Speech-to-Service: Using LLMs to Facilitate Recording of Services in Healthcare

Maj Smerkol
maj.smerkol@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Rok Susič
rs36117@student.uni-lj.si
University of Ljubljana, Faculty of Mathematics and Physics
Ljubljana, Slovenia

Mariša Ratajec
mr97744@student.uni-lj.si
University of Ljubljana, Faculty of Electrical Engineering
Ljubljana, Slovenia

Helena Halbwachs
h.halbwachs@senecura.si
SeneCura Kliniken- und Heimebetriedsgesellschaft m.b.H.
Vienna, Austria

Anton Gradišek
anton.gradisek@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Digital tracking of services is one of the main administrative burdens of the healthcare staff. Here, we present a proof-of-concept study of a so-called speech-to-service (S2S) system that is aimed at facilitating recording of activities, extracting information from the conversation between a healthcare provider and recipient. The system comprises of a speech recorder, a diarization component, an LLM to interpret the conversation, and a recommendation system integrated in a smart tablet that records completed activities and suggests possible other activities that may have still be required. We tested the system on 350 conversations and obtained 95% accuracy, 97% precision and 94% recall.

Keywords

healthcare, LLM, speech recognition, recommendation system

1 Introduction

Healthcare workers, including nurses, technicians, and care personnel form the backbone of the health system as they care for patients and tend to their needs. However, with the standardization and systematization of the healthcare professions and services often becomes a large bureaucratic burden, as healthcare workers have to record all the activities and services they provide to the patients. This process is of course needed as it provides traceability and ensures that all the required activities were taken care of, but the problem is that the interfaces designed for activity logging are often not user-friendly and require the users to choose the activities from a extensive lists of drop-down menus. In total, this amounts to substantial time required only for tedious administrative tasks, time that would be more beneficially spent otherwise.

With the aim to alleviate the administrative burden of activity logging, we explored the possibilities of novel technologies to assist the healthcare staff in their logging tasks. We developed and tested a proof-of-concept system that records the conversation between the healthcare worker and a patient, identifies the activities, and allows the healthcare worker to batch-confirm them on a dedicated smart tablet. Batch-confirmation saves a lot of time by significantly lowering the number of clicks required in the UI. The system is built using open-source or publicly accessible components, particularly a speech-to-text system that transcribes the recorded conversation, and a large language model (LLM) that leverages its natural language processing capabilities. The recommender system shows possible required tasks, serving as a reminder and to suggest tasks that are expected soon, which may lower the number of visits per patient. These recommended tasks are then suggested to the healthcare worker, who can review and confirm them using the LLM-assisted interface. LLMs, such as ChatGPT and Llama, have seen a surge in popularity in a wide variety of topics since their popularization in particular with the unveiling of ChatGPT3 in the autumn of 2022.

Several LLM based systems have been proposed recently, including administrative task automation [6], decision making process [10], improving existing automatic speech recognition (ASR) systems [1], and providing patients with needed information [9]. A recent study [11] concludes that utilising ASR to ease some administrative tasks leads to faster, more efficient work and even increase workers' moods.

2 System Architecture

This paper describes two early prototype systems, both aiming to alleviate the workload of healthcare workers by easing the task of documenting care actions performed. These are the ASR system that logs care actions based on captured dialogue between the healthcare worker and the patient, and a recommender system that predicts the required services at a specific time. This recommender system relies on the historical data, appropriate for long-term patient care facilities.

Both systems are limited in scope and only target the most common healthcare services in the dataset for detection or prediction respectively, which can still greatly easy the workload for medical workers, since the top 10 most common tasks out of around 200 care action types represent around 80% of all services performed.

The recommender system allows the care workers to anticipate tasks in advance and server as a reminder. This aims to lower the number of patient visits, which also alleviates the workload.

2.1 Speech-to-Service ASR

The ASR system consists of a speech diarization model, capable of segmenting the recorded speech based on who is currently speaking, a speech transcription model that transcribes the audio

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.4550

to text, and a LLM fine-tuned to extract specific information from the text. Figure 1 shows the architecture of the prototype system.

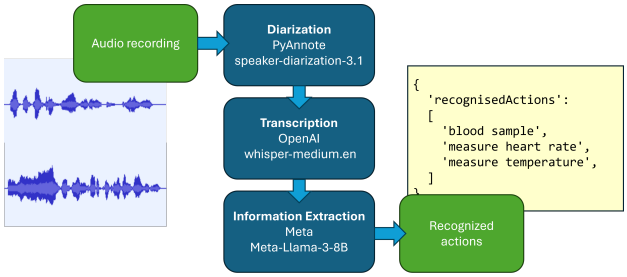

Figure 1: Overview of the ASR system.

We employ speaker-diarization pretrained model [1][4] for diarization, pyannote/speaker-diarization-3.1 pretrained model [7] for transcription, and fine-tuned Llama3 model[2][2] for information extraction via generating JSON formatted output.

2.2 Recommender System

The recommender system prototype is based on machine-learning prediction of events that are expected to occur in a certain time window for a specific patient with addition of tasks that commonly follow predicted tasks. Due to the sensitive nature of the data, we base our predictions only on the time window, patient ID and care type. Thus we consider multi-output binary classifiers that do not require large amounts of data for training. Additional tasks are added to the list based on a Markov chain model that commonly follow, e.g. the task 'clean table' follows the task 'lunch'.

The feature vector includes the time of day, day of week, week of month and month of year as numbers, allowing for capture of periodic events with different periods. Due to lack of patient data, we opted for personalized models, trained for each patient separately. We believe that results can be further improved by adding more patient-related attributes. The model training used five month period of data collected, with cross-validation, and the accuracy was evaluated on the data collected during the sixth month. Due to patients' medical states changing over time, some data drift is expected, which is reflected in our results.

3 Dataset

To fine-tune the information extraction model based on Llama3, we have created a dataset of conversations in text form and appropriate outputs for each of them, as the task on hand is very specific and we did not find any existing appropriate dataset. We automated the process and manually removed any bad examples. A real dataset, ideally recorded in the target environment, is needed for for final implementation - LLM generated datasets used for training LLMs are only appropriate in preliminary studies.

To generate the dataset, we prepared a BERT [3] LLM via prompting [5]. A training sample was generating by first randomly selecting 2 of the 10 target actions, and programmatically generating the target output JSON. The BERT model was then tasked with generating a conversation, in which these two tasks are mentioned as done during the conversation. We generated several hundred conversations that way, and manually checked for mistakes in the model output. Many conversations were removed due to selected actions not being mentioned or other reasons. Finally, the resulting dataset contains 350 conversations and JSON formatted lists of tasks.

For the prediction of services required during a visit, we have acquired a log of all services performed in one long-term patient care facility over a period of 6 months, with the next version expanding to data from six facilities. The tasks in dataset include measurements (body temperature, heart rate, blood pressure, ...), medical tasks (monitoring medicine intake, performing examinations, turning the patient in bed) and care tasks (breakfast, lunch, cleaning). There are over 200 different tasks mentioned. The dataset includes limited patient information—patient ID, care type, and a detailed chronological history of services received. Care types (CareType I, CareType II, CareType III/A, CareType III/B, CareType IIII) represent an estimate of how much assistance a person requires. Legal restrictions on accessing sensitive health data prevented us from obtaining more detailed patient records, so we developed prediction models based on these limited data points, balancing accuracy with regulatory constraints.

The data preprocessing involved determining each patient's presence in the facility by identifying the timestamps of their first and last recorded service. Patients with a stay of less than four months were excluded from the analysis to ensure sufficient data for reliable predictions.

4 Methods

This section describes the methodology used to develop the ASR system and the recommender system.

4.1 Clustering

The primary goal of the clustering process was to group patients with similar patterns in terms of the type and frequency of services they received, allowing us to predict relevant services more effectively for each cluster (since it was not clear, even among experts, whether care type and actual care provided were correlated).

The clusters, as shown in Figure 2, demonstrate that patients within the same care type tend to receive similar services. Some deviations, where multiple classifications appear within a cluster, are likely due to temporary conditions we could not fully exclude (for instance, an individual categorized under "Care Type II" may temporarily receive services typical of "Care Type III/A" (e.g. due to a broken arm), while their care type classification remains unchanged). Despite this, the care types differentiate well enough across clusters, leading us to use "CareType" as one of the key attribute for further service predictions.

In the clustering process, we excluded CareType IIII because this group is characterized by highly diverse healthcare needs due to specific diseases, and experts advised us to omit it for this part of the analysis.

4.2 Recommender System

To recommend the required services, we constructed the training dataset using a detailed log of care actions performed over a 6-month period. For each patient, the data was divided into consecutive 4-hour time windows. In each window, we examined whether specific care actions were performed, marking them as "positive" if they occurred within that time frame. This granular approach allowed us to capture the temporal dynamics of

[1] pyannote/speaker-diarization-3.1
[2] meta/meta-llama-3-8b
[3] google-bert/bert-base-multilingual-cased

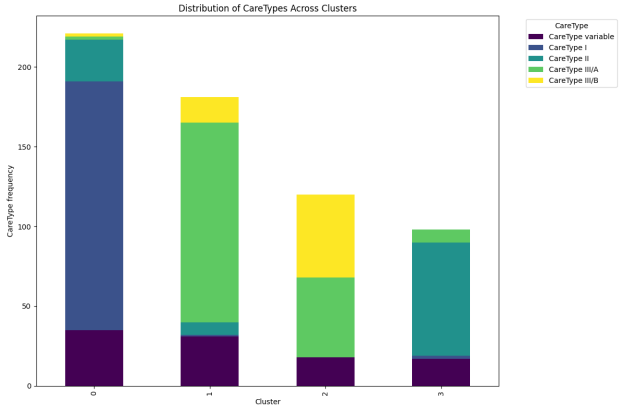

Figure 2: Clustering of patients closely aligns with pre-existing care type assignments, ranging from minimal personal assistance (CareType I) to moderate assistance (CareType II), and full or intensive personal assistance (CareTypes III/A and III/B) for those with more severe care needs.

service delivery, ensuring that for each time window, we had a clear record of the services provided. As a result, we generated over 1000 labeled examples per patient, with each example representing a specific time window and its associated care actions. This enabled the model to learn patterns in service requirements throughout the day and week.

To identify the best predictive model, we evaluated various classification algorithms, including Random Forest, Decision Tree, K-Neighbors, Support Vector Classifier (SVC), Gradient Boosting, and Naive Bayes. Each model was trained using a multi-output classification approach, with features including the frequency of the top services provided and the relevant time attributes. To ensure robust model evaluation, we implemented 5-fold cross-validation and subsequently tested the models on the sixth month's data to assess their predictive performance.

4.3 Speech Recognition and Information Extraction

Due to limited availability of training data, only the information-extraction model based on Llama3 was fine tuned using few-shot LoRA (low-rank adaptor) supervised training. The diarization and transcription models are used unchanged.

The diarization model used is speaker-diarization [4]. Initial experiments with few-shot LoRA fine tuning [3] did not improve the performance, hinting at the need for a larger training dataset. The model's performance is satisfactory at the task of segmentation, but less so at the task of identifying which segments belong to which speaker, especially for longer conversations. For a two-speaker situation, the model seems to assume the speakers take turns speaking, causing mistakes when a single speaker pauses before continuing to speak.

The transcription model used is whisper [8]. The model transcribes each segment separately. As mentioned above, the speakers are not robustly recognised, and we cannot reliably assign a speaker to each line of text. Still, labeling each line of text even with an ambiguous label improves the downstream task of information extraction. The transcribed lines of text are concatenated, and at the start of each utterance a label marking it as such is added. Thus, the transcribed text resembles a play with unknown characters speaking.

The information extraction model is Llama3 [2], and fine-tuned utilising a LoRA few-shot fine tuning. Our approach was to fine-tune the model for the task of extracting information about specific care actions and generate the output in a JSON format, providing structured data directly. A small training dataset was prepared as described in the section 3.

5 Results and Discussion

5.1 LLM Based Infromation Retrieval Model

The Llama3 based information extraction model is evaluated using a 5-fold cross validation, achieving **95% accuracy**, **97% precision**, and **94% recall**. For evaluation the model's JSON-formatted strings were deserialized to objects and tested against known correct objects to be able to interpret the results as multi-label binary classification.

The LLM infromation extraction model sometimes generates invalid JSON after fine-tuning, most commonly due to duplicated keys or getting stuck in a loop, generating same elements until maximum output size is generated. The generated strings are therefore post-processed to fix these mistakes via simple string manipulation, however this indicates that experiments with different output formats or avoiding generating the answers should be performed.

The whole ASR pipeline including diarization and transcription has not yet been evaluated and falls within the scope of future work.

5.2 Recommender System

Tables 1 and 3 present the classification results. Table 1 reports the average performance across all patients, including standard deviations for the different models, while Table 3 shows classification accuracy by care type, with averages and standard deviations across all patients within each care type, based on the model with the best results, which in this case is K-Neighbors (KNN).

Results are reported in two ways, tables 1 and 3 show accuracy considering all target attributes, only considering a prediction correct when all targets are predicted correctly. The table 2 show average of accuracies for each target attribute.

Table 1: Cross-validation and test accuracy (mean ± standard deviation) across all patients for various classification models.

Model	CV Accuracy	Test Accuracy
RandomForest	0.71 ± 0.14	0.66 ± 0.16
DecisionTree	0.65 ± 0.16	0.66 ± 0.16
KNeighbors	**0.73 ± 0.13**	**0.71 ± 0.16**
SVC	0.63 ± 0.12	0.63 ± 0.14
GradientBoosting	0.68 ± 0.12	0.66 ± 0.15
NaiveBayes	0.57 ± 0.17	0.55 ± 0.20

The K-Neighbors (KNN) classifier outperformed other models, achieving an average CV accuracy of 73%, a test accuracy of 71%, and R^2 score of 0.44. This made it the most effective model for predicting service plans. Random Forest also performed reasonably well, achieving a test accuracy of 66%, though it did not surpass KNN in overall performance.

Table 2: Majority Class Percentage and Task-wise Test Accuracy (mean ± standard deviation) across all patients for various classification models.

Model	Majority Class Percentage	Task-wise Accuracy
RandomForest	0.72 ± 0.19	0.89 ± 0.10
DecisionTree	0.72 ± 0.19	0.89 ± 0.11
KNeighbors	**0.72 ± 0.19**	**0.91 ± 0.10**
SVC	0.65 ± 0.16	0.88 ± 0.10
GradientBoosting	0.65 ± 0.16	0.89 ± 0.09
NaiveBayes	0.72 ± 0.19	0.85 ± 0.15

Table 3: Classification performance of K-Neighbors (KNN) by CareType, showing cross-validation and test accuracy (mean ± standard deviation), averaged across all patients within each care type.

CareType	CV Accuracy	Test Accuracy
CareType I	0.79 ± 0.12	0.76 ± 0.16
CareType II	0.79 ± 0.11	0.78 ± 0.13
CareType III/A	0.68 ± 0.13	0.66 ± 0.15
CareType III/B	0.70 ± 0.14	0.68 ± 0.17
CareType IIII	0.68 ± 0.10	0.67 ± 0.12

Since all predictive accuracy values exceed the 70% majority class baseline, this is an excellent result. In multi-label classification, where multiple services are predicted simultaneously, it's important to not only focus on overall accuracy but also on the accuracy of each individual task. By achieving 90% accuracy on the most common tasks, the model ensures that key services are reliably predicted.

The lower test accuracy compared to cross-validation can be explained by temporal changes in patient conditions, as the test set only included the last month of data. As patient care needs shift over time, predicting long-term patterns is more challenging than shorter-term cross-validation, where care remains more stable.

The test accuracy also reflected noticeable differences across care types. CareType I and CareType II showed higher accuracy rates, while more complex types, such as CareType III/A, III/B, and IIII, exhibited a drop in accuracy of around 10%. This is likely due to the more diverse and unpredictable care needs in these groups, making service prediction more challenging.

This approach, particularly with the strong performance of our K-Neighbors (KNN) model, demonstrated the potential of machine learning to enhance personalized planning in healthcare. In future work, including additional patient-specific features beyond time-based data, such as health-related attributes, could further improve accuracy, particularly for the more complex care types.

6 Conclusions

This is early work and further improvements are underway. The whole ASR pipeline needs to be evaluated and we expect noticeably worse performance comparing to only the information extraction model due to larger complexity and possibility for failure at each step. The information retrieval model itself is not inefficient considering computational time and memory required, but diarization and transcription steps are. The required service prediction should also be further improved. Using current dataset an alternative approach that may improve performance is using sequence modelling or event prediction approaches. Finally, the two models could work in tandem - predicting the required actions and using that information in the ASR pipeline could be beneficial.

Based on the proof-of-concept study, we conclude the suggested approach is in principle feasible and can be beneficial to healthcare providers. However, in view of regulations, special caution has to be paid during the implementation of any sort of such system in a real-world setting. Recording and diarizing conversations between healthcare staff and the patients is likely to include highly personal data, which falls under the EU relevant legislation, specifically the GDPR (*General Data Protection Regulation*) [4] and the EU AI Act (*Artificial Intelligence Act (Regulation (EU) 2024/1689)*) [5]. Furthermore, indiscriminately recording conversations and feeding them into an LLM will likely be considered as "high risk" in view of the AI Act. This means that implementing such services will require extensive screening, documentation, clear division of ownership and access roles, and other compliance with legal requirements.

Acknowledgements

We thank the healthcare provider organization for the dataset and for insightful discussions.

References

[1] Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*.

[2] AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[3] Shamil Ayupov and Nadezhda Chirkova. 2022. Parameter-efficient finetuning of transformers for source code. *ArXiv*, abs/2212.05901. https://api.semanticscholar.org/CorpusID:254564456.

[4] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*. Brno, Czech Republic, (Aug. 2021).

[5] 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. http://arxiv.org/abs/1810.04805 arXiv: 1810.04805.

[6] Senay A. Gebreab, Khaled Salah, Raja Jayaraman, Muhammad Habib ur Rehman, and Samer Ellaham. 2024. Llm-based framework for administrative task automation in healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 1–7. DOI: 10.1109/ISDFS60797.2024.10527275.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. (2022). DOI: 10.48550/ARXIV.2212.04356.

[8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. (2022). DOI: 10.48550/ARXIV.2212.04356.

[9] Prakasam S, N. Balakrishnan, Kirthickram T R, Ajith Jerom B, and Deepak S. 2023. Design and development of ai-powered healthcare whatsapp chatbot. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 1–6. https://api.semanticscholar.org/CorpusID:259280109.

[10] Raja Vavekanand, Pinja Karttunen, Yue Xu, Stephanie Milani, and Huao Li. 2024. Large language models in healthcare decision support: a review.

[11] Markus Vogel, Wolfgang Kaisers, Ralf Wassmuth, and Ertan Mayatepek. 2015. Analysis of documentation speed using web-based medical speech recognition technology: randomized controlled trial. *Journal of medical Internet research*, 17, 11, e247.

[4] https://gdpr-info.eu/

[5] https://artificialintelligenceact.eu/the-act/