Empirical Evaluation of ASR and NLU in a Multimodal Dialogue System for Survey Answering

Philipp L. Harnisch, Stefan Hillmann

Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
p.harnisch@tu-berlin.de

Abstract: PROM surveys, used to measure the effect of rehabilitation treatments, are typically filled out on paper, and often suffer from low response rates. Replacing it with a multimodal survey system, supporting touch and speech interaction, could lead to lower hurdles and therefore more data quantity. To do this, it requires task-specific training samples for the Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) to classify spoken answers into one of the standardized PROM answer options.
Due to the lack of training data for medical PROM surveys, we created augmented text samples with each answer option description, combined with different templates. To improve training capabilities, introduce a proper test set, and evaluate the ASR, we also collected 1,797 real voice samples within an empirical study. Further, we incorporate the contextual knowledge of the current question into our NLU architecture by implementing one classifier for every question scale.
Our results reveal that training with empirical data leads to better results than augmented data from templates and original answer option descriptions. Because of participant mislabeling of 33% due to the ambiguity of the task, we receive overall low NLU performances with up to 51.1% accuracy, and rank-1-accuracy up to 79.3%. We also find that our implementation of many scale-specific NLU classifiers significantly outperforms one NLU classifier for all labels, that incorporates the same contextual knowledge after the prediction, by 8 percent points.

1 Introduction

Patient-Reported Outcome Measures (PROM) surveys are collected to measure changes in the subjective health of patients in rehabilitation clinics, or the effect of other health treatments [1]. The traditional approaches of filling out surveys on paper or digitally without assistance, result in low data quality or response rates. To increase data quantity, we want to enable a multimodal dialogue system (see Figure 2), combined with an embodied conversational agent [2], to assist with this task. Therefore, we require an ASR component to translate voice inputs into text, and an NLU that is capable of classifying resulting texts into the possible answer options.

EXAMPLE PROM ITEM DESCRIPTION:
"Trotz meiner Beschwerden bin ich in der Lage schwierige Probleme zu lösen" ("Despite my complaints, I am able to solve difficult problems")

ANSWER OPTIONS FOR THIS ITEM:

1. "stimmt nicht" ("not true")
2. "stimmt wenig" ("true a little")
3. "stimmt mittelmäßig" ("true mediocre")
4. "stimmt ziemlich" ("pretty much true")
5. "stimmt sehr" ("very true")

Figure 1 – A health-related PROM item about the ability to solve difficult problems despite complaints.

DOI: 10.35096/othr/pub-7100, URN: urn:nbn:de:bvb:898-opus4-71007

In this paper, we focus on a 92 items PROM questionnaire, made with standardized PROM items [3]. The survey contains 13 different scales for the answer options, describing a range between the best and worst health condition. Among the scales, 11 consist of five textual descriptions (see Figure 1), one scale has only four textual descriptions, and another scale ranges from number 0 to 10. As PROM surveys are only designed to be filled out on paper or graphical user interfaces, there is a lack of spoken text examples of possible patient responses to survey items [4]. With an empirical study, we gathered a representative PROM answering dataset of 1,797 speech answer recordings from 20 participants. With these samples, we quantitatively assess the performance of the system's ASR and NLU on our survey in terms of word-error-rate and accuracy, respectively.

Further, we use this new data to compare many scale-specific NLUs with the standard approach of using one single NLU that covers all answer options at once. This reveals that the reduced complexity of incorporating contextual knowledge into the model architecture, increases prediction performance, and thus is not overshadowed by the overhead of implementing multiple classification models.

2 Contextual NLU

Besides the ASR evaluation, we investigate NLU implementations, with and without integration of contextual information. We introduce the approach *Many NLU* that uses one specific linear layer for each of the 13 question scales, decreasing the necessary complexity of a single model and simplifying the introduction or removal of questions from this survey (and other surveys in general). It takes the LaBSE sentence embedding [5] from transcripts of spoken user utterances and predicts the most likely answer option to a survey question. We compare this with an NLU consisting of one linear layer that spans all the answer options of all scales at once. In this paper, we denote this approach as *One NLU Naive*. Further, we denote the same approach that additionally incorporates the scale-specific context as *One NLU*. We implement this by setting all logits to negative infinity for labels outside the scope of the current scale. Our hypothesis is that *Many NLU* outperforms *One NLU* and *One NLU Naive*, because it simplifies the task to smaller sub-tasks.

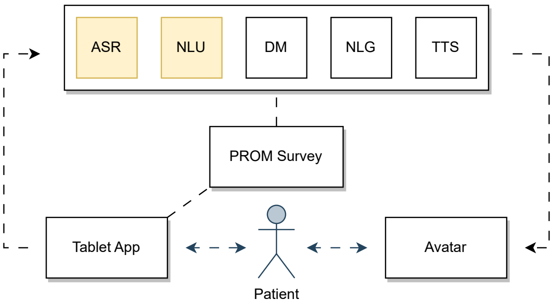

Figure 2 – Overview of the multimodal dialogue system. The patient can interact with a tablet app and an avatar, progressing through the prom survey. Spoken language is processed by ASR and NLU.

Linear layers from all approaches are fully-connected and trained for 20 epochs, with a batch size of 16, Cross-Entropy-Loss and the AdamW optimizer (lr=0.001). For evaluation, we select the best performing model of all epochs, in terms of training data accuracy. We used ten different random seeds to increase statistical confidence in comparison of approaches.

Because our use case is rather simple, our NLU does not include detection of entities, and we chose that all answer options are designed as intents. Thus, it is a traditional classification problem and our contextual approach of many small modular NLUs can also be applied to a variety of other domains.

3 Method

The two datasets, consisting of augmented and empirically collected data, are described in Section 3.1 and 3.2. The methodology to evaluate the different NLU approaches is described in Section 3.3.

3.1 Augmented Text Data

For the augmented dataset, we started with the original answer descriptions from the standardized PROM items, e.g., "stimmt nicht" ("not true"). We manually created 23 templates, like "Meine Antwort ist <*answer*>." ("My answer is <*answer*>."). Then, these templates are applied to all answer options of the 13 PROM scales, where <*answer*> is replaced by strings like "*stimmt nicht*" ("*not true*"), to generate additional training material.

3.2 Empirical Speech Data

To collect more data for training and evaluation of our models (see Figure 2), we designed an empirical study as a lab experiment in which 20 participants answered our PROM questions by spoken language, without seeing the answer options. Recordings were made in a hearing booth, with an application guiding through the survey with an optional reading function. At the beginning of every PROM item, the reading function automatically plays an audio of synthetic speech that reads the current PROM item description.

Table 1 – Information about audio respectively text of the empirical PROM answer dataset, collected from 20 participants, and spanning 13 answer scales.

EMPIRICAL PROM ANSWER DATASET (n=1,797)					
AUDIO	**total**	**mean**	**TEXT**	**total**	**mean**
seconds	16K	8.75	**words**	29K	16.34
speech rate	–	2.09	**characters**	175K	97.66

To prevent an uneven distribution of answer option choices, we randomly prime each participant to provide one specific answer out of five semantic answer options. The priming is done via a generic 5-point emoji face scale[1], where each emoji represents the well-being associated with the answer that should be given. We reduced noise in the labeling process by not relying on the priming for automatic label creation, but asking participants to select the correct one out of the standardized answer options after they have answered an item.

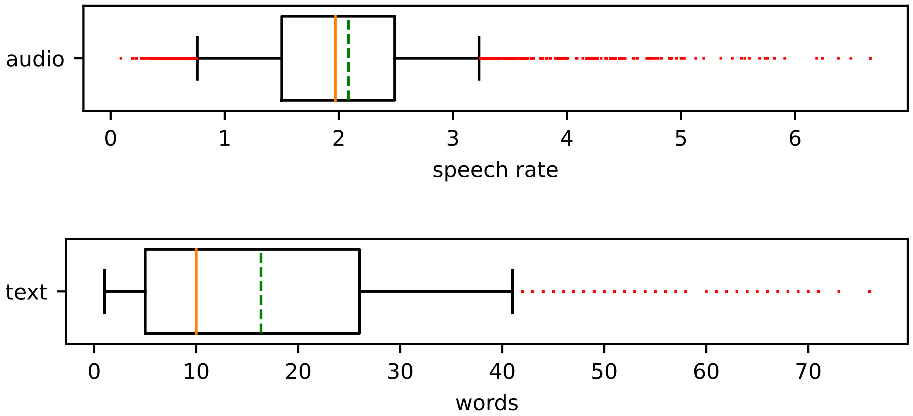

Figure 3 – Distribution of following data properties (from top to bottom): audio speech rate, and transcribed text length in words.

[1] https://commons.wikimedia.org/wiki/File:Emoji_face_rating_scale.png

18 study participants gave answers to 92 PROM questions, two persons were interrupted after 70 and 71 questions respectively due to time limitation. This results in a total of 1,797 voice recordings. The dataset contains diverse types of answers in terms of length and speech rate (words per second). An overview is provided in Table 1 and Figure 3. Before NLU training, all duplicates of audio transcriptions are removed for each answer type.

3.3 Evaluation

To assess the performance of our ASR component, we have manually transcribed 10% of all audio samples and normalized both, ASR predictions and the manually transcribed references, by removing special characters and additional white spaces. We calculated a variety of common metrics to quantify the performance: word error rate (WER), match error rate (MER), word information lost (WIL), word information preserved (WIP), and character error rate (CER) [6].

To assess the contribution of augmented and empirical data towards better NLU performance, and compare the different NLU approaches, we use three setups for training and testing as described in the following.

AUG Train, EMP Test ($n_{train} = 1,526$, $n_{test} = 1,733$): Training is done solely with the augmented data (100%, $n_{train} = 1,526$), and testing solely with the empirically collected data (100%, $n_{test} = 1,733$).

EMP Cross-Validation ($n_{train} = 1,386$, $n_{test} = 347$): Average over five divisions: Training is done with 80% ($n_{train} = 1,386$) of the empirically collected data, and testing with the other 20% ($n_{test} = 347$) of it.

AUG + EMP Cross-Validation ($n_{train} = 2,912$, $n_{test} = 347$): Average over five divisions: Training is done with 80% ($n^{emp}_{train} = 1,386$) of the empirically collected data plus all augmented data (100%, $n^{aug}_{train} = 1,526$). Both together sum up to $n_{train} = 2,912$ samples. Testing is done with the other 20% ($n_{test} = 347$) of the empirical data.

Accuracy and range-1-accuracy are used to quantify the performances of our approaches. Range-1-accuracy measures the rate of predictions that were either correct, or one off, in terms of the ordered answer scale.

4 Results

Empirical data was collected from a diverse group of lab study participants, with an average age of 34.1 (see Figure 4). 55% (11) of participants were male, 40% (8) female and 5% (1) non-binary. Despite all participants speaking proper German, some of them had a different nationality or mother tongue (see Figure 5). Within Germany, participants were raised in Berlin (7), Brandenburg (3), Bavaria (3), North Rhine-Westphalia (2), and Lower Saxony (1). Only one participant had experience in filling out a PROM survey in a rehabilitation clinic, whereas 2 participants had prior working experience in the field of medical rehabilitation.

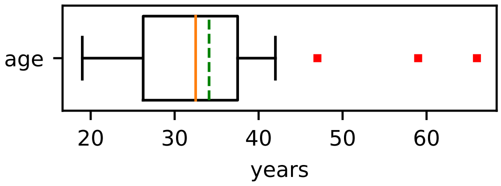

Figure 4 – Box plot describing the participant's age distribution.

We employ the collected samples to perform quantitative evaluation on the ASR and NLU component.

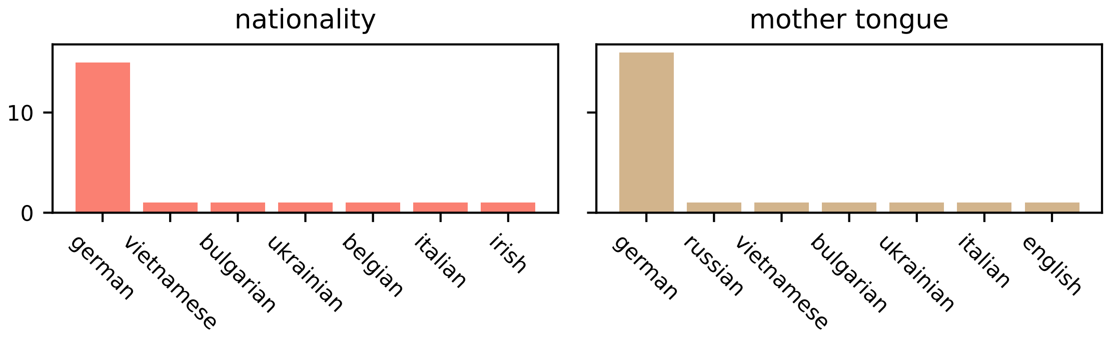

Figure 5 – Distribution of participant's nationalities and mother tongues.

4.1 ASR Performance

We manually transcribed the text and answer gold label of 14 samples for each of the 13 scales, leading to a total of 182 annotated samples (10% of the complete corpus). Table 2 shows similar results for all scales. Overall, the results describe with a WER of 17.1% that most of the spoken input is automatically transcribed correctly. Furthermore, a much lower CER of 6.5% indicates that many not matching words are probably a variation of the respective correct word.

Table 2 – Result scores of several ASR evaluation metrics, for each scale, and for all together.

Scale / Metric	S$_1$	S$_2$	S$_3$	S$_4$	S$_5$	S$_6$	S$_7$	S$_8$	S$_9$	S$_{10}$	S$_{11}$	S$_{12}$	S$_{13}$	*all*
WER	.180	.141	.203	.102	.141	.161	.221	.162	.150	.168	.190	.248	.124	.171
MER	.178	.138	.195	.101	.141	.158	.214	.161	.144	.168	.188	.227	.123	.167
WIL	.273	.233	.308	.183	.228	.255	.345	.276	.231	.288	.296	.324	.210	.269
WIP	.727	.767	.692	.817	.772	.745	.655	.724	.769	.712	.704	.676	.790	.731
CER	.088	.049	.055	.030	.084	.057	.084	.040	.046	.069	.060	.138	.040	.065

4.2 NLU Performance

Table 3 shows performance results for our different NLU approaches and dataset combinations, using augmented (AUG) and empirical (EMP) data (see sections 3.1 and 3.2). With an accuracy of 51.1%, *Many NLU* outperforms the *One NLU* accuracy of 43.1% by 8 percent points, for the best performing training dataset AUG + EMP Cross-Validation. *One NLU Naive* has an even lower accuracy of only 24%.

We received the best results from training with only empirical data, or a combination of empirical and augmented data. Thus, it seems important to use empirical data for the training of medical survey answer classification. Interestingly, adding the augmented data to the empirical collected does only modestly increase the performance (e.g., 0.505 to 0.511 for *Many NLU* in Table 3). Solely relying on it leads, as to be expected, to even worse results.

4.3 Further Analysis on Empirical Data

Table 4 shows the difference between labels made by participants and labels made by the author. A mean accuracy of 67% demonstrates the ambiguity of classifying natural language into PROM answer options. Because the scales are ordered, we use the range-1-accuracy to understand if there is only a small (at maximum 1 off), or a larger difference between both mappings. With this metric, we receive a matching of 96% for the labeling task.

Table 3 – Means and standard deviations of test accuracy and test range-1-accuracy for three different NLU variants, and three different train respectively test datasets.

METRIC	NLU Variant Datasets	One NLU Naive	One NLU	Many NLU
ACC.	**AUG Train, EMP Test**	0.082±0.004	0.327±0.005	**0.393**±0.002
	EMP Cross-Validation	0.246±0.009	0.410±0.013	**0.505**±0.009
	AUG + EMP Cross-Validation	0.240±0.005	0.431±0.012	**0.511**±0.012
RANGE-1-ACC.	**AUG Train, EMP Test**	0.120±0.006	0.625±0.006	**0.658**±0.004
	EMP Cross-Validation	0.383±0.006	0.710±0.011	**0.793**±0.008
	AUG + EMP Cross-Validation	0.362±0.005	0.718±0.012	**0.776**±0.009

Table 4 – Accuracy and Rank-1-Accuracy between manually created labels by participants and manually created labels by authors.

Scale Metric	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	*all*
Accuracy	0.71	0.50	0.79	0.79	0.71	0.71	0.64	0.86	0.43	0.79	0.86	0.71	0.21	0.67
Range-1-Acc.	1.00	1.00	0.93	1.00	0.86	0.86	1.00	1.00	1.00	1.00	0.93	1.00	0.86	0.96

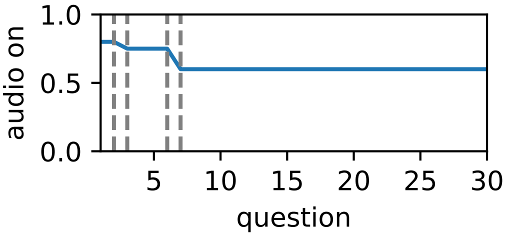

Figure 6 – Visualization of percentage of participants with reading function on per question. After the decline at question 6, there was no more change in the course of all experiments.

For further insights, we investigate the Pearson correlation coefficients of several study measurements with participant age, and find slightly significant results for the speech rate ($r = -0.37$, $p = 0.11$), which was also described by Jacewicz et al. [7]. For audio on, talking time, word count, or total interaction length, there is no significant correlation. Like previous work [7], we could not find any significant difference between gender for the investigated dimensions.

In total, the interaction time was only slightly higher for participants using the reading function, with 17.65 seconds, compared to 15.33 seconds for participants without it. This means that people required nearly as much time to read the item description by themselves, as the reading function takes.

When participants turned off the reading function, they mostly did it at item 6, which has by far the longest item description with 48 words, respectively 336 characters. On average, there are only 14 words, respectively 97 characters, per item description.

5 Conclusion

For evaluation of ASR and NLU of a multimodal system for filling of PROM surveys, we gathered audio input examples from 20 participants with diverse demographics, and 92 PROM items, leading to a total of 1,797 samples. We manually transcribed 10% of the corpus to assess the ASR performance, and check the answer labeling made from participants.

Our ASR seems to work properly on the corpus, with an acceptable WER of 17.1%, and CER of 6.5%.

Incorporating contextual knowledge into the NLU's model architecture with our approach *Many NLU*, significantly outperforms the approach *One NLU*, that incorporates the knowledge after prediction time. The performance suffers even more when the contextual knowledge is not incorporated at any point, like we did in the *One NLU Naive* variant. Further, *Many NLU* introduces the possibility for flexible removal or integration of new scales into a survey classifier.

From the evaluation of different training datasets, we found that empirical data is of benefit for good classification performance on PROM survey answers, in contrast to augmented data. Presumably, this results from our augmented data sticking more strict to the standardized answer within different phrases, and the empirical data being more diverse in formulation of the health status which is more difficult to map. In contrast to our study setting, people would see the textual answer options on a tablet in our multimodal application, and then probably use the standardized descriptions more often. Therefore, we suggest incorporating both, augmented and empirical data for the training of the NLU.

Nonetheless, even the highest accuracy of 51.1% seems too low for a proper application of the classifier in a medical dialogue system. This value can only partly be explained by errors in the ASR (see Section 4.1) and is probably caused by our setting which hides the written answer options to provoke different variants of answer utterances.

Many answer options are semantically very close, or use ambiguous terms. This is why we think the main challenge is grounded in the ambiguity of the task, also resulting in mislabeling (see Table 4). Examples for ambiguous terms are, “einige Schwierigkeiten” (“some difficulties”) where “einige” can mean “ein wenig” (“a little”) or “ziemlich viel” (“quite a lot”) in German. Additionally, participants use ambiguous words or phrases like “relativ zufrieden” (“relatively satisfied”), or do not give enough information for classification, like in “Schon öfter als ich zugeben mag.” (“More times than I care to admit.”). We found that the range-1-accuracy is a good tool to smooth out these ambiguities and understand the performance better.

Another important factor for the low accuracies is the number scale S_{13}, which only gets an accuracy up to 15% (compared to the best scale S_{10} with up to 64.4%). Precise mapping from natural language to these numbers is very ambiguous. For numeric scales, we suggest enforcing survey participants to use numbers within their answer directly.

6 Future Work

For future work, the annotation of the audio corpus and NLU gold labels is to be completed to increase the quality of ASR and NLU assessment. Further, different techniques for the generation of augmented data could be compared. That involves the use of large language models, e.g., as proposed by Stylianou et al. [8].

It could also be tested how models incorporating contextual knowledge via specific neuron connections (instead of a fully-connected layer) perform, although they would be less adaptable towards adding or removing option scales.

To increase accuracy of intent detection in dialogue systems, one common technique is to use some kind of confidence score to decide if a question should be re-asked, optionally with further instruction or narrowing down the answer options to the most likely ones. It should be evaluated which confidence scores, thresholds, and re-asking techniques could lead to an accuracy that is high enough for the application. We think that a range-1-accuracy of 95% would be adequate for our use-case.

More research is necessary to adequately assess prediction performance with new metrics, like range-1-accuracy, for ambiguously posed tasks, like with standardized PROM answer option scales.

Acknowledgement

The presented work has been funded by the Federal Ministry of Education and Research (Germany) under grant no. 16SV9018 for the joint research project MIA-PROM in the research program Interactive Technologies for health and Quality of Life.

References

[1] Kluzek, S., B. Dean, and K. A. Wartolowska: *Patient-reported outcome measures (PROMs) as proof of treatment efficacy*. *BMJ Evidence-Based Medicine*, 27(3), pp. 153–155, 2022. doi:10.1136/bmjebm-2020-111573.

[2] ter Stal, S., L. L. Kramer, M. Tabak, H. op den Akker, and H. Hermens: *Design features of embodied conversational agents in eHealth: a literature review*. *International Journal of Human-Computer Studies*, 138, p. 102409, 2020. doi:10.1016/j.ijhcs.2020.102409.

[3] Cella, D., S. Choi, D. Condon, B. Schalet, R. Hays, N. Rothrock, S. Yount, K. Cook, R. Gershon, D. Amtmann, D. Dewalt, P. Pilkonis, A. Stone, K. Weinfurt, and B. Reeve: *PROMIS ® adult health profiles: Efficient short-form measures of seven health domains*. *Value in Health*, 22, pp. 537–544, 2019. doi:10.1016/j.jval.2019.02.004.

[4] Boumans, R., F. Van Meulen, K. Hindriks, M. Neerincx, and M. Olde Rikkert: *A feasibility study of a social robot collecting patient reported outcome measurements from older adults*. *International Journal of Social Robotics*, 12, 2020. doi:10.1007/s12369-019-00561-8.

[5] Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang: *Language-agnostic BERT sentence embedding*. In S. Muresan, P. Nakov, and A. Villavicencio (eds.), *Proc. of 60th An. Meeting of ACL (Volume 1: Long Papers)*, pp. 878–891. Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.acl-long.62.

[6] Adegbegha, Y. E., A. Minocha, and R. Balyan: *Analyzing multilingual automatic speech recognition systems performance*. In F. Zhao and D. Miao (eds.), *AI-generated Content*, Communications in Computer and Information Science, pp. 191–204. Springer Nature, 2023. doi:10.1007/978-981-99-7587-7_16.

[7] Jacewicz, E., R. A. Fox, C. O'Neill, and J. Salmons: *Articulation rate across dialect, age, and gender*. *Language variation and change*, 21(2), pp. 233–256, 2009. doi:10.1017/S0954394509990093.

[8] Stylianou, N., D. Chatzakou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris: *Domain-Aligned Data Augmentation for Low-Resource and Imbalanced Text Classification*. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pp. 172–187. Springer Nature Switzerland, Cham, 2023. doi:10.1007/978-3-031-28238-6_12.