Leveraging Entailment Judgements in Cross-Lingual Summarisation

Huajian Zhang*   **Laura Perez-Beltrachini**
ILCC, School of Informatics
University of Edinburgh
huajian.zhang.98@gmail.com, lperez@ed.ac.uk

Abstract

Synthetically created Cross-Lingual Summarisation (CLS) datasets are prone to include document-summary pairs where the reference summary is unfaithful to the corresponding document as it contains content not supported by the document (i.e., hallucinated content). This low data quality misleads model learning and obscures evaluation results. Automatic ways to assess hallucinations and improve training have been proposed for monolingual summarisation, predominantly in English. For CLS, we propose to use off-the-shelf cross-lingual Natural Language Inference (X-NLI) to evaluate faithfulness of reference and model generated summaries. Then, we study training approaches that are aware of faithfulness issues in the training data and propose an approach that uses unlikelihood loss to teach a model about unfaithful summary sequences. Our results show that it is possible to train CLS models that yield more faithful summaries while maintaining comparable or better informativess.[1]

1 Introduction

A widely used method to create abstractive summarisation datasets is to crawl websites from which documents paired with reference summaries can be extracted. Examples of this are synthetic datasets created in the news (Grusky et al., 2018; Narayan et al., 2018; Scialom et al., 2020; Hasan et al., 2021) and instructional domains (Ladhak et al., 2020) and for descriptive summarisation (Liu et al., 2018; Perez-Beltrachini et al., 2019; Perez-Beltrachini and Lapata, 2021). The potential content misalignment in document-summary pairs created in this way raises concerns about the quality of training and evaluation data (Gehrmann et al., 2022).

Previous work carried out manual validation of document-summary pairs in automatically created datasets to assess content overlap thereof (Maynez et al., 2020; Hasan et al., 2021; Perez-Beltrachini and Lapata, 2021; Gao et al., 2023; Chen et al., 2023). This inspection aims at pinpointing whether summaries convey content that cannot be inferred from the document, i.e. *hallucinations*. (Maynez et al., 2020) found that 70% of the pairs in the XSum dataset (Narayan et al., 2018) contain summaries with hallucinations. For multi- and cross-lingual summarisation, (Hasan et al., 2021; Perez-Beltrachini and Lapata, 2021) found ~30% of the summaries to be unfaithful.

Taking a causal look into the hallucinations problem in automatically generated summaries, previous work (Mille et al., 2021; Nan et al., 2021; Liu et al., 2021; Cao and Wang, 2021; Goyal and Durrett, 2021; Choubey et al., 2022; Aharoni et al., 2022; Qiu et al., 2023) inspects whether reference summaries in a dataset contain hallucinations in an automatic way. They further exploit this information about the data quality in terms of faithfulness for training with a better signal. Their focus is on monolingual summarisation (the input document and output summary are in the same language) and the most studied language is English. In this work, we focus on cross-lingual summarisation where the document is written in one language (e.g., Czech) and the corresponding summary is written in a different language (e.g., English).

We propose to leverage cross-lingual natural language inference (Conneau et al., 2018) to supplement human validation and to automatically annotate synthetic cross-lingual datasets with hallucination judgements. Our study focuses on the XWikis corpus (Perez-Beltrachini and Lapata, 2021) consisting of descriptive summaries. It is extracted from Wikipedia by aligning articles in different languages (e.g., Czech and English) and re-combining the lead paragraph of one article with the body of

---

the other to form cross-lingual document-summary pairs. It includes English, French, German, and Czech and we extended it with Chinese.

We study simple training schemes that are aware of hallucinations occurring in reference summaries. These train with a smaller but cleaner dataset where highly unfaithful document-summary pairs are removed, include unfaithful document-summary pairs for training but ignore the unfaithful content or explicitly teach the model about unfaithful summary sub-sequences. Our experiments show that by simply fine-tuning with a smaller but more faithful training set it is possible to improve faithfulness on generated summaries while maintaining their informativeness in the context of CLS.

2 X-NLI Based Faithfulness Estimation

The XWikis Corpus We focus on the XWikis corpus which covers four European languages, English (en), French (fr), German (de), and Czech (cs). In addition, we follow the approach proposed in (Perez-Beltrachini and Lapata, 2021) to extend XWikis with Chinese (zh). In Appendix B.2 we show statistics about the XWikis summarisation task (Perez-Beltrachini and Lapata, 2021) and the added zh language. Thus, XWikis covers high resource pairs, e.g., fr-en and de-en, as well as lower resource and English-distant languages, i.e., cs-en and zh-en. Following previous work we focus our experiments on the All-to-English language pairs.

Efficacy of X-NLI We examine the performance of existing multi-lingual NLI models in our cross-lingual setting, i.e., where the premise and hypothesis are in different languages. We choose X-NLI (Conneau et al., 2018) as multi-lingual NLI dataset and an off-the-shelf model, namely mT5-large, fine-tuned on the X-NLI training set, i.e., pairs of English premise and English hypothesis.[2] The NLI model will transfer zero-shot to other languages, e.g., French, where the premise and hypothesis are in the same language (Conneau et al., 2018). The X-NLI test set includes 15 languages with a total of 5,000 premise-hypothesis pairs. However, we need to verify the zero-shot performance in our cross-lingual scenario where the premise and hypothesis are in different languages. To this end, we derive cross-lingual premise-hypothesis pairs from the X-NLI test set. We combine the premise

	en	fr	de	cs	zh
en	87.27	85.65	85.65	–	84.91
fr	84.97	82.83	82.44	–	81.84
de	85.81	83.41	83.05	–	81.58
cs	–	–	–	–	–
zh	84.49	81.86	82.02	–	81.20

Table 1: Accuracy of the X-NLI model on all language pairs of premise-hypothesis from the X-NLI test set for those languages present in XWikis. Czech (cs) is not included in X-NLI test set.

in one language (e.g., French) with the hypothesis in another language (e.g., English).

Table 1 shows results for the multi- and cross-lingual scenarios. The performance in the cross-lingual scenario is still competitive when compared to that of monolingual English and multi-lingual ones. Thus, following previous work on monolingual English summarisation evaluation, we make use of X-NLI to assess cross-lingual summary faithfulness. Appendix A provides details about the X-NLI dataset and examples of cross-lingual premise-hypothesis pairs.

Models	fr-en	de-en	cz-en	zh-en	AVG
FULLDOC	61.93	72.87	63.43	61.45	64.92
SUMMAC$_{zs}$	73.42	80.27	68.34	65.43	71.87
SENTLI	**80.32**	76.38	70.86	58.59	71.54
INFUSE	78.45	**83.78**	**79.20**	**69.58**	**77.75**

Table 2: Performance measured as ROC-AUC of cross-lingual NLI-based approaches on predicting faithfulness of reference summary sentences.

Benchmarking NLI-based Approaches We benchmark existing NLI-based approaches proposed for monolingual summarisation to assess reference summary faithfulness in our cross-lingual summarisation task. These approaches consider summary sentences as hypotheses and differ in what they consider as premise. FULL-DOC (Maynez et al., 2020; Honovich et al., 2022; Dziri et al., 2022) takes the entire input document as premise while SUMMAC (Laban et al., 2022) considers a single document sentence, the one that yields the highest entailment score. In a middle ground, SENTLI (Schuster et al., 2022) and IN-FUSE (Zhang et al., 2024a) consider a subset of document sentences as premise. Both rank document sentences according to the entailment score they yield when used as premise for summary sentences. SENTLI uses a fixed number of top-k ranked document sentences (together with top-k

---

[2]https://huggingface.co/alan-turing-institute/mt5-large-finetuned-mnli-xtreme-xnli

highest ranked by contradiction score) as premise. In contrast, INFUSE uses a variable input number of highest entailment scoring document sentences. Further details about these approaches and our implementation can be found in Appendix A.

For the comparison, we use the human annotated reference summaries from the XWikis development set in (Perez-Beltrachini and Lapata, 2021). We use the sentence level annotations where three bilingual speakers per language pair judge whether the content conveyed by summary sentences is supported by the corresponding document. The benchmark contains 238 annotated sentences in total. We aggregate annotators' judgements into a single yes (supported) / no (unsupported) class. In Appendix B.3, we provide details about the annotation, aggregation and statistics.

As Table 2 shows, INFUSE outperforms all other approaches. We attribute this to the fact that document-summary pairs in XWikis exhibit diversity (i.e., some summary sentences are simpler or extractive requiring only a few document sentences as supportive premise, while others are complex or abstractive necessitating a larger number of document sentences). Furthermore, the main type of hallucinations in reference summaries are extrinsic ones (i.e., there is no supportive evidence in the input document) and INFUSE shows the best performance on these (Zhang et al., 2024a).

3 Faithful Training

3.1 Cross-lingual Abstractive Summarisation

We formalise cross-lingual summarisation as follows. Given an input document Doc_X written in a language X represented as a sequence of tokens $x = (x_1 \cdots x_{|x|})$, the task is to generate a summary $\widehat{\text{Sum}}_Y$ in a target language Y. $\widehat{\text{Sum}}_Y$ consists of a sequence of tokens $(y_1 \cdots y_{|y|})$ and is generated token-by-token conditioning on x by a summarisation model p_θ as $\prod_{t=1}^{|y|} p_\theta(y_t | y_{1..t-1}, x)$.

Automatic Faithfulness Annotation We split reference summaries into sentences, i.e., $\text{Sum}_Y = (y^1 \dots y^M)$ where M is the number of sentences. We use INFUSE to automatically annotate each reference summary sentence y^j with a faithfulness judgement $F \in \{yes, no\}$. For training, the sentence level annotations are propagated to tokens in the sentence $y^j = (y_1^j \cdots y_{|y^j|}^j)$. Details about the annotation are provided in Appendix C.1.

3.2 Faithfulness Aware Approaches

We study approaches that differ in how they weight unfaithful document-summary pairs during fine-tuning. That is, approaches that range from using a hard weight where the pair is removed to applying a negative weighting scheme where the objective seeks to decrease the likelihood of unfaithful summary sequences. We propose the latter one following ideas in dialogue (Li et al., 2020) and computer vision (Zhu et al., 2022). All the approaches assume the initialization of the model parameters θ from a pre-trained multi-lingual language model.

Clean This variant fine-tunes on a subset from the original training set. We aggregate sentence labels to the summary level and any document-summary pair labeled with no will be removed.

Mask This method sets the loss to zero for tokens in $y^j = (y_1^j \cdots y_{|y^j|}^j)$ labelled with no (Goyal and Durrett, 2021). The training loss is formulated as:

$$\mathcal{L}_{\text{MASK}}(\theta) = -\sum_{t=1}^{|y|} F_t \log p_\theta(y_t | x, y_{<t}) \quad (1)$$

where $F_t = 0$ if y_t is labelled with no, 1 otherwise.

Unlike$_{PR}$ We propose to use unlikelihood loss (Welleck et al., 2019) to fine-tune a model together with MLE. The loss term is defined as:

$$\mathcal{L}_{\text{UL}}(\theta) = -\sum_{t=1}^{|y|} [\log p_\theta(y_t | x, y_{<t}) \quad (2)$$
$$- \alpha \sum_{c \in \mathcal{C}} \log(1 - p_\theta(c | x, y_{<t}))] \quad (3)$$

where α is a hyperparameter for the unlikelihood weight and $\mathcal{C}$ is the set of tokens in $y^j = (y_1^j \cdots y_{|y^j|}^j)$ that are labelled with no. We further encourage more faithful training via implicit prompt, which is similar to segment embedding (Devlin et al., 2019). The intuition is, instead of decreasing the probability of $p(y_t | y_{<t})$ when y_t is labeled as unfaithful, we learn $p(y_t | y_{<t}, F_t)$ instead to avoid impact on language modeling. We convert $\{yes, no\}$ labels to parameterized vectors $\mathbf{p}_1$ (yes) and $\mathbf{p}_0$ (no). We integrate either $\mathbf{p}_1$ or $\mathbf{p}_0$ for model learning, while using $\mathbf{p}_1$ at inference to encourage factual sequences. Additionally, we augment the reference summary Sum_Y with explicit prompts <h> and </h> to denote the start/end of a token sub-sequence labelled with no. At inference time, we mask out <h> and </h>.

4   Experimental Setup

We treat mBART50 (Tang et al., 2020) as our backbone model and benchmark those faithfulness aware approaches introduced in Section 3.2. **Vanilla** is our baseline model fine-tuned on the original training set assuming all instances are faithful. **Random** is a variant of Clean where we remove the same number of document-summary pairs but these are randomly selected. See details about model training in Appendix C.1.

5   Evaluation

Our evaluation focuses on complementary qualitative aspects, *faithfulness* and *informativeness*.

Automatic Evaluation   To evaluate faithfulness of generated summaries $\widehat{\text{Sum}}_{en}$ given input documents Doc_X, we rely on the X-NLI based approach INFUSE described in Section 2. For additional judgements on faithfulness, we use a monolingual metric to assess $\widehat{\text{Sum}}_{en}$ w.r.t. their corresponding input document Doc_{en}. Although Doc_X and Doc_{en} are only comparable, the intuition is that salient content will be present in Doc_{en} and that this verification assesses correctness of generated content. Note that the XWikis corpus provides cross-lingual pairs (Doc_X, Sum_{en}) but also their related document Doc_{en} and reference summary Sum_X. Concretely, we apply UniEval (Zhong et al., 2022) which has been shown to align better with human annotators when compared to ChatGPT-3.5 on various summarisation tasks Liu et al. (2023). UniEval frames faithfulness evaluation as a boolean question-answering task. The document and generated summary with a specific question targeting an evaluation dimension (e.g., *Is this a consistent summary of the document?*) are concatenated as input. The probability of getting *yes* as an answer is taken as the metric score. Here we only use the consistency dimension.

Ideally, increasing the faithfulness of summaries does not come at the cost of sacrificing their informativeness. To validate this, we assess the generated against the reference summaries. We exploit the additional multi-lingual reference summaries and compare $\widehat{\text{Sum}}_{en}$ against Sum_{en} with ROUGE-L (Lin, 2004) and against Sum_X with LaBSE (Feng et al., 2022). LaBSE relies on multi-lingual embeddings and provides cross-lingual similarity scores. This provides a more robust multi-reference evaluation for informativeness. Details about metric

implementations can be found in Appendix C.2.

Human Evaluation   We complemented the automatic evaluation with a human study on faithfulness and informativeness. Given that the evaluation of long document CLS with qualified bilingual speakers is costly, we carried out the evaluation on two representative language pairs, namely fr-en which is high resource in Wikipedia and close to English and zh-en which is low resource and distant. We randomly select 20 samples for each pair and hired two annotators per language pair. We ask the annotators to score candidate summaries on a Likert scale from 1 (low) to 4 (high) in terms of faithfulness (*Does the candidate summary contain any content that is not supported by or contradicts the given document?*) and informativeness (*Does the candidate summary cover the main points in the reference summary?*). Details about the instructions are provided in Appendix E.

6   Results and Discussion

Table 3 (left block) shows results for summarisation models fine-tuned with different signals from faithfulness judgements and evaluated on the XWikis test sets. For all language pairs, faithfulness aware approaches outperform the Vanilla variant trained on the original dataset in terms of faithfulness. The Clean variant outperforms the Random one across the board, showing the impact of faithfulness signal in removing training instances. As for informativeness, it is clear for high resource language pairs (fr-en and de-en) where RL and LaBSE increase. For lower resource and distant pairs (cs-en and zh-en) RL decreases slightly while LaBSE shows improvements. We speculate that test reference summaries for these language pairs contain more hallucinations. To verify this, we evaluate models in a smaller but more faithful subset from the original test set, namely Test_{Faith}, where we include only the highest scoring pairs in terms of INFUSE and LaBSE (i.e., reference summaries Sum_{en} that best align with the input document Doc_X and multi-lingual reference Sum_X). We provide details about the construction of Test_{Faith} in Appendix B.4. These results (Table 3 right block) show a similar overall trend, faithfulness aware approaches outperform the Vanilla and Random baselines. However, all, including baselines, perform better in terms of faithfulness. Two important observations should be drawn here. One is the importance of the quality of evaluation data in

	Model	RL	LaBSE	INFUSE	UniEval	RL	LaBSE	INFUSE	UniEval
				XWikis Test				XWikis Test$_{Faith}$	
fr-en	Vanilla	31.33	64.95	38.14	55.17	32.71	67.11	48.95	56.99
	Random	31.09	63.00	40.96	55.08	32.02	65.07	47.22	56.97
	Clean	31.27	65.20	46.70	56.86	32.93	67.54	58.63	59.02
	Mask	31.23	65.11	**47.54**	56.69	32.92	67.55	**59.34**	59.34
	Unlike$_{PR}$	**31.55**	**65.35**	47.49	**57.21**	**33.49**	**67.61**	59.19	**59.40**
de-en	Vanilla	32.15	66.21	41.73	57.60	34.28	67.58	50.73	59.29
	Random	32.01	64.23	41.28	56.47	33.92	64.27	50.21	56.27
	Clean	32.04	66.21	43.11	58.26	34.09	67.69	51.94	58.54
	Mask	32.23	66.53	**45.68**	58.35	34.50	68.23	**55.01**	**60.65**
	Unlike$_{PR}$	**32.31**	**66.65**	44.45	**58.61**	**34.63**	**68.26**	54.19	60.62
zh-en	Vanilla	**31.57**	64.37	43.91	53.21	33.49	67.98	55.23	57.10
	Random	30.43	61.97	43.11	52.92	33.02	64.24	53.92	55.36
	Clean	31.46	64.30	44.07	53.36	33.59	68.03	54.56	57.18
	Mask	31.36	64.08	47.85	**54.39**	33.45	68.04	57.43	**57.38**
	Unlike$_{PR}$	31.48	**64.52**	**49.22**	53.91	**33.87**	**68.86**	**58.77**	57.35
cs-en	Vanilla	**32.94**	67.33	37.53	57.51	34.47	69.01	48.51	60.04
	Random	31.81	64.82	37.36	57.19	33.25	66.41	48.06	58.12
	Clean	32.82	67.27	39.26	57.69	34.45	69.07	49.05	59.81
	Mask	32.70	67.24	39.04	57.45	34.50	69.15	49.93	59.78
	Unlike$_{PR}$	32.65	**67.39**	**41.30**	**58.55**	**34.89**	**69.54**	**50.74**	**61.89**

Table 3: Results on the XWikis test splits (the left block corresponds to the original test sets and the right block to the filtered higher faithfulness test set Test$_{Faith}$). Average ROUGE-L (RL), LaBSE, INFUSE, and UniEval scores. For all metrics higher is better and the best scores are shown in bold.

the obtained results (Gehrmann et al., 2022). The other is, that even a model trained on noisy data as Vanilla would perform better when the input document is more informative and contains the expected details to form the summary. Amongst the faithfulness aware approaches, both in the original test and Test$_{Faith}$, Clean and Mask perform the closest, while Unlike$_{PR}$ performs slightly better than these. The former ones train with less signal while Unlike$_{PR}$ explicitly decreases probabilities of unfaithful summary sentences.

Table 4 shows the results of our human study. Unlike$_{PR}$ (and to a lesser extent Mask) outperform the Vanilla baseline in terms of faithfulness while being comparable (zh-en) or better (fr-en) in terms of informativeness.[3]

7 Conclusion

We propose an NLI-based approach to faithfulness evaluation of cross-lingual document-summary pairs. We show that the cross-lingual NLI-based evaluation can approximate human judgements of summary faithfulness. We use it to automatically annotate the synthetically created XWikis

	Model	Informativeness	Faithfulness
fr-en	Vanilla	2.03	2.13
	Clean	2.00	2.28
	Mask	2.33	2.98
	Unlike$_{PR}$	2.38	3.10
zh-en	Vanilla	2.28	1.90
	Clean	1.96	2.05
	Mask	2.18	2.60
	Unlike$_{PR}$	2.30	3.08

Table 4: Averaged raw scores out of 4 assigned by human judges to the informativeness and faithfulness criteria on 20 samples from the XWikis test sets (fr-en and zh-en).

corpus and show that by considering faithfulness during training models can generate more faithful summaries while maintaining informativeness. Our cross-lingual faithfulness evaluation and faithfulness aware training approaches are relevant in the context of CLS with Large Language Models (LLMs). The evaluation is directly applicable to assess LLM generated summaries and it would be interesting to explore its use to automatically select good quality examples (Zhang et al., 2024b) for in-context learning or fine-tuning (Brown et al., 2020; Hu et al., 2022). It would also make sense to explore the unlikelihood loss for LLM alignment (Li et al., 2020).

[3] The Cohen Kappa inter-annotator agreement on fr-en is 0.25 for faithfulness and 0.43 for informativeness (0.21 and 0.39 on zh-en). We speculate that the low faithfulness agreement relates to annotators using different scoring ranges.

Acknowledgements

8 Limitations

Our experiments and conclusions are limited to mBART50 (Tang et al., 2020). It remains to be seen the effect of noisy document-summary pairs for training larger models which deliver superior performance (Kaplan et al., 2020).

In existing research that employs automatic faithfulness evaluation models to annotate document-summary pairs in a dataset, a challenge is the absence of a clear-cut threshold for deeming the pair faithful/unfaithful. Some studies (Aharoni et al., 2022) adopt a fixed threshold, such as 0.5 entailment score, where samples scoring above this value are considered faithful and those below are not. However, this one-size-fits-all threshold is not universally applicable across different datasets. Other studies (Qiu et al., 2023) forego setting a static threshold and instead use the evaluation score directly as a weight. With this approach it can be difficult to learn a downstream summariser if a faithfulness evaluation model scores' range over 0 to 0.7 (all training instances are penalised). Additionally, directly using the evaluation score as a weight can amplify the impact of inaccuracies in the faithfulness evaluation model. Even small differences in the evaluator model's score, possibly caused by noise, can disproportionately affect the performance of downstream tasks.

In this work, we regard a specific percentage of the training set as faithful; to this end, we grid search a percentage that proves an adequate data quality versus quantity trade-off showing improvements on the evaluation criteria. This method is similar to a fixed threshold and is also dataset dependent.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. mface: Multilingual summarization with factual consistency evaluation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yulong Chen, Huajian Zhang, Yijie Zhou, Xuefeng Bai, Yueguan Wang, Ming Zhong, Jianhao Yan, Yafu Li, Judy Li, Xianchao Zhu, and Yue Zhang. 2023. Revisiting cross-lingual summarization: A corpus-based study and a new benchmark with improved annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9332–9351, Toronto, Canada. Association for Computational Linguistics.

Prafulla Kumar Choubey, Alexander R. Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Fatema Rajani. 2022. Cape: Contrastive parameter ensembling for reducing hallucination in abstractive summarization.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 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.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Mingqi Gao, Wenqing Wang, Xiaojun Wan, and Yuemei Xu. 2023. Evaluating factuality in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12415–12431, Toronto, Canada. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

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.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR 2022*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

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, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

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.

Peter Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yang Liu, Sheng Shen, and Mirella Lapata. 2021. Noisy self-knowledge distillation for text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–703, Online. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer 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.

Simon Mille, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic Construction of Evaluation Suites for Natural Language Generation Datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. (NeurIPS 2021).

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating summaries with topic templates and structured convolutional decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5107–5116, Florence, Italy. Association for Computational Linguistics.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

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.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, page 1101–1109, USA. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024a. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024b. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and

Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deyao Zhu, Mohamed Zahran, Li Erran Li, and Mohamed Elhoseiny. 2022. Motion forecasting with unlikelihood training in continuous space. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 1003–1012. PMLR.

A X-NLI Based Faithfulness Estimation

X-NLI model The X-NLI (Conneau et al., 2018) dataset was constructed in the following way. The English premises are collected from MultiNLI (Williams et al., 2018) and corresponding English hypotheses are produced by employed workers. All premise-hypothesis pairs are further translated into 15 languages by professional translators. We derive cross-lingual premise-hypothesis pairs from the X-NLI test set by aligning the premise in one language (e.g., Chinese) with the hypothesis in another language (e.g., English). Table 5 shows examples of the original and derived premise-hypothesis pairs. Note also in Table 5 that we do not include results for the Czech language as it is not covered in the 15 languages of the X-NLI test. Nevertheless, the performance in the cs-en cross-lingual faithfulness evaluation setting is acceptable (see Table 2).

In Table 1, we find the performances on $\mathcal{D}_{X \to en}$ and $\mathcal{D}_{en \to X}$ are even better than the mono-lingual $D_{X \to X}$ direction in languages other than English. We guess the reason can be that the model is fine-tuned in monolingual English sentence pairs and our cross-lingual setting benefits from having premise or hypothesis sentences in English.

NLI-based Faithfulness Evaluation Approaches
SUMMAC$_{ZS}$ (Laban et al., 2022) operates under the assumption that each sentence within a summary is supported by one corresponding sentence from the document. Consequently, it selects the document sentence with the highest entailment score to serve as the context for each summary sentence.

SENTLI (Schuster et al., 2022) retrieves a fixed number of document sentences, denoted by k, to form its context for each summary sentence. This context is composed of the document sentences that have the highest scores for both entailment and contradiction. Following (Schuster et al., 2022), we set the value of $k = 5$.

INFUSE (Zhang et al., 2024a) incrementally selects entailment-ranked document sentences and evaluates summary sentences against the formed premise; if there is no further increase in the predicted neutral score, it will automatically stop yielding the entailment score of the current premise as faithfulness score. In our application of this method, we set the automatic stopping criterion as the upper bound, while requiring the premise to include at least 5 document sentences.

B Datasets

B.1 Ethics Consideration

We honor the ACL Code of Ethics. We use XWikis under CC BY-SA 4.0 International License. XWikis includes content from Wikipedia, which is under CC BY-SA 4.0 International License and GFDL. We ensure that the data was used only for academic purposes, which aligns with the intended use of the dataset. For data safety, it is not avoidable that some document/summary pairs can contain uncomfortable content, including hate, crimes, and wars.

For the human validation of the zh-en pair added to the XWikis corpus (see Section B.3 and Table 8), we recruited three Chinese-English bilingual annotators with degrees in English-speaking countries colleagues in our field. Annotators were asked to annotate 20 samples randomly extracted from the validation set. The annotation does not involve any personally sensitive information. Similarly, for the human study in Section 5 we recruited four volunteer bilingual annotators, two for fr-en and two for zh-en.

B.2 Adding Chinese to the XWikis Corpus

We follow the approach proposed in (Perez-Beltrachini and Lapata, 2021) to extend the XWikis corpus with the Chinese language. Table 6 shows the number of instances for each language pair in the XWikis corpus including monolingual (e.g., French) subsets. Note that these are taken from the recently published version of XWikis in the HuggingFace repository.[4] Figure 1 shows an XWikis example taken from (Perez-Beltrachini and Lapata, 2021) of document in French and corresponding reference summaries in all other languages where we include the zh summary.

Table 7 shows the analysis of the All-to-English XWikis CLS task carried out in (Perez-Beltrachini

---

[4] https://huggingface.co/datasets/GEM/xwikis

Language	Premise / Hypothesis	Label
English	There's so much you could talk about on that I'll just skip that.	Entailment
	I won't talk about that, even though there's a lot to cover.	
German	Es gibt so viel was ich darüber erzählen könnte, ich überspringe das einfach.	Entailment
	Ich werde nicht darüber reden, obwohl es viel zu decken gibt.	
French	Il y a tellement de choses dont vous pourriez parler que je vais juste m'en passer.	Entailment
	Je n'en parlerai pas, même s'il y a beaucoup à couvrir.	
Chinese	你可以讲的太多了，我就不提了。	Entailment
	即使需要说的很多，但我也不会谈论这个。	

Table 5: Example from the X-NLI test set. The semantics of premises/hypotheses in four languages are the same. We obtain cross-lingual NLI pairs by combining the premise in one language (e.g., *Il y a tellement de choses dont vous pourriez parler que je vais juste m'en passer.*) with the hypothesis in the other language (e.g., *I won't talk about that, even though there's a lot to cover.*).

and Lapata, 2021). We apply the same metrics to the newly extracted Chinese subsets. Grusky et al. (2018) introduced metrics to evaluate how much a summary borrows textual fragments from its source document. Coverage, calculates the average number of summary tokens that are part of an extractive fragment from the document, indicating how much of the summary is directly extracted. Density, measures the average length of these extracted fragments, showing the depth of direct borrowing. Compression, calculates the ratio of the number of tokens in the summary to the number of tokens in the corresponding document. These metrics help assess a summary's extractiveness. Additionally, we report % of novel ngrams (i.e., the percentage of novel tokens that appear in the summary but are not present in the document) as an additional indication of extractiveness. The observations drawn from this analysis for Chinese are the following. Chinese documents are relatively shorter than those in the other four languages but summaries are still close in length. As for diversity, the aspects covered in zh are closer to those in cs, much lower than de and fr. We speculate one reason may be due to the fewer number of document summary pairs for these languages. Generally, zh documents indeed contain multi-topic and sections per document (Sections/Doc) match the other four languages.

We also evaluate the performance of two extractive summarisation models, **LEAD** and **EXT-ORACLE**, on the validation set. We assume that these two extractive models should obtain good results if the dataset is extractive in nature. LEAD generates a summary by simply copying the first N tokens from the document, where N equals the

	en	de	fr	cs	zh
en		425,279	468,670	148,519	135,674
de	376,803		252,026	109,467	103,044
fr	312,408	213,425		91,175	99,301
cs	64,310	53,275	51,578		32,588
zh	75,524	73,969	81,847	43,281	

Table 6: Total number of document-summary pairs in the XWikis corpus considering all language pairs and directions.

length of the reference summaries. EXT-ORACLE is to generate a summary by selecting the portion of sentences to maximize ROUGE-2 with respect to the reference. Intuitively, when the salient information concentrates on the first several sentences of the document or so-called lead bias, LEAD performs well. And when the summarisation task is extractive, EXT-ORACLE should have a good performance (Perez-Beltrachini and Lapata, 2021). Rouge-L is used as the performance indicator.

The last two rows of Table 7 provide the results for the two extractive models. LEAD is below EXT-ORACLE by around 4 ROUGE-L points, suggesting no lead bias in the summaries. The performance of EXT-ORACLE is also not good. One reason can be that in the reference summaries, each sentence aggregates information from multiple sentences across the document. Another reason could be related to the high ratio of new paraphrases appearing in summary, which cannot be captured by an extractive model. The performance of these two models is extremely weak in Chinese. We further explore this phenomenon by translating target summaries and generated summaries (by LEAD/EXT-ORACLE) on the test set to English. They achieve

Huile d'Olive

Histoire. La consommation alimentaire d'olives sauvages date de la période préhistorique des chasseurs-cueilleurs du Néolithique. L'oléiculture (culture d'oliviers, d'oliveraie, et fabrication d'huile d'olive avec des moulins à huile) remonte à la période de l'invention de l'agriculture et de la culture de la vigne et du vin, il y a environ 8 000 ans, dans la région du croissant fertile au Levant au Proche-Orient et en Mésopotamie. L'huile d'olive est alors utilisée pour l'alimentation, la conservation des aliments, la cosmétique, la médecine, les lampes à huile... [...]. Durant la Renaissance du XVe siècle l'Italie devient le plus important producteur réputé d'huile d'olive du monde, avant d'être cultivée à ce jour par l'ensemble des pays du bassin méditerranéen en tant qu'un des fondements de la cuisine méditerranéenne. [...] **Utilisation.** L'huile d'olive est connue depuis la plus haute antiquité : les Grecs anciens, les Phéniciens, les Arabes, les Berbères et les Romains l'utilisaient déjà pour leur cuisine (à l'origine de la cuisine méditerranéenne) et pour leurs produits cosmétiques, ainsi que les Hébreux pour allumer leur chandelle. L'huile d'olive peut être utilisée aussi bien crue (dans des sauces pour salade ou à la place du beurre dans les pâtes par exemple) que cuite (pour la cuisson de viandes ou de légumes ou pour la friture). [...] L'huile d'olive peut également être utilisée pour le traitement du visage, comme le démaquillage des yeux, l'hydratant, l'apaisement des lèvres et la réparation des talons fissurés. Naturellement, l'huile d'olive regorge d'antioxydants anti-âge et de squalène hydratant, ce qui la rend superbe pour les cheveux, la peau et les ongles. Tout comme l'huile de noix de coco, c'est un élément essentiel de tout kit de beauté bricolage. L'huile d'olive est utilisée comme traitement capillaire depuis l'Antiquité égyptienne. [...] **Production.** L'obtention d'un litre d'huile nécessite 4 à 10 kg d'olives suivant la variété d'olive utilisée et son niveau de maturité. La méthode d'extraction utilisée a peu d'incidence. Cependant les moulins utilisant des presses ne peuvent pas utiliser des olives à très forte teneur en eau (à faible rendement en huile) à cause de la fluidité excessive de la pâte. Ceci peut laisser penser, à tort, que leurs rendements sont meilleurs. [...] **Santé.** Par rapport aux autres huiles contenant des acides gras insaturés, l'huile d'olive est assez stable à la cuisson et garde en ce cas ses effets bénéfiques relatifs sur le cholestérol. L'huile d'olive permet de lutter contre le cancer du sein. [...]

Olivenöl, auch Baumöl (von mittelhochdeutsch boumöl) und fachsprachlich Oleum olivarum genannt, ist ein Pflanzenöl aus dem Fruchtfleisch und aus dem Kern von Oliven, den Früchten des Ölbaums, das seit mindestens 8000 Jahren, ausgehend vom östlichen Mittelmeerraum, gewonnen wird. **De**

Olivový olej je rostlinný olej získaný z oliv (Olea europaea) lisováním či jinými mechanickými postupy. Je používán především při výrobě pokrmů, v kosmetickém a farmaceutickém průmyslu. Olivový olej je typickou součástí středomořské kuchyně. Konzumace olivového oleje je považována za zdraví prospěšnou, protože obsahuje mononenasycené mastné kyseliny a vitamin E. **Cs**

L'huile d'olive est une variété d'huile alimentaire, à base de matière grasse végétale extraite des olives (fruits d'oliviers cultivés en oliveraie d'oléiculture) lors de la trituration dans un moulin à huile. Elle est un des fondements de la cuisine méditerranéenne (et du paradoxe français) et peut être, sous certaines conditions, bénéfique pour la santé. **Fr**

橄榄油是植物油的一种, 由木犀科油橄 榄(学名：Olea europaea L.；又被称为"洋橄榄"或"齐墩果")的果实压榨而成, 是一种常用的食油, 也可用以制作化 妆品、药物及油灯燃料等。地中海国家的居民所说的"油", 指的总是"橄榄油"。西班牙语的"油"(aceite), 是来自阿拉伯语阿泽特(azzáyt)一词, 这个词的字面意思是"橄榄汁"。新鲜的橄榄油, 可直接用于凉拌菜, 比如沙拉。 **Zh**

Olive oil is a liquid fat obtained from olives (the fruit of Olea europaea; family Oleaceae), a traditional tree crop of the Mediterranean Basin, produced by pressing whole olives and extracting the oil. It is commonly used in cooking, for frying foods or as a salad dressing. It is also used in cosmetics, pharmaceuticals, and soaps, and as a fuel for traditional oil lamps, and has additional uses in some religions. The olive is one of three core food plants in Mediterranean cuisine; the other two are wheat and grapes. Olive trees have been grown around the Mediterranean since the 8th millennium BC. **En**

Figure 1: Example document in French and reference summaries in German, French, English, Czech, and Chinese.

17.14 and 24.57, respectively, which can be comparable with the other three languages. One explanation is that the evaluation of ROUGE score on the other three languages is done after stemming, which is not supported for Chinese, thus leading to a drop in performance. There potentially exist other reasons such as noise introduced by the Chinese word segmentation tool in the data pre-processing stage.

We use the following model configurations and tools for the analysis presented in Table 7. We apply PTBTokenizer from Stanford CoreNLP (Manning et al., 2014) to tokenize references. We evaluate the generated outputs with files2rouge package[5]. For extractive methods applied to Chinese, we uti-

lize HanLP (He and Choi, 2021) for segmenting sentences into words, removing stopwords, and performing tokenization. For all other models, we use the same SentencePiece (Kudo and Richardson, 2018) model used by (Perez-Beltrachini and Lapata, 2021).

B.3 Human Annotation of Hallucinations in XWikis Reference Summaries

To assess the performance of faithfulness evaluation approaches, we use the human annotated reference summaries from the XWikis validation set in (Perez-Beltrachini and Lapata, 2021). We also reproduce this evaluation for the newly added Chinese subset, zh-en. It contains sentence level annotations where three bilingual speakers per language pair judge whether the content conveyed by sum-

---

[5]https://github.com/pltrdy/files2rouge

	XWikis (comp)				XWikis (Test)	
	de	fr	cs	zh	de/fr/cs	zh
Words/Doc	906	1040	890	769	972	890
Sents/Doc	41	38	42	36	42	40
Sections/Doc	5	7	6	6	6	6
Words/Sum	56	59	65	61	61	70
Sents/Sum	3	2	3	3	3	3
Aspects	253,425	248,561	65,151	75,796	9,283	11,722
Coverage	65.53	72.23	55.97	55.05	65.41	62.21
Density	1.23	1.51	0.99	0.94	1.23	1.14
Compression	17.44	20.16	15.12	15.17	18.35	10.89
% novel unigrams	33.30	26.77	42.29	39.77	33.25	33.38
bigrams	80.70	73.19	85.17	84.37	79.51	80.01
trigrams	93.60	90.25	95.19	95.32	93.17	93.44
4-grams	97.98	95.68	97.98	98.13	97.11	97.17
LEAD	19.09	23.51	20.21	12.24	20.88	13.19
EXT-ORACLE	24.59	28.38	24.25	16.48	25.95	17.43

Table 7: XWikis statistics (number of words and sentences per document (/Doc) and summary (/Sum)) and task characterisation metrics. Statistics for de/fr/cs-en are taken from (Perez-Beltrachini and Lapata, 2021) and zh-en are computed in this work.

mary sentences is supported by the corresponding input document. As we need a single label per sentence, we aggregate human judgements as follows. We map each judgement to a numerical value and distinguish when a summary sentence is partially supported, that is, $no \rightarrow 0$, $partial \rightarrow 1$, and $yes \rightarrow 2$. We then take the sum of the three values. If the sum is in the interval $[0 - 2]$ we assign a *not-entail* label, if it falls in the interval $[3 - 6]$ we assign an *entail* label. Table 8 shows the percentage of *not-entail*. We also show the inter-annotator agreement measured by Fleiss's Kappa on yes/no judgements (yes includes partial) before the aggregation (all values, except zh-en that we add here, are taken from (Perez-Beltrachini and Lapata, 2021)). The overall moderate agreement highlights the difficulty of the task even for humans.

B.4 Faithful Test Subsets

In order to create higher quality evaluation sets, we utilize LaBSE and INFUSE for evaluating reference English summaries from the original XWikis

	de-en	fr-en	cs-en	zh-en
% not-entail	48.33	37.74	47.69	36.67
Fleiss's Kappa	0.48	0.55	0.59	0.47

Table 8: Proportion of no labels, i.e., summary sentences whose content is not supported by the input document, and inter-annotator agreement.

language pairs test sets. We aggregate their results for each sample. Those samples with top 10% scores (700 samples) are then selected to create faithful Test$_{Faith}$ sets (each language pair has its test data).

C Implementation Details

C.1 Automatic Faithfulness Annotation and Model Training

We use the mBART50 checkpoint provided as mMBART-50-finetuned-many-to-many. We follow (Perez-Beltrachini and Lapata, 2021) setting, fine-tuning for a total of 20K updates with a batch size

	Model	RL	INFUSE
fr-en	Vanilla	32.59	43.73
	Random	32.53	43.05
	Clean	32.79	47.21
	Mask	33.01	**49.25**
	Unlike$_{PR}$	**33.22**	48.77
de-en	Vanilla	33.24	43.54
	Random	33.05	43.42
	Clean	33.14	46.40
	Mask	33.32	**48.90**
	Unlike$_{PR}$	**33.64**	47.44
zh-en	Vanilla	**32.52**	43.53
	Random	31.50	43.31
	Clean	32.09	44.12
	Mask	32.44	50.20
	Unlike$_{PR}$	32.36	**50.98**
cs-en	Vanilla	**34.29**	38.16
	Random	34.14	37.53
	Clean	34.23	38.79
	Mask	34.15	41.09
	Unlike$_{PR}$	34.23	**41.44**
en-en	Vanilla	35.62	44.74
	Random	35.43	44.16
	Clean	35.58	46.44
	Mask	35.07	47.93
	Unlike$_{PR}$	**35.69**	**48.63**

Table 9: Results on the XWikis validation splits. Average ROUGE-L (RL) and INFUSE scores.

	Model	RL	LaBSE	INFUSE	UniEval
en-en	Vanilla	31.12	68.32	49.21	78.48
	Random	30.94	65.87	48.58	75.06
	Clean	31.21	**68.48**	51.21	**79.71**
	Mask	30.72	68.14	51.38	78.72
	Unlike$_{PR}$	**31.28**	68.43	**51.98**	78.91

Table 10: Results on the XWikis en-en test split. Average ROUGE-L (RL), LaBSE, INFUSE, and UniEval scores.

C.2 Automatic Evaluation

LaBSE (Feng et al., 2022) is a dual-encoder model based on the pre-trained BERT model. It is further fine-tuned on translation ranking tasks. It is trained on mono-lingual and bi-lingual data. The monolingual data contains CommonCrawl4 and Wikipedia. The translation corpus is constructed from Web pages using a bitext mining system (Uszkoreit et al., 2010). We utilize its PyTorch implementation available at `https://github.com/yang-zhang/labse-pytorch`. For INFUSE (Zhang et al., 2024a), we access the code from `https://github.com/HJZnlp/infuse`. For UniEval (Zhong et al., 2022), we employ the implementation at `https://github.com/maszhongming/UniEval`, specifically downloading UniEval-Sum, which is designed for evaluating summarisation tasks. UniEval reports evaluation scores across four dimensions, but we only focus on consistency, which measures the factual alignment between a summary and the corresponding document.

D Additional Results and Analyses

D.1 The Effect of Faithfulness Aware Training in Monolingual English Summarisation

Additionally, we report results for the monolingual English subset (en-en) from XWikis in Table 10. Interestingly, in this setting it is also possible to train with faithfulness judgements on document-summary pairs and obtain models that generate more faithful summaries without loss in informativeness. As this is a monolingual summarisation task, it enables us to inspect the relation between performance improvements in terms of faithfulness and informativeness, amount of training data, and extractiveness of generated summaries. Table 11 shows statistics about performance versus extractiveness for the Clean approach for different amounts of annotated document-summary pairs considered (i.e., removed). For informativeness metrics, RL and LaBSE, the optimal performance

of 80. We used 2 A6000 GPUs, and the fine-tuning time cost is 45 hours.

We employ INFUSE to automatically annotate each reference summary sentence y^j with a faithfulness score. We regard X% of the sentences with the lowest score as unfaithful and leave the rest without annotations (faithful). For the Clean model variant, we average sentence level scores to sample level, and take X% of samples with the lowest score as unfaithful. The determination of X for different languages was informed by an analysis of average entailment scores obtained from the cross-lingual NLI model across the five languages pairs. This inspection revealed lower entailment scores for English and Chinese, suggesting a higher prevalence of unfaithful content in summaries in these language pairs. Based on these findings, we chose a threshold value of 10 for de-en, fr-en, and cs-en; and 20 for zh-en and en-en. We also performed a grid search on the validation set (with X ranging from 10% to 50%) to confirm that this thresholds, maintain informativeness while improving faithfulness. Table 9 shows RL and INFUSE results for the different model variants (Vanilla, Random, Clean, Mask, and Unlike$_{PR}$) on the validation set.

Ratio	RL	LaBSE	INFUSE	UniEval	Coverage	Density	Compression	unigrams	bigrams	trigrams	4grams
10%	31.06	68.00	47.69	78.09	0.83	5.43	12.93	0.17	0.49	0.66	0.73
20%	**31.21**	**68.48**	51.21	79.71	0.83	6.51	12.45	0.17	0.47	0.63	0.70
30%	30.88	67.79	50.99	77.91	0.83	4.83	13.03	0.17	0.49	0.67	0.75
40%	30.68	68.36	**55.44**	**80.95**	0.85	8.32	11.78	0.15	0.43	0.58	0.65
50%	30.69	68.09	55.21	80.58	0.85	7.40	12.29	0.15	0.44	0.59	0.66

Table 11: Results on the en-en XWikis test split for the Clean approach. Ratio shows the percentage of removed data examples from the training set. Average INFUSE, ROUGE-L (RL), LaBSE, and UniEval scores are the evaluation metrics introduced in Section 5. The other metrics are those extractive metrics described in Section B.2.

is observed upon the removal of 20% of noisy (i.e., with hallucinations) document-summary pairs, whereas, for faithfulness metrics, INFUSE and UniEval, the performance generally improves as the ratio increases up to 40%. We can also observe that summaries generated by Clean variants fine-tuned with less but higher faithfulness scoring document-summary pairs tend to become more extractive (i.e., incorporating fewer novel n-grams). A more extractive summary, which largely copies content directly from the document, is indeed more faithful but potentially less informative. This observation suggests that although simply filtering out noisy document-summary pairs can lead to better summaries in terms of both faithfulness and informativeness, it is crucial to maintain a balance between the two (Ladhak et al., 2022).

D.2 Ablation Experiments on the Unlike$_{PR}$ Model Variant

We further inspect the role of segment embeddings in Unlike$_{PR}$. We remove segment embeddings from Unlike$_{PR}$ and name the resulting model **Unlike**. Table 12 shows the performance of Unlike$_{PR}$ versus Unlike. We observe that Unlike$_{PR}$ outperforms Unlike in terms of informativeness (RL and LaBSE). For faithfulness, INFUSE and UniEval, in low-resource (smaller training sets and data in the base pre-trained language model) language pairs (cs-en and zh-en), Unlike$_{PR}$ also performs better. However, in high-resource languages (de-en and fr-en), Unlike obtains higher INFUSE scores (higher than all approaches in Table 3). One possible reason for this is that fr-en and de-en are far larger datasets and thus they receive a stronger signal to adhere to the input document from Unlike.

D.3 Faithfulness in Out-of-distribution Summarisation

We additionally evaluate the faithfulness aware training approaches in a cross-lingual out-of-

	Model	RL	LaBSE	INFUSE	UniEval
fr-en	Unlike	31.37	64.22	**48.30**	56.77
fr-en	Unlike$_{PR}$	**31.55**	**65.35**	46.49	**57.21**
de-en	Unlike	32.20	65.14	**45.53**	57.33
de-en	Unlike$_{PR}$	**32.31**	**66.65**	44.45	**58.61**
zh-en	Unlike	31.09	63.90	49.03	53.10
zh-en	Unlike$_{PR}$	**31.48**	**64.52**	**49.22**	**53.91**
cs-en	Unlike	32.03	65.78	40.10	57.65
cs-en	Unlike$_{PR}$	**32.65**	**67.39**	**41.30**	**58.55**

Table 12: Results on the XWikis test splits. Average ROUGE-L (RL), LaBSE, INFUSE, and UniEval scores.

distribution setting. This experiment aims at assessing whether models that are trained in higher quality (i.e., more faithful) data will exhibit better faithfulness when evaluated in out-of-distribution. To this end, we follow (Perez-Beltrachini and Lapata, 2021) and evaluate zero-shot the different model variants trained in XWikis monolingual English over the Voxeurop dataset on cross-lingual news summarisation. Results are shown in Table 13.

Zero-shot results are generally lower than those shown in Table 3, indicating that zero-shot performance remains a challenge. In terms of faithfulness (INFUSE and UniEval), the Unlike$_{PR}$ method significantly outperforms others. However, regarding informativeness (RL and LaBSE), all model variants show small differences. Models trained on Wikipedia can capture the gist of news documents and all faithfulness aware variants show good transfer ability. However, the summary patterns differ between Wikipedia and the news domain. Consequently, these models achieve similarly low scores on metrics requiring reference summaries (RL and LaBSE).

E Human evaluation

We conduct a human evaluation to verify the effectiveness of faithfulness-aware approaches. To this end, we hire two bilingual annotators for fr-en and two for fr-en. We present them with 20 document-

	Model	RL	LaBSE	INFUSE	UniEval
fr-en	Vanilla	20.62	**41.84**	31.30	53.90
	Random	20.21	41.09	30.83	53.57
	Clean	20.51	41.12	31.54	55.14
	Mask	20.82	41.65	39.48	57.98
	Unlike$_{PR}$	**20.95**	41.03	**42.69**	**62.50**
de-en	Vanilla	21.14	42.16	31.86	52.88
	Random	20.85	41.70	31.06	52.29
	Clean	21.10	42.08	34.68	55.00
	Mask	21.14	**42.55**	38.45	55.21
	Unlike$_{PR}$	**21.42**	41.98	**44.70**	**60.34**
cs-en	Vanilla	21.23	44.85	35.00	52.41
	Random	20.10	44.35	34.19	52.91
	Clean	20.51	**44.88**	39.26	54.68
	Mask	20.51	44.67	39.32	55.31
	Unlike$_{PR}$	**21.60**	44.52	**45.15**	**59.12**

Table 13: Results on the Voxeurop test splits. All models are trained in monolingual english setting. Average ROUGE-L (RL), LaBSE, INFUSE, and UniEval scores.

summary samples of the corresponding language pair. Each sample includes a document and four summaries generated by the Vanilla, Clean, Mask, and Unlike$_{PR}$ models. The annotators are tasked to score each summary first in terms of faithfulness and then in terms of informativeness. The scores are on a Likert scale from 1 (low) to 4 (high). The full instructions are shown below.

Annotation Instructions To run this annotation task you should be native (or near native) of French (Chinese) and have a good level of English.

A cross-lingual summary expresses relevant information from a document in a different language than the language of the document. For instance, an English summary from the French (Chinese) document for the Wikipedia title Olive Oil is shown in the example below.[6] Your task is to assess the quality of cross-lingual summaries in terms of content overlap with the underlying document. To judge this, you will read a document and corresponding candidate summaries. Then, you will score the candidate summaries in terms of consistency (low consistency 1 to high consistency 4) and then score them in terms of informativeness (low informativeness 1 to high 4).

Criteria

Consistency: Does the candidate summary contain any content that is not supported by or contra-

---

[6]The French (Chinese) and Englih document-summary pair as the one in Figure 1 was shown in the interface; together with example annotated candidate summaries.

dicts the given document? You should **ONLY** refer to the document when evaluating this aspect.

Informativeness: Does the candidate summary cover the main points in the reference summary? (That is, which candidate summary covers better/more content expressed in the reference). To respond to this question you should **ONLY** look at the reference summary.

F Examples of Model Outputs

Tables 14, 15, 16, and 17 show examples of the baseline (Vanilla) and the studied faithfulness aware approaches. Hallucinated text spans are highlighted in gray. We manually verify their content against the corresponding documents. We can see that across all language pairs Unlike$_{PR}$ has fewer hallucinated spans.

Model	
Reference	The canton of St. Gallen, also canton of St Gall (; dialectally , , ,), is a canton of Switzerland. The capital is St. Gallen.
Vanilla	The canton of Saint-Gall () is a canton in Switzerland. It is located in the north-eastern part of Switzerland and is bordered by the Rhine valley to the south by the cantons of Grisons, Glaris and Schwytz. The population of the canton is .
Clean	The canton of Saint-Gall () is an administrative division of Switzerland. It is located in the northeastern part of the state of Saxony . The administrative divisions of the canton are: Its seat is in St. Gallen.
Mask	The canton of Saint-Gall () is a canton in Switzerland. It was created on 19 February 1803. Its territory covers the following territories: It is bordered on the north by Lake Constance and Thurgovia. On the east it is separated from Austria and Liechtenstein by the Rhine Valley, on the south by the cantons of Grisons, Glaris and Schwytz and on the west by Zurich.
Unlike$_{PR}$	The canton of Saint-Gall () is an administrative division of Switzerland. It is located in the north-eastern part of the country. The administrative center of the canton is the city of St. Gallen.

Table 14: Example comparing the Reference summary with the generated ones by the baseline (Vanilla) and the faithfulness aware approaches (Clean, Mask, and Unlike$_{PR}$) from the fr-en test set.

Model	
Reference	Piaget SA () is a Swiss luxury watchmaker and jeweller. Founded in 1874 by Georges Piaget in the village of La Côte-aux-Fées, Piaget is currently a subsidiary of the Swiss Richemont group.
Vanilla	La Côte-aux-Fées () is a Swiss luxury watch manufacturer. Founded in 1874 by Georges Edouard Piaget, the company produces watches and jewellery. The company's headquarters are in Neuchâtel , Switzerland. It is one of the world's best-selling luxury brands.
Clean	Piaget is a Swiss luxury watch manufacturer based in La Côte-aux-Fées, Switzerland. The company was founded in 1874 by Georges Edouard Piaget, son of Georges Piaget . It is best known for its watches and watchmaking machinery. Piaget's watches are made in a variety of shapes and sizes.
Mask	Piaget is a Swiss luxury watch manufacturer based in La Côte-aux-Fées. Founded in 1874 by Georges Edouard Piaget, the company produces watches and jewellery. The company's name is derived from the French word "Piaget", which means "horsemaker".
Unlike$_{PR}$	Piaget is a Swiss luxury watch manufacturer based in the village of La Côte-aux-Fées. Founded in 1874 by Georges Edouard Piaget, the company produces watches and jewellery. The company is known for its high precision watches, including the "Emperador" watches.

Table 15: Example comparing the Reference summary with the generated ones by the baseline (Vanilla) and the faithfulness aware approaches (Clean, Mask, and Unlike$_{PR}$) from the cz-en test set.

Model	
Reference	Giorgio Vasari (, , ; 30 July 1511 – 27 June 1574) was an Italian painter, architect, writer, and historian, best known for his "Lives of the Most Excellent Painters, Sculptors, and Architects", considered the ideological foundation of art-historical writing. He was also the first to use the term "Renaissance" in print.
Vanilla	Giorgio Vasari (30 July 1511 – 27 June 1574) was an Italian painter, sculptor, and architect. He is best known for his portraits of the walls and ceilings of the Palazzo della Cancelleria in Florence , Italy, and for his murals in Rome.
Clean	Giorgio Vasari (30 July 1511 – 27 June 1574) was an Italian painter, painter and painter of the Renaissance. He is best known for his portraits of the walls and ceilings of the Palazzo Vecchio, the Palazzo della Cancelleria, and the Palazzo dei Cento Giorni.
Mask	Giorgio Vasari (30 July 1511 – 27 June 1574) was an Italian painter, sculptor, and art theorist. He is best known for his paintings on the walls and ceilings of the Palazzo della Cancelleria in Florence , and for the frescoes in the Palazzo Vecchio in Rome .
Unlike$_{PR}$	Giorgio Vasari (30 July 1511 – 27 June 1574) was an Italian Renaissance painter and sculptor. He is best known for his frescoes in the Palazzo della Cancelleria (Palazzo dei Cento Giorni) in Florence and in Rome, as well as for his work on the walls and ceilings of the Palazzo Vecchio.

Table 16: Example comparing the Reference summary with the generated ones by the baseline (Vanilla) and the faithfulness aware approaches (Clean, Mask, and Unlike$_{PR}$) from the zh-en test set.

Model	
Reference	Rudolf Karl Bultmann (; 20 August 1884 – 30 July 1976) was a German Lutheran theologian and professor of the New Testament at the University of Marburg. He was one of the major figures of early-20th-century biblical studies. A prominent critic of liberal theology, Bultmann instead argued for an existentialist interpretation of the New Testament. His hermeneutical approach to the New Testament led him to be a proponent of dialectical theology.
Vanilla	Rudolf Bultmann (November 17, 1855 – January 22, 1951) was a German Protestant theologian and philosopher. He was born in Oldenburg, Germany. His father was a liberal theologians, and his mother was a pietist.
Clean	Rudolf Bultmann (February 1, 1855 – March 2, 1951) was a German Protestant theologian and philosopher. He was born in Oldenburg, Germany, and died in Marburg, Germany. His father was a Lutheran pastor.
Mask	Rudolf Bultmann (October 22, 1855 – February 26, 1951) was a German Protestant theologian and philosopher. He was born in Oldenburg, Germany. His father was a liberal theologians, and his mother was a pietist. In 1895 he attended the gymnasium at Oldenburg. During this time he was a member of the student association Camera obscura Oldenburgensis.
Unlike$_{PR}$	Rudolf Bultmann (born Arthur Kennedy Bultman in Oldenburg, Germany; died in Marburg, Germany) was a German Protestant theologian and philosopher. He is best known for his work on the New Testament and its mythology, which he published in 1921.

Table 17: Example comparing the Reference summary with the generated ones by the baseline (Vanilla) and the faithfulness aware approaches (Clean, Mask, and Unlike$_{PR}$) from the de-en test set.