Detecting Critical Errors Considering Cross-Cultural Factors in English-Korean Translation

Sugyeong Eo[1], Jungwoo Lim[1], Chanjun Park[2], Dahyun Jung[1], Seonmin Koo[1], Hyeonseok Moon[1], Jaehyung Seo[1], Heuiseok Lim[1]
[1]Korea University, [2]Upstage
{djtnrud,wjddn803,dhaabb55,fhdahd,glee889,seojae777,limhseok}@korea.ac.kr
chanjun.park@upstage.ai

Abstract

Recent machine translation (MT) systems have overcome language barriers for a wide range of users, yet they still carry the risk of critical meaning deviation. Critical error detection (CED) is a task that identifies an inherent risk of catastrophic meaning distortions in the machine translation output. With the importance of reflecting cultural elements in detecting critical errors, we introduce the culture-aware "Politeness" type in detecting English-Korean critical translation errors. Besides, we facilitate two tasks by providing multiclass labels: critical error detection and critical error type classification (CETC). Empirical evaluations reveal that our introduced data augmentation approach using a newly presented perturber significantly outperforms existing baselines in both tasks. Further analysis highlights the significance of multiclass labeling by demonstrating its superior effectiveness compared to binary labels.

Keywords: Quality estimation, Critical error detection, Neural machine translation, Large language model

1. Introduction

Recent studies have exhibited remarkable achievements in machine translation (MT), overcoming language barriers for a broad spectrum of users. Yet, MT output inevitably is under a risk of catastrophic meaning deviations (Sharou and Specia, 2022; Tang et al., 2022).

Critical error detection (CED) task aims at identifying the meaning distortions that arise during the translation process by referring to the source sentence and its MT output (Specia et al., 2021; Zerva et al., 2022; Chen et al., 2021; Rubino et al., 2021). While the task is designed to classify binary labels, it includes error cases on whether the example belongs to critical errors associated with *toxicity, safety, named entity, sentiment, and number*. Although those risky cases rarely emerge, a single fatal semantic deviation within the scope of CED error attributes may incur devastating consequences in daily life. To prevent MT output from causing serious ethical, social, financial, or legal issues, the significance of the task is increasingly emphasized (Sharou and Specia, 2022; Eo et al., 2022; Jiang et al., 2021a).

Despite the significance of CED, errors related to cultural elements remain overlooked. Language in each country is infused with its distinct culture, and translation serves as a medium to facilitate communication between speakers of these languages. Since various countries around the world have their distinct cultural expression or standards based on their history and way of life, translation aligned with the cultural property is essential. Especially, the honorific expressions in Korean, reflecting a culture intrinsically emphasizing courtesy and respect, serve as one of the representative cases for this viewpoint. Among various honorific forms in Korean, the speaker has to adopt an adequate level considering the subject or object's hierarchical or social status. As this is distinguishing from English, the translation results are often regarded as informal or disrespectful expressions.

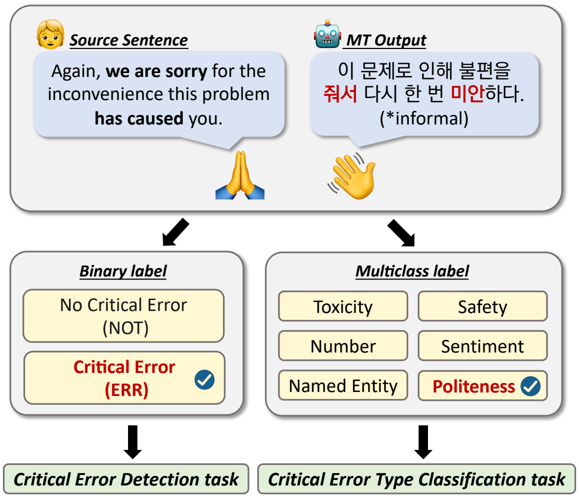

Source Sentence
Again, we are sorry for the inconvenience this problem has caused you.
MT Output
이 문제로 인해 불편을 줘서 다시 한 번 미안하다.
(*informal)
Binary label
No Critical Error (NOT)
Critical Error (ERR) ✓
Multiclass label
Toxicity
Safety
Number
Sentiment
Named Entity
Politeness ✓
Critical Error Detection task
Critical Error Type Classification task

Figure 1: Illustration of the KNOTICED dataset example

To this end, this paper introduces KNOTICED, a critical error detection dataset for English-Korean MT [1]. Considering the Korean cultural aspect, we additionally include a new "Politeness" error type;

* Corresponding Author

[1] KNOTICED dataset is available at: https://github.com/sugyeonge/KNOTICED.

LREC-COLING 2024, pages 4705–4716
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

examples with this type address honorific issues that may arise in English-Korean translation. Furthermore, we additionally provide the fine-grained error type labels for detailed error detection, enabling two types of tasks: (1) **Critical error detection (CED)** (binary classification) and (2) **Critical error type classification (CETC)** (multiclass classification) task.

The construction process for the dataset begins with the selection and refinement of source data. Annotators then inject critical errors into the data in accordance with the provided guidelines. In the process, the definition, scope, and injection samples for each type have been meticulously crafted and refined based on applying a pilot annotation test. With a particular emphasis on ensuring high quality, we further implement quality evaluation after the error injection, wherein humans additionally annotate the inspection result. Figure 1 shows the composition of our dataset: it comprises source sentences, MT output, binary critical error detection labels, and multiclass critical error type labels.

In the experiment, this study present a generative model named perturber that produces MT outputs containing critical errors. ChatGPT (OpenAI-Blog, 2022) is employed to assign labels to the perturbed samples, which are subsequently used for data augmentation. Experimental results on the KNOTICED dataset demonstrate the effectiveness of the data augmentation method via perturber, outperforming pre-trained language model (PLM) and large language model (LLM) baseline by a considerable margin in both of the tasks. We also reveal that exploiting multiclass labels in model training is more effective than utilizing binary labels in the CED task, highlighting the necessity of introduced additional critical error type labeling. Our contributions are three-fold:

- This paper presents KNOTICED, a dataset for the critical error detection task in English-Korean MT. Cultural aspect is considered in the dataset, by introducing a new "Politeness" error type to reflect Korean honorific etiquette.

- We additionally annotate fine-grained error types in the dataset, enabling both critical error detection and critical error type classification task. Further analysis demonstrates the benefits of using multiclass labels.

- Our simple but effective data augmentation method through perturber outperforms the baseline by a significant margin in both tasks.

2. Related Work

Critical error detection (CED) is the task of detecting cases whether there is a catastrophic meaning distortion by referring to the source sentence and the MT output sentence (Specia et al., 2021; Zerva et al., 2022; Costa-jussà et al., 2023; Chen et al., 2021). It is first introduced in the 6th Conference on Machine Translation in 2021 (WMT21)[2], with the aim of detecting fatal meaning deviations. Although instances of critical errors are not frequently observed, their occurrence may lead to considerable disruptions in daily life. Even errors can incur serious problems, resulting in a loss of user trust and disuse of the translation engine, thereby emphasizing the importance of detecting such errors (Sharou and Specia, 2022; Eo et al., 2022; Jiang et al., 2021a).

The existing CED dataset contains examples with five common error types that are agnostic to the language, while they do not provide labels: *toxicity, safety, named entity, number, and sentiment*. These errors appear in three forms: *deletion*, where an expression in the source sentence disappears; *hallucination*, where an expression not present in the source is added; and *mistranslation*, where an expression is translated into a different meaning than in the source sentence. While all the error types are language-agnostic, those that are culturally aware remain unaddressed, despite their significance in the MT field. A comparison of our dataset with existing datasets is presented in Table 1. KNOTICED introduces a language-dependent politeness error type along with the five language-agnostic error types. Given that honorific etiquette plays a pivotal role in Korean communication, the dataset allows for an inspection of its distinct property. Moreover, additional annotation of multiclass labels supports a CETC task.

The task shares similarities with the domain of hallucination (Xu et al., 2023; Wang and Sennrich, 2020; Zhou et al., 2021; Raunak et al., 2021). Hallucination is defined as MT output completely disconnected from the original source, resulting in content that is not grounded in the source sentence (Lee et al., 2019; Ji et al., 2023; Guerreiro et al., 2023a,b; Voita et al., 2021). CED distinguishes itself by focusing on detecting catastrophic semantic shifts that induce real-life, ethical, social, economic, legal, or safety issues. In the context of CED, hallucination is characterized as one form of critical error. The area of offensive language detection (Zampieri et al., 2019; Deng et al., 2022; Jeong et al., 2022) also has relevance. CED is restricted to the translation task, and the issues addressed within offensive language detection can be included under the *toxicity* type errors.

[2] https://www.statmt.org/wmt21/

Dataset	Pair	# Training Data	# Error Type	# Binary Label	# Multiclass Label
WMT 21 (Specia et al., 2021)	En-De	7,878	5	✓	✗
	En-Cz	7,476	5	✓	✗
	En-Zh	6,859	5	✓	✗
	En-Ja	7,658	5	✓	✗
WMT 22 (Zerva et al., 2022)	En-De	155,511	5	✓	✗
	Pt-En	39,925	5	✓	✗
KNOTICED (Ours)	En-Ko	7,265	**6**	✓	✓

Table 1: Comparison of KNOTICED to previous CED datasets. The binary class label is to detect critical error existence, and the multiclass label is used for classifying the critical error type.

3. KNOTICED

The dataset construction process is partitioned into three processes: 1) source data selection and refinement, 2) critical error injection, and 3) quality evaluation. The constructed datasets consist of the source sentence, its MT outputs, binary labels indicating the critical error existence, and multiclass labels indicating the critical error type.

3.1. Design Considerations

As a step before constructing KNOTICED dataset, we establish the settings guided by the following two design considerations. First, the source data should be selected from domains that are most likely to be adversely affected by critical error types. Second, since the standard of critical error's definition, range, and fine-grained types can be varied, our standard should be consistent with previous research. Consequently, we adhere to the overall standard and instructions presented by WMT21.

3.2. Source Data Selection and Refinement

Source data We adopt the "Daily Conversation and Colloquial Translation Corpus[3]" released by AIHUB on July 29, 2022, as our source text. Following the design consideration, we assume that daily life scenarios are the most exposed to various risks from severe semantic change so that, and this data well-covers such domain. The dataset consists of English-Korean parallel sentences with MT output.

In line with the criteria set by WMT21, neither general translation errors nor accurate translations are considered critical errors. However, to distinguish critical errors from a diverse range of translation results, we decided to include both translation results as examples. To facilitate this, the source data is segmented into four categories.

1. Source sentence + MT output (Error injection)
2. Source sentence + Reference translation (Error injection)
3. Source sentence + MT output (No error injection)
4. Source sentence + Reference translation (No error injection)

Categories 1. and 2. consist of 1,000 examples each, into which we plan to inject critical errors. Category 3. has 3,500 examples, encapsulating correct or incorrect translations but not considered a critical error case. Category 4., totaling 3,300 cases, is a correct translation. Note that errors are not inserted into all instances of this type; rather, critical errors are selectively injected by human annotators.

Data refinement MT output in the source data may include critical translation errors itself. Accordingly, a preliminary inspection is conducted for both Categories 1. and 3., which include MT output, to retain only those instances without critical errors. For clarification of the differences between the four categories, we filter out instances in Category 4. where unexpected fragments not appeared in the source sentence. From these procedures, the examples in each category exhibit distinctive characteristics.

3.3. Schema for KNOTICED

- **Toxicity (TOX):** Toxicity error refers to cases where the meaning distortion poses a risk related to religion, gender, and race. This includes slurs that have a high risk of causing ethical and social issues, such as the N-word, profanity, insults, etc.
- **Safety and health (SAF):** Safety and health error is a case where meaning distortion could potentially lead to safety or health problems. For instance, this case includes the semantic shift of usage instructions for medicines directly related to human life or safety, dietary instructions, etc.
- **Named entity (NAM):** Named entity error refers to situations where the errors impede the restoration of named entities, such as names, organizations, places, etc.
- **Number (NUM):** Numeric error involves distortion in quantities or units in numbers, time, dates, etc. This could pose potential issues in economic, financial, and legal scenarios as well as day-life timing and measurements.
- **Sentiment (SEN):** Sentiment error occurs when the polarity of the sentiment is changed due to the meaning distortion. This includes cases where strong polarity changes to weak polarity and vice versa.
- **Politeness (POL):** This study introduces a new culture-aware type, politeness. Due to the cul-

[3] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71265

Type	Source Sentence	MT Output with Critical Error	Reference Translation	Description
TOX	The maintenance fee includes garbage pick-up, gardening, and snow removal.	유지비에는 병신 (a**hole) 수거, 정원 가꾸기, 제설 비용이 포함됩니다.	유지비에는 쓰레기 수거, 정원 가꾸기, 제설 비용이 포함됩니다.	The ambiguity of the word leads to its translation into an offensive term.
SAF	People who are pregnant and breast-feeding need to use products that are free of toxic chemicals .	임신 중이거나 모유 수유를 하는 사람들은 독성 화학 물질 (toxic chemicals) 제품을 사용해야 합니다. 없는 (free of) .	임신 중이거나 모유 수유를 하는 사람들은 독성 화학 물질이 없는 제품 을 사용해야 합니다.	Meaning is deviated into encouraging the use of products containing potentially toxic substances.
NAM	I could spend a few weeks in Jordan just eating the local dishes there.	나는 요던 (Yodan) 에서 현지 요리를 먹는 것만으로도 몇 주를 보낼 수 있습니다.	나는 요르단 에서 현지 요리를 먹는 것만으로도 몇 주를 보낼 수 있습니다.	The country name is replaced with an out-of-vocabulary word. It hinders the overall understanding of the sentence's context.
NUM	However, I've encountered ads at least 10 times while I was streaming music over three days.	그러나 3일 동안 음악을 스트리밍하는 동안 광고를 10배 (multiplication) 이상 접했습니다.	그러나 3일 동안 음악을 스트리밍하는 동안 광고를 10번 이상 접했습니다.	The word is incorrectly translated, resulting in a substantial numerical discrepancy.
SEN	The invisible blue light of ultraviolet light is a concern for all photographers.	자외선의 보이지 않는 푸른 빛은 모든 사진 작가의 우려 (worry) 입니다.	자외선의 보이지 않는 푸른 빛은 모든 사진 작가의 관심사 입니다.	The term is mistranslated due to its ambiguity, entirely reversing the polarity.
POL	Again, we are sorry for the inconvenience this problem has caused you.	이 문제로 인해 불편을 줘서 (*informal) 다시 한 번 미안하다 (*informal) .	이 문제로 인해 불편을 드려 다시 한 번 죄송합니다 .	A formal apology is mistranslated into an informal expression used towards someone of a lower age or status, which may be perceived as impolite.

Table 2: Examples of injected errors by type in KNOTICED dataset. The highlighted tokens indicate erroneous part.

tural emphasis on courtesy, Korean consists of specified levels of honorific expressions depending on the situation. Therefore, when translating English sentences into Korean, it is necessary to properly judge the degree of honorifics according to the context. We classify only a case where the translated sentence is perceived as disrespectful and impolite due to the use of informal expressions in contexts that require honorifics. We also classify neologisms used by many young generations as critical errors when they are used inappropriately in the situation.

3.4. Annotation Process

Annotator recruitment To inject critical errors into the examples, we employ three human annotators. Given the necessity to reference both the source and MT output, we specifically recruit human annotators proficient in both English and Korean. We require that at least one of these languages should be their native language. To create high-quality data, the definition and purpose of the CED task are provided to annotators. This clarifies the directions for a CED task in the labeling process.

Guidelines Our guideline formulates the definition, scope, and examples of error types. From the perspective of the overall annotation process, we encourage annotators to avoid repeating specific word selections or similar patterns. Also, when there is an example that can be tagged over two labels, we request to tag the one that is perceived as more critical. In individual label instructions, we adhere to the guidelines by WMT21[4]. Concerning the 'POL' error, we add a condition that errors should be inserted primarily when the context of the sentence necessitates adherence to etiquette.

We apply a pilot annotation test on 30 randomly selected samples with the guidelines. This is to ensure that the guidelines cover all kinds of error injection by correcting inappropriate or misrepresented parts.

Error injection and labeling procedure As described in Section 3.2, we ask human annotators to inject and annotate errors for Categories 1. and 2. Initially, annotators select the example that is suitable for the desired error types. Following this, they inject critical errors in the form of mistranslation, hallucination, or deletion into the source or MT output. In the final stage, they annotate error labels to the corresponding sample, tagging the 'ERR' label for binary class and fine-grained error types for multiclass. Any examples without injected errors are automatically assigned a 'NOT' label.

3.5. Quality Evaluation

To eliminate inaccurately injected error instances for the high-quality dataset, we additionally recruit three annotators and provide them with the same guidelines. These annotators are requested to refer to the error-injected examples with labels and annotate '0' if they disagree with the generated example-label pair or '1' otherwise. To assist in establishing clear boundaries for critical errors, we randomly insert 200 examples that do not have critical errors. To include all the potential concerns regarding critical errors, we filter out instances where

[4] We leverage the same guidelines presented in the following link: https://statmt.org/wmt21/quality-estimation-task_critical-error-examples.html

Attribute	Training	Eval	Test
Size	7,265	500	1,000
Avg SRC toks	14.26	14.22	13.95
Avg MT toks	10.10	10.00	9.86
Min/max SRC toks	8/39	8/31	8/40
Min/max MT toks	3/30	4/25	5/25
Binary label			
# NOT	6,606	444	924
# ERR	659	56	76
% ERR	9.98%	12.61%	8.23%
Multiclass label			
# TOX	133	6	7
# SAF	122	15	10
# NAM	95	12	20
# NUM	116	6	12
# SEN	110	12	14
# POL	83	5	13

Table 3: Statistics of the KNOTICED dataset

all annotators disagree with the example and accept the remaining examples.

3.6. Data Analysis

We present samples and statistics of the KNOTICED dataset in Table 2 and Table 3, respectively.

Data Samples Errors injected by type in Table 2 are all cases considered critical errors due to semantic changes in the translation process. The **TOX** type example illustrates a case where the term "garbage" has been mistranslated as "a**hole" instead of "trash". The ambiguity of the word "garbage" leads to its translation into an offensive term. In the **SAF** type, a phrase expressing the absence of toxic chemicals is split into "toxic chemicals" and "free of", with "free of" being attached after the sentence had ended. This converts the sentence's polarity, potentially encouraging the use of a product containing toxic substances. For the **NAM** label, the country name "Jordan" is replaced with "Yodan," an out-of-vocabulary word. As this word doesn't exist in Korean, it hinders the overall understanding of the sentence's context. In the **NUM** type, the word "times" is incorrectly translated to "multiplication," resulting in a substantial numerical discrepancy. For the **SEN** type, the term "concern" is mistranslated due to its ambiguity. The word is translated from a positive meaning, "interest," to a negative one, "worry," entirely reversing the polarity. For the **POL** type, a situation requiring a formal apology for the inconvenience caused is mistranslated into an informal expression used towards someone of lower age or status. This can be perceived as impolite and potentially problematic.

Data Statistics Table 3 details the statistics of our dataset. We generate a size of 8,765 examples, out of which we randomly separate 500 for the evaluation set and 1,000 for the test set. The overall error ratio ranges between approximately 8% and 12%. Reflecting that the critical errors are a long-tailed problem, our dataset intentionally maintains this imbalanced setting among labels. We also present the count of each multiclass label. The distribution of error types that could arise in actual translations is unknown. Therefore, we did not equalize the label distribution for the training, evaluation, and test datasets.

4. Experiments

We experiment with two tasks with the KNOTICED dataset: critical error detection (CED) and critical error type classification (CETC). The input for both tasks is source sentence and MT output without reference translation. The model should classify one of the binary labels or multiclass labels depending on the task. During the experiments, we propose a data augmentation method through perturber.

4.1. Baseline

Pre-trained language model As baseline models for the KNOTICED dataset, we exploit representative multilingual pre-trained models (PLMs) that hold general knowledge of English and Korean languages. Specifically, we conduct experiments with models of multilingual BERT (mBERT) (Devlin et al., 2019), multilingual BART25 (mBART25) (Liu et al., 2020), multilingual BART50 (mBART50) (Tang et al., 2021), and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) from HuggingFace (Wolf et al., 2020)[5]. We use the last hidden state at position `[CLS]`. Then, linear projection with a classifier and softmax function are sequentially applied to classify the final label. The model is trained for 30 epochs using one A100 GPU, setting a maximum sequence length of 128 and a learning rate of 2e-5. As for batch size, we set mBART to 64 and all others to 128, without any additional gradient accumulation.

Large language model To investigate the ability of ChatGPT to identify critical errors, we also conduct experiments on the CED and CETC tasks. Given that the performance of ChatGPT vary significantly based on the prompting, we design the prompt into three according to the information utilized: plain, demonstration, and description.

- **Plain**: We omit any explanations describing the task. The prompt is simply designed to predict the existence of critical errors or fine-grained labels for the given input.
- **Demonstration**: One example for each label is provided as a part of the prompt. The demonstration examples per label are randomly selected

[5] https://huggingface.co/models

from the training dataset. Subsequently, it is prompted to classify the appropriate labels for the given input.

- **Description**: Task descriptions for CED provided by WMT21, or descriptions for each type, are included in the prompt. The model is prompted to select the appropriate labels for the provided input.

4.2. Augmentation via Perturber

To enhance the performance of identifying critical errors, we propose a simple and effective data augmentation method through perturber. The perturber is a generative model that produces MT output containing critical errors from an input source sentence. The sentences generated in this manner are then labeled and augmented into the training dataset for further use in fine-tuning the model.

Perturber model training We extract instances with the 'ERR' label from KNOTICED, creating a dataset that contains only instances with critical errors. We then train the Transformer (Vaswani et al., 2017b) model by feeding the source sentence as input and training the model to generate MT output containing critical errors.

Considering the insufficient size of data for perturber training, we opt not to train it from scratch. Instead, we leverage the translation model learned on a 1.6M English-Korean parallel corpus[6]. Perturber training is performed on this model using the fairseq framework[7] on a single GPU. We set hyperparameters to 32,000 vocab size, 4096 maximum token count, 5e-4 learning rate, and five epochs of early stopping. As even minor sentence alterations can result in critical cases, we set the loss as the best checkpoint metric for training instead of the BLEU score.

Label assigned to the perturbed data Depending on the task, we devise different label annotation methods for the augmented data via perturber. For the CED task, we utilize two labeling strategies. We first assign all samples to the 'ERR' label since the purpose of the perturber is to create an MT output containing critical errors. Secondly, it is not definitively secure that all generated samples are critical. Therefore, we utilize ChatGPT (gpt-3.5-turbo) (OpenAI-Blog, 2022) to perform labeling through prompting [8].

For the CETC task, ChatGPT is used to predict critical error types through prompting to attach one of the six error types for each sample. Additionally, since the distribution is one of the factors considered in label assignment Chen et al. (2022); Ye et al. (2019), we employed two different distribution-aware labeling strategies. Firstly, we assign six label types to augmented data with an equal distribution. Secondly, we apply a distribution-aware method that attaches labels based on the label distribution by type in the training data. Note that the two distribution-aware strategies are introduced for comparison with the ChatGPT method, given that ChatGPT is not deterministic.

4.3. Evaluation Metrics

For the CED task, we adopt Matthew's correlation coefficient (MCC) as the primary metric. Additionally, we measure and report the f1-score for both the 'ERR' and 'NOT' labels individually, as well as their combined performance. We utilize the evaluation script presented by WMT[9]. Regarding the CETC task, we set classification accuracy as the main metric. We also report the f1-score for each label to find which labels exhibit higher classification performance in general. The scikit-learn package is used for measuring the accuracy and f1-score.

5. Results and Analysis

5.1. TASK1: Critical Error Detection

Table 4 presents the CED results. Alongside the baseline, we include existing balancing methods as comparison groups, aiming to alleviate label imbalance. Additionally, we present the results of our approach, which incorporates augmentation using perturber.

Baseline results Among the five baseline models, XLM-R-large reports the best performance. It achieves the highest MCC score in evaluation and testing, particularly showing a distinct performance difference in F1-Good. mBART ranks second based on MCC. Notably, mBART25 demonstrates better performance than mBART50. As Korean has its unique language family and character, the performance gained through learning with similar high-resource languages is not observed, in line with Conneau et al. (2020).

As a baseline result for ChatGPT, the description strategy outperforms the plain and demonstration in the test set. Notably, even though ChatGPT has not

[6] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=126

[7] https://github.com/facebookresearch/fairseq.git

[8] We employ the description prompt during the label annotation process, which outperforms the other two prompt designs outlined in § 4.1.

[9] https://github.com/sheffieldnlp/qe-eval-scripts

Method	Model	Eval				Test			
		MCC	F1-Bad	F1-Good	F1-Multi	MCC	F1-Bad	F1-Good	F1-Multi
Baseline	mBERT (Devlin et al., 2019)	0.1520	0.9042	0.2478	0.2240	0.1227	0.9160	0.2000	0.1832
	mBART25 (Liu et al., 2020)	0.5605	0.9597	0.5542	0.5319	0.4793	0.9636	0.4742	0.4570
	mBART50 (Tang et al., 2021)	0.5433	0.9587	0.5250	0.5033	0.4459	0.9672	0.4364	0.4220
	XLM-R-base (Conneau et al., 2020)	0.2917	0.9377	0.3294	0.3089	0.2484	0.9566	0.2679	0.2562
	XLM-R-large (Conneau et al., 2020)	0.7239	0.9726	0.7126	0.6931	0.5497	0.9719	0.5470	0.5316
	ChatGPT-Demo	0.2088	0.9398	0.2000	0.1880	0.2428	0.9583	0.2476	0.2373
	ChatGPT-Plain	0.2596	0.9368	0.2927	0.2742	0.3544	0.9654	0.2826	0.2728
	ChatGPT-Description	0.0962	0.9050	0.1905	0.1724	0.3633	0.9653	0.3125	0.3017
Ours	XLM-R-large (+Pert (All Err))	**0.7352**	**0.9736**	**0.7391**	**0.7196**	0.5709	0.9735	0.5455	0.5310
	XLM-R-large (+Pert (ChatGPT))	0.6598	0.9670	0.6667	0.6447	**0.6019**	**0.9752**	**0.5607**	**0.5468**

Table 4: MCC and F1-score results for the CED task

Method	Model	Eval ACC	Test ACC	F1-score per label						
				TOX	SAF	NAM	NUM	SEN	POL	NOT
Baseline	mBERT (Devlin et al., 2019)	0.872	0.915	0.00	0.00	0.00	0.29	0.18	0.00	0.96
	mBART25 (Liu et al., 2020)	0.908	0.934	0.33	0.25	0.17	0.35	0.32	0.56	0.97
	mBART50 (Tang et al., 2021)	0.882	0.914	0.40	0.12	0.00	0.22	0.11	0.12	0.96
	XLM-R-base (Conneau et al., 2020)	0.886	0.926	0.31	0.00	0.24	0.24	0.38	0.50	0.96
	XLM-R-large (Conneau et al., 2020)	0.918	0.945	**0.67**	0.15	**0.31**	0.40	0.58	0.61	0.97
	ChatGPT-Demo	0.886	0.910	**0.05**	0.17	0.00	0.15	0.00	0.00	0.96
	ChatGPT-Plain	0.902	0.924	0.00	**0.31**	0.00	0.15	0.12	0.00	0.96
	ChatGPT-Description	0.878	0.928	0.00	**0.31**	0.00	**0.40**	**0.13**	**0.12**	**0.97**
Ours	XLM-R-large (+Pert (ED))	0.912	0.949	0.36	**0.43**	0.24	0.67	0.57	0.58	**0.98**
	XLM-R-large (+Pert (TD))	0.916	0.949	0.62	0.17	**0.31**	0.70	0.54	0.60	**0.98**
	XLM-R-large (+Pert (ChatGPT))	**0.930**	**0.953**	0.55	0.35	0.26	**0.74**	**0.72**	**0.67**	**0.98**

Table 5: Accuracy and F1-score per label results for the CETC task

been trained on the CED data, ChatGPT demonstrates its powerful capabilities by outperforming mBERT and XLM-R-base.

Perturber results We show the results of our proposed method, which utilizes a perturber for data augmentation. The experimental results demonstrate that our method outperforms the XLM-R-large baseline in both approaches: labeling the entire sample as the 'ERR' label (+Pert (All Err)) and label assigning using ChatGPT (+Pert (ChatGPT)). Particularly, we observe a significant performance improvement in terms of test MCC. There is an increase from 0.5497 to 0.5709 for the +Pert (All Err) and a further improvement to 0.6019 for the +Pert (ChatGPT). This justifies the positive impact of our data augmentation methodology on CED training. The higher performance of ChatGPT compared to +Pert (All Err) indicates that there may be cases where not all perturbed examples are error cases. By incorporating ChatGPT's linguistic knowledge, we achieve additional performance gains.

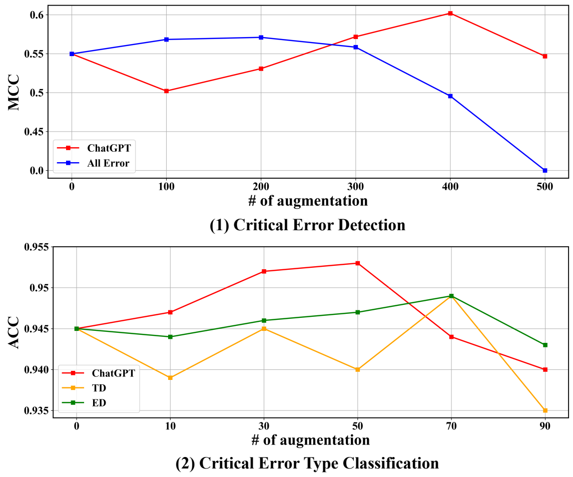
5.2. TASK2: Critical Error Type Classification

Table 5 is the results of the CETC task. In this experiment, we compare the performance of our proposed methodology with the baseline.

Baseline results Consistent with the results from the CED task, the XLM-R-large model exhibits the highest accuracy among the baselines. Particularly, when comparing f1-scores across types, the XLM-R-large model reveals balanced results as well as outperforming performance compared to other multilingual PLMs. Conversely, while models like mBERT and mBART50 record comparable results surpassing 0.9 in accuracy, their f1-scores per label are solely high around the 'NOT' label. We interpret this as a consequence of label bias, with the models potentially being trained to select the 'NOT' label primarily. This is not regarded as proper learning.

According to the ChatGPT results, the description strategy achieves the highest test accuracy, consistent with the CED results. Even the performance is comparable to the PLM baseline, with a marginal gap. This is quite surprising as the model is not trained with the CED dataset.

However, the F1-score for each error type displays a completely different aspect. In particular, the model's correctly predicted labels are limited to 'SAF', 'NUM', and 'NOT' error types, while the performance is markedly poor in classifying other types. Concurrently, it is evident that the system



(1) Critical Error Detection

(2) Critical Error Type Classification

Figure 2: Performance change according to the amount of augmented data using perturber

struggles immensely in distinguishing the 'POL' type. Although ChatGPT shows distinguishing performance compared to a great number of other models, it appears to lack the ability to discern critical errors that violate cultural nuances in translations. This underscores the heightened necessity for our datasets that address cultural factors.

Perturber results All the fine-tuning results with perturber-augmented data reveal outperforming performance than the baseline. In particular, we find performance improvements when utilizing ChatGPT over evenly distributed labels (ED) and training data distribution-aware labeling methods (TD). We analyze that ChatGPT may release the insufficiency and uncertainty of label assignment compared to the other approaches.

Focusing on the f1-score per label, we verify the performance of the newly introduced 'POL' error. Experimental results for the type show comparable performance compared to other fine-grained labels. Secondly, among the types, the 'SAF' type shows the most inferior performance. We speculate that the precise scope and boundaries threatening safety and health can be considered ambiguous to the model, resulting in low performance. Moreover, for the 'NAM' label, any instances where the user can restore the entity should be excluded from the critical error. This result implies that the task is challenging enough for the current introduced models.

Method	Used Label	MCC	F1-Bad	F1-Good	F1-Multi
XLM-R-large	Binary Class	0.5497	0.9719	0.5470	0.5316
	Multi Class	**0.5835**	**0.9740**	**0.5664**	**0.5517**
XLM-R-large (+Pert (ChatGPT))	Binary Class	0.6019	0.9752	0.5607	0.5468
	Multi Class	**0.6540**	**0.9777**	**0.6379**	**0.6237**

Table 6: Performance comparison between experiments leveraging binary and multiclass labels. For comparison, we convert the critical error type predictions into binary classes by changing all non-'NOT' error types to the 'ERR' label.

	XLM-R-large			XLM-R-large (+Pert (ChatGPT))		
	Binary Class	Multi Class	Δ	Binary Class	Multi Class	Δ
TOX	42.86	71.43	+28.57	42.86	57.14	+14.29
SAF	50.00	40.00	-10.00	50.00	50.00	0.00
NAM	10.00	20.00	+10.00	20.00	15.00	-5.00
NUM	58.33	25.00	-33.33	33.33	58.33	+25.00
SEN	57.14	57.14	0.00	57.14	71.43	+14.29
POL	53.85	61.54	+7.69	53.85	61.54	+7.69
NOT	99.03	99.46	+0.43	98.38	99.68	+1.30

Table 7: Accuracy comparison results by critical error type for binary and binary-transformed multiclass experiments. For the comparison, we separate the test dataset according to error types and measure the accuracy for each type.

6. Analysis

6.1. Impact on the Amount of Perturbed Data

To explore the influence of the amount of augmented data, we gradually append data into the training set. In this context, the units of data increment were determined empirically.

The upper part of Figure 2 represents the results of the CED task. The performance tendency consistently shows an increase as the data size is scaled to 400 with the perturber augmentation using ChatGPT, and to 200 in the case of All error, which assigns an 'ERR' label for all augmented data. However, the performance drops beyond a certain amount in both cases.

The lower part represents the results of the CETC task, where all three models show improved performance upon augmentation. Especially, labeling the perturbed data with ChatGPT shows the most significant performance improvement when 50 data are added. The models assigning labels in an ED manner and TD manner show the most performance gain when 70 data are added. Similarly, the CETC model also shows a performance decrease beyond a certain threshold.

6.2. Comparison for the Experiments using Binary and Multiclass

We explore which types of labels are more effective in detecting critical errors: binary labels or multiclass labels. To induce the two performances to be comparable, we convert the predictions from the CETC task into a binary class. Namely, regardless of classified error types, we change all predictions that are not predicted as 'NOT' to 'ERR'. We then compare MCC performance with the prediction from the CED task. The comparison is conducted under two settings, XLM-R-large and XLM-R-large (+Pert (ChatGPT)).

As reported in Table 6, the results indicate that the MCC shows remarkable improvements from 0.5497 to 0.5835 in XLM-R-large, 0.6019 to 0.6540 in XLM-R-large (+Pert (ChatGPT)) when multiclass labels are utilized for model training than binary labels. For a more in-depth analysis, we divide the test set by critical error type and measure the accuracy for each type, as in Table 7. Consistent with the results in Table 6, predictions from models trained with multiclass labels generally yield outperforming results. In the case of the experiment with XLM-R-large, even if 'SAF' and 'NUM' label performance shows a trade-off, four ('TOX', 'NAM', 'POL', and 'NOT') out of seven types achieve a performance improvement, and one reports the same performance. For the XLM-R-large experiment with data augmentation, all performances except for the 'NAM' type show outperforming performance. From the results, we analyze that informing the model by specifying the label, rather than assigning a single 'ERR', can reduce label ambiguity and increase explainability. Accordingly, the results demonstrate its necessity of providing fine-grained critical error types in the KNOTICED.

7. Conclusion

We proposed KNOTICED, a dataset for detecting critical meaning distortion in English-Korean MT. This dataset introduced a novel culture-aware politeness label reflecting Korean etiquette culture. Moreover, the dataset included fine-grained type labels in addition to the binary labels, enabling two tasks: CED and CETC. In the experiments, we observed that our proposed data augmentation method via perturber remarkably outperformed the baseline performance. We hope this dataset will be utilized to mitigate a variety of risks that may arise in the MT field in the future.

Limitations

Our dataset introduces five language-agnostic critical error types, along with a newly incorporated culture-aware politeness type. However, there exists a possibility for further error types, either language-dependent or independent. We are positively considering the investigation and development of more granular error types. Furthermore, our experiments are limited to leverage ChatGPT. For the benefit of easy replication and reproduction, we opted for ChatGPT because of its accessibility without further billing. However, we are open to exploring and integrating other models such as GPT-4 in our future work. We believe that this paper lays the groundwork, and subsequent studies can expand on our methodology by utilizing a range of models.

Ethics Statement

Portions of this study and KNOTICED dataset encompass descriptions and instances of error types that potentially raise ethical issues. Notably, examples related to the toxicity type in Table 2 could evoke offensive or upsetting emotions. These instances are inevitably included due to the inherent purpose of the dataset, which is to detect such cases and prevent them from being produced in translations. However, considering this, we partially obscure portions containing unpleasant content in the examples.

Moreover, in the process of creating this dataset, we unavoidably request human annotators to inject potentially discomforting content. In view of this, we notified human annotators about this aspect during recruitment. We ensure that if annotators experience any unsetting emotions during the data creation, they can immediately stop and take a break before resuming. In terms of annotator recruitment, we have paid an amount exceeding the legally mandated wage, and all personal information beyond the minimum requirement was immediately discarded.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425) and this work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge) and this work was supported by the NRF grant funded by MSIT(No. 2022R1A5A7026673).

8. References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Lucien Brown. 2011. Korean honorifics and 're-vealed','ignored'and 'suppressed'aspects of korean culture and politeness. *Politeness across cultures*, pages 106–127.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, pages 2778–2788.

Yimeng Chen, Chang Su, Yingtao Zhang, Yuxia Wang, Xiang Geng, Hao Yang, Shimin Tao, Guo Jiaxin, Wang Minghan, Min Zhang, Yujia Liu, and Shujian Huang. 2021. HW-TSC's participation at WMT 2021 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 890–896, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, Christophe Ropers, Eric Michael Smith, Daniel Licht, Carlos Escolano, and Javier Ferrando. 2023. Toxicity in multilingual machine translation at scale.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2022. KU X upstage's submission for the WMT22 quality estimation: Critical error detection shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 606–614, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023a. Hallucinations in large multilingual translation models. *arXiv preprint arXiv:2303.16104*.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Genze Jiang, Zhenhao Li, and Lucia Specia. 2021a. ICL's submission to the WMT21 critical error detection shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 928–934, Online. Association for Computational Linguistics.

Genze Jiang, Zhenhao Li, and Lucia Specia. 2021b. Icl's submission to the wmt21 critical error detection shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 928–934.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

OpenAI-Blog. 2022. Chatgpt: Optimizing language models for dialogue.

Telmo Pessoa Pires, Robin M Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. Learning language-specific layers for multilingual machine translation. *arXiv preprint arXiv:2305.02665*.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. NICT Kyoto submission for the WMT'21 quality estimation task: Multimetric multilingual pretraining for critical error detection. In *Proceedings of the Sixth Conference on Machine Translation*, pages 941–947, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. *arXiv preprint arXiv:1905.11901*.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Joël Tang, Marina Fomicheva, and Lucia Specia. 2022. Reducing hallucinations in neural machine translation with feature attribution. *arXiv preprint arXiv:2211.09878*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Thomas Wolf, 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, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Qinyuan Ye, Liyuan Liu, Maosen Zhang, and Xiang Ren. 2019. Looking beyond label noise: Shifted label distribution matters in distantly supervised relation extraction. 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 3841–3850, Hong Kong, China. Association for Computational Linguistics.

Kyong-Ae Yu. 2011. Culture-specific concepts of politeness: indirectness and politeness in english, hebrew and korean requests.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.