Plain Language Summarization of Clinical Trials

Polydoros Giannouris, Theodoros Myridis, Tatiana Passali, Grigorios Tsoumakas
School of Computer Science, Aristotle University of Thessaloniki
{polydoros,tmyridis,scpassali,greg}@csd.auth.gr

Abstract

Plain language summarization, or lay summarization, is an emerging natural language processing task, aiming to make scientific articles accessible to an audience of non-scientific backgrounds. The healthcare domain can greatly benefit from applications of automatic plain language summarization, as results that concern a large portion of the population are reported in large documents with complex terminology. However, existing corpora for this task are limited in scope, usually regarding conference or journal article abstracts. In this paper, we introduce the task of automated generation of plain language summaries for clinical trials, and construct CARES (Clinical Abstractive Result Extraction and Simplification), the first corresponding dataset. CARES consists of publicly available, human-written summaries of clinical trials conducted by Pfizer. Source text is identified from documents released throughout the life-cycle of the trial, and steps are taken to remove noise and select the appropriate sections. Experiments show that state-of-the-art models achieve satisfactory results in most evaluation metrics.

Keywords: plain language summarization, lay summarization, clinical trials, simplification, summarization

1. Introduction

Lay summarization, also known as plain language summarization, is the process of distilling intricate information into clear, concise and easily digestible summaries (Vinzelberg et al., 2023). In an era of abundant specialized knowledge and technical jargon, lay summarization plays a vital role in making complex ideas, scientific findings, or technical concepts accessible and comprehensible to individuals who may lack expertise in a particular field.

Lay summarization is particularly important in communicating scientific articles to the general public, especially in the field of medicine. This became apparent during the COVID-19 pandemic when millions of scientific articles with medical content were published (Islam et al., 2020). These articles were comprehensible to only a few, resulting in misinterpretations to the extent that it led to misinformation and fake news (Brennen et al., 2020).

The state-of-the-art in lay summarization is constantly evolving, driven by cutting-edge NLP approaches, as well as by the development of appropriate datasets for supervised learning (Chandrasekaran et al., 2020; Guo et al., 2021, 2024; Goldsack et al., 2022; Attal et al., 2023). Existing datasets concern scientific publications. Another important type of scientific information is clinical trials, which comprise several long documents, including a study protocol, a statistical analysis plan and a report synopsis, as well as related scientific publications. Lay summarization of clinical trials is not only important for the general public, but also a requirement for pharmaceutical companies by regulations such as EU Regulation No 536/2014 and US Public Health Service Act 2007. However, to the best of our knowledge, this task has not been considered by the NLP community before and no relevant public datasets exist.

This work takes some first steps to fill this gap, by introducing lay summarization of clinical trials as a task, and constructing a corresponding dataset, CARES (Clinical Abstractive Result Extraction and Simplification), which pairs publicly available plain language summaries (PLSs) with relevant pieces of text from the associated documents. Table 1 shows a sample of such a pair.

Source: Programmed death ligand 1 (PD-L1, also called B7-H1 or CD274) has a known role in the suppression of T-cell responses. The PD-1 receptor is expressed on activated CD4+ and CD8+ T cells. By interaction with its ligands, PD-L1 and PD-L2, PD-1 delivers a series of strong inhibitory signals to inhibit T-cell functions. Avelumab*(MSB0010718C), a fully human antibody of the immunoglobulin G1 (IgG1) isotype, specifically targets and blocks PD-L1, the ligand for PD-1 receptor. In preclinical studies, the combination of avelumab with chemotherapy (gemcitabine, oxaliplatin, 5FU) showed improved anti-tumor activity over single-agent chemotherapy ...
Summary: Avelumab is a medicine that may work by targeting a protein called programmed death-ligand 1 (pd-l1) found on the cancer cell. Pd-l1 is involved in the bodys immune system response to cancer. When this study was started, avelumab was being tested for use in women with advanced ovarian cancer. Although avelumab is approved in other types of cancer, it is not approved for use ...

Table 1: Sample of a source and summary pair from the CARES dataset.

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 60–67

The rest of the paper is structured as follows: Section 2 presents related work in this field. Section 3 discusses the developed dataset. Section 4 covers the experiments conducted on this dataset and finally, Sections 5 and 6 introduce the conclusions and limitations of the dataset, respectively.

2. Related Work

In this section, we delve into the existing research and methodologies regarding lay summarization, particularly focusing on datasets available for the task and methods employed for generating simplified summaries from scholarly documents.

2.1. Datasets

One of the first resources in the field of lay summarization was a corpus of 572 full-text papers accompanied by lay summaries, in a variety of domains, including archaeology, hematology, and engineering, which was made available by Elsevier in the context of the 1st Workshop on Scholarly Document Processing (Chandrasekaran et al., 2020).

In biomedicine, (Guo et al., 2021) developed a dataset pairing 7,805 systematic reviews from the Cochrane database with plain language abstracts written by domain experts. (Goldsack et al., 2022) introduced two datasets: the Public Library of Science (PLOS) and eLife, each containing biomedical articles along with PLSs written by experts, the first having over 27k examples. Recently, (Attal et al., 2023) presented PLABA a dataset containing 750 abstracts from PubMed from 75 different health-related topics and expert-created adaptations at the sentence level. Lastly, (Guo et al., 2024) describe CELLS, the largest dataset of over 62k examples of parallel scientific abstracts and the corresponding expert-authored lay language summaries.

The dataset developed for our study differs from prior efforts in that CARES is the first dataset tailored specifically to *clinical trials*, instead of scientific publications.

2.2. Methods

There have been several efforts to develop models and methods for lay summarization. Specifically, (Chaturvedi et al., 2020), in their attempt to tackle CL-LaySumm20, which requested the development of non-technical summaries from scholarly documents, introduced a two-step divide-and-conquer technique. This approach involves extracting sentences from plain sections of the inputs using an unsupervised network and then performing abstractive summarization and merging them.

Furthermore, during CL-LaySumm 2020 in SDP workshop at EMNLP 2020, (Kim, 2020) achieved the top performance on the task of generating simplified summaries for scientific papers. They employed the PEGASUS (Zhang et al., 2019) model for producing the initial lay summaries, which were improved by appending important sentences to the summary of which the number of words was under a certain threshold, using a Presumm (Liu and Lapata, 2019), a BERT-based (Devlin et al., 2018) extractive summarization model.

Lastly, (Shaib et al., 2023) utilized GPT-3 in the zero-shot setting to summarize and simplify articles describing trials. They also applied this approach to the summarization of meta-analyses involving multiple documents. Despite also working on the lay summarization of clinical trials, our approach differs in that we aim to reproduce a particular document and not provide a general lay summary.

Although there is existing literature on lay summarization tasks for scientific publications and articles, we are the first to apply the generation of simplified summaries to whole clinical trials.

3. CARES Dataset

Motivated from the crucial role of high-quality parallel corpora in developing biomedical simplification models (Ondov et al., 2022), we introduce CARES, the 1st dataset for plain language summarization of clinical trials[1]. Although summaries (referred to as targets or golden summaries hereafter) are readily available, there exists no single respective technical text (source) for the entire summary. In this section, we outline our methodology for creating the dataset, as well as the process of identifying the suitable document and subsection for each component of the PLS.

3.1. Target Extraction

We start the construction of CARES from the *Plain Language Study Results Summaries* repository of Pfizer[2]. We collected the PDF files of the 176 summaries that existed in this repository, up to March 3rd, 2023. Next, we extracted their text, making sure artifacts are not introduced in the form of page numbers or identifiers present in the margins.

We found that their length often exceeds 1,200 words, which surpasses the capacity of most state-of-the-art models such as BART and PEGASUS. To address this issue, we exploited their discourse structure, inspired by the divide-and-conquer paradigm in (Gidiotis and Tsoumakas, 2020). Authors follow a question-answer structure, aimed at addressing different aspects of the clinical

[1] https://github.com/PolydorosG/CARES

[2] https://www.pfizer.com/science/clinical-trials/plain-language-study-results-summaries

trial, from its conception to its results. The respective titles of these sections are as follows:

- Q1: “Why was this study done?"
- Q2: “What happened during $\{i\}$ study?", $i \in \{$the, this$\}$
- Q3: “What were the results of the study?"
- Q4: “What $\{i\}$ did $\{j\}$ have during the study?", $i \in \{$medical problems, side effects$\}$ $j \in \{$participants, patients, children, boys, volunteers, infants$\}$
- Q5: “Were there any serious medical problems?" $\vee$ "Did $\{i\}$ have any serious $\{j\}$?", $i \in \{$(study) participants, study infants$\}$, $j \in \{$side effects, medical problems$\}$

Table 2 shows the number of examples per section, along with their average word counts. It is evident that Q1 and Q5 appear consistently in each of the initial 176 summaries, in contrast to Q2-Q4. Missing data are attributed to two factors: a) the existence of studies with different objectives, leading to certain sections being deemed irrelevant to the specific analysis being conducted and therefore not being included by the summary authors, and b) the introduction of noise during the text extraction process, despite our measures to prevent this. In such cases, portions of the text become corrupted, rendering certain sections irrecoverable.

Section	Summaries	Average # of words
$Q1$	176	325
$Q2$	175	555
$Q3$	166	287
$Q4$	171	267
$Q5$	176	130
Total	864	-

Table 2: Section headers identified in the PLSs.

3.2. Source Selection

Every clinical trial has a set of documents that describe each of its parts, from its design to the analysis of the results. Clinical studies begin with a *study protocol*. This is a detailed description of the plan that explains the objective of the clinical trial, as well as how it will be conducted. The protocol is usually accompanied by a *statistical analysis plan*. At the same time, scientific journal articles may be published for certain studies, mainly of new drugs. Finally, after the end of the trial, a *clinical study report synopsis* is created, which analyzes the results as well as the events that occurred during the study. Many of these documents exceed 100 pages of text, rendering summarization impossible for models without selecting a small portion of each document. Next, we will describe the document selected for each section, with the exception of Q3 for which no appropriate section was found.

The content of Q1 is the most general of all. It usually includes research-independent elements, such as general information about the disease and results of previous studies. Concerning the trial itself, the questions that will be answered, as well as the motivation of the researchers are described. As these are determined in advance, the most appropriate document is the study protocol. When available, we keep the study protocol’s summary, often referred to as *synopsis*. Otherwise, we use its *introduction* section, as it was found to contain most of the necessary information.

Q2 concerns the design of the clinical study. It analyzes data on the population and the separation of patients into groups. Afterward, the strategy followed regarding the administration of the substance is mentioned, as well as the type of the study, such as whether the groups are randomly selected (randomization), whether a control group is included, or whether it is single-blind or double-blind. This information is located in the *study design* part of the study protocol. Since section titles are not consistent across study protocols, we use regular expressions to isolate the particular segment.

Finally, Q4 and Q5 both refer to the side effects and medical problems experienced by the trial participants. Their difference lies in the severity, as they are analyzed separately in Q5 if they were life-threatening, required medical attention, or caused permanent damage. Due to the thematic similarity of the two questions, the source of both is found in the safety results section of the clinical study report synopsis.

Despite an initial choice of both document and section within the selected document, we find that source lengths remain prohibitively large for neural models. A major reason for this was the introduction of noise during text extraction. Despite preprocessing steps, including cropping margins and automatically identifying and removing text from tabular data, errors in these steps may persist, introducing a large volume of artifacts. For this reason, we proceed to evaluate each sentence of the large initial source section with regard to its similarity to the target.

Let $\mathcal{Y} = \{Y_1, Y_2, ..., Y_n\}$ be a set of golden summaries, and $\mathcal{X} = \{X_1, X_2, ..., X_n\}$ be a set of the respective candidate, uncleaned sources. We tokenize each initial source into a sequence of m sentences, $X_i = [x_1, x_2, ..., x_m]$. We then quantify the similarity of each sentence with the summary using the ROUGE-L (Lin, 2004) recall score R_{LCS}:

$$R_{LCS}(x_i, Y_i) = \frac{LCS(x_i, Y_i)}{l}, \quad (1)$$

where l is the length of sentence x_i and $LCS(x_i, Y_i)$

is the length of the longest common subsequence between x_i and Y_i.

Although ROUGE was used in previous work to match sentences of the summary with parts of a document and automatically create source-target pairs for training (Gidiotis and Tsoumakas, 2020), no special care has been taken for entities. Given the simplified vocabulary of the summaries, we propose the addition of entity matches between candidate sources and targets as an anchor between complex and simplified text. In the context of factual consistency of summarization, Nan et al. (2021) proposed the use of named entity recall and precision. We adapt this concept in order to measure entity-level recall R_E. Specifically, we extend the proposed method to also include numbers and percentages. Since decimals are not included in PLSs, we identify such digits in the source. These numbers are then rounded to the nearest integer to best align the source and target formats. The final similarity score is thus defined as:

$$S(x_i, Y_i) = \alpha * R_{LCS}(x_i, Y_i) + (1 - \alpha) * R_E(r(NE_{x_i}), NE_{Y_i}) \quad (2)$$

where α is a hyper parameter, NE_{x_i}, NE_{Y_i} the named entities detected in source sentence x_i and PLS Y_i respectively, and $r(\cdot)$ the simplification function applied to source entities.

Finally, after evaluating each sentence's similarity to the respective target, we select sentences in descending order until scores reach a threshold or the maximum length is exceeded. Sentences are reordered before forming the dataset to best replicate the document's structure. In case of no appropriate source sentences (i.e. no sentences pass the fixed similarity threshold), the examples are removed completely. An example of the source extraction pipeline is provided in Figure A.1.

The final dataset consists of 478 source-target pairs. The final word counts, along with the number of examples in each split are presented in Table 3. Note that to best reproduce real-world settings, we make sure each summary's subsections are not present in different splits. Lastly, to aid future research, we publish both the selected sources as well as the entire source documents and targets.

	Length		Summaries		
Section	**Source**	**Target**	**Train**	**Val.**	**Test**
Q1	713	330	87	13	9
Q2	665	570	79	12	8
Q4	406	279	102	18	13
Q5	405	129	104	19	14

Table 3: Source and summary length after our similarity-based filtering method, along with number of examples in train, test and validation splits.

4. Experiments

To facilitate future work, we benchmark our dataset using state-of-the-art summarization models BART and PEGASUS. We run all experiments on an Nvidia Tesla T4 with 16 GB of memory, using the open-source Hugging-Face implementations (Wolf et al., 2019) for a maximum of 10 epochs. We monitor the models' performance on the validation set for each epoch and select the best model according to ROUGE-L score. All BART models were initialized from the "facebook/bart-large" model, and PEGASUS from "google/pegasus-large". Finally, entity recognition was performed using spaCy (Honnibal and Montani, 2017).

Since Q4 and Q5 of the same PLS may have the same source, we train models under two settings. The first consists of training separate models for each section, referred to as BART and PEGASUS, treating them as a distinct summarization task. For the second approach, in order to utilize the whole training set in a single model we prepend the section's title to each source employing special tokens to tag it, before feeding the document to the model (Passali and Tsoumakas, 2022). These models are referred to as BART$_{TAG}$ and PEGASUS$_{TAG}$.

We evaluate generated summaries using both ROUGE and named entity recall and precision, to evaluate entity-level factual consistency. The experimental results reported in Table 4 are highly promising, with the BART models outperforming PEGASUS on most sections. However, determining what constitutes a *good* ROUGE score can vary depending on the domain and the specific task at hand. In our investigation, we observed that the ROUGE scores of models trained on our dataset align with those reported in similar studies on analogous datasets. It is worth noting that while individual models exhibit superior performance compared to the tagging method, this enhanced performance is achieved at the expense of requiring four times as many models.

Regarding the somewhat subpar performance in Q2, we attribute it to the open-endedness of the question rather than our regular expressions. To further investigate this we calculate the ROUGE scores between the selected sources and golden summaries. As can be seen in Table 5, contrary to the model performance, our retrieval approach appears to be most successful in Q2. Therefore we ascribe the relatively bad performance, to Q2 being a harder section to simplify and summarize.

Finally, we notice that models trained on the entire dataset (BART$_{TAG}$ and PEGASUS$_{TAG}$), despite generally showcasing lower ROUGE scores, are able to more accurately generate entities. This observation is consistent with previous claims that ROUGE alone is inadequate to quantify factual con-

	Model	ROUGE – 1	ROUGE – 2	ROUGE – L	$\text{Precision}_{\text{NE}}$	$\text{Recall}_{\text{NE}}$
Q1	BART	**51.8**	**25.3**	**31.3**	46.79	35.34
	BART_{TAG}	50.7	24.0	30.6	**47.36**	**44.60**
	PEGASUS	34.0	11.7	22.7	35.51	42.34
	PEGASUS_{TAG}	35.0	10.9	22.4	43.46	41.73
Q2	BART	**47.1**	19.3	25.7	69.98	**44.13**
	BART_{TAG}	46.5	**19.5**	**26.8**	**70.05**	38.77
	PEGASUS	34.0	13.0	25.4	58.32	39.94
	PEGASUS_{TAG}	35.0	12.4	24.2	69.50	26.85
Q4	BART	**73.9**	**62.4**	67.0	53.90	48.74
	BART_{TAG}	70.7	58.6	63.8	**58.71**	**57.09**
	PEGASUS	66.2	57.0	60.8	54.86	46.85
	PEGASUS_{TAG}	69.7	61.1	**68.7**	60.61	48.64
Q5	BART	**62.3**	**47.4**	**53.7**	30.52	56.22
	BART_{TAG}	61.0	46.6	53.1	35.42	**58.60**
	PEGASUS	50.4	39.4	46.0	**39.05**	48.97
	PEGASUS_{TAG}	56.3	44.8	51.2	31.59	37.24

Table 4: ROUGE F1 and named entity results of BART and PEGASUS models on our dataset. We mark the best performances with bold. BART_{TAG} and PEGASUS_{TAG} are trained on the entire dataset.

Section	R-1	R-2	R-L
$Q1$	28.00	6.55	13.48
$Q2$	32.91	8.34	14.42
$Q4$	27.34	5.83	13.84
$Q5$	18.47	4.78	10.89

Table 5: ROUGE scores between selected source segments and golden summaries.

sistency (Kryściński et al., 2019b).

Following previous work on lay summarization (Guo et al., 2022), we report the average Coleman-Liau readability score (Coleman and Liau, 1975) for the source, gold summary and model-generated summary for BART_{TAG} in Table 6. This score evaluates the simplicity of a passage, by providing an estimate of the years of education required to understand it. A lower score suggests a simpler writing style. We confirm that PLSs offer greater readability than the respective source segments. We also find that the $BART_{TAG}$ consistently exhibits readability levels are consistently closer to the desired target, reflecting its effectiveness in producing simplified versions of the source.

Section	Source	Summary	Model Summary
$Q1$	14.5	11.6	11.1
$Q2$	12.0	10.2	11.0
$Q4$	13.5	11.7	12.1
$Q5$	13.2	12.0	12.6
Average	13.3	11.4	11.7

Table 6: Coleman-Liau readability scores for source, golden and BART_{TAG} summaries.

Despite impressive ROUGE scores, we note the factual inconsistency of generated summaries, which has previously been reported by several authors as a problem in abstractive summarization (Kryściński et al., 2019b; Cao et al., 2018; Kryściński et al., 2019a). Qualitative analysis shows that this problem can be largely attributed to three reasons: i) Missing information, where identified sources do not contain all necessary information to accurately produce summary entities, ii) Typos, where entities are "mistyped", due to the model's dictionary (e.g. letters missing from a substance's name), iii) Hallucinations, where entities are made up due to biases present in the training set (e.g. stating that a study was performed in the US rather than the UK). We present representative examples for some identified causes of factual inconsistencies in Table 7 of Appendix A.2.

5. Conclusion

This work introduced the task of automatic generation of lay summaries for *clinical trials* and constructed the first related dataset to support training and evaluation. To enable the use of transformer models for this task, we proposed the division of each golden summary into thematic subsections with appropriate length. Additionally, we located the source of each section from an array of documents and proposed similarity measures as a means of improving source quality. To facilitate future research, we benchmarked our dataset with popular summarization models using several metrics and found that BART performs well on all thematic sections. Finally, we noted challenges in the form of factual inconsistency of generated summaries, attributable to both model biases and source imperfections.

6. Limitations

Although CARES utilizes all publicly available PLSs by Pfizer, it remains smaller than datasets available for other summarization tasks. This is largely attributed to plain summaries being made mandatory in recent years. Another limitation of CARES is the inclusion of summaries by a single sponsor. Although the general format is similar between trials of different sponsors, we cannot guarantee models trained on CARES will generalize well across different sponsors.

7. Bibliographical References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation*. Ph.D. thesis, University of Oxford.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Rochana Chaturvedi, Saachi ., Jaspreet Singh Dhani, Anurag Joshi, Ankush Khanna, Neha Tomar, Swagata Duari, Alka Khurana, and Vasudha Bhatnagar. 2020. Divide and conquer: From complexity to simplicity for lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 344–355, Online. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *arXiv preprint arXiv:2211.03818*.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):160–168.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, S. M. Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621 – 1629.

Seungwon Kim. 2020. Using pre-trained transformer for better lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 328–335, Online. Association for Computational Linguistics.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

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. *arXiv preprint arXiv:2102.09130*.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Tatiana Passali and Grigorios Tsoumakas. 2022. Topic-aware evaluation and transformer methods for topic-controllable summarization.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Oliver Vinzelberg, Mark David Jenkins, Gordon Morison, David McMinn, and Zoe Tieges. 2023. Lay text summarisation using natural language processing: A narrative literature review.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

A. Appendix

A.1. Source extraction

Figure A.1 presents an example of the source selection pipeline. Initially, target sections are identified based on their titles. Subsequently, the most relevant document and subsection within it are extracted. Given that this text may contain irrelevant sentences or noise artifacts, the proposed similarity score, as defined in Equation 2, is utilized to assess the alignment between the candidate source and the target. Finally, we obtain the clean source segment by filtering out sentences that fail to reach a similarity threshold.

A.2. Hallucination examples

Table 7 contains examples for each of the identified types of model hallucinations. In the first example, the model incorrectly calculated that 5 out of 17 equates to 17%, which is inaccurate. The second example highlights a typographical error where the drug "palbociclib" was mistakenly spelled as "palbocciclib". Finally, in the third case, the model erroneously stated that a vaccine had been approved both in the United States and the European Union when, in reality, it was only approved in Europe. These errors demonstrate the importance of carefully assessing the outputs of NLP models, as they can sometimes produce inaccuracies or hallucinate information that differs from the factual reality.

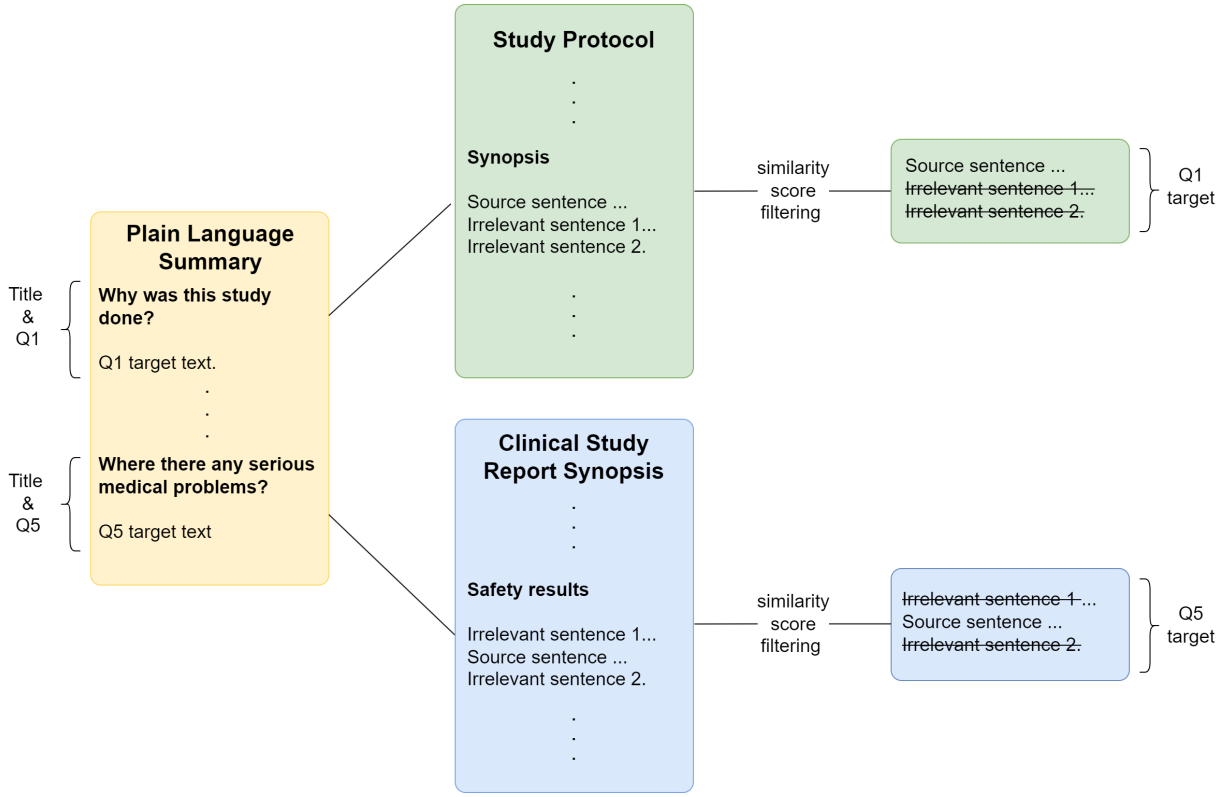

Figure 1: An example of source extraction.

Type	Example
Numerical Error	In this study, 5 out of 17 (17%) participants who received pregabalin 5 mg/kg/day had at least 1 medical problem ...
Typo	This study compared 2 groups of patients to find out if patients taking palbociclib in combination with letrozole had their cancer get better compared to patients taking a placebo ... The patients and researchers did not know who took palbocciclib...
Hallucinations	**Target summary** However, invasive meningococcal disease may be prevented with a vaccine. A vaccine is a type of medicine that helps people fight off germs. Meningococcal disease is caused by the meningococcus germ. There are different types of this germ. For example, meningococcal type a disease is caused by the meningococcus a germ. Menacwy-tt (nimenrix) is a vaccine approved in Europe for the prevention of meningococcal disease. **Model Generated Summary** However, invasive disease may be prevented with a vaccine. A vaccine is a type of medicine that helps people fight off germs. Menacwy-tt (nimenrix) is a vaccine approved in the United States, the US, and the European Union for the prevention of invasive disease.

Table 7: Error examples from BART$_{TAG}$ generated summaries. Model mistakes and hallucinations are marked in red, while the corresponding correct information is highlighted in blue.