LTRC-IIITH at PerAnsSumm 2025: SpanSense - Perspective-specific span identification and Summarization

Sushvin Marimuthu, Parameswari Krishnamurthy
LTRC, International Institute of Information Technology, Hyderabad, India
sushvin.marimuthu@research.iiit.ac.in
param.krishna@iiit.ac.in

Abstract

Healthcare community question-answering (CQA) forums have become popular for users seeking medical advice, offering answers that range from personal experiences to factual information. Traditionally, CQA summarization relies on the best-voted answer as a reference summary. However, this approach overlooks the diverse perspectives across multiple responses. Structuring summaries by perspective could better meet users' informational needs. The PerAnsSumm shared task addresses this by identifying and classifying perspective-specific spans (Task_A) and generating perspective-specific summaries from question-answer threads (Task_B). In this paper, we present our work on the PerAnsSumm shared task 2025 at the CL4Health Workshop, NAACL 2025. Our system leverages the RoBERTa-large model for identifying perspective-specific spans and the BART-large model for summarization. We achieved a Macro-F1 score of 0.9 (**90%**) and a Weighted-F1 score of 0.92 (**92%**) for classification. For span matching, our strict matching F1 score was 0.21 (**21%**), while proportional matching reached 0.68 (**68%**), resulting in an average Task A score of 0.6 (**60%**). For Task B, we achieved a ROUGE-1 score of 0.4 (**40%**), ROUGE-2 of 0.18 (**18%**), and ROUGE-L of 0.36 (**36%**). Additionally, we obtained a BERTScore of 0.84 (**84%**), METEOR of 0.37 (**37%**), BLEU of 0.13 (**13%**), resulting in an average Task B score of 0.38 (**38%**). Combining both tasks, our system achieved an overall average score of **49%** and ranked 6th on the official leaderboard for the shared task.

1 Introduction

In PerAnsSumm shared task 2025 at the CL4Health Workshop, NAACL 2025 (Agarwal et al., 2025), the goal is to identify and classify perspective-specific spans (Task_A) and generate summaries tailored to specific perspectives from question-answer threads (Task_B) (Naik et al., 2024).

Span identification is the task of identifying and extracting a continuous range of words from a given text that correspond to a specific piece of information (Fu et al., 2021). This span is a subset of text, usually defined by its starting and ending positions within a sentence. Perspective-specific span identification is the task of finding parts of the text that are relevant to a particular perspective in a given context (Xu et al., 2023). TASK_A involves identifying the specific spans in user answers that reflect distinct perspectives and classifying each span into the appropriate perspective.

For TASK_A, we fine-tuned BERT-large (Devlin et al., 2018) and RoBERTa-large (Liu et al., 2019) models to identify relevant spans within the text. Initially, BERT-large achieved an accuracy of **45%**, while RoBERTa-large performed slightly better at **47%**. To improve their performance, we first pre-trained both models using Masked Language Modeling (MLM) for better domain adaptation before fine-tuning them for span identification. This additional pre-training helped—BERT-large improved to **50%**, and RoBERTa-large improved to **51%**. Further, we optimized the RoBERTa-large model by implementing gradual training, where we fine-tuned the model while keeping some layers frozen for a few epochs. Then, we froze the already fine-tuned layers, unfroze the remaining layers, and fine-tuned them separately. Finally, we fine-tuned the entire model. This step-by-step strategy significantly improved performance, raising accuracy to **60%**.

Summarization is the task of generating a concise and meaningful summary of a longer text while preserving its key information. It helps in reducing large amounts of text into a shorter version while retaining its core meaning (Allahyari et al., 2017). Perspective-specific summarization is a technique that generates summaries focused on a particular aspect of a topic, highlighting information relevant to that perspective instead of providing a general

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 409–414
May 4, 2025 ©2025 Association for Computational Linguistics

summary (Tan et al., 2020). TASK_B involves generating a concise summary that captures the underlying perspective present across all identified spans in the user answers.

For Task_B, we fine-tuned the BART-large (Lewis et al., 2019) and Pegasus-large (Zhang et al., 2019) models to summarize the perspective spans identified and extracted in Task_A. Initially, Pegasus-large achieved TASK_B relevance score of **29%**, while BART-large performed slightly better at **31%**. To enhance their performance, we pre-trained both models using Masked Language Modeling (MLM) for better domain adaptation before fine-tuning them for summarization. This additional pre-training boosted BART-large to **38%** and Pegasus-large to **35%**.

In our proposed solution, we use RoBERTa-large for perspective-specific span identification (TASK_A) and BART-large for perspective-specific summarization (TASK_B).

2 Related Work

Several approaches have been proposed for span identification tasks, focusing on detecting meaningful spans and classifying them into predefined categories. Early works (Chiu and Nichols, 2016) framed SpanID as a sequence tagging problem, where spans were identified token by token using contextual embeddings. Recent research has shifted towards Machine Reading Comprehension (MRC)-based methods (Li et al., 2020), that make use of category-specific queries to extract relevant spans. To address challenges like overfitting and data scarcity, PeerDA (Xu et al., 2023) introduces a peer relation (PR) along with the subordinate relation (SUB), enriching training data and improving generalization. The contrastive learning (Gunel et al., 2021) strategy further enhances the model's ability to distinguish spans across different categories, making PeerDA a promising approach for perspective-based SpanID tasks.

Recent research on fine-grained text analysis has explored span extraction as an alternative to clause-level classification for more precise identification of relevant information. Emotion-cause span extraction (ECSE) (Li et al., 2021) refines emotion cause identification (ECI) by focusing on extracting targeted cause spans rather than entire clauses, improving interpretability and usability. Multi-attention mechanisms have been used to enhance cause-span extraction by leveraging context-sensitive representations, a method that could be adapted for perspective identification (Bi and Liu, 2020). Additionally, position-aware learning has been found to enhance token-level representations, improving the ability to capture key spans within longer texts (Xia and Ding, 2019).

Recent advancements in text summarization have explored span-based extraction and contrastive learning (CL) to improve content selection and representation. In medical question summarization (MQS), CL-enhanced methods have been used to capture key focus words, making sure that the summaries accurately reflect the core intent of the input text (Ma et al., 2022). Similarly, perspective-based summarization benefits from identifying and preserving essential spans that convey underlying viewpoints. Studies on Seq2Seq-based models and reinforcement learning (RL)-enhanced approaches demonstrate the importance of maintaining both syntactic accuracy and semantic coherence in summaries (Keneshloo et al., 2019).

3 Dataset

The dataset (Naik et al., 2024) provided for the shared task is the PUMA dataset, a perspective-aware summary annotated corpus of medical question-answer pairs. It consists of 3,167 CQA threads with approximately 10,000 answers filtered from the Yahoo! L6 corpus. Each answer in the dataset is annotated with five perspective spans: 'cause', 'suggestion', 'experience', 'question', and 'information'. These annotations create concise summaries for each identified perspective, which captures the core idea reflected in the spans across all answers. Each CQA thread may contain up to five perspective-specific summaries.

The data is provided in JSON format. Each entry in the training and validation datasets includes fields such as uri [1], question, context, answers, labelled_answer_spans, labelled_summaries, and raw_text. The labelled_answer_spans contains the span text and the index positions indicating where the span starts and ends within the raw_text. The labelled_summaries provide concise summaries corresponding to each identified perspective.

In the test dataset, only the fields uri, question, context, and answers are available, with no annotations for answer spans or summaries. The dataset was split into 2,236 instances for training, 959 for validation, and 50 for testing.

[1] Unique resource identifier

Dataset	Size
Train	2236
Valid	959
Test	50

Table 1: Dataset Splits

4 System Description

Our system is made of fine-tuned RoBERTa and BART models, where RoBERTa is used for precise token classification tasks, efficiently identifying and labeling specific information within the text. BART, on the other hand, is fine-tuned for summarization, enabling it to generate coherent, contextually relevant summaries by compressing complex input into concise representations.

4.1 Data Pre-Processing

In the pre-processing step, for span identification, we focus on the "answers" and "labelled_answer_spans" fields. The "labelled_answer_spans" field provides perspective spans, where each span contains indices referring to the "raw_text" field. To handle this, we merged all the answers and compared each perspective span to the merged answer, labeling the corresponding tokens as perspective spans (e.g., "I-INFORMATION", "I-SUGGESTION", "I-CAUSE", "I-QUESTION", "I-EXPERIENCE", "O") and marking the rest as "O". We experimented with three token classification formats: BIO (Beginning-Inside-Outside), IO (Inside-Outside), and BIOES (Beginning-Inside-Outside-End-Single). For summarization, we treated the merged spans as context and the "labelled_summaries" as the corresponding summaries.

4.2 Fine-Tuning

We fine-tuned BERT-large and RoBERTa-large models for span identification.

With the BERT-large model, we achieved a Macro-F1 score of 0.83 **(83%)** and a Weighted-F1 score of 0.86 **(86%)** for classification. However, for span matching, the strict matching F1 score was 0.0 **(0%)**, while proportional matching reached 0.47 **(47%)**, resulting in an average Task A score of 0.45 **(45%)**. For the RoBERTa-large model, we obtained a Macro-F1 score of 0.84 **(84%)** and a Weighted-F1 score of 0.88 **(88%)** for classification. Similarly, in span matching, the strict matching F1 score was 0.0 **(0%)**, while proportional matching achieved 0.54 **(54%)**, yielding an average Task A score of 0.47 **(47%)**.

The results indicate that the RoBERTa-large model outperforms the BERT-large model. To further improve performance, we fine-tuned both models for domain adaptation using Masked Language Modeling (MLM) and then retrained them for span identification. After domain adaptation, both models showed improvement.

For the domain-adapted BERT-large model, we achieved a Macro-F1 score of 0.87 **(87%)** and a Weighted-F1 score of 0.9 **(90%)** for classification. In span matching, the strict matching F1 score was 0.0 **(0%)**, while proportional matching reached 0.59 **(59%)**, resulting in an average Task A score of 0.5 **(50%)**.

With the domain-adapted RoBERTa-large model, we achieved a Macro-F1 score of 0.88 **(88%)** and a Weighted-F1 score of 0.92 **(92%)** for classification. For span matching, the strict matching F1 score was 0.01 **(1%)**, while proportional matching reached 0.62 **(62%)**, yielding an average Task A score of 0.51 **(51%)**.

The results now show that the domain-adapted RoBERTa-large model outperforms the domain-adapted BERT-large model. To further enhance performance, we applied a gradual training approach over 10 epochs to both domain-adapted models, each consisting of 24 layers. Initially, during the first 2 epochs, we froze all layers except for the first 4. In the next 2 epochs, we unfroze the subsequent 4 layers while keeping the earlier layers frozen. Over the following 2 epochs, we continued to unfreeze 4 additional layers, leaving the previously trained ones frozen. Finally, during the last 4 epochs, we unfroze all remaining layers and trained the entire model.

Despite using this gradual training method, BERT-large did not show any significant improvement. In contrast, the domain-adapted and gradually trained RoBERTa-large model achieved better results. For classification, we obtained a Macro-F1 score of 0.9 (**90%**) and a Weighted-F1 score of 0.92 (**92%**. For span matching, the strict matching F1 score was 0.21 (**21%**), while the proportional matching F1 score reached 0.68 (**68%**), yielding an average Task A score of 0.6 (**60%**).

We fine-tuned both BART-large and Pegasus-large models for summarization. Using the Pegasus-large model, we achieved a Rouge-1 score of 0.3 (**30%**), Rouge-2 score of 0.12 (**12%**), Rouge-

Pre	GT	Model	Macro-F1	Weighted-F1	strict matching F1	proportional matching F1	Average
✗	✗	BERT-large	0.83	0.86	0.0	0.47	0.45
✗	✗	RoBERTa-large	0.84	0.88	0.0	0.54	0.47
✓	✗	BERT-large	0.87	0.9	0.0	0.59	0.5
✓	✗	RoBERTa-large	0.88	0.92	0.01	0.62	0.51
✓	✓	RoBERTa-large	**0.9**	**0.92**	**0.21**	**0.68**	**0.6**

Table 2: Performance comparison of BERT-large and RoBERTa-large models with and without pre-training (Pre) and gradual training (GT) across different evaluation metrics. The table presents Macro-F1, Weighted-F1, strict matching F1, and proportional matching F1 scores, along with their average performance.

Pre	Model	Rouge-1	Rouge-2	Rouge-L	BERTScore	Meteor	BLEU	Average
✗	Pegasus-large	0.3	0.12	0.27	0.73	0.26	0.1	0.29
✗	BART-large	0.33	0.12	0.29	0.77	0.28	0.09	0.31
✓	Pegasus-large	0.37	0.16	0.33	0.81	0.33	0.12	0.35
✓	BART-large	**0.4**	**0.18**	**0.36**	**0.84**	**0.37**	**0.13**	**0.38**

Table 3: Performance comparison of Pegasus-large and BART-large models for summarization, with and without pre-training (Pre). The table presents performance across various metrics, including Rouge-1, Rouge-2, Rouge-L, BERTScore, METEOR, BLEU, and the overall average score.

L score of 0.27 (**27%**), BERTScore score of 0.73 (**73%**), METEOR score of 0.26 (**26%**), and BLEU score of 0.1 (**10%**). This resulted in an average Task B score of 0.29 (**29%**). For the BART-large model, we achieved a Rouge-1 score of 0.33 (**33%**), Rouge-2 score of 0.12 (**12%**), Rouge-L score of 0.29 (**29%**), BERTScore score of 0.77 (**77%**), METEOR score of 0.28 (**28%**), and BLEU score 0.09 (**9%**), giving an average Task B score of 0.31 (**31%**).

The results indicate that the BART-large model outperformed the Pegasus-large model. To boost performance even further, we fine-tuned both models for domain adaptation using Masked Language Modeling (MLM) and retrained them for span identification. Following domain adaptation, both models showed noticeable improvements.

For the pre-trained Pegasus-large model after domain adaptation, we achieved a Rouge-1 score of 0.37 (**37%**), Rouge-2 score of 0.16 (**16%**), Rouge-L score of 0.33 (**33%**), BERTScore of 0.81 (**81%**), METEOR score of 0.33 (**33%**), and BLEU score of 0.12 (**12%**), resulting in an average Task B score of 0.35 (**35%**).

Similarly, the pre-trained BART-large model showed improved results, we obtained a Rouge-1 score of 0.4 (**40%**), Rouge-2 score of 0.18 (18%), Rouge-L score of 0.36 (**36%**), BERTScore of 0.84 (**84%**), METEOR score of 0.37 (**37%**), and BLEU score of 0.13 (**13%**), resulting in an average Task B score of 0.38 (**38%**).

After domain adaptation, both models improved, with BART-large still outperforming Pegasus-large.

4.3 Inference

4.3.1 Span Identification Module

To identify spans, we process the dataset by extracting the "uri," "answers," and "labelled_answer_spans" fields. The model is then applied to predict spans based on the "answers" field. The predicted spans are stored in a JSON format, where each "uri" is associated with a dictionary containing the identified spans. If no spans are predicted for a given category, an empty array is used for that category. For example, if a dataset entry discusses newborn care, a recommendation such as "So you might want to check your baby in daylight in a sunny room" would be classified under "SUGGESTION," while a factual statement like "Jaundice is an illness that can occur within the first few days of a baby's life" would be categorized under "INFORMATION."

4.3.2 Summarization Module

We use the BART-large model to generate summaries based on the predicted spans. The generated summaries are then stored in the "summaries" dictionary, corresponding to each perspective span, such as "EXPERIENCE," "INFORMATION," "CAUSE," "SUGGESTION," and "QUES-

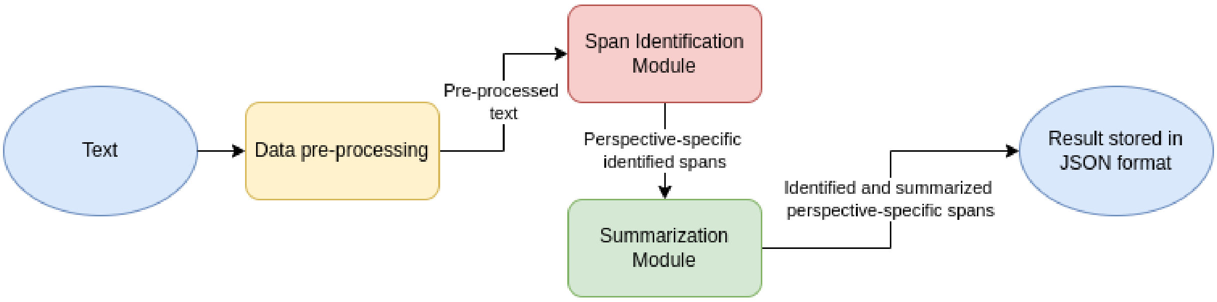
Text
Data pre-processing
Pre-processed text
Span Identification Module
Perspective-specific identified spans
Summarization Module
Identified and summarized perspective-specific spans
Result stored in JSON format

Figure 1: System Workflow

TION." Each category holds a relevant summary derived from the respective spans.

5 Evaluation Metrics

For Task A (Span Identification and Classification), performance is assessed using a macro-averaged F1 score for classification, which ensures balanced evaluation across all classes. For span identification, two matching strategies are employed: Strict matching, which requires an exact span match, and proportional matching, which allows partial matches to account for variability in span boundaries.

For Task B (Summarization), a comprehensive set of evaluation metrics is utilized to measure the quality of generated summaries. These include ROUGE (R1, R2, and RL), which captures the overlap between generated and reference summaries, BLEU, which evaluates n-gram precision, Meteor, which accounts for synonymy and stemming, and BERTScore, which leverages contextual embeddings to assess semantic similarity. These metrics collectively provide a robust evaluation framework for summarization performance.

6 Results

The evaluation results for the different experiments are presented in Table 2 and Table 3. For Task A (Span Identification and Classification), we submitted the RoBERTa-large model, while for Task B (Summarization), we used the BART-large model. Our system achieved an average score of **60%** for TASK_A and **38%** for TASK_B, leading to an overall average score of **49%**. Based on these scores, we secured 6th place on the leaderboard.

7 Conclusion

Our study demonstrates the effectiveness of fine-tuning large language models for perspective-specific span identification and summarization. By leveraging domain-adaptive pre-training and optimization techniques such as gradual training, we significantly improved performance in both tasks. For TASK_A, RoBERTa-large proved to be the most effective model, achieving a final accuracy of **60%** through gradual fine-tuning. For TASK_B, BART-large outperformed Pegasus-large, reaching **38%** accuracy after additional pre-training. These results highlight the importance of targeted pre-training and optimization strategies in enhancing model performance for specialized NLP tasks. Our approach provides a reliable method for identifying and summarizing perspective-specific information, contributing to more advanced and context-aware text processing applications.

Limitations

While our approach improves performance, it still depends on manually annotated training data for TASK_A and TASK_B. We used a gradual training method, but exploring alternative approaches could further enhance results. Moreover, our method requires extensive high-quality annotated data, making scalability challenging, especially in new domains where annotation is costly and time-consuming. Another challenge is handling overlapping or implicit perspectives, where multiple viewpoints exist within the same span or are only implied rather than explicitly stated. This makes it harder for the model to extract distinct perspectives, potentially leading to incomplete or biased summaries. Additionally, while our approach effectively extracts and summarizes perspective-specific information, it does not verify factual accuracy or neutrality, which may impact real-world use. Future improvements could optimize training, better handle ambiguous perspectives and integrate fact-checking mechanisms.

References

Siddhant Agarwal, Md Shad Akhtar, and Shweta Yadav. 2025. Overview of the peranssumm 2025 shared task on perspective-aware healthcare answer summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey. *Preprint*, arXiv:1707.02268.

Hongliang Bi and Pengyuan Liu. 2020. Ecsp: A new task for emotion-cause span-pair extraction and classification. *Preprint*, arXiv:2003.03507.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

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.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. *Preprint*, arXiv:2011.01403.

Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, and Chandan K. Reddy. 2019. Deep reinforcement learning for sequence to sequence models. *Preprint*, arXiv:1805.09461.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Min Li, Hui Zhao, Hao Su, Yurong Qian, and Ping Li. 2021. Emotion-cause span extraction: a new task to emotion cause identification in texts. *Applied Intelligence*, 51(10):7109–7121.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Comput. Surv.*, 55(5).

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! perspective-aware healthcare answer summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15919–15932, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Weiwen Xu, Xin Li, Yang Deng, Wai Lam, and Lidong Bing. 2023. PeerDA: Data augmentation via modeling peer relation for span identification tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8681–8699, Toronto, Canada. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.