Evaluating Factuality in Cross-lingual Summarization

Mingqi Gao[*,1,2,3], **Wenqing Wang**[*,4], **Xiaojun Wan**[1,2,3], **Yuemei Xu**[4]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Center for Data Science, Peking University
[3]The MOE Key Laboratory of Computational Linguistics, Peking University
[4]School of Information Science and Technology, Beijing Foreign Studies University
{gaomingqi,wanxiaojun}@pku.edu.cn
{19190010,xuyuemei}@bfsu.edu.cn

Abstract

Cross-lingual summarization aims to help people efficiently grasp the core idea of the document written in a foreign language. Modern text summarization models generate highly fluent but often factually inconsistent outputs, which has received heightened attention in recent research. However, the factual consistency of cross-lingual summarization has not been investigated yet. In this paper, we propose a cross-lingual factuality dataset by collecting human annotations of reference summaries as well as generated summaries from models at both summary level and sentence level. Furthermore, we perform the fine-grained analysis and observe that over 50% of generated summaries and over 27% of reference summaries contain factual errors with characteristics different from monolingual summarization. Existing evaluation metrics for monolingual summarization require translation to evaluate the factuality of cross-lingual summarization and perform differently at different tasks and levels. Finally, we adapt the monolingual factuality metrics as an initial step towards the automatic evaluation of summarization factuality in cross-lingual settings. Our dataset and code are available at https://github.com/kite99520/Fact_CLS.

1 Introduction

Cross-lingual summarization, the task of generating a summary in different languages from the source documents, aims to help people efficiently grain the main point of the original document. It is recognized as a challenging task that combines the difficulties of text summarization as well as machine translation. Traditional pipeline methods first translate the document and then summarize it in the target language or vice versa (Leuski et al., 2003; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015). Currently, modern neural cross-lingual summarization models have witnessed rapid growth in recent research (Shen et al., 2018; Duan et al., 2019; Zhu et al., 2019, 2020; Cao et al., 2020).

Document
海关总署10日发布的数据显示，2月我国进出口总值为2604.3亿美元，增长29.4%。其中出口1144.7亿美元，增长18.4%；进口1459.6亿美元，增长39.6%。当月贸易逆差314.9亿美元，为近10年来单月贸易逆差最大值。(Data released by the General Administration of Customs on the 10th show that the total value of China's imports and exports in February was 260.43 billion U.S. dollars, up by 29.4%. Among them, the export was U.S. dollars 114.47 billion, up by 18.4%; Imports reached 145.96 billion U.S. dollars, up by 39.6 %. The trade deficit of 31.49 billion U.S. dollars was the largest in nearly a decade.)
Summaries
TNCLS: China's exports exceeded *100 billion* US dollars in February, the biggest trade deficit in nearly 10 years. ✗
CLSMS: China's trade deficit *in the past 10 years* is the largest in nearly 10 years. ✗
CLSMT: In February, the total import and export value of China's foreign trade increased by 29.4 % compared with the same period last year. ✔
ATS: China's foreign trade deficit in February was *26.4 billion* US dollars, the biggest in 10 years. ✗

Table 1: A real example from Chinese-to-English dataset. The Spans of factual errors are marked in red.

Factuality, a crucial dimension, is absent from the current evaluation of cross-lingual summarization approaches. ROUGE (Lin, 2004) is the main automatic evaluation metric. Informativeness, fluency, and conciseness are the dimensions of human evaluation. However, many case studies have pointed out that the summaries generated by neural cross-lingual summarization models have factual errors (Zhu et al., 2019, 2020; Bai et al., 2021). Table 1 also shows the state-of-the-art cross-lingual summarization models generate factually incorrect summaries. A variety of factuality evaluation metrics have drawn close attention in monolingual summarization (Kryscinski et al., 2020; Maynez et al., 2020; Goyal and Durrett, 2021), yet so far no study has comprehensively studied the factuality of cross-lingual summarization.

To fill the gap, we collect summaries from six models on a cross-lingual summarization dataset

[*]Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12415–12431
July 9-14, 2023 ©2023 Association for Computational Linguistics

proposed by Zhu et al. (2019) and obtain the human judgments of fine-grained factuality. The result of human evaluation suggests that over half of the generated summaries and over 27% of reference summaries contain at least one factual error. During the annotation process, we identify the peculiarity of factual errors in cross-lingual summaries, such as translation-related errors. Further, since the existing monolingual factuality metrics require the aid of translation to use in cross-lingual settings, after analyzing their performance, we explore the challenging automatic evaluation of factuality in cross-lingual summarization. In summary, our contributions are as follows:

- We propose a cross-lingual factuality dataset by collecting fine-grained human annotations over references as well as the outputs of six cross-lingual summarization systems at both summary level and sentence level. The dataset will be released and contribute to future cross-lingual summarization research.

- We introduce a typology of factual errors and conduct a fine-grained analysis of the factuality of the summaries and the performance of existing metrics. To the best of our knowledge, this is the first work to analyze the factuality of cross-lingual summarization.

- We adapt the monolingual factuality metrics as an initial step towards automatic factuality assessment in cross-lingual summarization.

2 Related Work

2.1 Cross-lingual Summarization

Early explorations on cross-lingual summarization are pipeline methods that simply integrate machine translation into monolingual summarization and achieve some improvement through incorporating bilingual parallel information. (Leuski et al., 2003; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015). Recently, neural-based methods have been applied to cross-lingual summarization (Shen et al., 2018; Duan et al., 2019; Zhu et al., 2019, 2020; Cao et al., 2020). Shen et al. (2018) first propose the neural-based cross-lingual summarization system with a teacher-student framework. Similarly, Duan et al. (2019) improve the teacher-student framework by using genuine summaries paired with the translated pseudo source sentences for training. Zhu et al. (2019) propose a multi-task learning framework, which incorporates monolingual summarization or machine translation into cross-lingual summarization training process. A concurrent work by Zhu et al. (2020) improves the performance by combining the neural model with an external probabilistic bilingual lexicon. Cao et al. (2020) propose a multi-task framework with two encoders and two decoders that jointly learns to summarize and align context-level representations.

2.2 Factuality Evaluation in Summarization

There are many analyses and meta-evaluations for factuality in monolingual summarization (Maynez et al., 2020; Pagnoni et al., 2021; Gabriel et al., 2021). Cao et al. (2017) reveal nearly 30% of the outputs from a state-of-the-art neural summarization system contain factual errors. Similarly, Falke et al. (2019) conduct the initial crowdsourcing of binary factual annotations and find that nearly 25% of the generated summaries are factually inconsistent.

In terms of evaluation metrics, the most commonly used ones based on n-gram overlap like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and METEOR (Agarwal and Lavie, 2008) are insufficient to measure the factual consistency of summaries and fail to correlate with the human judgments of factuality (Falke et al., 2019; Kryscinski et al., 2019). Yuan et al. (2021) convert the evaluation task to a conditional generation task and utilize the generation probability of the pre-trained language model BART (Lewis et al., 2020) to estimate the quality of system output, including faithfulness. Further, several works have explored using natural language inference (NLI) models to evaluate the factuality of summaries (Falke et al., 2019; Kryscinski et al., 2020; Maynez et al., 2020). In addition, Durmus et al. (2020) and Wang et al. (2020) evaluate factual consistency through question generation and question answering models. All the above metrics can not be used directly in cross-lingual settings.

3 Typology of Factual Errors

We define a typology of ten factual errors by analyzing both reference summaries and generated summaries. An example for each error type is shown in Table 2.

Hallucination Error (HalE): This occurs when the events not directly inferable from the input doc-

Document: 去年新昌一批企业因铬超标胶囊被查处。沃州公司状告省药监局将于明日开庭，认为当时处罚失当。原告公司认为，他们从未使用过工业明胶，铬含量仅超过国家标准1PPM的轻微超标情形，产品又已召回，未有实际危害后果，其情形不构成吊证。(Last year, a number of enterprises in Xinchang were investigated and punished for exceeding the chromium standard capsules. The WoZhou company's lawsuit against the Provincial Drug Administration will be heard tomorrow, arguing that the punishment was improper at the time. The plaintiff company argued that they had never used industrial gelatin, the chromium content only exceeded the national standard of 1 PPM slightly exceeding the standard, the product has been recalled without actual harmful consequences, and the situation does not constitute the certificate revocation.)
HalE: A group of enterprises were fined RMB 20,000 for chromium capsules exceeding the standard.
ParE: A group of enterprises in Beijing were investigated for Alum exceeding the standard.
PreE: A group of enterprises were commended for chromium capsules exceeding the standard.
EntE: A group of enterprises in Xinchang will sue the Food and Drug Administration in court.
CorE: WoZhou Company believes that she didn't cause harmful consequences.
IncE: A group of enterprises were investigated for [UNK].
TenE: Wozhou Company's case against the Food and Drug Administration went to trial.
PluE: An enterprise was investigated for chromium capsules exceeding the standard.
TerE: They never used bright colloid and has no actual harmful consequences.

Table 2: An illustration of the taxonomy on factual error types. Not from a real dataset.

ument are added to the summaries.

Particulars Error (ParE): This occurs when the summary contains the major events of the source document, but some details are inaccurate or mistaken, like time, location and direction.

Predicate Error (PreE): This occurs when the predicate in the summary is contradictory to the source document.

Entity Error (EntE): This occurs when the entity of an event is wrong, including substitution, addition and deletion cases.

Coreference Error (CorE): This occurs when pronouns and other references to the aforementioned entities are either incorrect or ambiguous.

Incompleteness Error (IncE): This occurs when the word [UNK] is presented in the summary.

The above factual error types are from monolingual summarization (Pagnoni et al., 2021; Wang et al., 2022). Considering the specificity of cross-lingual summarization, three error types in machine translation (denoted as **MTE**) are added after some case studies.

Tense Error (TenE): This occurs when the tense of the summary is inconsistent with the source document, which is common in machine translation as the natural differences in tense expression between Chinese and English. Tenses in English can be directly indicated through predicate verbs, while they are not clearly marked in Chinese (Shi, 2021).

Plural Error (PluE): This occurs when the summary changes the singular or plural forms of nouns in the source document. English nouns focus on the concept of singular and plural, while nouns in Chinese are usually replaced by flexible and fuzzy quantitative expressions (Xiao, 2013).

Terminologies Error (TerE): This occurs when the terminologies in the source document cannot be expressed professionally or accurately in summary. Li and Feng (2020) find that terminologies error ranks first among the high-frequency errors in machine translation.

Finally, we add the additional type **Other Error (OthE)** to ensure the completeness of the typology. This occurs when the error does not correspond to any of the above types.

4 Data Annotation

4.1 Dataset and Model

We annotate samples from the cross-lingual summarization datasets released by Zhu et al. (2019), which includes an English-to-Chinese (En-to-Zh) dataset and a Chinese-to-English (Zh-to-En) dataset. The Chinese summaries of the En-to-Zh dataset are translated correspondingly from the union set of CNN/DM (Hermann et al., 2015) and MSMO (Zhu et al., 2018). The Zh-to-En dataset is constructed from the Chinese LCSTS dataset (Hu et al., 2015).

Based on the dataset, we collect generated summaries from six models. Since Wan et al. (2010) have shown that summarize-then-translate is preferable to avoid both the computational expense of translating more sentences and sentence extraction errors caused by incorrect translations, we first use PGN (See et al., 2017), a monolingual summarization model to generate the summaries [1], and a

[1] CNN/Dailymail (https://github.com/abisee/pointer-generator) and LCSTS (https://github.com/LowinLi/Text-Summarizer-Pytorch-Chinese). Prior studies had trained PGNs on the original monolingual

translation model [2] is then applied to translate the summary. To compare the impact of different translation systems on summarization, we also use a commercial translator Youdao [3], during the process of translation. We refer to these two pipeline methods as **Pipe-ST** and **Pipe-ST*** respectively. We also collect outputs from four neural cross-lingual summarization models. **TNCLS** (Zhu et al., 2019) trains a standard Transformer model on the parallel corpora in an end-to-end manner. **CLSMS** (Zhu et al., 2019) combines the cross-lingual summarization task with monolingual summarization and calculates the total losses. Similarly, **CLSMT** (Zhu et al., 2019) combines cross-lingual summarization with machine translation. They both use one encoder and multiple decoders for multi-task frameworks. **ATS** (Zhu et al., 2020) is another Transformer-based model that utilizes a pointer-generator network to exploit the translation patterns in cross-lingual summarization.

100 documents are randomly sampled from the test set of En-to-Zh and Zh-to-En corpus respectively with the corresponding model-generated summaries. Each summary is manually split into sentences.

4.2 Annotation Procedure

We recruit eight college students with qualification certificates who are fluent in both English and Chinese languages, with Chinese as their mother tongue. They are provided with an annotation guideline. Further, we design a qualification test consisting of 10 document-summary pairs, only annotators who pass the test are considered to be qualified and are allowed to continue annotation. To ensure the annotation quality, we set the number of tasks each annotator needs to complete each day. After receiving the results of the day, we check the results and provide feedback.

In the annotation interface, we show the full source document on the left and a summary sentence by sentence on the right. Seven summaries are listed for each document in random order, including one translated reference summary and six model-generated summaries. The fine-grained annotations are a three-step process: For each sentence in a summary, annotators first determine whether it is factual or not. If a sentence is marked as not factual, annotators identify the error types based on our typology. A sentence can be annotated with multiple types. Finally, a Likert Scale from 1-5 is used to rate the overall factuality of the summary.

Each sample is annotated by two distinct annotators. For the sentence-level binary label, a third annotator from us makes the final decision if they are in disagreement. For the error types of each sentence, the intersection and union of two annotators are both collected. For the summary-level annotation, we take the average score of two annotators as the final result.

4.3 Inter-annotator Agreement

Table 3 shows the nearly perfect inter-annotator agreement on two datasets. For sentence-level annotation, we obtain an average agreement of Cohen's kappa (Cohen, 1960) with κ=0.891. For the summary level, we obtain the agreement of Krippendorff's alpha (Krippendorff, 1970) with α=0.903 on average. More annotation details can be found in Appendix F.

	κ	α
En-to-Zh	0.925	0.906
Zh-to-En	0.856	0.900
Average	0.891	0.903

Table 3: Cohen's Kappa κ at sentence level and Krippendorff's alpha α at summary level of the samples.

5 Fine-grained Factuality Analysis

5.1 Factuality of Reference Summaries

The two cross-lingual summarization datasets were originally constructed in a two-step way: (1) Given a source document, a reference summary in the same language was written by humans or crawled from their titles. (2) Then the summary was translated into another language by an automatic translator [4], maybe followed by a manual correction. The following analysis shows that the constructed datasets are error-prone and the factual errors can be introduced at both steps.

Error Proportion. Table 4 reports the annotation results on reference summaries from cross-lingual summarization datasets. We discover that 27%-50% single sentences contain at least one factual

summarization datasets. We just got the generated summaries.

[2] https://huggingface.co/Helsinki-NLP/opus-mt-zh-en/tree/main and https://huggingface.co/Helsinki-NLP/opus-mt-en-zh/tree/main

[3] https://fanyi.youdao.com

[4] http://www.anylangtech.com

	AvgScore ↑	%Error ↓	%Error Type									
			HalE	ParE	PreE	EntE	CorE	IncE	TenE	PluE	TerE	OthE
En-to-Zh	3.89	26.98	46.0	20.0	4.0	0.0	2.0	0.0	4.0	4.0	4.0	2.0
			60.0	40.0	10.0	8.0	2.0	0.0	4.0	6.0	8.0	4.0
Zh-to-En	3.46	50.00	25.3	27.3	7.1	8.1	1.0	0.0	5.1	1.0	4.0	0.0
			28.3	47.8	17.2	11.1	1.0	0.0	5.1	1.0	7.1	4.0

Table 4: Summary-level average score (left), proportion of sentences with at least one factual error (middle) and distribution of error types (right). For each error type, upper and lower part show the intersection and union by two annotators. ↑ indicates that the larger values, the better factual consistency results are.

error. However, the most popular n-gram based evaluation metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) only utilize the reference summary to evaluate the quality of the model-generated summary. The reference with poor quality will undermine the reliability of their evaluation under cross-lingual settings. We encourage future researchers to be aware of the issue, especially for evaluation.

Error Types. Both the intersection and union of two annotators are in a similar distribution that HalE and ParE occur most frequently. Since references are abstractive and do not simply copy sentences from the documents, it is natural to incorporate the author's background knowledge (van Dijk and Kintsch, 2014; Brown and Day, 1983), e.g.:

> **Document:** 失踪的女生身高158厘米左右，体重约90斤，皮肤白皙。9月3日失踪当天手提两大袋东西。当日她从厦门去福州，中午12点36分和她通话后再也联系不上。警方调取福州金山公交总站附近监控发现，她在附近上了一辆出租车。(The missing girl is about 158 cm tall, weighing about 90 pounds, with fair skin. She was carrying two large bags of stuff on the day she went missing, September 3. That day she went from Xiamen to Fuzhou, but could not be reached after calling her at 12:36 PM. The police retrieved the surveillance near Fuzhou Jinshan bus terminal and found that she got into a cab in the vicinity.)
> **Reference:** Xiamen *23-year-old* girl disappeared after she went to Fuzhou *to find her classmates* with a taxi.
> *(Zh-to-EnSum, HalE)*

Particularly, such hallucination may not always be erroneous. The information not entailed by the source in the above example is consistent with the relevant introduction in Wikipedia. It remains controversial whether this kind of hallucination should be allowed (Maynez et al., 2020) because it is difficult to verify whether it is factual outside of the source document.

Task Comparisons. We notice the difference between the two tasks of summarization: The factuality of references in En-to-Zh task is better than Zh-to-En task on both average score and error proportion. The reasons are two-fold: (1) In En-to-Zh task, the references of original English dataset are manually-written highlights offered by the news providers (Hermann et al., 2015; Zhu et al., 2018). While in Zh-to-En task, the Chinese dataset is constructed from a microblogging website and the crawled references are headlines or comments, which are generally more error-prone as they usually contain rhetoric to attract readers. An example is shown in Table 10 in Appendix G. (2) In En-to-Zh task, the dataset belongs to the news domain and existing machine translation for news reports has reached human parity (Hassan et al., 2018). While the dataset in Zh-to-En task comes from social media, the proportion of abbreviations, omitted punctuation, and catchphrases in the text is much higher than in news, resulting in lower translation quality. An example is shown in Table 11 in Appendix G.

5.2 Factuality of System Outputs

Error Proportion. Figure 2 visualizes the proportion of summaries with factual errors for each model and dataset. We observe that over 50% summaries contain factual errors, with the best model (**CLSMT**) generating 52.6% inconsistent summaries in En-to-Zh and 53.0% in Zh-to-En task. Similar observations have been made in monolingual summarization where 30%-80% of generated texts are factually inconsistent (Cao et al., 2017; Pagnoni et al., 2021).

Error Types. Error distributions in system outputs are shown in Figure 1. As in the reference, models also produce HalE and ParE error types with highest proportion.

For three error types that occur frequently in machine translation, we notice the proportion in Zh-to-En task (18.49%) is higher than that of En-to-Zh task (3.24%) showing the natural differences between the two languages. The comparison in IncE is more apparent. In Zh-to-En task, models

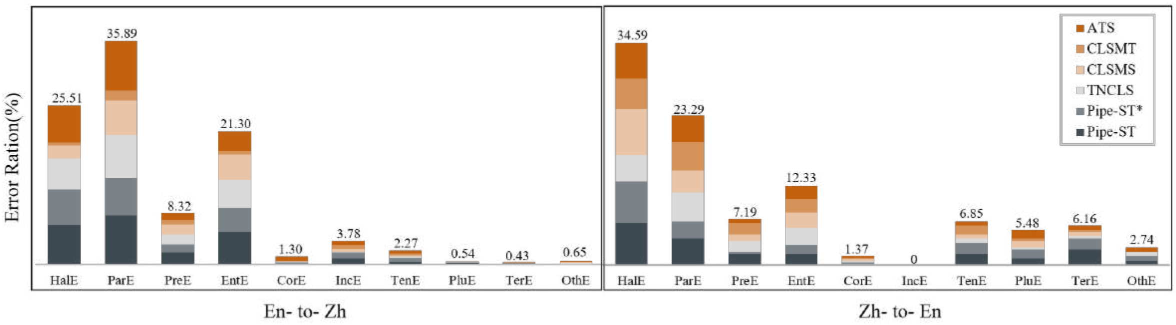

Figure 1: The proportion of generated summaries with different types of factual errors. The height of the model in the bar chart indicates the relative percentage of errors it makes on this error category compared to other models. Here we show the intersection by two annotators for each type and union results are detailed in Appendix A.

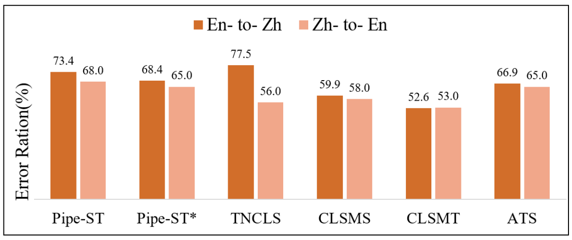

Figure 2: Proportion of single sentences in generated summaries with at least one factual error.

seldom generate UNK because the subword-based tokenization makes the vocabulary list smaller (Zhu et al., 2019). Additionally, OthE makes up a very small percentage (less than 3%) of errors showing that our typology is relatively complete.

It is worth noting that compared with the reference, the model-generated summaries contain more EntE, suggesting that models tend to confuse the role of each entity in an event, such as the subject and object, e.g.:

Document Fragment: (CNN) – *Rafa Benitez*'s turbulent reign as *Chelsea* manager took another battering on the day his supposed successor, *Pep Guardiola*, agreed a deal to become the new manager of *Bayern Munich*. *Benitez*, who was appointed as *Chelsea* interim manager in November following the dismissal of *Roberto Di Matteo*, was left stunned after his team squandered a two-goal lead to draw 2-2 with lowly *Southampton*. European champion *Chelsea* is now 13 points adrift of Premier League leader *Manchester United* with 16 games remaining and has failed to win any of their past three games at Stamford Bridge. *Guardiola* agrees three-year deal with *Bayern*. [. . .]

Summary: 拉法·贝尼特斯同意与拜仁慕尼黑签署一份为期三年的合同。切尔西以2 - 2 战平南安普顿，贝尼特斯被解雇。西班牙人现在在英超联赛中落后切尔西13分。(*Rafa Benitez* agreed to sign a three-year contract with Bayern Munich. Chelsea drew 2-2 with Southampton and *Benitez* was dismissed. European champion Chelsea is now 13 points adrift of Premier League leader. *The Spanish team* is now 13 points adrift of *Chelsea* in Premier League.)

(En-to-ZhSum, EntE)

Four entities mismatch in the generated summary. Since the document contains 40 entities in total, it is challenging for models to accurately locate the logical correlation between different parts when multiple entities appear.

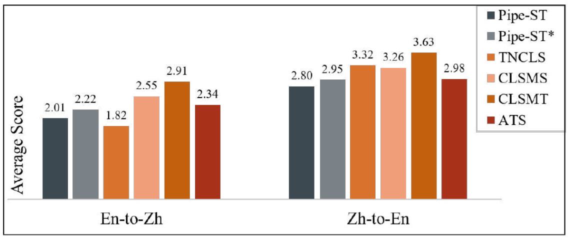

Figure 3: Summary-level average score ranging from 1 to 5 of summaries generated by models.

Task Comparisons. Figure 2 shows that most models perform almost equally at sentence level on Zh-to-En and En-to-Zh tasks except TNCLS. In contrast, the summary-level average scores in Zh-to-En task are generally higher than that in En-to-Zh task for each model as shown in Figure 3. One possible reason is that the summary-level average scores are influenced by the relationship between sentences. Note that En-to-Zh summaries are longer, with three sentences on average, while Zh-to-En summaries only contain one single sentence. Factual errors of conjunctions may exist in Eh-to-Zh summaries.

Model Comparisons. For traditional pipeline-based methods, **Pipe-ST*** outperforms **Pipe-ST** in both En-to-Zh and Zh-to-En tasks, suggesting the impact of different translators on factuality. We also notice that the two pipeline-based methods account for nearly half of the translation-related errors, TenE, PluE, and TerE (50.0% in En-to-Zh and 53.7% in En-to-Zh), probably because machine translation is one step of the pipeline. Table

3 shows that Modern neural-based cross-lingual summarization models generate more factual summaries than pipeline methods, but the average is not great. Table 6 in Appendix B reports the inconsistencies in ROUGE and factuality scores of the models.

6 Automatic Evaluation of Factuality

6.1 Existing Evaluation Metrics

ROUGE (Lin, 2004) is the most commonly used reference-based evaluation metric in summarization. The ROUGE score is usually used as a measure of overall quality.

The following monolingual factuality metrics use the source document and the summary to be evaluated as the inputs. To apply them in the cross-lingual scenarios, we use an automatic translator to translate the summaries in En-to-Zh tasks and the source documents in Zh-to-En tasks from Chinese into English. Please see Appendix C for more details.

BARTScore (Yuan et al., 2021) use the probability that BART (Lewis et al., 2020) generates the hypothesis given the source text to measure the faithfulness of the hypothesis.

FactCC (Kryscinski et al., 2020), a BERT-based model trained on a synthetic dataset to classify text pairs as factually consistent or inconsistent at sentence level.

DAE (Goyal and Durrett, 2021), an ELECTRA-based model that classifies each dependency arc in the model output as entailing the source text or not at dependency level.

QuestEval (Scialom et al., 2021), a QA-based metric that introduces question weighting and negative sampling into the training process to evaluate the factuality of text.

6.2 Exploration in Cross-lingual Settings

The existing monolingual metrics require the use of a translator, which is inconvenient. To make them directly usable in cross-lingual settings, we make the following attempts:

(1) For **BARTScore**, we replace the original checkpoint with mBART-50 (Tang et al., 2020) and use the multilingual version of tokenizer to allow multilingual inputs. We call it **mBARTScore**.

(2) For **FactCC**, we adapt its data augmentation methods of constructing synthetic data to generate cross-lingual data: In En-to-Zh task, sentences from the source document in English are extracted. The sentences themselves are used as positive claims, and the pronouns, entities, dates, numbers, and negatives in them are replaced by other words of the same type in the source document as negative claims. The positive and negative claims are translated into Chinese by an automatic translator. Finally, the translated claims are combined with the source document in English. In Zh-to-En task, the source documents in Chinese are first translated into English, and then data augmentation is applied to the translated source documents to construct claims in the same way. Finally, the claims are combined with the source document in Chinese. The source document and sentences are concatenated and fed to mBERT (Devlin et al., 2019) to train binary classification. The models trained on the synthetic data of the two tasks separately are denoted as **mFactCC-split**. The models trained by mixing the synthetic data of the two tasks are denoted as **mFactCC-mix**. More details can be found in Appendix D.

6.3 Correlation with Human Evaluation

To measure the correlation between metrics and human judgments, we compute the Pearson correlation coefficients r and Spearman's rank correlation coefficients ρ respectively at both system level and summary level. Furthermore, we also compute the binary classification accuracy of single sentences. we have the following findings from Table 5:

The performance of the metrics varies considerably across tasks and levels. All the existing metrics exhibit a higher correlation with human judgments in En-to-Zh task than that in Zh-to-En task. The correlations of the metrics are lower at summary level than at system level. The performance of the metrics differs relatively little at sentence level.

Compared to ROUGE, there is an advantage of the factuality metrics but it is not significant. Although ROUGE does not obtain the best performance at any level except system level of Zh-to-En task, its system-level correlation is close to the best results. Considering the existing factuality metrics need to be used with the help of translators and the translation process also introduces uncertainties and errors, it is challenging to introduce monolingual factuality metrics in cross-lingual summarization. However, this does not mean that the current evaluation mechanism does not need improvement, as we have illustrated the shortcomings of refer-

	En-to-Zh Summarization					Zh-to-En Summarization				
Metrics	**System-level**		**Summary-level**		**Sentence-level**	**System-level**		**Summary-level**		**Sentence-level**
	Pearson	Spearman	Pearson	Spearman	Accuracy	Pearson	Spearman	Pearson	Spearman	Accuracy
Rouge-1	0.91	0.71	0.29	0.27	0.57	0.44	0.75	0.20	0.22	0.56
Rouge-2	0.90	0.79	0.35	0.29	0.59	0.43	0.46	0.18	0.25	0.61
Rouge-L	0.91	0.71	0.28	0.27	0.57	0.44	**0.79**	0.20	0.24	0.56
BARTScore	**0.93**	**0.93**	**0.53**	**0.55**	0.62	0.02	0.11	0.27	0.25	0.57
DAE	0.83	**0.93**	0.39	0.39	**0.68**	**0.58**	0.34	0.16	0.16	**0.63**
FactCC	0.67	0.89	0.24	0.23	0.60	0.26	0.21	0.07	0.07	0.60
Questeval	0.82	**0.93**	0.39	0.40	0.63	0.32	0.32	**0.34**	**0.36**	0.56
mBARTScore	-0.25	-0.21	-0.04	-0.01	0.43	-0.29	-0.14	0.05	0.09	0.47
mFactCC-split	-0.34	-0.04	-0.03	-0.03	0.54	0.13	0.09	-0.03	-0.04	0.51
mFactCC-mix	0.20	0.34	0.02	0.01	0.53	-0.02	-0.13	0.002	0.001	0.47

Table 5: Pearson correlation and Spearman's rank correlation coefficients between human evaluation and metric scores at system level and summary level as well as accuracy at sentence level. Upper, middle and lower part show results for ROUGE, monolingual factuality metrics and cross-lingual factuality metrics respectively. The best values of each column are bolded. For metrics with continuous output, the score of a single sentence is truncated to a binary classification label based on the average score of all single sentences generated by the same model.

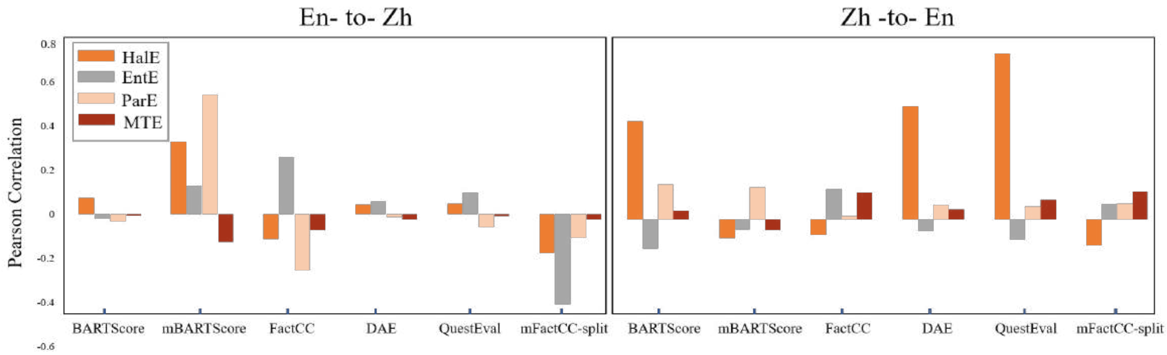

Figure 4: Change in Pearson correlation at summary level when the summaries with an error type are ignored.

ence summaries in Section 5.1. The reason for ROUGE's good performance may be related to how the dataset is constructed. Specifically, the training and test sets are not constructed in exactly the same way: the reference summaries of the training set are obtained through automatic translation, while the reference summaries of the test set are obtained through a combination of automatic translation and manual post-editing. The manual post-editing is likely to correct some obviously incorrect phrases. Summarization models fit the distribution of the training set, and errors in their outputs may be easier to capture when compared to the reference summaries.

mBARTScore and mFactCC adapted by us perform poorly. This suggests that it is challenging to evaluate the factuality of summaries in cross-lingual settings. We observe that there is a big difference between the synthetic claims and model-generated summaries. For future work, it is possible to fine-tune mBARTScore or design other data augmentation approaches.

6.4 Identification of Factual Error Types

To inspect capabilities of metrics identifying different types of factual errors, we use the result of the original correlation minus the correlation after ignoring the summaries with an error type as the measure. For each metric, we consider the three most frequent types HalE, ParE, ObjE, as well as MTE, and plot the contribution of error types to the overall correlation in Figure 4. A higher value indicates the better capabilities of the metric to capture the corresponding error types.

Similar to our discovery in Section 6.3, each metric exhibits a great difference between the two tasks. For example, almost all metrics correlate well with MTE in Zh-to-En task but have a negative correlation with MTE in En-to-Zh task. Figure 4 also reveals great limitations of factuality metrics in detecting different types of factual errors. Taking the entailment-based metrics as examples, DAE shows better ability at identifying HalE while FactCC has a negative correlation. Nevertheless, FactCC has the highest correlation with EntE in both tasks, showing the effectiveness of entity swapping transformation of its data augmentation to capture entity errors.

7 Conclusion

In this work, we comprehensively evaluate and analyze the factuality of reference summaries and model-generated summaries in cross-lingual summarization, showing that there are special factual errors in them. Automatic evaluation of cross-lingual summarization is yet to be addressed due to the shortcomings of reference summaries and the limitations of monolingual factuality metrics. Moreover, our exploration of the automatic factuality evaluation in cross-lingual settings illustrates its challenging nature.

Limitations

The scenarios we studied are limited to Chinese to English and English to Chinese. For other languages, the factual characteristics may be different. The genre of the source documents we study is news or blog post. For other genres, such as dialogue, our conclusion may not apply.

The number of systems we selected is limited, so there is some chance of system-level evaluation of evaluation metrics.

Ethics Statement

We recruit annotators from a college campus. They are completely free to decide whether or not to participate in our annotation. The payment is 9 dollars per hour, higher than the local minimum wage. There is no personal information in our collected dataset. The information which may be used to identify the participants is deleted after the annotation.

The model-generated summaries may contain toxic language, which can make annotators uncomfortable. We reviewed the data before annotation and found no problematic samples.

We check the licenses of the artifacts used in this study and do not find conflicts. The license of the dataset we will release is CC BY-NC 4.0.

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFF0901502), National Science Foundation of China (No. 62161160339), State Key Laboratory of Media Convergence Production Technology and Systems and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio. Association for Computational Linguistics.

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. 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 6910–6924, Online. Association for Computational Linguistics.

Ann L. Brown and Jeanne D. Day. 1983. Macrorules for summarizing texts: the development of expertise. *Journal of Verbal Learning and Verbal Behavior*.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization. *ArXiv*, arXiv:1711.04434.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of* *the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan H. Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv: Computation and Language*, arXiv:1803.05567.

Karl Moritz Hermann, Tomá Koiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *neural information processing systems*, arXiv:1506.03340.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Bivariate agreement coefficients for reliability of data. *Sociological methodology*, 2:139–150.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c* st* rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):245–269.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yamin Li and Li Feng. 2020. Pre-editing and post-editing in human-machine cooperation translation. *The Border Economy and Culture*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, Mao-song Sun, et al. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Zhaojia Shi. 2021. *Machine Translation Limitations and Post-Editing Solutions- A Case Study on The Global City Translation Project*. Ph.D. thesis, Shanghai International Studies University.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Computing Research Repository*, arXiv:2008.00401.

Teun A. van Dijk and Walter Kintsch. 2014. Cognitive psychology and discourse: Recalling and summarizing stories. pages 61–80. de Gruyter Berlin.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1546–1555, Portland, Oregon, USA. Association for Computational Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. *ArXiv*, arXiv:2210.11777.

Jun Xiao. 2013. A brief talk on the plural nouns translation. *Secondary school curriculum guidance (teacher newsletter)*.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 118–127, Lisbon, Portugal. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *arXiv: Computation and Language*, arXiv:1904.09675.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. 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 3054–3064, Hong Kong, China. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321, Online. Association for Computational Linguistics.

A Union proportion of model generated error types

It can be seen from Figure 5 that the union annotation of error types have the same distribution with intersection result, i.e., HalE and ParE are the most frequent, followed by EntE, PreE. Particularly, the proportion of OthE in union(10.99%) is much higher than that in intersection(3.39%), suggesting the influence of subjective factors on the determination of error types.

B ROUGE F1 and factuality scores of cross-lingual summarization models

Table 6 reports ROUGE F1 scores and the manually annotated factuality score of each model.

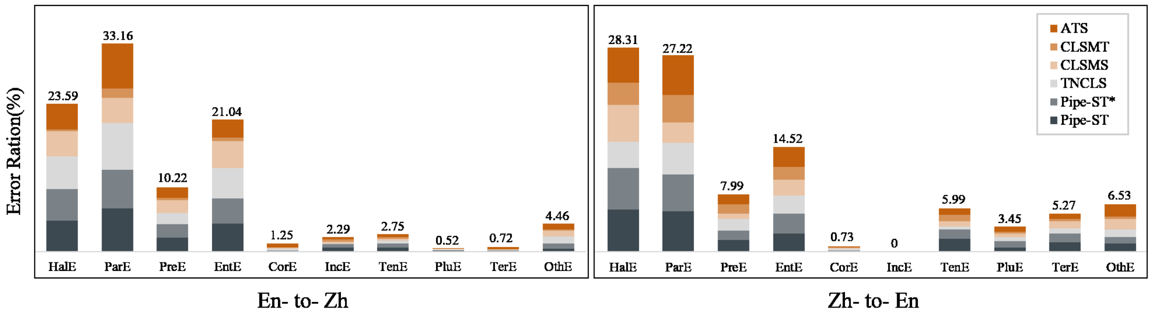

Figure 5: The union proportion of different error types in the generated summary. The height of the model in the bar chart indicates the relative proportion of errors it makes on this error category compared with other models.

En-to-Zh	R1	R2	RL	Fac.	**Zh-to-En**	R1	R2	RL	Fac.
Pipe-ST	26.53	17.34	27.98	2.01	Pipe-ST	32.24	12.60	33.27	2.80
Pipe-ST*	27.65	18.90	29.11	2.22	Pipe-ST*	37.31	18.96	38.79	2.95
TNCLS	21.62	14.04	22.93	1.82	TNCLS	38.45	**20.34**	**39.69**	3.32
CLSMS	26.32	16.91	27.58	2.55	CLSMS	**38.63**	19.52	**39.69**	3.26
CLSMT	30.14	**22.64**	31.65	**2.91**	CLSMT	37.42	18.90	39.19	**3.63**
ATS	**33.11**	22.42	**34.25**	2.34	ATS	37.40	19.20	39.04	2.98

Table 6: ROUGE F1 scores (%) as well as the average of the manually annotated summary-level factuality score of each model on cross-lingual tasks.

Name	**Value**
limit_length	True
length_limit	100
length_limit_type	'words'
apply_avg	True
apply_best	False
alpha	0.5
weight_factor	1.2
stemming	True

Table 7: Parameters of ROUGE.

C Details of existing metrics

For **ROUGE**, we use the Py-rouge package [5]. All parameters are listed in Table 7. For Chinese text, we insert spaces between Chinese characters as pre-processing.

For monolingual factuality metrics, we use the same translator[6] as used in Section 4.1 to translate the summaries or the source documents.

For **FactCC**[7] and **DAE** [8], We use NLTK [9] to split a summary in English into sentences. For Chinese, we use regular expressions to slice sentences based on Chinese punctuation. Each sentence is classified as factually correct or incorrect. The factual score of a summary is measured as the ratio of sentences classified as correct.

For **QuestEval** [10], we use the reference-less mode. For **BARTScore** [11], we use the $s \rightarrow h$ mode and the checkpoint trained on Parabank2, which is available at the GitHub repository.

D Details of the model implementation in cross-lingual settings

The checkpoint and tokenizer used for **mBARTScore** is available at https://huggingface.co/facebook/mbart-large-50. We do not fine-tune it.

The En-to-Zh dataset contains 370K English documents, with training set of 364687 items, validation set of 3000 items and test set of 3000 items.

[5] https://github.com/Diego999/py-rouge

[6] https://huggingface.co/Helsinki-NLP/opus-mt-zh-en/tree/main
https://huggingface.co/Helsinki-NLP/opus-mt-en-zh/tree/main

[7] https://github.com/salesforce/factCC

[8] https://github.com/tagoyal/factuality-datasets

[9] v3.7, https://www.nltk.org/

[10] https://github.com/ThomasScialom/QuestEval

[11] https://github.com/neulab/BARTScore

The Zh-to-En dataset is split into training set, validation set, and test set with 1693713, 3000, and 3000 items.

For the data augmentation of **mFactcc**, we use the same translator[12] as used in Section 4.1. For the En-to-Zh dataset, we randomly select 100000 samples from the training set and 2500 samples from the validation set and they are used to construct synthetic data. Finally, we construct a synthetic dataset with 200000 items as the training set and 5000 items as the validation set, where the ratio of the positive and negative items is 1:1. For the Zh-to-En dataset, the data is sampled in the same way resulting in the same size and ratio of the positive and negative items. When mixing the data, we randomly sample 100000 items and 2500 items from the training set and validation set of the above two synthetic datasets. The size of the mixing synthetic training set and validation set is 200000 and 5000. The positive items account for 50.41% in the training set and 50.64% in the validation set.

In all settings, the pre-trained checkpoint[13] is used to initialize parameters and we train the model for 10 epochs with a learning rate of 2e-5 and a max sequence length of 512. We try two batch sizes, 6 and 12. The best checkpoint is chosen according to the classification accuracy on the validation set.

We use two GeForce GTX 1080 Ti with 12GB memory for training and inference. Each single training session costs 24-36 hours.

E p-value of the correlation

In Table 8, we supplement the p-values corresponding to Pearson correlation and Spearman's rank correlation coefficients in Section 6.3, showing the significance level of the correlation between two variables.

F Annotation Details

In detail, the participants we recruited are from Asia. There are 5 females and 3 males, with an average age of around 24. We first conduct a qualification test before the formal annotation. 10 document-abstract pairs are randomly sampled with 5 in Zh-to-En and 5 in En-to-Zh datasets respectively. We annotated them first. Finally, we calculated the accuracy of each participant based on our annotation. Higher accuracy means a more consistent understanding of our guidelines. Annotators who achieve at least 80% accuracy are considered qualified to continue the annotation.

We conducted the annotation procedure two times and measure the inter-annotator agreement through two metrics as reported in Section 4.3. For the first annotation, we obtain average moderate agreement of Cohen's Kappa with κ=0.597 and substantial agreement of Krippendorff's alpha with α=0.762. However, we notice the low inter-annotators agreement in En-to-Zh task with fair agreement (κ=0.338) of Cohen's Kappa and moderate agreement (α=0.624) of Krippendorff's alpha compared with good inter-agreement in Zh-to-En task. Moreover, we find that some of the annotators may not take the work seriously, and label most sentences as factual although the errors are obvious.

To achieve high-quality annotations, we replaced the annotators who have a low agreement with others. After retraining and re-evaluating, we ask the annotator to annotate 10 items a day. Inspired by Pagnoni et al. (2021), we continuously evaluate annotators during the task as described in Section 4.2 to alleviate human-made disagreement. Finally, we achieve almost perfect agreement on both the two metrics in the second annotation. Here we show a sample from the annotated document-summary pairs on two tasks in Table 9.

G Additional Examples

Table 10 and Table 11 show two additional examples.

[12] https://huggingface.co/Helsinki-NLP/opus-mt-zh-en/tree/main
https://huggingface.co/Helsinki-NLP/opus-mt-en-zh/tree/main

[13] https://huggingface.co/bert-base-multilingual-cased

	En-to-Zh Summarization				Zh-to-En Summarization			
Metrics	**System-level**		**Summary-level**		**System-level**		**Summary-level**	
	Pearson-p	Spearman-p	Pearson-p	Spearman-p	Pearson-p	Spearman-p	Pearson-p	Spearman-p
Rouge-1	0.0051	0.0713	0.0000	0.0000	0.3287	0.0522	0.0000	0.0000
Rouge-2	0.0051	0.0362	0.0000	0.0000	0.3358	0.2939	0.0000	0.0000
Rouge-L	0.0049	0.0713	0.0000	0.0000	0.3223	0.0362	0.0000	0.0000
BARTScore	0.0027	0.0025	0.0000	0.0000	0.9607	0.8192	0.0000	0.0000
DAE	0.0221	0.0025	0.0000	0.0000	0.1726	0.7599	0.0000	0.0000
FactCC	0.1025	0.0068	0.0000	0.0000	0.5695	0.6445	0.0584	0.0481
Questeval	0.0233	0.0025	0.0000	0.0000	0.4826	0.4821	0.0000	0.0000
mBARTScore	0.5900	0.6445	0.2404	0.7876	-0.2900	-0.1400	0.0500	0.0178
mFactCC-split	0.4557	0.9377	0.4131	0.3847	0.7807	0.8448	0.3562	0.3194
mFactCC-mix	0.6752	0.4523	0.5482	0.6977	0.9597	0.7876	0.9594	0.9875

Table 8: P-values of Pearson correlation and Spearman's rank correlation coefficients reported in Section 6.3.

Document (English)		
Furniture giant IKEA has banned people from playing one of the most-loved childhood games - hide and seek. More than 33,000 shoppers have signed up on Facebook to participate in the giant maze-like store in Tempe, inner west of Sydney on Saturday, May 23. [. . .] But the Swedish retailer has put a stop to the unofficial event after attracting tens of thousands of participants, claiming the game 'raises security issues for both customers and co-workers.' [. . .]		
Summary 1:	**Annotation 1:**	**Annotation 2:**
家具巨头宜家禁止人们玩电子游戏。(Furniture giant IKEA banned people from playing video games.)	0 (ParE)	0 (ParE)
超过33，000名购物者报名参加了悉尼的一家商店。(More than 33,000 shoppers signed up for the store in Sydney.)	1	1
但这家瑞典零售商声称，这款游戏将“引发安全问题”。(But the Swedish retailer claimed that the game would “raises security issues”.)	1	1
Summary-level Score	4	4
Summary 2:	**Annotation 1:**	**Annotation 2:**
家具巨头宜家禁止人们玩隐藏的游戏。(Furniture giant IKEA banned people from playing the hidden games.)	0 (TerE)	0 (TerE)
超过33，000名购物者在脸书上签署了这项活动。(More than 33,000 shoppers signed up for the event on Facebook.)	1	1
瑞典零售商表示，游戏将吸引成千上万的参与者。(The Swedish retailer said that the game will attract tens of thousands of participants.)	0 (HalE, TenE)	0 (EntE, TenE)
Summary-level Score	2	3
Document (Chinese)		
盖洛普调查显示：6月份，55%的美国人通过电视获取新闻资讯，互联网以21%的份额排在第二位。令人感到意外的是，2%的受访者通过社交网络获取新闻，表明了Facebook和Twitter等服务在获取新闻资讯方面日趋提高的重要性。(Gallup survey shows that 55% of Americans get news information through TV, and the Internet ranked second with 21% in June. Surprisingly, 2% of the respondents get news through social networks, which shows the increasing importance of services like Facebook and Twitter in obtaining news information.)		
Summary 1:	**Annotation 1:**	**Annotation 2:**
55 % of Americans get news through social networking.	0 (ParE)	0 (ParE)
Summary-level Score	2	3
Summary 2:	**Annotation 1:**	**Annotation 2:**
Are you still using social media?	0 (HalE)	0 (OthE)
Summary-level Score	1	1

Table 9: A real example from the annotated document-summary pairs on two tasks.

Document with Original Reference: 【马云最后的演讲：商人没有得到应该得到的尊重】马云在淘宝十周年之际辞去了阿里巴巴集团CEO一职。马云表示，今天人类已经进入了商业社会，但很遗憾，这个世界商人没有得到他们应得到的尊重。我想我们像艺术家、教育家、政治家一样，我们在尽自己最大的努力，去完善这个社会。([Ma Yun’s Last Speech: Merchants don’t receive the respect they deserve.] Ma Yun resigned as CEO of Alibaba Group on the 10th anniversary of Taobao. He said that human beings have entered the commercial society today, but unfortunately, merchants have not received the respect they deserve. I think we merchants, like artists, educators, and politicians, are trying our best to improve the society.)
Translated Reference: Ma Yun ’s *Last Speech*: Businessmen are not respected as they deserve.

(Zh-to-EnSum, HalE)

Table 10: An example shows the reference summary with rhetoric in Zh-to-En task.

Document with Original Reference: 【刘强东中欧化身“吐槽哥”】刘强东在评论苹果时说道，科技领域日新月异，任何消费电子公司都不可能一直占优势，即便是颠覆了手机行业的苹果，“这不是诅咒，但我不认为苹果还能再活10年”。他还说，在中国长期来讲，所有的服务行业，加盟的都不看好，包括快递行业。([Liu Qiangdong’s sarcasm in China Europe International Business School]When commenting on Apple, Liu Qiangdong said that science and technology is changing increasingly, and it is impossible for any consumer electronics company to always take the advantage, even if it subverts the mobile phone industry. "This is not a curse, but I don’t think Apple can live for another 10 years." He added that all the service industries in China, including the express delivery industry, are not optimistic in the long run.)
Translated Reference: Liu Qiangdong ’s incarnation *“tucao ge”*

(Zh-to-EnSum, TerE)

Table 11: An example shows the reference summary with a catchphrase improperly translated in Zh-to-En task. The catchphrase 吐槽(sarcasm) is simply translated in pinyin without expressing its meaning. Moreover, China Europe International Business School, as the location of the report is abbreviated as 中欧(China Europe) in the original reference and omitted in the translated reference, probably because it is difficult for the automatic translator to understand the context.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the "Limitations" Section.

☑ A2. Did you discuss any potential risks of your work?
Yes, in the "Ethics Statement" Section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes. We summarize our main claims in the abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes. We use artifacts in Section 4 and Section 6. We create artifacts in Section 4.

☑ B1. Did you cite the creators of artifacts you used?
Yes. We cite the authors in the corresponding sections.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Yes, in the "Ethics Statement" Section.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, in the "Ethics Statement" Section.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Yes, in the "Ethics Statement" Section.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in Section 4 and Appendix F.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, in Section 3, Section 4, and Appendix D.

C ☑ Did you run computational experiments?

Yes, in Section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, in Appendix D.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in Appendix D.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes, in Section 6 and Appendix D.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes, in Appendix D.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Yes, in Section 3.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Yes, in Section 3 and Appendix F.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Yes, in Section 3, Appendix F, and the "Ethics Statement" Section.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Yes, in the "Ethics Statement" Section.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. There is no formal ethics committee in our institution, but our plan was discussed internally. Our data collection adheres to the relevant code of ethics.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Yes, in Appendix F.