Estimation of the Factual Correctness of Summaries of a Ukrainian-language Silver Standard Corpus

Oleksandr Bauzha [1], Artem Kramov [2] and Oleksandr Yavorskyi [2]

[1] *Taras Shevchenko National University of Kyiv, Volodymyrska Street 64/13, 01601, Kyiv, Ukraine*
[2] *Seraf AI LLC, PO Box 3978, Lisle, Illinois 60532, United States*

Abstract

In this paper, different metrics for estimating the factual correctness of summaries of a Ukrainian-language silver standard summarization corpus have been analyzed. The different state-of-the-art methods of detecting the factually inconsistent document-summary pairs have been considered first; moreover, the types of errors in current summarization datasets have been analyzed too. It has been shown that suggested metrics can be used for the discrimination of correct/incorrect document-summary pairs that may be useful for the automatic generation of a summarization corpus. The results obtained for the ground-truth samples may indicate the availability of many erroneous summaries: more than 50% of the test subset can contain factually inconsistent samples. Further analysis of the factual correctness of model-generated summaries showed better factual consistency between documents and summaries than the ground-truth summaries. However, due to the availability of noisy ground-truth samples, the generated summaries can still contain hallucinated information; applying the suggested metrics may allow filtering out erroneous samples, which should also increase the summarization model's performance.

Keywords [1]

Natural language processing, factual correctness, abstractive summarization, low-resource languages, multilingual models

1. Introduction

Abstractive text summarization falls into the category of sequence-to-sequence natural language processing (NLP) tasks. The development of the self-supervised methods of the training of language models on large corpora [1, 2] with further fine-tuning of the corresponding model on the summarization dataset allows for achieving remarkable success in the domains of the abstractive summarization of articles [3, 4] and dialogues [5, 6]. However, the aforementioned advances in the abstractive text summarization task are mostly connected with the analysis of high-resource languages (English, Chinese, etc.). Unfortunately, the research on the abstractive summarization of Ukrainian documents is still in the initial stage. Similarly to other NLP issues that are presented for the low-resource languages, *the lack of human-written datasets* remains a key problem for the investigation of the summarization of Ukrainian corpora [7, 8]: while the summarization models themselves can potentially be created by the projection of the corresponding English models into the Ukrainian-language space (e.g., the Ukrainian GPT-2 model has been recently created according to the paper [9]), the verification of the quality of the summaries that are generated by the produced models remains a challenging task. One of the possible solutions for the generation of a summarization dataset consists in the web-scrapping of news portals [10, 11]. Namely, the well-known XSum dataset [12] was created by treating the headline of a news article as the corresponding summary. However, such an approach cannot be reliable as far as the headlines can contain extra information that is not presented in the article for the attention attraction of a reader.

Information Technology and Implementation (IT&I-2023), November 20-21, 2023, Kyiv, Ukraine
EMAIL: asbauzha@gmail.com (O. Bauzha); artemkramov@gmail.com (A. Kramov); yaotianjiu@gmail.com (O. Yavorskyi)
ORCID: 0000-0002-4920-0631 (O. Bauzha); 0000-0003- 3631-1268 (A. Kramov); 0009-0001-5175-3825 (O. Yavorskyi)

To overcome this problem, the authors of the paper [13] suggested extracting the summary of the article from the short description of the article of a BBC news portal resulting in **a multilingual silver standard XL-Sum summarization dataset**. The statistics of the Ukrainian-language subset (article-summary pairs) in the XL-Sum dataset are presented in Table 1. Moreover, the corresponding Ukrainian-language summarization model was trained as well.

Table 1

Number of samples for the Ukrainian-language part of the XL-Sum dataset according to the performed train-dev-test split

Train	Dev	Test	Total
43201	5399	5399	53999

The aforementioned automatic generation of the document-summary pairs requires the answer to the following question: ***how to verify the quality of the collected summaries automatically?*** While the coherency and fluency of summaries should be preserved (texts were written by editors), the **factual consistency** between the document and the summary should be estimated. The authors of the XL-Sum dataset conducted the human evaluation of the summaries of 10 languages from a small subset (around 250 article-summary pairs). According to the results [13], up to 42% of the selected summaries contained extra information. The availability of such factual errors complicates the usage of the dataset for the verification of the quality of any summarization model; moreover, the training of the model on such samples can lead to the generation of hallucinated summaries by the last one. Thus, the detection of factual errors in summaries is a relevant problem for the analysis of the automatically generated dataset and the estimation of the performance of the summarization model.

In this paper, the factual consistency metric for a Ukrainian-language document-summary pair is suggested. Namely, different cross-lingual approaches that can be applied to a wide range of languages are considered with the following analysis of their effectiveness. Moreover, the factual correctness of the Ukrainian-language summaries of the XL-Sum dataset is considered due to the retrieved metrics. In addition, the performance of the already trained Ukrainian summarization model in terms of the factual consistency of generated summaries is analyzed as well.

Before the creation of the metric for the estimation of the factual consistency of a document-summary pair, it was decided to consider existing approaches and current issues within this subject area. The next section is devoted to the analysis of the different state-of-the-art methods of the detection and correction of factual mistakes in a summary given an input document.

2. Related work

One of the key concepts in the factual consistency analysis consists in the generation of the corresponding dataset that defines the types of factual errors that are presented in an erroneous summary. According to the paper [14], two approaches for dataset generation are mostly used: *entity-centric approach (Ent-C)* and *generation-centric approach (Gen-C)*.

The Ent-C approach implies the transformation of a ground-truth summary into an erroneous summary by applying different modification operations on its entities and noun phrases: entity swap, pronoun swap, negation, etc. The corresponding dataset (K2019) was first presented in the paper [15] and was later used as a baseline for other methods. The ground-truth samples were taken from the CNN/DM dataset. The authors of the dataset also presented a FactCC method for detecting factual errors in a summary. The main idea consists in the fine-tuning of the uncased BERT model [16] with the further binary classification of a document-summary pair (consistent/inconsistent) on the training dataset. As was shown, the FastCC method outperformed the MNLI-based approach [17] that consisted in the interpretation of an entailment measure between a document and a summary as a factual consistency metric. In the paper [18], it was suggested to fine-tune the sequence-to-sequence BART model [4] to generate the corrected version of a summary. Namely, a document and an inconsistent summary were concatenated and passed to the input of an encoder; the entire model was trained to generate the corrected consistent summary. The authors of the paper [19] proposed to mask each entity of a summary with the further usage of the BERT model (BertForQuestionAnswering architecture) for the prediction of answer spans in a source document. In contrast to this paper, the method QAGS [20]

consists in generating questions to the entities of a summary automatically; then the question-answering model is used to find answers in both a source document and a summary to verify their match.

Unlike the Ent-C approach, **the Gen-C approach** [21] consists in the transformation of a ground-truth summary by applying the paraphrasing model. The following assumption is made: the bottom-placed candidates of the beam search (e.g., the 10th best paraphrase) potentially contain error facts. In contrast to the Ent-C-related methods, the authors [21] considered the factual consistency problem at the level of dependency arcs retrieved from a syntactic parser: the dependency arc (fact) is entailed by a source document if a semantic relation between the corresponding head and the child word is also entailed by the document. Elaborating on this assumption, the Dependency Arc Entailment (DAE) model was designed and trained to estimate the entitlement of dependency arcs by a source document. In order to extend the consideration of the dependency arcs as a representation of facts in a more general way, the FactGraph method [22] was recently proposed. The main idea of the FactGraph method consists in decomposing the document and the summary into structured meaning representations. Such meaning representations define semantic concepts and their relations by generating a semantic graph for both a document and a summary. Following the idea of the entailment of dependency arcs, the factual consistency was calculated based on the probability of establishing edges between the semantic concepts of a summary.

As mentioned in the papers [15, 18], the NLI-based models showed worse results than their counterparts. However, in the paper [23], the usage of the NLI models was reconsidered by presenting a SummaC method. Namely, while the previous attempts were focused on estimating the entailment of a document and a summary entirely, the SummaC method is based on the consideration of their factual consistency at the level of sentences. The SummaC method outperformed FastCC, DAE, and QA-based methods, thus, confirming the ability of the usage of the NLI models for the estimation of the factual correctness of summaries. In parallel with our work, the factual consistency evaluation method for multilingual corpora based on the usage of the NLI model was recently suggested [24]. The NLI model was created by fine-tuning the mT5-XXL model [25] for the binary classification of a document-summary pair: the input data are represented as the concatenation of a document and a summary; the output binary value indicates whether the given pair is consistent or not. This classification model was later used for filtering inconsistent samples in the XL-Sum dataset and re-training models. Such an approach allowed for better results in ROUGE scores and human scores (the Ukrainian language was not considered during those experiments). However, according to the conclusion of annotators, only 52% of retrieved summaries (or even more for some languages) were factually consistent with documents. Moreover, the estimation of the entailment of a document-summary pair entirely can contradict recent results shown by the sentence-level SummaC method [23]. We assume that the consideration of the entailment of a document and a summary at the level of sentences may be crucial for the XL-Sum dataset: collected summaries can potentially contain additional information (references, full names, positions, etc.) that may be revealed by increasing the granularity of the analysis of the document parts.

Finally, before applying the aforementioned methods or creating a new one, the following question should be answered: *which types of factual errors are most expected in the XL-Sum dataset?* In order to get insights, the corresponding statistics for the XSum dataset [12] that was also generated automatically can be considered. In the paper [14], the authors conducted an error analysis of the summaries of the XSum. Namely, the errors were classified into four main categories:

- Entity-related (conflating two different entities, hallucinated entities).
- Event-related (incorrect event description, agents, new event).
- Noun phrase related (incorrect NP or NP modifiers, new NP, etc.).
- Others (grammar, noise).

In addition, each category was divided into 2 subcategories: extrinsic (hallucination) and intrinsic (incorrect data interpretation) errors. According to the results [14], most of the errors are actually connected with the appearance of extrinsic errors of all categories. The ratio of intrinsic entity-related errors which are typical for the aforementioned K2019 dataset is relatively small. Thus, it was decided to rely on NLI-based approaches that can be useful for detecting relevant types of errors. The next section describes the corresponding selected methods and metrics.

3. Factual consistency estimation metrics

According to the previous section, the usage of the NLI-based metrics seems useful for analyzing the different types of errors. Taking into account the findings of the SummaC zero-shot method [23], it was decided to process document-summary pairs at the level of sentences. Namely, given a pair of a document and a summary (doc,summary), let us represent both of them (D and S correspondingly) as a list of sentences:

$$D = \left\{ s_1^{doc}, s_2^{doc}, ..., s_M^{doc} \right\}$$
$$S = \left\{ s_1^{sum}, s_2^{sum}, ..., s_N^{sum} \right\} \quad (1)$$

The next step consists in the calculation of the entailment matrix $Ent_{N \times M}$. Each element of the matrix Ent_{ij} corresponds to the consistency score of a document sentence s_i^{doc} and a summary sentence s_j^{sum}:

$$Ent_{ij} = \text{Score}(s_i^{doc}, s_j^{sum}) \quad (2)$$

In order to calculate the $\text{Score}(s_i^{doc}, s_j^{sum})$ measure, it was decided to consider two different approaches:

- the semantic similarity between the sentences (Score^{Emb} notation);
- the probability value of the entailment of sentences (Score^{Ent} notation).

The Score^{Emb} value is calculated as the cosine similarity between the sentences embedding vectors:

$$\text{Score}^{Emb}(s_i^{doc}, s_j^{sum}) = \frac{\mathbf{s}_i^{doc} \cdot \mathbf{s}_j^{sum}}{\left\| \mathbf{s}_i^{doc} \right\| \left\| \mathbf{s}_j^{sum} \right\|}, \quad (3)$$

where $\mathbf{s}_i^{doc}$ and $\mathbf{s}_j^{sum}$ - embedding vectors of sentences s_i^{doc} and s_j^{sum}. Such vector representation can be retrieved from different embedding models; the choice of the corresponding models is described in the next section. The Score^{Ent} measure is represented as the probability value of the entailment label "e" among the possible NLI categories $\{\text{e,c,n}\}$:

$$\text{Score}^{Ent}(s_i^{doc}, s_j^{sum}) = P(\text{e} \mid s_i^{doc}, s_j^{sum}), \quad (4)$$

where NLI categories $\{\text{e,c,n}\}$ correspond to the entailment, contradiction, and neutral labels.

After the generation of the matrix Ent_{ij}, it is reduced to the vector representation by taking a maximum value across all columns:

$$\mathbf{EntRed} = \max(Ent, axis = col) \quad (5)$$

In other words, the retrieved vector **EntRed** contains information about the best consistency score for each summary sentence. Then an output factual consistency score $FactCons^{Ent}$ is calculated as the mean value of the vector **EntRed**:

$$FactCons^{Ent} = \text{mean}(\mathbf{EntRed}) \quad (6)$$

The aggregation of the consistency scores for summary sentences as an average value allows reducing the $FactCons^{Ent}$ in cases when some summary sentences are not consistent with any of the document sentences. Taking into account the potential big ratio of hallucinated summaries in the XL-Sum dataset, such an approach may help to reveal erroneous samples. Figure 1 demonstrates an example of the detection of a factually inconsistent hallucinated sentence.

The summary sentence (s2) which describes the source of information in a news article is not consistent with any of the document sentences; thus, its maximum consistency value is low. The availability of such consistency outlier decreases the final factual consistency score $FactCons^{Ent}$.

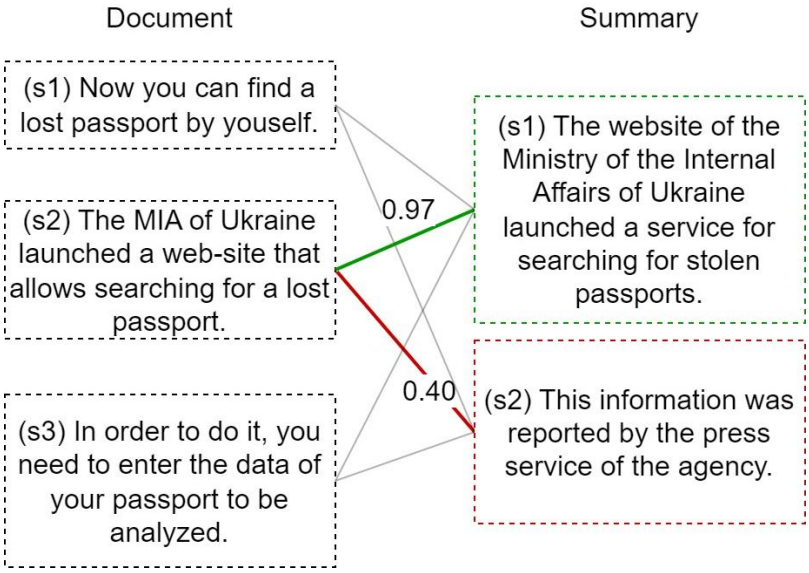


Figure 1: Detection of a factually inconsistent summary sentence. The edge values indicate the maximum consistency score for each summary sentence. As far as the summary sentence (s2) is not entailed with any of the document sentences, its consistency score is lower.

4. Experimental part

4.1. Inconsistent summaries discrimination

Before the calculation and analysis of the values of metrics for the Ukrainian part of the XL-Sum dataset, it was decided to verify the ability of the different methods to discriminate between factually consistent and inconsistent summaries. This inconsistent summaries discrimination task consists in the following: given two document-summary pairs with a common document where one pair contains a correct summary, and another one contains an incorrect one, it is necessary to predict which pair contains a factually consistent summary. The accuracy is calculated as the ratio of correctly processed pairs to a general number of them.

4.1.1. Dataset

The test part of the Ukrainian-language XL-Sum dataset was analyzed. In order to generate a factually inconsistent sample for each document, the following rules were applied:

- an inconsistent summary should belong to another document;
- the ROUGE-1 F1 measure between the document and inconsistent summary should be higher than the corresponding value between the document and the consistent summary.

The aforementioned rules allowed picking inconsistent summaries that can relate to the same topic as a document, but contain other information to make the discrimination task more challenging. Half of the test dataset was analyzed resulting in 1619 data points. The statistics of the dataset are available in Table 2. The Stanza package [26] was used for the tokenization; the stemming process was performed with the usage of the Ukrainian Stemmer library [27].

4.1.2. Metrics configurations

According to the previous section, it was suggested to use the NLI-based metric (SummaC). SummaC metric ($SummaC^{Emb}$) was calculated with the usage of sentence embedding models that are mentioned below:

- **paraphrase-multilingual-mpnet-base-v2** [28] - multilingual sentence embedding model based on the MPNet [29] model;

- **distiluse-base-multilingual-cased-v2** [28] - multilingual knowledge distilled version of multilingual Universal Sentence Encoder [30].

SummaC metric ($SummaC^{Ent}$) was implemented based on the usage of the NLI model **xlm-roberta-large-xnli** - XLM-RoBERTa model [31] fine-tuned on the multilingual XNLI dataset [32].

All pre-trained models were taken from the Huggingface repository [33]. It was decided to use the chosen multilingual models for the SummaC-based metric as far as they were pre-trained on Ukrainian parallel data as well.

Table 2
Statistics of the generated dataset for the inconsistent summaries discrimination task: a number of samples, an average number of sentences per a document, an average number of sentences per summary

Samples number	Doc sentences	Summary sentences
1619	24.40 ± 17.92	1.43 ± 0.65

4.1.3. Results

Table 3 shows the results of solving the inconsistent summaries discrimination task using different metrics. Except for the accuracy of the discrimination of incorrect/correct samples, the Pearson correlation coefficient (PCC) between metrics and the ROUGE-1 score is also provided. As can be seen, the $SummaC^{Emb}$ metric showed the best accuracy results. The usage of the $SummaC^{Emb}$ metric based on the model **paraphrase-multilingual-mpnet-base-v2** (the best option due to accuracy results) may be useful especially for the automatic construction of a summarization dataset when it is necessary to map a document with a potential summary. For instance, the BookSum [34] summarization dataset (namely, its chapter-level subset) was constructed by the mapping of the chapter of a book with sentences of a summary that relates to an entire book; we assume that the analyzed metrics can be used for the construction of a similar Ukrainian or even multilingual dataset as well.

Let us consider the Pearson correlation coefficient values between metrics and ROUGE-1 scores. As far as a higher ROUGE-1 score should imply the lower value of a metric (incorrect summaries have higher ROUGE-scores than correct ones), the PPC value should be low. As can be seen, the lowest (and a negative) PPC value was retrieved for the $SummaC^{Ent}$ metric indicating the possibility of the usage of the metric for the detection of the factually inconsistent summaries by setting up some threshold value. Thus, this metric was later used to analyze the Ukrainian-language part of the XL-Sum dataset and the summarization itself.

Table 3
Results of solving the inconsistent summaries discrimination task using different metrics: accuracy of the discrimination of correct/incorrect summaries and the Pearson correlation coefficient (PCC) between the metrics and the ROUGE-1 score of samples

Metric	Model	Accuracy, %	PCC
$SummaC^{Emb}$	**paraphrase-multilingual-mpnet-base-v2**	85.855	0.168
	distiluse-base-multilingual-cased-v2	81.655	0.273
$SummaC^{Ent}$	**xlm-roberta-large-xnli**	75.664	-0.075

4.2. XL-Sum dataset analysis

Firstly, let us analyze the Ukrainian test part of the dataset. The value of the $SummaC^{Ent}$ metric across the dataset was calculated. The density of the distribution of the retrieved metric value is shown in Figure 2. As can be seen, the distribution is skewed, and the 50th percentile equals 0.845. Thus, referring to the paper [24] where the threshold value 0.5 for the NLI model allowed filtering almost a half of incorrect samples (but approximately 50% of left summaries were judged by human evaluation as factually inconsistent), it can be concluded that a higher threshold value for the $SummaC^{Ent}$ has to

be taken as well. Indeed, the probability mass peak that starts from the 70th percentile value can potentially indicate the threshold for filtering incorrect summaries; however, this hypothesis should be later verified by an appropriate human evaluation.

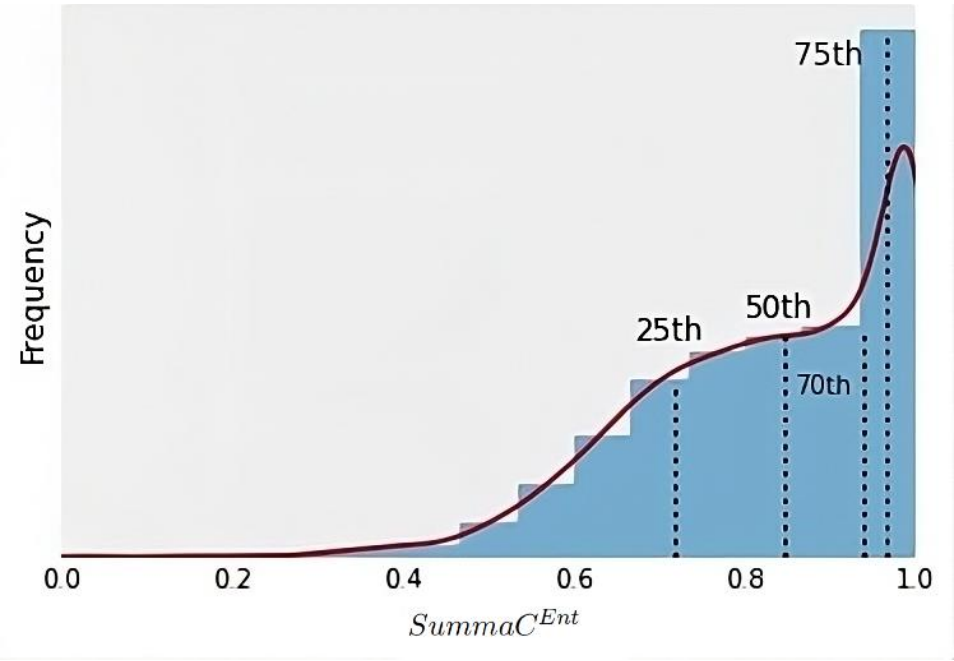


Figure 2: Density of the distribution of the $SummaC^{Ent}$ metric value across the Ukrainian test dataset for ground-truth summaries. The red line represents the kernel density estimation; dashed lines describe the 25th, 50th, 70th, and 75th percentiles

It should be mentioned that the main advantage of considering factual correctness at the level of document and summary sentences could be treated as a disadvantage because if the summary contains a high level of abstractiveness then the output $SummaC^{Ent}$ can be pretty low. In order to verify *if all low values of the metric correspond to error cases*, the future steps of the research are to involve an appropriate human evaluation of the factual correctness of summaries with the further estimation of the correlation of the suggested metric with human-labeled data.

4.3. Summarization model analysis

As the test dataset may contain many erroneous samples, it is hard to rely on the estimated ROUGE metrics. Thus, it was decided to calculate the $SummaC^{Ent}$ metric for the summaries generated by the summarization model on the test dataset. The summaries were picked from the set provided by the authors [13]. Figure 3 shows the retrieved distribution. As can be seen, the distribution of $SummaC^{Ent}$ scores is skewed too.

In order to compare the results between ground-truth and model-generated summaries, it was decided to take a median value as an average score, and the interquartile range (IQR) value for the measurement of the deviation of the metric. Table 4 demonstrates the retrieved results. The median value of the metric for the model-predicted summaries is higher; moreover, its IQR value is lower. Thus, *the summaries that were generated by the model are considered to be even more factually correct than the ground-truth summaries*.

As can be seen from Figure 3, there are some document-summary pairs whose $SummaC^{Ent}$ value is close to zero. Moreover, as can be expected from the noisy hallucinated dataset, the summarization model learned some pattern relations available in the dataset (e.g., the positions of persons) that led to the generation of hallucinated content (see Figure 4 and Figure 5 for such examples that were revealed by the low values of the metric). The removal of such dataset samples by the suggested metric can allow

for avoiding such a situation and provide a more robust summarization model in terms of its ability to generalize the knowledge of a source document.

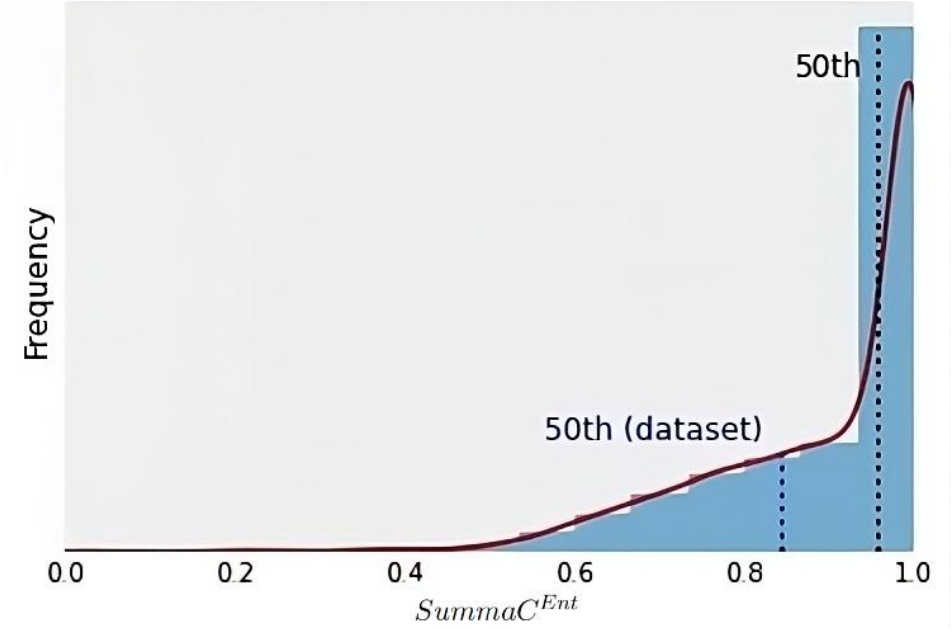


Figure 3: Density of the distribution of the $SummaC^{Ent}$ metric value across the Ukrainian test dataset for summaries generated by the model. The red line represents the kernel density estimation; the 50th (dataset) dashed line means the 50th percentile value for the ground-truth summaries

Table 4

Statistics of the $SummaC^{Ent}$ metric for ground-truth and model-predicted summaries

Summaries	Median	IQR
Ground-truth	0.848	0.248
Model-predicted	0.958	0.186

Document

This became known after Thursday's meeting of the finance ministers of these countries did not bring a breakthrough in solving this problem. The head of the Eurogroup, Dutch Finance Minister Jeroen Dijsselbloem, said that "too little" progress had been made and "no agreement is currently in sight." Greece has less than two weeks to reach an agreement with its creditors or face the threat of default. Mr. Dijsselbloem emphasized that Greece "has very little time left." Earlier, the managing director of the IMF, Christine Lagarde, said that the fund would not give Greece any more debt payments.

Summary

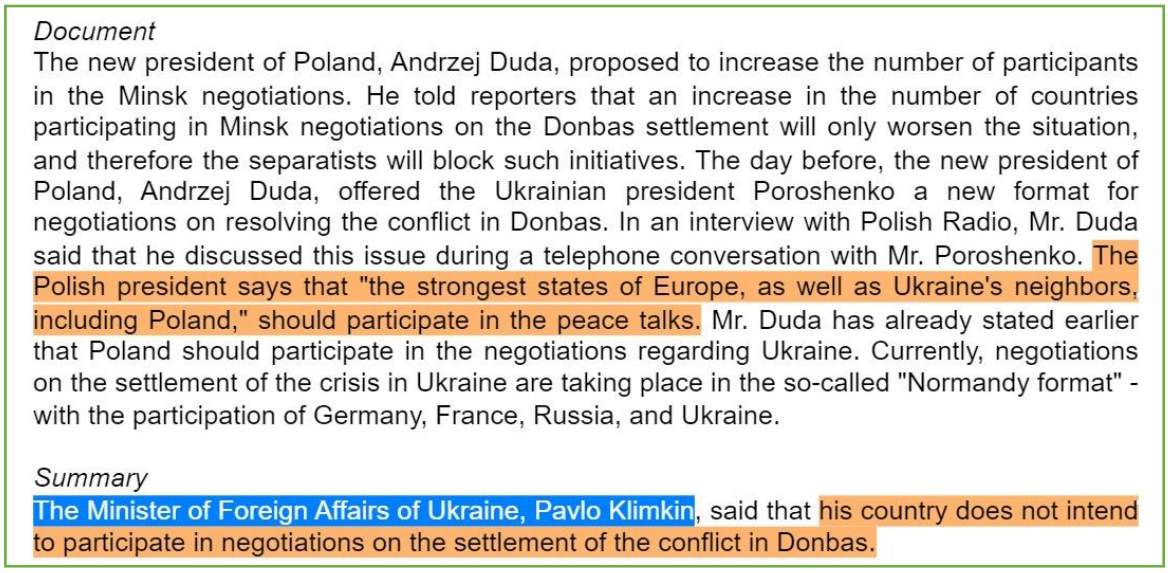
The finance ministers of the Eurogroup member states announced that they had reached an agreement with Greece regarding its debt payments.

Figure 4: A summary is inconsistent with a document in terms of events: an entire summary statement contradicts the facts from a document (both are highlighted in orange color)

5. Conclusions

In this paper, several metrics for estimating the factual consistency of documents and summaries were analyzed for processing the Ukrainian-language part of the XL-Sum corpus. Moreover, the experimental verification of the effectiveness of the chosen SummaC metric was performed on the Ukrainian-language part of the XL-Sum corpus using different configurations and models. According to the results obtained from the evaluation of the discrimination of factually correct/incorrect document-

summary pairs, the best accuracy was achieved with the usage of the multilingual sentence embedding model. Such a result may indicate the advisability of the utilizing of the aforementioned model for related tasks as the automatic construction of document-summary pairs for the generation of a silver standard Ukrainian summarization corpus. Moreover, the configuration of the metric SummaC with an NLI model showed the lowest expected correlation with a ROUGE score that can underline the possibility of the usage of this model for further detailed analysis of factual mistakes.


Document

The new president of Poland, Andrzej Duda, proposed to increase the number of participants in the Minsk negotiations. He told reporters that an increase in the number of countries participating in Minsk negotiations on the Donbas settlement will only worsen the situation, and therefore the separatists will block such initiatives. The day before, the new president of Poland, Andrzej Duda, offered the Ukrainian president Poroshenko a new format for negotiations on resolving the conflict in Donbas. In an interview with Polish Radio, Mr. Duda said that he discussed this issue during a telephone conversation with Mr. Poroshenko. The Polish president says that "the strongest states of Europe, as well as Ukraine's neighbors, including Poland," should participate in the peace talks. Mr. Duda has already stated earlier that Poland should participate in the negotiations regarding Ukraine. Currently, negotiations on the settlement of the crisis in Ukraine are taking place in the so-called "Normandy format" - with the participation of Germany, France, Russia, and Ukraine.

Summary

The Minister of Foreign Affairs of Ukraine, Pavlo Klimkin, said that his country does not intend to participate in negotiations on the settlement of the conflict in Donbas.

Figure 5: A summary contains two types of errors: hallucinated entity (person name and his position) that is marked in a blue color, and the contradiction of facts (the document states that the person suggests participating in a negation process, but the summary states an opposite fact)

The analysis of the values of the chosen NLI-based metric for the ground-truth samples of the XL-Sum dataset may indicate the availability of at least 50% of erroneous summaries that match the results of the previous research. Moreover, the retrieved distribution of metric values may indicate the presence of even more than 70% of error samples; however, the search for an appropriate threshold value for the considered metric still requires the usage of a more general human evaluation.

Finally, it was shown that the metrics retrieved from evaluating the factual consistency of model-generated summaries are higher than those of ground-truth summaries. Nevertheless, the availability of generated summaries with an almost zero metric score may indicate the big impact of the hallucinated dataset on the trained model. Further filtering of erroneous samples from the dataset using the considered metrics may allow learning the model to generate more factually consistent summaries.

6. References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: http://arxiv.org/abs/1907.11692. doi:10.48550/arXiv.1907.11692.

[2] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving pre-training by representing and predicting spans, Transactions of the Association for Computational Linguistics 8 (2020) 64–77. URL: https://aclanthology.org/2020.tacl-1.5. doi:10.1162/tacl_a_00300.

[3] J. Zhang, Y. Zhao, M. Saleh, P. J. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019. arXiv:1912.08777.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, 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, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[5] X. Feng, X. Feng, B. Qin, X. Geng, Dialogue discourse-aware graph model and data augmentation for meeting summarization, in: International Joint Conference on Artificial Intelligence, 2020.

[6] C.-S. Wu, L. Liu, W. Liu, P. Stenetorp, C. Xiong, Controllable abstractive dialogue summarization with sketch supervision, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 5108–5122. URL: https://aclanthology.org/2021.findings-acl.454. doi:10.18653/v1/2021.findings-acl.454.
[7] A. Kramov, S. Pogorilyy, Evaluation of the coherence of Ukrainian texts using a transformer architecture, in: 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), 2020, pp. 296–301. doi:10.1109/ATIT50783.2020.9349355.
[8] S. Pogorilyy, A. Kramov, Coreference resolution method using a convolutional neural network, in: 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 2019, pp. 397–401. doi:10.1109/ATIT49449.2019.9030596.
[9] B. Minixhofer, F. Paischer, N. Rekabsaz, WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3992–4006. URL: https://aclanthology.org/2022.naacl-main.293. doi:10.18653/v1/2022.naacl-main.293.
[10] S. Pogorilyy, A. Kramov, Automated extraction of structured information from a variety of web pages, PROBLEMS IN PROGRAMMING (2018) 149–158. URL: https://pp.isofts.kiev.ua/ojs1/article/view/277. doi:10.15407/pp2018.02.149.
[11] A. Anisimov, S. Pogorilyy, D. Vitel, About the issue of algorithms formalized design for parallel computer architectures, Applied and computational mathematics 12 (2013) 140–151.
[12] S. Narayan, S. B. Cohen, M. Lapata, 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, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1797–1807. URL: https://aclanthology.org/D18-1206. doi:10.18653/v1/D18-1206.
[13] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, R. Shahriyar, XL-sum: Large-scale multilingual abstractive summarization for 44 languages, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 4693–4703. URL: https://aclanthology.org/2021.findings-acl.413. doi:10.18653/v1/2021.findings-acl.413.
[14] T. Goyal, G. Durrett, 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, Association for Computational Linguistics, Online, 2021, pp. 1449–1462. URL: https://aclanthology.org/2021.naacl-main.114. doi:10.18653/v1/2021.naacl-main.114.
[15] W. Kryscinski, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9332–9346. URL: https://aclanthology.org/2020.emnlp-main.750. doi:10.18653/v1/2020.emnlp-main.750.
[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[17] A. Williams, N. Nangia, S. Bowman, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. URL: https://aclanthology.org/N18-1101. doi:10.18653/v1/N18-1101.
[18] M. Cao, Y. Dong, J. Wu, J. C. K. Cheung, Factual error correction for abstractive summarization models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6251–6258. URL: https://aclanthology.org/2020.emnlp-main.506. doi:10.18653/v1/2020.emnlp-main.506.
[19] Y. Dong, S. Wang, Z. Gan, Y. Cheng, J. C. K. Cheung, J. Liu, Multi-fact correction in abstractive text summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9320–9331. URL: https://aclanthology.org/2020.emnlp-main.749. doi:10.18653/v1/2020.emnlp-main.749.
[20] A. Wang, K. Cho, M. Lewis, Asking and answering questions to evaluate the factual consistency of summaries, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,

Association for Computational Linguistics, Online, 2020, pp. 5008–5020. URL: https://aclanthology.org/2020.acl-main.450. doi:10.18653/v1/2020.acl-main.450.
[21] T. Goyal, G. Durrett, Evaluating factuality in generation with dependency-level entailment, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 3592–3603. URL: https://aclanthology.org/2020.findings-emnlp.322. doi:10.18653/v1/2020.findings-emnlp.322.
[22] L. F. R. Ribeiro, M. Liu, I. Gurevych, M. Dreyer, M. Bansal, FactGraph: Evaluating factuality in summarization with semantic graph representations, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3238–3253. URL: https://aclanthology.org/2022.naacl-main.236. doi:10.18653/v1/2022.naacl-main.236.
[23] P. Laban, T. Schnabel, P. N. Bennett, M. A. Hearst, SummaC: Re-visiting NLI-based models for inconsistency detection in summarization, Transactions of the Association for Computational Linguistics 10 (2022) 163–177. URL: https://aclanthology.org/2022.tacl-1.10. doi:10.1162/tacl_a_00453.
[24] R. Aharoni, S. Narayan, J. Maynez, J. Herzig, E. Clark, M. Lapata, mface: Multilingual summarization with factual consistency evaluation, 2022. URL: https://arxiv.org/abs/2212.10622. doi:10.48550/ARXIV.2212.10622.
[25] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.
[26] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 101–108. URL: https://aclanthology.org/2020.acl-demos.14. doi:10.18653/v1/2020.acl-demos.14.
[27] Vladislav Klim, Ukrainian stemmer, 2019. URL: https://github.com/Desklop/Uk_Stemmer.
[28] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERTnetworks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.
[29] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 16857–16867. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.
[30] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic retrieval, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 87–94. URL: https://aclanthology.org/2020.acl-demos.12. doi:10.18653/v1/2020.acl-demos.12.
[31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.
[32] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, V. Stoyanov, XNLI: Evaluating cross-lingual sentence representations, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2475–2485. URL: https://aclanthology.org/D18-1269. doi:10.18653/v1/D18-1269.
[33] Clément Delangue, Hugging face, 2023. URL: https://huggingface.co/models.
[34] W. Kryscinski, N. Rajani, D. Agarwal, C. Xiong, D. Radev, Booksum: A collection of datasets for long-form narrative summarization, 2021. URL: https://arxiv.org/abs/2105.08209. doi:10.48550/ARXIV.2105.08209.