.2.1 as pre-training data. The training regime adopted is to train the mT5 model on the pre-training data until convergence and then to fine-tune the model on the gold Unbabel data until convergence.

The results are in Table 6. It is noticeable, as expected, that there is quite a big jump in the model's performance. Despite having very little data, only 1k sentences for the Portuguese data were able to boost the results quite significantly.

4.2.3 Using Publicly Available Data

As a subsequent experiment, we tried the regime of pre-training first on publicly available GEC data and then fine-tuning on the Unbabel gold data. The public data used is the Falko-Merlin dataset (Boyd, 2018), comprised of 250K sentences. This is only possible to experiment with German data since there is no publicly available Portuguese GEC data. The results are in Table 6.

When only pre-training is done, the best results are obtained using publicly available GEC data and not the data generated with direct-noise. When we then fine-tune both of these models, the model pre-trained with the data generated with direct-noise achieves better results.

This indicates that despite the data generated with direct-noise being of lower quality, it provides the model with data that better matches the domain of the test set, which seems to be more important as pre-training than having better quality.

4.2.4 Using Data Generated by the Backtranslation Model

Our last data augmentation experiment, as mentioned in Section 4.1, is to take an mT5 model and train on the reversed pre-training data. Then this

[4] https://github.com/Joao-Maria-Janeiro/GEC-data-augmentation

Table 6: Training regimes results.

Experiment	DE ($F_{0.5}$)	PT ($F_{0.5}$)
pre-train	35.33	18.33
pre + fine	51.12	31.19
pre$_{falko-merlin}$	41.08	-
pre$_{falko-merlin}$ + fine	46.16	-
backtrans	36.98	18.70

Table 7: Re-ranking experiments with publicly available data.

experiment	$F_{0.5}$
baseline	51.12
COMET	37.82
TransQuest	35.73
W-TransQuest	46.68
mT5 + COMET + TransQuest	51.46
mT5 + COMET + W-TransQuest	**51.85**

model is fine-tuned on the reversed training data of the Unbabel gold data.

Unfortunately, no more data was available for us to use, so we generated new errors for the pre-train data, distilling our reverse model's knowledge into it. Taking that newly generated dataset, we trained a GEC mT5 model on that generated data. The results are in Table 6.

These results are more in-line with the pre-training results, being slightly better.

4.2.5 Reranking With Our Method

As a last experiment, we leverage our re-ranking method, now to re-rank the hypotheses of the model developed in Section 4.2.2.

We conduct the experiments using both publicly available datasets and in-domain data. Both COMET and TransQuest models were trained with hypothesis generated by the model from section 4.2.2 for the Falko-Merlin dataset. We chose the dataset with 250K sentences. Since there is no public GEC dataset for Portuguese, we only conducted this experiment with German data. The results for the publicly available data are in Table 7.

We now apply the same re-ranking solutions, but training the re-ranker models with in-domain Unbabel data. The data used to train these re-rankers models is a subset of the *pre-training data* used in Section 4.2.1, where we randomly select 250K sentences, to match the size of the dataset used in the experiments with publicly data. The results obtained are shown in Table 8. Taking into account previous results, it is not surprising that using in-domain data improves upon using public data.

Table 8: Reranking experiments with German and Portuguese in-domain data.

experiment	DE($F_{0.5}$)	PT($F_{0.5}$)
baseline	51.12	31.19
COMET	36.89	18.84
TQ	41.35	29.03
mT5 + COMET + TQ	**51.55**	**33.37**

Table 9: Results from using GEC as cleanup step for MT, computed by COMET-QE-MQM.

Lang	SRC	REF	mT5	reranked
de	0.1176	0.1201	0.1188	0.1190
pt	0.1061	0.1086	0.1071	0.1073

4.3. GEC as pre-processing for MT

In order to assess whether our GEC models could improve machine translation quality, we preprocess our inputs by correcting them with our model before feeding them to the MT model. The German and Portuguese data were translated with Ng et al. (2020) and Lopes et al. (2020), respectively. The German model was one of the best performing models from WMT19, and the Portuguese model was one of the best scoring models from WMT20. These models are already strong, and able to handle slightly noisy inputs, due to the data used to train them, which makes it hard to get large further improvements with cleaner inputs. Since we did not have access to a dataset with *source*, *source corrected* and *reference translation*, we could not use a reference-based metric for our translation evaluation. The model used to score these translations was the COMET-QE-MQM (Rei et al., 2021), which is a QE, i.e. referenceless, metric trained to predict MQM scores. The dataset used to test the translations was the Unbabel test set.

The results are presented in Table 9, where it is possible to see that using the GEC model is helpful in both languages, and that our state-of-the-art reranker works better than the mT5, as expected.

5. Conclusions

One of the goals of this thesis was to develop GEC models for a new domain, not previously explored, the CS domain. Another goal was to further extend the existing work on multilingual models for the GEC task, by developing the first public GEC research work for Portuguese, as well as a new iteration of German models, in our case for a different domain. A necessary first step to achieve these goals was to reproduce the results from the state of the art approach from Rothe et al. (2021). We successfully developed models for this new domain, by exploring data augmentation techniques, training regimes and re-ranking approaches. We define the baselines for this new domain, that future work can take as reference for comparison.

The most ambitious goal of this thesis, was to develop a new state of the art re-ranking system. We were able to achieve this goal by introducing several novel method for the GEC field. We utilized existing QE models mainly developed for MT, and introduced a novel method with a word-level QE model that we leverage to generate a sentence score, to the best of our knowledge it is the first time such an approach is taken in the GEC field. It is also the first time a multilingual re-ranker is developed for the GEC task. We observed that errors in sentences are sparse, and based on this observation we hypothesised that it would be advantageous to work more locally, at the word level, instead of at the sentence level. Taking that into account, we developed our word-level QE, and the results of our experiments confirmed our hypothesis. Our re-ranking method was also advantageous in the new domain, which helps demonstrate its robustness, even in languages other than English.

Our developed GEC multilingual models were also able to improve MT quality, being used as an additional pre-processing steps to improve the quality of the input segments.

6. Future Work

In this work, we developed a word-level quality estimation model. For future work, it would be interesting to develop an edit-level quality estimation model. This would likely be better than word level since we would be classifying the edits at once; meanwhile, when we classify edits at the word level, it is confusing which words are affected by an edit, and we have multiple tags for a single edit. Having a single tag for an edit is identical to the M^2 metric, which might be able to give a better correlation.

For our word-level model, we used a multilingual word tag generator, which had some flaws when working monolingual. For future work, it would be interesting to develop tags based on ERRANT extracting from the edits, which would be better aligned with the task. It would also be fruitful to perform a search to find the best way to get a sentence score from the word level tags.

Concerning the data, for future work, the pre-training data generated with *direct noise* could have the same error distribution as the test set to better align with the new domain.

Unfortunately, we did not have the resources to train a T5-XXL, following Rothe et al. (2021), and re-rank it with our technique, this would be a logical next step, if the resources to train such a model are met.

References

C. Amrhein and R. Sennrich. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for comet. *arXiv preprint arXiv:2202.05148*, 2022.

A. Boyd. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6111. URL https://aclanthology.org/W18-6111.

C. Bryant, M. Felice, and T. Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1074. URL https://aclanthology.org/P17-1074.

C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4406. URL https://aclanthology.org/W19-4406.

Y. J. Choe, J. Ham, K. Park, and Y. Yoon. A neural grammatical error correction system built on better pre-training and sequential transfer learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4423. URL https://aclanthology.org/W19-4423.

S. Chollampatt and H. T. Ng. A multilayer convolutional encoder-decoder neural network for grammatical error correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), Apr. 2018a. doi: 10.1609/aaai.v32i1.12069. URL https://ojs.aaai.org/index.php/AAAI/article/view/12069.

S. Chollampatt and H. T. Ng. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels, Belgium, Oct.-Nov.

2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1274. URL https://aclanthology.org/D18-1274.

D. Dahlmeier and H. T. Ng. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/N12-1067.

D. Dahlmeier, H. T. Ng, and S. M. Wu. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-1703.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

P. Fernandes, A. Farinhas, R. Rei, J. G. C. de Souza, P. Ogayo, N. Graham, and A. F. T. Martins. Quality-Aware Decoding for Neural Machine Translation. In *Proceedings at the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, Washington, july 2022. Association for Computational Linguistics. URL https://openreview.net/pdf?id=aYNjzZJP9qa.

M. Freitag, D. Grangier, Q. Tan, and B. Liang. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *arXiv preprint arXiv:2111.09388*, 2021.

J. Gu, Z. Lu, H. Li, and V. O. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1154. URL https://aclanthology.org/P16-1154.

S. M. Jayanthi, D. Pruthi, and G. Neubig. NeuSpell: A neural spelling correction toolkit. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 158–164, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.21. URL https://aclanthology.org/2020.emnlp-demos.21.

M. Junczys-Dowmunt, R. Grundkiewicz, S. Guha, and K. Heafield. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1055. URL https://aclanthology.org/N18-1055.

H. Kim, H.-Y. Jung, H. Kwon, J.-H. Lee, and S.-H. Na. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1), sep 2017. ISSN 2375-4699. doi: 10.1145/3109480. URL https://doi.org/10.1145/3109480.

S. Kiyono, J. Suzuki, M. Mita, T. Mizumoto, and K. Inui. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1119. URL https://aclanthology.org/D19-1119.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

J. Lichtarge, C. Alberti, S. Kumar, N. Shazeer, N. Parmar, and S. Tong. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1333. URL https://aclanthology.org/N19-1333.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Z. Liu, X. Yi, M. Sun, L. Yang, and T.-S. Chua. Neural quality estimation with multiple hypotheses for grammatical error correction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5452, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.429. URL https://aclanthology.org/2021.naacl-main.429.

A. Lopes, R. Nogueira, R. Lotufo, and H. Pedrini. Lite training strategies for Portuguese-English and English-Portuguese translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 833–840, Online, Nov. 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.90.

T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand, Nov. 2011. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I11-1017.

J. Náplava and M. Straka. Grammatical error correction in low-resource scenarios. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 346–356, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5545. URL https://aclanthology.org/D19-5545.

H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-3601.

H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, 2014.

N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov. Facebook fair's wmt19 news translation task submission. In *Proc. of WMT*, 2020.

K. Omelianchuk, V. Atrasevych, A. Chernodub, and O. Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.bea-1.16. URL https://aclanthology.org/2020.bea-1.16.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

T. Ranasinghe, C. Orasan, and R. Mitkov. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

T. Ranasinghe, C. Orasan, and R. Mitkov. An exploratory analysis of multilingual word level quality estimation with cross-lingual transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

R. Rei, A. C. Farinha, C. Zerva, D. van Stigt, C. Stewart, P. Ramos, T. Glushkova, A. F. T. Martins, and A. Lavie. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.111.

S. Rothe, J. Mallinson, E. Malmi, S. Krause, and A. Severyn. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association* *for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.89. URL https://aclanthology.org/2021.acl-short.89.

J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.240. URL https://aclanthology.org/2020.acl-main.240.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

F. Stahlberg and S. Kumar. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online, Apr. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.bea-1.4.

Z. Xie, G. Genthial, S. Xie, A. Ng, and D. Jurafsky. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1057. URL https://aclanthology.org/N18-1057.

H. Yannakoudakis, T. Briscoe, and B. Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1019.

W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1014. URL https://aclanthology.org/N19-1014.