Prototype of a Care Documentation Support System Using Audio Recordings of Care Actions and Large Language Models

Matthias Hirschmanner
matthias.hirschmanner@tuwien.ac.at
TU Wien, Vienna, Austria

Reinhard Grabler
reinhard.grabler@tuwien.ac.at
TU Wien, Vienna, Austria

Helena Anna Frijns
helena.frijns@tuwien.ac.at
TU Wien, Vienna, Austria

Evelyn Mayer-Haas
evelyn.haas@caritas-wien.at
Caritas Wien, Vienna, Austria

Markus Vincze
markus.vincze@tuwien.ac.at
TU Wien, Vienna, Austria

ABSTRACT

Care documentation is an essential but time-consuming part of nursing practices. We present a first prototype to support care workers by generating summaries from audio recorded during standard nursing interactions. The audio is transcribed with Automatic Speech Recognition (ASR), and a summary is generated by a Large Language Model (LLM), both running locally. For evaluation, we recorded four mock care interaction scenarios with a training manikin. We compare different local LLMs with GPT-3.5 and GPT-4. We find that most of the important topics relevant to care documentation were present in the resulting summaries.

KEYWORDS

Automatic Documentation, Electronic Health Records, Large Language Models, Robots in Healthcare

1 INTRODUCTION

Robots in care are seen as a promising technology to alleviate the workload of care workers in an aging society. However, robots are hardly used in care facilities despite multiple research projects over the last decade [15]. Reasons for the lack of robots in healthcare are, among others, cost, perceived threats to professional roles, absence of personal benefits of staff, technical limitations [26]. We conducted 13 participatory design workshops in two residential care homes with care workers and care recipients to investigate how robotic technology can support care needs and meet workplace concerns. A dominant topic among care workers was the possibility of robots to support documentation work [9].

We propose a documentation support system that records audio of German nurse-patient dialogues during standard care interactions with a wireless microphone worn by the caregiver. An LLM generates a summary of essential information such as performed care actions and the condition of care recipients (Figure 1). Due to the sensitivity of the data, we investigate three state-of-the-art offline models applicable to German: EM German [11], Mixtral [13] and Sauerkraut-Mixtral [27]. The resulting text is presented to care workers when they do their routine documentation. This approach

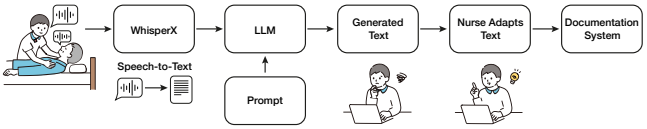

Figure 1: Pipeline of the documentation support system.

was chosen to be frictionless and agnostic to documentation systems while maintaining the nurses' agency and responsibility. By using only a microphone, we follow recommendations by [26] to first introduce the least disruptive application to showcase advantages, reducing stakeholders' resistance to innovation. In a later step, this system is envisioned to be integrated into a multi-purpose care robot that accompanies care workers. Our main contributions are:

- A pipeline to support care workers with documentation using local off-the-shelf ASR and LLMs.
- A qualitative evaluation of a documentation support prototype using data recorded with a training manikin.
- A comparison of three open-source and two proprietary models for summarization of nurse-patient dialogues.

2 RELATED WORK

Care documentation is essential to nursing practices to ensure quality and continuity of care [20], but is time-consuming, with studies reporting between 25% and 41% of nurses' working hours spent on documentation [5, 24, 25, 29]. Multiple studies have been conducted to introduce automatic documentation systems in healthcare settings [7, 8]. Knoll et al. [14] conducted a user study on medical note generation software from patient-doctor dialogues for telehealth and refined the system introduced in [21] for real-time application. Ben Abacha et al. [2] and Yim et al. [30] conducted two challenges for generating medical notes from ASR transcribed doctor-patient conversation with diarized speakers. For this task, multiple authors found GPT-4 with in-context examples to outperform other methods, such as fine-tuned T5 models, few-shot prompting, and multi-stage prompting [10, 19, 28]. All these methods tackle doctor-patient dialogues, which usually consist of very directed questions to arrive at a diagnosis. In our use case, nurses talk to a patient while performing care actions to describe what they are doing. The dialogues often contain small talk that still includes useful information for care documentation [18]. Additionally, doctors tend to take notes during the conversation. Nurses write documentation at a later point, after multiple care interactions. To the best of our

Workshop on Human — Large Language Model Interaction HRI '24, March 11–15, 2024,

knowledge, there is currently no attempt to extract care documentation from nurse-patient dialogues during care actions using LLMs. This includes different challenges compared to doctor-patient dialogues, such as detecting performed care actions, patient's physical condition, and mental state (e.g., feeling anxious, lonely, sad).

3 DOCUMENTATION SUPPORT SYSTEM

The documentation support system needs to handle German conversations and run locally due to the sensitivity of the data. The audio is recorded by a wireless Lavalier microphone worn by the caregiver, which picks up the nurse-patient dialogue. The recording is transcribed with WhisperX [1] using Whisper's `large-v3` model [23] in a post-processing step. The transcription is not changed, and no speaker diarization is performed as we found it unreliable. We compare GPT-3.5 [4] and GPT-4 [22] to multiple models that adhere to our limitations to run locally and support German:

- EM German [11]: a unilingual model based on Mistral7B [12] and fine-tuned for German.
- Mixtral-8x7b-Instruct [13]: a multilingual Mixture of Experts (MoE) model by MistralAI.
- Sauerkraut-Mixtral-8x7b-Instruct [27]: a variant of Mixtral fine-tuned with German data.

4 EXPERIMENTS AND RESULTS

We recorded four staged care scenarios with the "Nursing Anne" training manikin [16]. The manikin has a built-in speaker, which enabled a nurse in another room to perform the role of a care recipient. The different scenarios involved body care, an accident, mobilization, medication intake, and description of health conditions such as pain and mood. Using a Lavalier microphone worn by the care giver, we recorded the German nurse-patient dialogues. After each care scenario, the researchers sat down with the participating care workers for a semi-structured post-interview about how they would document the nursing interaction. Based on the recordings and interviews, three of the authors independently created mock care documentation in two different formats (i.e., complete sentences and bullet lists). We use the ROUGE-1 metric [17] for evaluation similar to [2, 3]. All six human-generated summaries were used as a reference pool for the evaluation, taking the top-scoring match as described in [17]. Moreover, we report the percentage of core topics that should be documented as identified by care workers in post-interaction interviews. We used the same prompt for all models, which included a general system description, a transcript of the patient-nurse dialogue, and the query (translated from German): "*Write a summary of the conversation as a list for the care documentation. Pay particular attention to activities that were carried out, the person's condition, and possible discomfort. Answer in German!*"

According to the ROUGE-1 score (Table 1), Mixtral seems to be superior in most situations. However, this shows the imperfections of the measure as described in [6] rather than indicating the superior model. ROUGE checks for the appearance of the exact words between the generated summary and reference summaries instead of semantic similarities. Additionally, it punishes longer summaries that tend to be generated by GPT-3.5 and GPT-4. As another measure, we identified 24 relevant topics for the care documentation over all scenarios from the care workers' interviews and checked if they appeared in the generated summaries. GPT-3.5 and GPT-4 were superior in this measure, mentioning 21 of 24 (87.5%) relevant topics in the created summaries, while Sauerkraut-Mixtral performs best of the local LLMs with 83.3% (Table 1).

Table 1: ROUGE-R1 scores for different models and the four scenarios. Topics refers to the percentage of relevant topics present in the summaries for each scenario (S_N).

	S_1	S_2	S_3	S_4	Avg.	Topics
EM-German [11]	.178	.217	.288	.164	.212	50.0%
Mixtral [13]	**.341**	**.341**	**.372**	.196	**.313**	75.0%
SK-Mixtral [27]	.216	.293	.335	**.294**	.284	83.3%
GPT-3.5 [4]	.338	.207	.357	.242	.286	**87.5%**
GPT-4 [22]	.243	.223	.307	.136	.227	**87.5%**

We discuss scenario S_1 in more detail to showcase the differences between models. The interaction was a standard morning routine, which lasted for ~22 minutes. The nurse-patient dialogue during performing care actions included today's planned visit of the daughter, pain in different body parts, refusal to eat due to nausea, and biographical information involving the resident's life at a farm. The care recipient repeated questions multiple times, which could indicate dementia. From the interviews, we identified six main topics that should go into the care documentation: daughter visit, unlocatable pain, refusal to eat, nausea, limited movement abilities, and impaired short-term memory. None of the models was able to extract the impaired short-term memory from the conversation. Both GPT variants and Sauerkraut-Mixtral have all other topics in the summaries. Mixtral ommited the daughter visit. The smaller EM-German model does not extract the topics of nausea and refusal to eat. In scenario S_2, the refusal to take medication was only detected by the GPT models. All models output biographical information, which were not mentioned as relevant topics in the interviews, but are a specific subsection of care documentation in the studied care home. Small changes of the prompt can change the outputs and resulting scores significantly. In some cases, both Mixtral variants output English text instead of German.

5 DISCUSSION AND FUTURE STEPS

We introduce a first prototype for automatic care documentation generation from nurse-patient dialogues recorded during care interactions. It shows promising results summarizing most of the relevant topics for care documentation of staged care interactions with a training manikin. However, we encountered high variability in the quality of results depending on the specific prompt.

As a next step, we will record real dialogues of nurse-patient care interactions in a residential care home and observe the ensuing care documentation process. We plan to further refine the care documentation support system with promising approaches such as in-context examples, prompt optimization, multimodal models, and multi-stage prompting to increase reliability and consistency. Open challenges are privacy concerns, user interface design, care recipients' and nurses' attitudes to the system (e.g., mistrust, overreliance, inconvenience). Another interesting aspect is that LLMs could facilitate translation since many nurses do not have German as their first language. We will involve care workers continuously during the iterative development process.

ACKNOWLEDGMENTS

This research is supported by the Austrian Science Foundation (FWF) project Caring Robots // Robotic Care (CM 100-N) and project No. I6114-N iChores. We want to thank the care workers at the residential care home Caritas Haus St. Barbara for participating in the data recording and especially Brankica Cegar for organizing. The vector graphics are by Soco St in CC Attribution License.

REFERENCES

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023* (2023).

[2] Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop* (Toronto, Canada). Association for Computational Linguistics, 503–513. https://doi.org/10.18653/v1/2023.clinicalnlp-1.52

[3] Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 2575–2588.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Kim De Groot, Anke JE De Veer, Anne M Munster, Anneke L Francke, and Wolter Paans. 2022. Nursing documentation and its relationship with perceived nursing workload: a mixed-methods study among community nurses. *BMC nursing* 21, 1 (2022), 1–12.

[6] Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 6038–6052. https://doi.org/10.18653/v1/2022.naacl-main.442

[7] Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating Medical Reports from Patient-Doctor Conversations Using Sequence-to-Sequence Models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations* (Online). Association for Computational Linguistics, 22–30. https://doi.org/10.18653/v1/2020.nlpmc-1.4

[8] Frederico Soares Falcetta, Fernando Kude De Almeida, Janaína Conceição Sutil Lemos, José Roberto Goldim, and Cristiano André Da Costa. 2023. Automatic Documentation of Professional Health Interactions: A Systematic Review. *Artificial Intelligence in Medicine* 137 (2023), 102487. https://doi.org/10.1016/j.artmed.2023.102487

[9] Helena Anna Frijns, Ralf Vetter, Matthias Hirschmanner, Reinhard Grabler, Laura Vogel, and Sabine Theresia Koeszegi. 2024. Co-design of Robotic Technology with Care Home Residents and Care Workers. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '24)*.

[10] John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. Wanglab at Mediqa-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations Using Large Language Models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*. 323–334.

[11] Jan Philipp Harries. 2023. EM German (V01). https://github.com/jphme/EM_German.

[12] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[13] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[14] Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anja Belz, and Aleksandar Savkov. 2022. User-Driven Research of Medical Note Generation Software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 385–394.

[15] Maria Kyrarini, Fotios Lygerakis, Akilesh Rajavenkatanarayanan, Christos Sevastopoulos, Harish Ram Nambiappan, Kodur Krishna Chaitanya, Ashwin Ramesh Babu, Joanne Mathew, and Fillia Makedon. 2021. A Survey of Robots in Healthcare. *Technologies* 9, 1 (Jan. 2021), 8. https://doi.org/10.3390/technologies9010008

[16] Laerdal Medical. 2024. *Nursing Anne*. https://laerdal.com/gb/products/simulation-training/nursing/nursing-anne/

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[18] Lindsay M Macdonald. 2016. Expertise in Everyday Nurse–Patient Conversations: The Importance of Small Talk. *Global qualitative nursing research* 3 (2016), 2333393616643201.

[19] Yash Mathur, Sanketh Rangreji, Raghav Kapoor, Medha Palavalli, Amanda Bertsch, and Matthew Gormley. 2023. SummQA at MEDIQA-Chat 2023: In-context Learning with GPT-4 for Medical Summarization. In *The 61st Annual Meeting of the Association for Computational Linguistics*.

[20] Bridie McCarthy, Serena Fitzgerald, Maria O'Shea, Carol Condon, Gerardina Hartnett-Collins, Martin Clancy, Agnes Sheehy, Suzanne Denieffe, Michael Bergin, and Eileen Savage. 2019. Electronic Nursing Documentation Interventions to Promote or Improve Patient Safety and Quality Care: A Systematic Review. *Journal of nursing management* 27, 3 (2019), 491–501.

[21] Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Aleksandar Savkov, and Anja Belz. 2022. Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation. In *ACL 2022: 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5739–5754.

[22] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[24] Nadia Roumeliotis, Geneviève Parisien, Sylvie Charette, Elizabeth Arpin, Fabrice Brunet, and Philippe Jouvet. 2018. Reorganizing Care with the Implementation of Electronic Medical Records: A Time-Motion Study in the PICU. *Pediatric Critical Care Medicine* 19, 4 (2018), e172–e179.

[25] Elizabeth Schenk, Ruth Schleyer, Cami R Jones, Sarah Fincham, Kenn B Daratha, and Karen A Monsen. 2017. Time Motion Analysis of Nursing Work in ICU, Telemetry and Medical-Surgical Units. *Journal of Nursing Management* 25, 8 (2017), 640–646.

[26] David Silvera-Tawil. 2024. Robotics in Healthcare: A Survey. *SN Computer Science* 5, 1 (2024), 189.

[27] VAGO Solutions. 2023. SauerkrautLM-Mixtral-8x7B-Instruct. https://huggingface.co/VAGOsolutions/SauerkrautLM-Mixtral-8x7B-Instruct.

[28] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2023. *Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts*. arXiv:2309.07430 [cs] http://arxiv.org/abs/2309.07430

[29] Po-Yin Yen, Marjorie Kelley, Marcelo Lopetegui, Jacqueline Loversidge, Esther M Chipps, Lynn Gallagher-Ford, and Jacalyn Buck. 2018. Nurses' Time Allocation and Multitasking of Nursing Activities: A Time Motion Study. In *AMIA Annu Symp Proc*. 1137–1146.

[30] Wen-wai Yim, Asma Ben Abacha, Neal Snider, Griffin Adams, and Meliha Yetisgen. 2023. Overview of the MEDIQA-Sum Task at ImageCLEF 2023: Summarization and Classification of Doctor-Patient Conversations. In *CLEF 2023 Working Notes* (Thessaloniki, Greece) *(CEUR Workshop Proceedings)*. CEUR-WS.org.

Received 07 March 2024