NLG4Health 2022

The First Workshop on Natural Language Generation in Healthcare

Proceedings of the Workshop

July 18, 2022

The NLG4Health organizers gratefully acknowledge the support from the following sponsors.

Gold

ARRIA
NATURAL LANGUAGE GENERATION

Google

Silver

Bronze

aiXplain

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-63-6

Organizing Committee

Program Chair

Ehud Reiter

Organizers

Emiel Krahmer, Tilburg University
Kathy McCoy, University of Delaware
Ehud Reiter, University of Aberdeen

Program Committee

Erkan Basar, Radboud University, The Netherlands
Tibor Bosse, Radboud University, The Netherlands
Daniel Braun, Twente University, The Netherlands
Barbara di Eugenio, University of Illinois, USA
Albert Gatt, Utrecht University, The Netherlands
Aki Harma, Philips Research, The Netherlands
Saar Hommes, Tilburg University, The Netherlands
Saad Mahamood, Trivago, Germany
Francesco Moramarco, University of Aberdeen, UK
Natalie Parde, University of Illinois, USA
Stefan Pauws, Philips Research, The Netherlands
Thomas Schaaf, CMU, USA

Program Committee

Program Chairs

Emiel Krahmer, Tilburg University
Kathleen McCoy, University of Delaware
Ehud Reiter, University of Aberdeen

Table of Contents

DRIVINGBEACON: Driving Behaviour Change Support System Considering Mobile Use and Geo-information

Jawwad Baig[♡], **Guanyi Chen**[♠], **Chenghua Lin**[♢] and **Ehud Reiter**[♡]
[♡]Department of Computing Science, University of Aberdeen
[♠]Department of Information and Computing Sciences, Utrecht University
[♢]Department of Computer Science, University of Sheffield
r04jb18@abdn.ac.uk, g.chen@uu.nl,
c.lin@sheffield.ac.uk, e.reiter@abdn.ac.uk

Abstract

Natural Language Generation has been proved to be effective and efficient in constructing health behaviour change support systems. We are working on DRIVINGBEACON, a behaviour change support system that uses telematics data from mobile phone sensors to generate weekly data-to-text feedback reports to vehicle drivers. The system makes use of a wealth of information such as mobile phone use while driving, geo-information, speeding, rush hour driving to generate the feedback. We present results from a real-world evaluation where 8 drivers in the UK used DRIVINGBEACON for a period of 4 weeks. Our preliminary results are promising but not conclusive.

1 Introduction

There has been a long tradition of adopting Natural Language Generation (NLG) techniques in health care (Cawsey et al., 1997; Portet et al., 2009; Schneider et al., 2013; Enarvi et al., 2020). One line of work focus on building Behaviour Change Support Systems (BCSSs) to help people live healthier and more safely. These include systems for encouraging people to stop smoking (Reiter et al., 2003), for ecological driving (Endres et al., 2010; Boriboonsomsin et al., 2010; Tulusan et al., 2012), and for safer driving (Braun et al., 2015, 2018)[1]. Such BCSSs can generate feedback automatically based on users' current behaviours by employing NLG techniques.

Within the domain of safe driving, personalised feedback via postal mail has proved to be useful to improve users' driving habits (Ouimet et al., 2004; Lefèvre et al., 2015). For example, `DriveSafe` (Bergasa et al., 2014) is a mobile application that utilises data from vehicle cameras

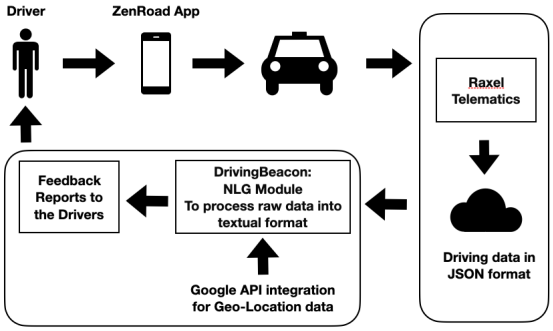

Figure 1: DRIVINGBEACON System Design

combined with GPS and audio data from the mobile phone to identify unsafe driver behaviours. `DriveSafe` estimate a driving score for each driver and then provides alerts when the score crosses a certain threshold. Eco-Driving (Allison and Stanton, 2019) is a study about reducing gas emissions arising from bad driving styles. Minimising unnecessary acceleration and braking can improve eco-driving and fuel consumption and eventually reduce emissions. The study also highlights that despite the benefits of eco-driving, drivers also require feedback about their actions in order to promote long-term behavioural change. Another study by Jannusch et al. (2021) investigated the high fatality rate amongst Young Novice Drivers and their use of mobile phones while driving. A survey among 700 young drivers was conducted, where they compared distracted driving behaviour. They focused on participants' use of smartphones during driving and found that most of those uses are music-related activities (e.g., playing the next song or increasing the volume).

Braun et al. (2018) built SAFERDRIVE, the first NLG based driving BCSS, which generates weekly textual driver feedback from telemetric data[2] and the feedback is delivered through mobile phones.

[1]The United Nations considers unsafe driving to be a health issue and lists the target of fewer road traffic accidents as a health goal for sustainable development. See: `https://sdgs.un.org/goals/goal3`.

[2]The data was gathered by a mobile phone app to track individual driving styles.

Proceedings of the First Workshop on Natural Language Generation in Healthcare (NLG4Health), pages 1 - 8

It was reported that the generated textual feedback is more helpful to drivers than the traditional score-based and map-based feedback, especially to learners and young drivers (Braun et al., 2015). This is because textual feedback gives drivers a more concrete idea of how to change their driving behaviours. For example, for a speeding incident, SAFERDRIVE could generate feedback such as "*You speeded 7 times on roads with 20 mph speed limit and 12 times on roads with 30 mph speed limit*".

This paper introduces DRIVINGBEACON which is able to make use of richer information compared to SAFERDRIVE aiming at generating better feedback reports. Vital additional information includes (i) the mobile phone use information of drivers during driving; and (ii) geofencing, which highlights driving incidents that take place near crowded places such as schools, mosques, superstores, etc.

To assess our DRIVINGBEACON, we conduct a real-world A/B test on 8 drivers in UK. This is not a lot of drivers, but it is more than the 6 drivers used by Braun et al. (2018). Concretely, we generated basic feedback (not considering rich information) and enhanced feedback (considering rich information) using DRIVINGBEACON. We divided our 8 drivers into two groups, one of which is sent the basic feedback while the other is sent the enhanced feedback. The experiment lasted for 4 weeks, during which we monitor the change in their driving behaviours. To summarise, the key contributions of our work are two-fold:

1. We designed and implemented the DRIVINGBEACON system which makes use of the mobile use information and Geo-information in addition to telemetric data;

2. We evaluated DRIVINGBEACON through a four-week period A/B test on 8 real drivers.

2 System Design of DRIVINGBEACON

We implemented DRIVINGBEACON using Java and connected it to two third-party APIs to acquire required information: the Google Map API[3] and the Damoov API[4]. Figure 1 shows our system architecture.

ID	Driving Behaviour
1	Brake and Acceleration
2	Speeding
3	Speeding near crowd areas
4	Using mobile while driving

Table 1: Driving behaviours that DRIVINGBEACON monitors

Specifically, Damoov Telematics uses a mobile phone application called ZenRoad[5] to collect driving behaviour related information. It collects driving data from embedded sensors in the mobile device, such as gyroscope, GPS and accelerometer. This data is then uploaded to the data hub of Rexel Telematics.

DRIVINGBEACON pulls raw data from the data-hub using the Damoov API and extracts related Geo-information using Google MAP API. With these data, we use a rule-based NLG system to generate feedback reports.

3 Feedback Generation

We classify the information we obtain into two sets: one contains the information that also has been used in SAFERDRIVE (Braun et al., 2018) (henceforth basic information), including information such as location, speed, speed limit, time, etc. The other contains additional information, including mobile phone usage, geo-fencing (driving near crowded places), traffic law and penalty points. Basic feedback reports include only the first set, while Enhanced feedback reports include both sets.

We also list the driving behaviours that DRIVINGBEACON captures in Table 1. DRIVINGBEACON will detect driving behaviours based on the information it collects and generates feedback accordingly.

3.1 Basic Feedback Report

Without the additional information, we generate what we call the *basic feedback*. It uses similar parameters as Braun et al. (2018). Since the system access only basic information, it detects limited types of driving behaviours (i.e., only the first and second driving behaviours in Table 1). Based on the detected behaviours, it generates basic feedback,

[3] https://developers.google.com/maps; Terms of Service: https://developers.google.com/maps/terms-20180207

[4] Damoov - Mobile Telematics as a service, www.telematicssdk.com; License: https://docs.damoov.com/docs/license

[5] tinyurl.com/ZenRoadApp

Basic Report	Enhanced Report
Last week, your total number of driving incidents was nine, including speeding on Low Rd, Grantham. Your speed was 51mph on a 30mph road. Remember that fast driving can cause serious accidents and will lead to points on your driving licence and fines of up to 150% of your weekly income. When driving, a few miles per hour can mean the difference between life and death. The total number of braking incidents was two; your braking counts are less than five; Well done! Acceleration incidents were two; your acceleration counts are less than five, Keep it up! Unnecessary acceleration and harsh braking can impact fuel costs and car maintenance costs.	Last week, your total number of driving incidents was nine;including mobile phone usage on Tuesday, 4th May at 8:56 AM. You used a mobile phone while driving near Helmsley Rd, Grantham. It was during rush hour where distracted driving could have caused a serious accident with up to 6 penalty points and a £1,000 fine. On Wednesday, 5th May at 3:27 PM, you exceeded the speed limit near a crowded place; the location was Barrowby Preschool, Low Rd, Grantham. Schools, mosques, train stations and superstores are sensitive and often crowded zones. Your speed was 51mph on a 30mph road. Remember, driving fast near a crowded place can cause a serious accident and may lead to points on your driving licence and fines of up to 150% of your weekly income. On Wednesday, 5th May at 8:30 AM, you drove at extreme speed near Alberic cottage, Low road, Grantham. Your speed was 50mph on a 30mph road. Remember, when driving, a few miles per hour can mean the difference between life and death. Unnecessary acceleration and harsh braking can impact fuel costs and car maintenance costs. Last week, your total number of braking incidents was two, the total count is less than five; Well done! You did acceleration near a crowded place on Wednesday 5th May at 3:27 PM, however, your acceleration counts are less than five, Keep it up!

Table 2: Example of basic feedback and enhance feedback (difference highlighted in blue)

which tells drivers about speeding, road speed limits, unnecessary acceleration and harsh braking. For example, in the example basic feedback in Table 2, with the information about location, speed and speed limit, the system detected that the driver overspeeded on Low Rd, Grantham and generated a message about both the detail of this poor driving behaviour (i.e., "*... speeding on Low Rd, Grantham. Your speed was 51mph on a 30mph road.*") and its consequences (i.e., "*Remember that fast driving can cause serious accidents and will lead to points on your driving licence and fines of up to 150% of your weekly income ...*")

3.2 Enhanced Feedback

With both the basic and additional information, DRIVINGBEACON generates *enhanced feedback* reports, an example of which is shown in Table 2. This additional information can help the system detect more kinds of poor driving behaviours (see Table 1) and can be useful to drivers to understand where and when they did drive unsafely and the potential impact on them and others of their unsafe driving. It highlights the dangers of the incidents.

We use geofencing (illustrated in Figure 2) to identify regions near schools and other sensitive or crowded areas (e.g., shopping malls, hospitals). We highlight to drivers unsafe driving within a geofenced area as it is more likely to result in incidents in these areas compared to less crowded places. For example, in addition to tell the driver that s/he was speeding, the enhanced feedback add a message that the speeding happened near a school

Figure 2: Geo-fencing around sensitive areas

(i.e., "*you exceeded the speed limit near a crowded place; the location was Barrowby Preschool*").

Mobile phone use is classed as distracted driving and a major cause of serious accidents (Jannusch et al., 2021). Due to the high usage of mobile phones these days, we included mobile phone usage in our feedback reports, as shown in Table 2: "*You used a mobile phone while driving near ...*".

Additionally, the Enhanced report also adds the detail of when each incident happened (e.g., "*... mobile phone usage on Tuesday, 4th May at 8:56 AM.*").

3.3 Hypothesis

Having two types of feedback reports, we show them to two groups, i.e. the Basic Group with basic feedback and the Enhanced Group with enhanced feedback report. We define our hypotheses for this experiment as follows:

1. Across basic and enhanced groups, there will be fewer incidents per mile of bad driving[6] at

[6]"Bad driving" means behaviour such as over speeding,

the end of the experiment (week 4) compared to the beginning (week 1).

2. Bad driving incidents per mile will reduce more in the Enhanced group than the Basic group, looking at all weeks (not just week 4).

4 Evaluation

In order to evaluate our system in a real-world scenario, we conducted a short longitudinal study where we evaluated the system with real drivers.

4.1 Materials and Participants

On 30th April 2021, we started a field experiment that lasted for 30 days. Eight participants (including both males and females) between the age of 20 and 45 were given the ZenRoad app. There were no incentives given related to driving performance. On the consent form, we explained that we would collect their driving data through the ZenRoad app and generate textual feedback reports with the intention of helping drivers to improve their driving habits. To protect the anonymity of the users, Damoov provided us with Data-Hub and Device and Track IDs (a unique number created for each driver and each trip), where all personal identifiers were removed. This anonymised data was then used for our analysis and feedback reports.

We divided the eight drivers into two groups (four drivers in each group): the Basic Group who received the basic feedback report and the Enhanced Group who received the enhanced feedback report. They drove a total of 3,179 miles around the UK with 239 trips in total. From driving data, we calculated the driving incidents per mile (I/M), an indicator for measuring drivers' relative performance. Basic group drove a total of 963 miles and did an average of around 0.098 driving I/M, while the Enhanced group drove a total of 2216 miles (enhanced group drivers went on long drives over the weekend hence more mileage as compared to basic group)[7] and did an average of 0.014 driving I/M. Throughout the experiment, we have noticed a decline in I/M in both groups but the enhanced group improved in their driving behaviours more quickly as compared to the basic group. Driver's feedback about Enhanced reports shows that it has made an impact on their driving behaviours during this experiment (see Section 4.3 below).

Our system monitored the following types of incidents: speeding, harsh braking, acceleration, mobile use while driving, unsafe driving near sensitive zones. These were highlighted in the Enhanced reports for the drivers.

Group	Week1	Week4	p-value
Basic	0.22	0.07	.321051
Enhanced	0.05	0.01	.243089

Table 3: Numbers of I/M; Week 1 vs Week 4

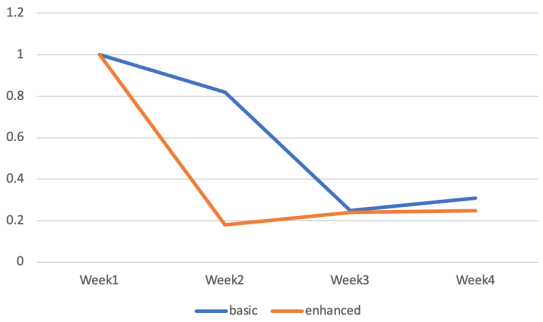

Figure 3: The number of incidents per mile (I/M) was normalised by that of week 1 in all four weeks.

4.2 Driving Behaviour Change

Table 3 shows the I/M of the drivers in the two groups. We conduct paired t-test to compare the incident per mile ratio in week 1 (i.e., the beginning of our experiment) and week 4 (i.e., the end of our experiment) for both Basic and Enhanced groups, where the results are shown in Table 3. It can be observed that numerically, there were fewer incidents in Week 4 than in Week 1. However, the difference is not significant and hence our Hypothesis 1 is not supported. We also like to mention here that, coincidentally, the drivers in the Enhanced group did fewer incidents in Week 1 as compared to the Basic group, which again might be attributed to the scale of our experiment.

To validate the second hypothesis, we quantified how much was I/M reduced by normalising the I/M of each week using that of week 1. The results are presented in Figure 3. We conducted a t-test on the results, but, unfortunately, there is no significant difference between the Enhanced group and the Basic group ($p > 0.05$). This embodies that our second hypothesis is also rejected. Nonetheless, we found that the enhanced report affects drivers much

harsh braking, speeding near crowded places etc.

[7]Reason for high miles driven by the Enhanced group is two out of four drivers went onto long journeys over the weekends.

Metric	Basic	Enhanced
Usefulness	4.00	4.17
Readability	4.50	4.17
Intervention	3.00	3.33

Table 4: Average scores for the human evaluation.

faster than the basic report since, as we can see from Figure 3, the largest decline of the enhanced group happened in the second week whereas that of the basic group happened in the third week. This suggests that enhanced reports are more efficient basic reports. More importantly, higher efficiency often results in fewer accidents in total.

4.3 Human Evaluation of the Feedback

With the weekly feedback report, we asked the participants to rate each generated report in the following categories on a scale from 1 (Strongly disagree) to 5 (Strongly agree): (1) **Usefulness**: *the feedback is useful to you*; (2) **Readability**: *the content is easy to understand*; and (3) **Intervention**: *the feedback has intervened your bad driving behaviour*.

Table 4 charts the results of the human evaluation. We found that the enhanced reports were rated higher in usefulness and intervention whereas the basic reports achieve higher readability, although no significant difference can be established on any criterion.

4.4 Feedback of the Generated Reports

At the end of the trial, we showed both the basic and enhanced feedback reports to all participants and asked about their opinion of this experiment and approach. The feedback was overall positive. Participants understood the idea and liked the approach where they can see their driving styles with details of their journeys, day and time when they make mistakes and most importantly the locations. Two drivers explicitly said they preferred the Enhanced report and no drivers said they preferred the Basic report. All comments are shown in Appendix A. Encouragingly, some of the subjects told us months after the experiment that they are still driving more carefully because of their experience with DRIVINGBEACON, even though they no longer use the system.

5 Conclusion

We presented DRIVINGBEACON, a behaviour change support system which can generate enhanced feedback reports by utilising the mobile use information and Geo-information in addition to telemetric data. Experimental results suggest that enhanced reports are more effective than the basic reports, although the difference is not significant. In the future, we plan to make the enhanced reports more effective by personalising feedback reports for individual drivers based on their interests, background, and driving history. We also plan to test our system in a larger-scale experiment with regard to both the number of participants and the duration of the experiment.

Acknowledgements

We thank Damoov for their help in setting up the experiment. We also thank the anonymous reviewers for their helpful comments.

Ethical Considerations

The current study has been approved by the ethics review board of the University of Aberdeen. The use of each API and Software (i.e., Google Map API, Damoov API, and ZenRoadAPP) is consistent with its licence or terms of use.

References

Craig K. Allison and Neville A. Stanton. 2019. Eco-driving: the role of feedback in reducing emissions from everyday driving behaviours. *Theoretical Issues in Ergonomics Science*, 20(2):85–104.

Luis M. Bergasa, Daniel Almería, Javier Almazán, J. Javier Yebes, and Roberto Arroyo. 2014. Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 240–245.

Kanok Boriboonsomsin, Alexander Vu, and Matthew Barth. 2010. Eco-driving: pilot evaluation of driving behavior changes among u.s. drivers. *UC Berkeley: University of California Transportation Center*.

Daniel Braun, Ehud Reiter, and Advaith Siddharthan. 2015. Creating textual driver feedback from telemetric data. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 156–165.

Daniel Braun, Ehud Reiter, and Advaith Siddharthan. 2018. Saferdrive: An nlg-based behaviour change support system for drivers. *Natural Language Engineering*, 24(4):551–588.

Alison J Cawsey, Bonnie L Webber, and Ray B Jones. 1997. Natural language generation in health care. *Journal of the American Medical Informatics Association*, 4(6):473–482.

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*, pages 22–30, Online. Association for Computational Linguistics.

Christoph Endres, Jan Miksatko, and Daniel Braun. 2010. Youldeco-exploiting the power of online social networks for eco-friendly driving. In *Adjunct proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2010)*, page 5.

Tim Jannusch, Darren Shannon, Michaele Völler, Finbarr Murphy, and Martin Mullins. 2021. Smartphone use while driving: An investigation of young novice driver (ynd) behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour*, 77:209–220.

Stéphanie Lefèvre, Ashwin Carvalho, Yiqi Gao, H Eric Tseng, and Francesco Borrelli. 2015. Driver models for personalised driving assistance. *Vehicle System Dynamics*, 53(12):1705–1720.

MC Ouimet, TG Brown, JP Bédard, and J Bergeron. 2004. Impact of mailed feedback on speeding behaviours of convicted male drivers: A brief intervention. In *International Conference on Traffic and Transport Psychology*. Citeseer.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Ehud Reiter, Roma Robertson, and Liesl M Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58.

Anne Schneider, Alasdair Mort, Chris Mellish, Ehud Reiter, Phil Wilson, and Pierre-Luc Vaudry. 2013. MIME - NLG in pre-hospital care. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 152–156, Sofia, Bulgaria. Association for Computational Linguistics.

Johannes Tulusan, Thorsten Staake, and Elgar Fleisch. 2012. Providing eco-driving feedback to corporate car drivers: what impact does a smartphone application have on their fuel efficiency? In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 212–215.

A Feedback Examples

In what follows, we list some feedback examples, where feedback from the basic group is marked as B and feedback from the enhanced group is marked as E.

Group	**Give us your thoughts about the feedback approach ?**	**Will you use mobile phone app to improve your driving behaviours?**
B	*"I think the consistent feedback is encouraging; I have 4 kids of different age groups and in different schools so my speeding incidents are due to my rush hour driving as I have to drop my kids to three different schools. By September, the youngest will join the same school as my other kids and my daughter will join the secondary school which will reduce a lot of extra driving in the morning and afternoons."*	*"yes I like the idea. I am very busy but an extra app on my mobile which can show my driving behaviours will not harm. After being part of this experiment, I am thinking of changing my car insurance to a company which calculates annual premiums based on driving styles. I think having an incentive attached to this process can definitely change my driving behaviours."*
B	*"Its a good method, it tells me how I drive. Reading textual paragraph gives you a good idea; however, it should be shorter or may be only regarding extreme speeding related incidents which could fit in a mobile phone notification or an SMS. This process regularly will help me improve my driving behaviours. We receive M&S and Next clothing related promotional SMS messages every weekend, why not if an app on my phone generate an SMS or a notification with a report about the most dangerous driving incident of the week with location and time. I think it will definitely encourage me to change my style over time."*	*"Yes, definitely. I want to improve."*
B	*"The feedback report process is like a reminder to me to behave. But it didn't manage to connect me to this process so that it can sit in the back of my mind all the time when I drive. There should be some kind of incentive attached to this process. Like a reward or make me feel like I am part of saving the world. If there is a week when I did not have any incidents or bad driving, the app or process should share this to my friends and family that I am part of some noble cause or I receive a certificate or title or money."*	*"yes I will use a mobile phone app to improve my behaviours."*
B	*"I did not completely agree that the feedback report gives me a good summary of my driving. A enhanced version with lot more details and locations will be better. The basic reports which I have received were good but I was confused when and where did I do that speeding. Definitely, detailed summary of driving but not too lengthy."*	*"Yes, I like the idea and I think I have improved a lot in the last 4 weeks. I might keep the ZenRoad app for looking at my driving scores."*

Group	**Give us your thoughts about the feedback approach ?**	**Will you use mobile phone app to improve your driving behaviours?**
E	*"For me, feedback approach did work. Its a reminder for me to be careful while driving. In the last 4 weeks, I have tried my best to understand my driving behaviours through these reports and noticed that I should be extra careful at motorway. The reports shows that I did a lot of speeding during long journeys and I should change that habit."*	*"Yes, I will use mobile phone app to improve my driving behaviours."*
E	*"The textual feedback was concise and to-the-point and that was informative so it was quite a useful part of the feedback. The eco score was also simple enough to understand, and was useful. Risk score chart can be simplified a bit by adding some more explanation around it, or a simple kind of pie chart or something similar can be used instead. The rest was good."*	*"Maybe"*
E	*"Possibly weekly prompts on the app like, "Here is your weekly progress report on your driving". Similar to how screen-time reports work on IOS."*	*"Yes, it's instant and I can get instant feedback."*
E	*"I like the enhanced reports but its bit lengthy though."*	*"Yes, I want to see my driving insights."*

In-Domain Pre-Training Improves Clinical Note Generation from Doctor-Patient Conversations

Colin A. Grambow **Longxiang Zhang** **Thomas Schaaf**
3M Health Information Systems
cgrambow,lzhang28,tschaaf@mmm.com

Abstract

Summarization of doctor-patient conversations into clinical notes by medical scribes is an essential process for effective clinical care. Pre-trained transformer models have shown a great amount of success in this area, but the domain shift from standard NLP tasks to the medical domain continues to present challenges. We build upon several recent works to show that additional pre-training with in-domain medical conversations leads to performance gains for clinical summarization. In addition to conventional evaluation metrics, we also explore a clinical named entity recognition model for concept-based evaluation. Finally, we contrast long-sequence transformers with a common transformer model, BART. Overall, our findings corroborate research in non-medical domains and suggest that in-domain pre-training combined with transformers for long sequences are effective strategies for summarizing clinical encounters.

1 Introduction

Necessitated by electronic health records (EHR), physicians spend a large amount of time on documentation work (Sinsky et al., 2016), which contributes significantly to burnout (Wright and Katz, 2018; Kumar and Mezoff, 2020), may result in lower job satisfaction (Shanafelt et al., 2016), and can even increase the likelihood of errors and reduce the quality of patient care (Panagioti et al., 2017). To alleviate some of the burden on physicians, medical scribes are often used to summarize recordings or transcriptions of doctor-patient conversations into clinical notes. While this essential, yet tedious process may enable more effective clinical care, it shifts the burden onto medical scribes. Furthermore, the continued reliance on human experts is expensive and only scalable to a limited degree.

Natural language generation models, such as the ones developed in this paper, have the potential to significantly reduce the documentation burden by providing suggested clinical notes to physicians or scribes nearly instantaneously. While still somewhat error-prone and not yet fully automated, these models are able to focus on much of the relevant information in doctor-patient conversations and distill it into a human-readable format for further review by trained medical professionals.

Pre-trained transformer models have revolutionized the field of natural language processing (Radford et al., 2019; Devlin et al., 2019; Lewis et al., 2020) and have already been applied to various medical tasks (Lee et al., 2019; Li et al., 2020; Zhang et al., 2021; Yalunin et al., 2022). Nonetheless, medical conversation summarization continues to present challenges due to its idiosyncrasies, foremost of which is the requirement to contain all relevant medical information rather than summarizing every part of a conversation. Additionally, specialized medical vocabulary renders the use of conventional pre-trained models difficult.

Additional phases of in-domain pre-training have shown to be useful across a wide variety of domains and tasks (Gururangan et al., 2020), but limited work has been done on in-domain pre-training using unlabeled doctor-patient conversations. To address this, we leverage a doctor-patient conversation dataset described in Section 3 to investigate two different pre-training methodologies using BART, LED, and DialogLED transformer models (Section 4). We fine-tune all models on a subset of medical conversations with human-written summaries (Section 4.2) and contrast them with a baseline of models that are not pre-trained in the medical domain using several different evaluation methods, including a transformer-based model for clinical concept extraction (Sections 4.3 and 5). We show that our methods improve the performance on the medical summarization task and also evaluate the additional benefit of using models designed to work with long sequences (Section 6).

Proceedings of the First Workshop on Natural Language Generation in Healthcare (NLG4Health), pages 9 - 22

2 Related Work

Medical summarization Recent research has devoted significant attention to the problem of summarizing medical encounters and documents in an automated fashion. Finley et al. (2018) describe a fully automated medical scribe using a combination of RNN and rule-based approaches to automatically recognize speech, convert it into a transcript, extract the relevant information, and convert it to a report. However, they omit any examples and results and mention that the scribe is still limited in its utility.

Since then, several deep learning approaches have been developed to summarize doctor-patient conversations. Joshi et al. (2020) develop a modified pointer-generator (PG) network to summarize local snippets. Furthermore, they explicitly model negation, which can cause difficulties for automatic approaches. Interestingly, they report that transformer models did not work well, which is contrary to the findings in a lot of subsequent research. Yim and Yetisgen (2021) also use a PG model to perform the similar task of sentence alignment and snippet summarization. Notably, they achieve good results using only a very small dataset. Krishna et al. (2021) take on the challenging task of generating complete clinical summaries (SOAP notes) using various LSTM, PG, and transformer models. They extract important utterances, cluster them, and then generate single-sentence summaries of each cluster. Enarvi et al. (2020) use a large dataset of doctor-patient conversations generated using automatic speech recognition to train a combined transformer-PG model from scratch. They are able to handle somewhat longer input because they do not rely on pre-trained transformer models. As an alternative approach to handle long conversations, Zhang et al. (2021) use a pre-trained BART model with a two-stage chunking approach to generate summaries for a section of the clinical notes.

Related to the summarization of doctor-patient conversations, other research has explored the summarization of clinical notes and clinical history. Zhang et al. (2018) use a PG network to summarize radiology findings and found that incorporating additional information in the form of background information about the patient improves the results. Yalunin et al. (2022) construct a model using a Longformer encoder with a BERT decoder to generate parts of discharge notes from the patient history. They pre-train BERT and Longformer on domain-specific data and create a custom tokenizer, which yields strong results.

Domain shift An intrinsic problem with using pre-trained models is that the domain of the pre-training data is often significantly different from that of the target medical domain. PG networks during fine-tuning can be helpful because they are able to copy words from the new vocabulary, but starting from a model in a domain that is closely related to that of the fine-tuning task would provide additional benefit. Gururangan et al. (2020) show that a second round of pre-training in a domain related to the fine-tuning task can provide significant benefit even if the continued pre-training only uses the unlabeled training set for a given task. They investigate this across a broad range of domains and classification tasks. Similarly, Hsu et al. (2021) find that in-domain pre-training improves learning speech representations. Zhong et al. (2021) show that improved summarization results are possible by continuing pre-training in the (non-medical) conversation domain. As already mentioned previously, Yalunin et al. (2022) use in-domain pre-training very successfully for generating discharge notes from patient histories.

Instead of pre-training all model parameters in the new domain, there has been some investigation into learning small extension modules instead, which can be helpful if there are limited pre-training data or if complete model training is too costly. Tai et al. (2020) adapt BERT to the medical domain by creating an additional vocabulary and adding a corresponding embedding layer. They compose their extension module as a weighted summation of the embedding vectors from the original and the extension layers and demonstrate that this method is very effective at adapting to the new domain.

3 Dataset

The dataset we use has already been described by Zhang et al. (2021) and is composed of 83 605 clinical encounters involving doctors from many different specialties, patients, and potentially other speakers, *e.g.*, nurses and caregivers. For each encounter, we use the de-identified doctor-patient conversation transcribed by a human. The median number of tokens in a conversation is 2040 (using the BART byte-pair encoding from Lewis et al., 2020), and there are a total of 203M tokens in the

entire dataset.

Annotations are available for a subset of 1342 conversations in the form of medical summaries across internal medicine and primary care specialties. Each conversation was summarized by multiple professional medical scribes into several sections, of which we only use the History of Present Illness (HPI) section for this paper due to its complexity and because it is usually written in complete sentences. There are an average of 17 reference summaries per doctor-patient conversation. The median conversation length in the summarization subset is 1334 tokens for a total of 2.5M tokens. The 95$^{\text{th}}$ percentile corresponds to approximately 5120 tokens, which is the length limit we use for our long-sequence models.

For pre-training, we exclude the entire subset of data that we have summaries for in order to avoid any data leakage and potentially biased results. In addition, we split off a random 5% of the remaining conversations as the validation dataset to monitor during pre-training.

For fine-tuning, we attempt to remove poor summaries for a given conversation using an in-house rule-based system to extract medical concepts from the training summaries and only keeping the summary with the most concepts. Even though this results in fewer labeled data, we have not observed a significant drop in performance. Nonetheless, we keep all reference summaries for the test data. After splitting and removing extraneous summaries from training and validation data, we end up with 939, 201, and 202 conversations; and 939, 201, and 3450 summaries in the training, validation, and test sets, respectively.

4 Methods

All methods are based on pre-training and/or fine-tuning of BART (Lewis et al., 2020), Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), and DialogLED (Zhong et al., 2021). BART is a pre-trained encoder-decoder transformer model designed for fine-tuning on text generation tasks, such as summarization. However, it can only encode up to 1024 tokens in both its encoder and decoder, which is less than the median sequence length in our fine-tuning dataset. LED and DialogLED can handle significantly longer input sequences (we use 5120 and 1024 tokens for their encoders and decoders, respectively) by employing a combined global and local attention mechanism which scales linearly with sequence length. The LED architecture is almost identical to that of BART except that the position embeddings of BART are copied 16 times to enable longer input. The parameters of LED are initialized from BART and no additional pre-training was done. DialogLED is initialized from LED and further pre-trained on long dialog data using a window-based denoising task specifically designed for conversations, which results in significant improvement for long-dialog summarization.

We initialize and train all of our models using the pre-trained BART, LED, and DialogLED models available in the Hugging Face Transformers library (Wolf et al., 2020). We use the corresponding tokenizers (all of which use the BART/GPT-2 byte-pair encoding with a vocabulary size of 50 265), but we add additional speaker tokens, *e.g.*, `[DR]:`, `[PT]:`, etc. We investigate both the base and large models (140M vs. 400M parameters, respectively). Except for the additional position embeddings, all base models and all large models have the same number of parameters.

4.1 Pre-training

We investigate two types of pre-training with doctor-patient conversations: BART-style denoising using the entire input as described by Lewis et al. (2020) and DialogLED-style window-based denoising as described by Zhong et al. (2021). We found that sentence and turn permutation are always detrimental to the downstream summarization of our doctor-patient conversations as measured by a decrease in ROUGE scores, so we only perform text infilling for BART-style pre-training and we only use speaker masking, turn splitting, turn merging, and text infilling for window-based pre-training. The other denoising hyperparameters are identical to those used in the original papers. For BART-style denoising, we discovered that it is beneficial to allow the attention mechanism to attend to the additional padding tokens that are added as a result of the text infilling. We hypothesize that this could imply that adding noise to the entire input is too "difficult" of a pre-training task so that some additional information is necessary in the form of the padding tokens, but we leave the further investigation of this observation to a future study.

For all models, we split the conversations into chunks of 1024 tokens for BART-style pre-training,

and we simply truncate long conversations at 5120 tokens for window-based denoising with LED and DialogLED. The number of epochs that each model is pre-trained for is chosen to achieve optimal performance on the downstream summarization task. For the large models, this results in less than one full pass across the pre-training dataset being required. All of our pre-training hyperparameters are shown in the Appendix in Table A.1.

4.2 Fine-tuning

Our fine-tuning task is training a text generation model to summarize doctor-patient conversations into coherent HPI summaries containing all relevant medical information. As with pre-training, we use a decoder sequence length of 1024 tokens and encoder sequence lengths of 1024 tokens for BART and 5120 tokens for LED and DialogLED. 5120 tokens corresponds to the 95$^{\text{th}}$ percentile of conversations in the summarization dataset, which allows us to encode the full length of the majority of the conversations when using LED and DialogLED. Other than that, we maintain consistency across all other fine-tuning hyperparameters (see Table A.2) for all of our models. We train for a maximum of 30 epochs with a batch size of 8 and evaluate every 50 steps. We perform evaluation by using the validation data input to generate text using beam search and monitor the geometric mean of ROUGE-1 F1, ROUGE-2 F1, and ROUGE-L F1 scores on the validation data. We stop training if the validation score has not improved over the last five evaluation calls and save the best model checkpoint.

4.3 Evaluation

In order to rapidly estimate performance across all reference and generated summaries, we employ several automatic evaluation methods. In addition to ROUGE and UMLS concept-based evaluation, which have been used previously in the literature, we also suggest a named entity recognition model as a second form of concept-based evaluation due to the ease of fine-tuning such a model on publicly available data.

4.3.1 ROUGE

We use the `rouge-score` package[1] to compute ROUGE scores, which aims to replicate results from Lin (2004). While there are some issues with using ROUGE for abstractive summarization (Kryscinski et al., 2020), especially with regard to hallucination (Maynez et al., 2020), it is a useful metric to assess the degree of overlap between reference and generated summary. As there are multiple reference summaries per conversation in the fine-tuning test set, we first compute the ROUGE scores of a generated summary with all of its corresponding reference summaries for a single doctor-patient conversation and then average each score. To obtain an aggregate ROUGE score, we can then average the scores across all conversations.

4.3.2 Clinical concepts

As ROUGE measures word overlap indiscriminately, it takes into account unimportant words and is not as suitable for measuring semantic overlap. Therefore, it is beneficial to quantify additional metrics that are not as prone to these issues and focus more on the relevant medical content of a summary.

UMLS concept extraction The methodology described in this paragraph is largely identical to the evaluation described by Zhang et al. (2021). The Unified Medical Language System (UMLS) (Bodenreider, 2004) is a large database of medical concepts and relations between them. We use the approximate string matching algorithm implemented in QuickUMLS (Soldaini and Goharian, 2016) to extract strings from our summaries and match them to concepts in the UMLS database. However, this approach sometimes mislabels irrelevant strings as medical concepts. To mitigate this somewhat, we first aggregate and filter concepts from all reference summaries for a given conversation by only keeping a concept if it occurs in at least three reference summaries or if it occurs in all reference summaries if there are fewer than three. We then extract the UMLS concepts for the generated summary and compute precision, recall, and F1-score. Aggregate scores are averaged across all conversations.

Transformer-based clinical concept extraction (NER) To further deal with the limitations of QuickUMLS, such as the extraction of irrelevant strings from a summary, we train a deep learning model to extract clinical concepts instead. For this, we follow the clinical concept extraction approach by Yang et al. (2020). We use their RoBERTa (Liu et al., 2019) model pre-trained on MIMIC-III clinical notes (Johnson et al., 2016) to fine-tune a named entity recognition (NER) model on the i2b2 2010 dataset (Uzuner et al., 2011), which is a large col-

[1] https://github.com/google-research/google-research/tree/master/rouge

lection of clinical notes annotated with three types of medical concepts.

First, we mostly reproduce the strong classification performance reported by Yang et al. (2020) using the conventional i2b2 2010 train-test split and a conditional random field layer on top of the transformer model (see Table B.1 in the Appendix). After verifying that this approach is successful, we train our final clinical concept extraction model on all i2b2 2010 data for use on our summaries.

To automate the NER-based concept evaluation, we map the extracted entities to UMLS concept unique identifiers (CUIs) using QuickUMLS (there are frequently multiple CUIs per entity) and drop any entities that cannot be mapped. We combine entities that are of the same type (as predicted by the NER model) and have overlapping sets of UMLS CUIs. Similar to the QuickUMLS-only approach, we only keep reference summary entities if they occur in at least three reference summaries for a given conversation. Finally, we compute precision, recall, and F1-score. For this, we define a true positive as a concept extracted from the generated summary where its predicted type matches that of a concept extracted from the reference summaries and there exists an intersection between the sets of UMLS CUIs corresponding to the concepts. False positives and false negatives are defined accordingly.

5 Experiments

We establish baselines by fine-tuning base and large versions of vanilla BART, LED, and DialogLED models on the doctor-patient conversation summarization dataset as described in Section 4.2, *i.e.*, using the versions of those models that are pre-trained as described in their original papers. To assess whether a second round of pre-training on in-domain data is beneficial, we continue pre-training the models on our doctor-patient conversation dataset as described in Section 4.1 followed by fine-tuning on the summarization dataset.

For BART, window-based denoising results in a negative impact on ROUGE scores, so we only investigate normal text infilling denoising, whereas for LED and DialogLED, we consider both BART-style text infilling and window-based denoising. The results of performing all types of evaluation described in Section 4.3 on the summarization test set are shown in Table 1. Furthermore, we report the median length of generated summaries in Table 2.

6 Qualitative Analysis

In-domain pre-training Across all models and pre-training objectives, ROUGE *F1* scores always improve with additional in-domain pre-training (Table 1), clearly indicating that pre-training leads to improved overlap between the generated and reference summaries. For the sake of completeness, it should be mentioned that we find that ROUGE *precision* generally decreases with increasing sequence length (Table 2) whereas ROUGE *recall* generally increases; however, we see no such correlation for ROUGE F1 so that we continue to use ROUGE F1 for the discussion here. The full evaluation results, including precision and recall can be found in Table C.1 in the Appendix. There exists some research into removing the length bias from ROUGE score calculations (*e.g.*, Sun et al., 2019), but this is out of scope for our current study.

Overall, we find that pre-training LED with the window-based denoising task leads to the strongest models in terms of ROUGE scores. For LED-large, in-domain pre-training improves the summarization performance of doctor-patient conversations by 1.59 points for ROUGE-1, 1.13 points for ROUGE-2, and 1.10 points for ROUGE-L relative to the vanilla LED-large baseline (Table 1).

Similarly, in-domain pre-training almost always improves both of our concept-based evaluation metrics with the only noticeable outlier being BART-large. We note that we observe a slightly stronger dependence of precision and recall on sequence length (Table 2 and Appendix Table C.2) than with ROUGE. Nonetheless, in-domain pre-training leads to the best-performing models as measured by concept-based F1 scores even if the pre-trained version does not generate longer sequences on average.

Overall, we find that pre-training DialogLED with the BART-style text infilling task leads to the strongest models in terms of concept-based scores, which is contrary to the performance of DialogLED when measured with ROUGE. This could imply that while DialogLED generates extraneous text (also shown by its long generation length in Table 2) which results in lower ROUGE scores, it is better at generating the relevant medical concepts, which might make it more useful for medical summarization.

Comparing across denoising tasks used for pre-training, there seems to be no significant difference in terms of ROUGE between BART-style

Model	Pre-train	ROUGE F1			UMLS F1	NER F1
		R-1	*R-2*	*R-L*		
BART-base	—	35.19	13.28	25.43	24.27	36.22
BART-base	BART	36.83	**14.13**	26.53	28.63	40.24
LED-base	—	36.01	13.49	25.99	26.40	38.59
LED-base	window	**36.96**	14.12	**26.72**	27.45	39.96
LED-base	BART	36.65	13.60	26.31	28.47	41.39
DialogLED-base	—	36.07	13.14	25.13	31.22	42.66
DialogLED-base	window	36.85	13.79	25.74	31.62	41.87
DialogLED-base	BART	36.79	13.59	25.88	**33.33**	**42.72**
BART-large	—	38.25	14.78	26.65	35.19	47.52
BART-large	BART	38.47	14.89	**27.37**	27.77	43.45
LED-large	—	37.29	13.83	26.09	30.45	43.97
LED-large	window	**38.88**	**14.96**	27.19	32.03	46.78
LED-large	BART	38.07	14.56	26.82	35.33	47.15
DialogLED-large	—	37.04	13.74	25.55	32.36	47.23
DialogLED-large	window	37.26	14.15	25.73	34.05	45.86
DialogLED-large	BART	37.73	14.56	25.69	**38.90**	**51.57**

Table 1: Evaluation results on the summarization test set. In the "Pre-train" column, "BART" refers to BART-style pre-training without sentence permutation (text infilling across the entire input) and "window" refers to window-based denoising (without turn permutation). The metrics from left to right are ROUGE-1 F1, ROUGE-2 F1, ROUGE-L F1, QuickUMLS concept-based F1, and NER concept-based F1.

Model	Pre-train	Median summary length
BART-base	—	53
BART-base	BART	62
LED-base	—	62
LED-base	window	65
LED-base	BART	65
DialogLED-base	—	71
DialogLED-base	window	73
DialogLED-base	BART	71
BART-large	—	89
BART-large	BART	70
LED-large	—	98
LED-large	window	81
LED-large	BART	88
DialogLED-large	—	108
DialogLED-large	window	110
DialogLED-large	BART	115
Training set reference summaries		114
Test set reference summaries		81

Table 2: Median sequence length in number of tokens of generated summaries and of summaries in the training (with validation) and test data.

text infilling and window-based denoising, whereas concept-based scores improve with BART-style pre-training compared to window-based denoising. Even though Zhong et al. (2021) designed the window-based denoising task for conversation data, it seems that it is not always beneficial to use over more conventional pre-training. The most important thing is simply the process of pre-training on medical conversations itself, regardless of pre-training objective used.

One benefit of using automatic metrics is that they may quantify smaller improvements which would not be as visible with small-scale human evaluation. In particular, Table 3 compares example output from vanilla LED-large and from LED-large pre-trained on doctor-patient conversations using window-based denoising. While there are some differences in the output, it is not immediately evident that the output from the pre-trained model is better. Both models produce fluent summaries and include all of the important concepts mentioned in the reference summary. In the conversation, the patient incorrectly refers to the mitral valve as "microvalve", and no explicit mention of the correct term occurs (the relevant conversation snippet is shown in Table 4). Naturally, a trained

Source	Summary
Reference	The patient is a female presenting today for routine follow-up. She states that she is doing well and continues to take Fosamax as prescribed. She continues to experience pain to her back. She is also requesting to have her heart checked as she was diagnosed with mitral valve prolapse several years ago.
BART-large (vanilla)	The patient is a female presenting to the clinic today for a follow up visit concerning her hypercholesterolemia. She reports that she has been on Fosamax for at least 4 years. She has been taking it weekly for the past 9 years. Her last bone density test was in June of last year. She is still taking Prilosec every day.
LED-large (vanilla)	The patient is a female presenting to the clinic today for a follow-up visit. She agreed to a virtual scribe. Back Pain - She has noticed that her back has been bothering her for the last month. It does not hurt to push, but it is bothersome. She takes Tylenol or Advil if it is really annoying. She has been taking Fosamax for the past 4 years. She would like to have a stress test of her arteries to see if they are strong. She is currently taking Prilosec every day. Heart Failure - Her last heart exam was 5 years ago. She had a microvalve prolapse at that time.
LED-large (window-based pre-trained)	The patient is a female presenting to the clinic today for a follow up visit. She has a history of hyperlipidemia and hypercholesterolemia. She states that she has been taking Fosamax for the past 4 years. She reports that her back has been bothering her for the last month. She denies any fractures or fractures in her bones. She is taking Prilosec every day. She would like to know how her arteries are doing and if she needs more vitamin D. She also wants to know if she has a microvalve prolapse.

Table 3: Comparison of reference summary and several generated summaries for a conversation with 2088 tokens from the test set.

medical scribe uses the correct term in the summary, whereas the LED models are not able to perform this line of complex reasoning without additional information, so they copy the term used by the patient. The vanilla LED model makes another error by stating that the patient takes Tylenol or Advil; however, the doctor is the one to suggest this in the conversation, the patient never made such a statement. A small error also occurs in the pre-trained LED model, which mentions that the patient is inquiring about vitamin D, but this is also something said by the doctor, not the patient.

Long conversations Vanilla BART-large is a strong baseline that cannot always be outperformed by the long-sequence models (Table 1). In fact, DialogLED is noticeably weak in terms of ROUGE which might imply that a non-trivial amount of information was lost during the first round of continued pre-training (on non-medical long-dialog data). Such a direct comparison between DialogLED and BART is possible because DialogLED is a further pre-trained version of LED, which is itself initialized from BART. However, as mentioned earlier, concept-based evaluation of DialogLED shows strong performance, indicating that ROUGE alone may not be sufficient for quantifying the utility of a model.

For a different reason than DialogLED, vanilla LED is also a weak baseline. We observed difficulty during fine-tuning of vanilla LED on our small dataset and hypothesize that this could be a result of non-ideal initialization of its copied position embeddings (Beltagy et al., 2020). As the position embeddings for positions greater than 1024 never underwent their own additional pre-training, their parameters are not necessarily optimal at the start of fine-tuning. However, in-domain pre-training results in a suitable initialization for the position embeddings prior to fine-tuning, which manifests itself in good performance compared to pre-trained BART after fine-tuning. Still, LED is not significantly better than BART after in-domain pre-training even though it can process much longer input. One possible reason is that most of the relevant information for the HPI section might be contained at the beginning of long conversations.

[PT]:	And, uh, I want to have my heart checked out because my heart, um, I think it was five years ago -
[DR]:	Um-hum.
[PT]:	We did it. I have a microvalve prolapse.
[DR]:	Um-hum.
[PT]:	But they said at that time it wasn't that bad -
[DR]:	Um-hum.
[PT]:	But, um, I feel like I need to check up on that again. And can they do the, uh, also can they do the arteries? Can they check your arteries?
[DR]:	They do that with the stress test. The stress test is a way of, um, the stress test has a way of looking at the arteries. You don't want to actually have the dye put in your arteries because that is dangerous.

Table 4: Snippet of the conversation corresponding to Table 3 revolving around the heart valve prolapse.

Another reason is that our median conversation length in the fine-tuning dataset (see Section 3) is not much longer than the maximum input size BART can process, so there may not be enough long conversations for the difference in models to make a large difference.

If we bin the conversations by their number of tokens and compare BART-large to LED-large, we observe less of a drop in ROUGE for longer conversations with LED-large than with BART-large (Figure 1), suggesting that LED does extract additional useful information from long inputs. The improved performance on long conversations with LED-large is even more evident when analyzing the concept-based metrics across different conversation lengths as shown in Figure 2. LED-large is very effective at extracting relevant concepts from long conversations.

The example in Table 3 corroborates this finding: The summary generated by BART-large fails to mention the back pain and heart valve prolapse, whereas LED-large correctly includes both of these concepts. Both concepts are only mentioned in the latter half of the conversation, which, with a length of 2088 tokens, is significantly longer than the maximum BART sequence length. Unrelated to conversation length, the BART-large model is seemingly confused by the duration for which the patient has been taking Fosamax. However, the BART output is actually more accurate than the LED output, which states a duration of four years. In the conversation, the doctor is briefly confused about the Fosamax duration and initially assumes "at least four years", but then corrects that estimate to "at least nine years" over the course of several subsequent sentences.

Generated summary lengths We can observe several trends in the lengths of generated summaries in Table 2. First, large models generate longer summaries than base models, and while good performance is possible using base models (Table 1), this might hint at an inadequate intrinsic capacity of small models to model complex abstractive summarization, suggesting that one would be better served by using the large models. Second, pre-trained base models generate longer summaries than their corresponding vanilla versions with the exception of DialogLED-base, which could be a result of it already having been pre-trained on long-dialog data. Interestingly, this effect seems to be reversed for the large models: pre-trained BART-large and LED-large generate shorter summaries than their vanilla versions while pre-trained DialogLED-large generates slightly longer text. Third, DialogLED always generates the longest summaries compared to BART and LED even if these have been pre-trained on in-domain data. Again, this could be due to the round of pre-training on (non-medical) long-dialog data that DialogLED underwent.

On average, the generated summaries are shorter than those in the fine-tuning training set, although they happen to correspond well in length to those in the test set. As described in Section 3, the training set summaries are longer on average because they only contain the references with the most concepts extracted using our in-house rule-based system. Overall, these results indicate that there might be a need to bias the models toward longer generation length. However, we do not add any sort of length penalty here because our goal was to compare what the models learn in an unbiased fashion.

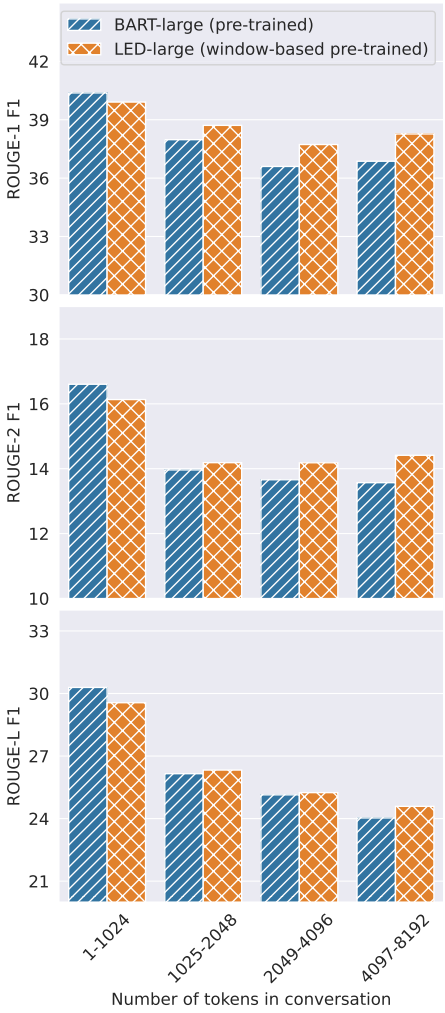

Figure 1: ROUGE score comparison binned by conversation length.

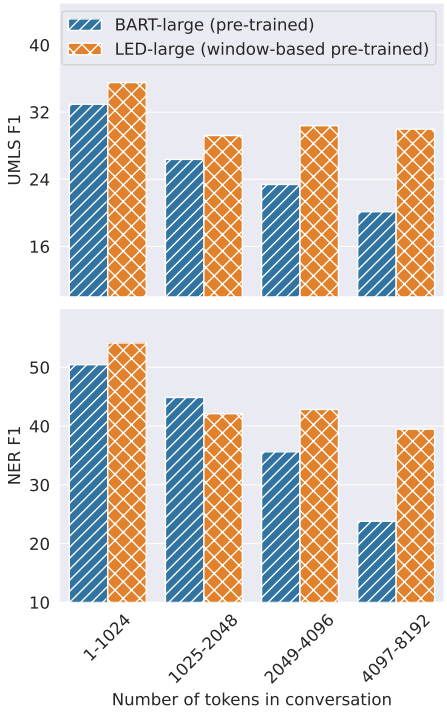

Figure 2: Concept-based score comparison binned by conversation length.

7 Conclusion

We showed that in-domain pre-training improves abstractive summarization of long doctor-patient conversations into HPI notes across several models based on the BART architecture and across two different pre-training objectives. To measure the improvement, we used conventional evaluation methods like ROUGE and UMLS concept-based evaluation and also trained a neural clinical concept extraction model to better extract relevant concepts. We also demonstrated the benefit of using models that can deal with long conversations intrinsically, especially for ensuring that relevant medical concepts are present.

While unlabeled doctor-patient conversations are a useful source of pre-training data, we hope to investigate additional types (*e.g.*, clinical notes) in the future. Similar research has already shown that other types of pre-training data can be very effective, *e.g.*, pre-training on patient histories (Yalunin et al., 2022) or pre-training on clinical notes for named entity recognition (Yang et al., 2020). Additionally, we can explore combining and contrasting our holistic pre-training approach with methods that only pre-train a small amount of additional parameters (Tai et al., 2020).

Lastly, given the varying lengths of generated summaries, we are considering methods to control generation length as another future research direction (Kikuchi et al., 2016).

Ethical Considerations

The models developed in this paper may omit important information or incorrectly include misleading details in the output they generate. Due to this, we stress the importance of not using the generated outputs unsupervised. In all cases, medical experts should review and edit the generated summaries. Nonetheless, we expect that our models can act as virtual assistants to alleviate some of the documentation burden.

The data used for pre-training and fine-tuning inherently contain sensitive medical information. To protect private health information, the data were manually de-identified by medical experts and no private information was used in the methods described in this paper.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *Computing Research Repository*, arXiv:2004.05150. Version 2.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, pages 22–30, Online. Association for Computational Linguistics.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. *Computing Research Repository*, arXiv:2104.01027. Version 2.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. 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 4958–4972, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Gogi Kumar and Adam Mezoff. 2020. Physician Burnout at a Children's Hospital: Incidence, Interventions, and Impact. *Pediatric Quality & Safety*, 5(5):e345.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):7155.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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. *Computing Research Repository*, arXiv:1907.11692.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Maria Panagioti, Efharis Panagopoulou, Peter Bower, George Lewith, Evangelos Kontopantelis, Carolyn Chew-Graham, Shoba Dawson, Harm van Marwijk, Keith Geraghty, and Aneez Esmail. 2017. Controlled Interventions to Reduce Burnout in Physicians: A Systematic Review and Meta-analysis. *JAMA Internal Medicine*, 177(2):195–205.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Tait D. Shanafelt, Lotte N. Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P. West. 2016. Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction. *Mayo Clinic Proceedings*, 91(7):836–848.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*, 165(11):753–760.

Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a Fast, Unsupervised Approach for Medical Concept Extraction. In *Medical Information Retrieval (MedIR) Workshop, SIGIR 2016*.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pretrained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alexi A. Wright and Ingrid T. Katz. 2018. Beyond Burnout — Redesigning Care to Restore Meaning and Sanity for Physicians. *New England Journal of Medicine*, 378(4):309–311.

Alexander Yalunin, Dmitriy Umerenkov, and Vladimir Kokh. 2022. Abstractive summarization of hospitalisation histories with transformer networks. *Computing Research Repository*, arXiv:2204.02208.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. DialogLM: Pretrained Model for Long Dialogue Understanding and

Summarization. *Computing Research Repository*, arXiv:2109.02492. Version 2.

Appendix

A Hyperparameters

The pre-training hyperparameters are listed in Table A.1 and the fine-tuning hyperparameters are listed in Table A.2. Each model was trained using a single NVIDIA V100 GPU. Mixed precision training and gradient checkpointing were used as needed in order to fit the larger models into memory.

B NER Model Performance

We fine-tune RoBERTa (pre-trained on MIMIC-III) on the i2b2 2010 dataset using the approach of Yang et al. (2020) in order to use it as a clinical concept extraction model for concept-based evaluation. We show our performance on the fine-tuning dataset in Table B.1 and compare it to theirs. While we were not able to fully match their results, we believe this is due to the fact that the i2b2 2010 dataset is no longer available in its original form. Nonetheless, we also achieve strong results that are suitable for our purposes.

Model	*P*	*R*	*F1*
Yang et al. (2020)	89.63	90.26	89.94
Ours	87.80	88.58	88.19

Table B.1: Comparison of clinical named entity recognition models.

C Additional Evaluation

The complete ROUGE evaluation results are shown in Table C.1, which shows precision and recall in addition to the F1 score. Similarly, Table C.2 shows precision and recall for the two concept-based evaluation methods.

Parameter	BART		LED		DialogLED	
	base	*large*	*base*	*large*	*base*	*large*
Maximum encoder length	1024	1024	5120	5120	5120	5120
Maximum decoder length	1024	1024	1024	1024	1024	1024
Text infilling ratio[a]	0.3	0.3	0.3	0.3	0.3	0.3
Window ratio	0.1	0.1	0.1	0.1	0.1	0.1
Maximum window size	512	512	512	512	512	512
Text infilling ratio[b]	0.15	0.15	0.15	0.15	0.15	0.15
Speaker mask ratio	0.5	0.5	0.5	0.5	0.5	0.5
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch size	8	8	8	8	8	8
Epochs	3	0.6	3 (1)[c]	0.4	3 (1)[c]	0.4 (0.2)[c]
Warm-up ratio	0.01	0.01	0.01	0.01	0.01	0.01
Weight decay	0.001	0.001	0.001	0.001	0.001	0.001
Maximum gradient norm	1.0	1.0	1.0	1.0	1.0	1.0

Table A.1: Hyperparameters used for continued pre-training. We differentiate between BART-style noise, which uses text infilling across the entire input (*a*), and window-based denoising, which only performs text infilling within the window (*b*) and masks speakers separately. Both types of denoising are investigated for LED and DialogLED. LED and DialogLED sometimes use different number of epochs during training for BART-style and window-based denoising (*c*). No sentence or turn permutation is used.

Parameter	BART		LED		DialogLED	
	base	*large*	*base*	*large*	*base*	*large*
Maximum encoder length	1024	1024	5120	5120	5120	5120
Maximum decoder length	1024	1024	1024	1024	1024	1024
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch size	8	8	8	8	8	8
Maximum epochs	30	30	30	30	30	30
Warm-up steps	200	200	200	200	200	200
Weight decay	0.001	0.001	0.001	0.001	0.001	0.001
Maximum gradient norm	0.1	0.1	0.1	0.1	0.1	0.1
Steps between evaluation	50	50	50	50	50	50
Early-stopping patience	5	5	5	5	5	5
Number of beams	5	5	5	5	5	5
Maximum generation length	512	512	512	512	512	512
No repeat n-gram size	3	3	3	3	3	3

Table A.2: Hyperparameters used for fine-tuning.

		ROUGE-1			ROUGE-2			ROUGE-L		
Model	**Pre-train**	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
BART-base	—	**44.71**	34.73	35.19	**17.05**	13.29	13.28	**31.73**	25.81	25.43
BART-base	BART	43.12	38.44	36.83	16.71	14.87	**14.13**	30.51	28.43	26.53
LED-base	—	43.26	37.03	36.01	16.36	14.03	13.49	30.65	27.46	25.99
LED-base	window	42.68	38.97	**36.96**	16.39	15.11	14.12	30.27	28.95	**26.72**
LED-base	BART	42.29	38.65	36.65	15.82	14.44	13.60	29.85	28.50	26.31
DialogLED-base	—	39.09	40.39	36.07	14.25	14.94	13.14	26.67	29.08	25.13
DialogLED-base	window	38.98	**42.04**	36.85	14.64	**15.89**	13.79	26.62	**30.45**	25.74
DialogLED-base	BART	39.56	41.49	36.79	14.67	15.53	13.59	27.24	30.14	25.88
BART-large	—	38.30	46.09	38.25	14.86	18.01	14.78	26.24	33.17	26.65
BART-large	BART	**42.73**	41.80	38.47	**16.64**	16.27	14.89	**29.83**	30.60	**27.37**
LED-large	—	37.00	46.11	37.29	13.70	17.44	13.83	25.43	33.40	26.09
LED-large	window	40.50	44.62	**38.88**	15.69	17.35	**14.96**	27.83	32.16	27.19
LED-large	BART	38.28	46.05	38.07	14.69	17.81	14.56	26.56	33.41	26.82
DialogLED-large	—	33.83	49.98	37.04	12.53	18.81	13.74	22.83	35.95	25.55
DialogLED-large	window	34.06	50.27	37.26	12.89	19.37	14.15	23.01	36.20	25.73
DialogLED-large	BART	33.34	**53.29**	37.73	12.88	**20.72**	14.56	22.25	**37.79**	25.69

Table C.1: Complete ROUGE evaluation results on the summarization test set.

		UMLS			NER		
Model	**Pre-train**	*P*	*R*	*F1*	*P*	*R*	*F1*
BART-base	—	56.65	18.28	24.27	**76.52**	26.66	36.22
BART-base	BART	**57.05**	22.93	28.63	70.93	32.56	40.24
LED-base	—	52.77	21.21	26.40	69.62	31.61	38.59
LED-base	window	53.04	21.88	27.45	70.29	32.43	39.96
LED-base	BART	53.29	22.72	28.47	66.81	34.98	41.39
DialogLED-base	—	51.99	26.63	31.22	66.34	**38.02**	42.66
DialogLED-base	window	53.51	26.35	31.62	65.27	36.31	41.87
DialogLED-base	BART	55.51	**28.36**	**33.33**	69.48	36.64	**42.72**
BART-large	—	**53.95**	30.52	35.19	66.39	42.92	47.52
BART-large	BART	46.97	23.49	27.77	**66.84**	37.42	43.45
LED-large	—	43.80	28.01	30.45	51.84	44.27	43.97
LED-large	window	51.15	27.13	32.03	65.93	41.39	46.78
LED-large	BART	52.93	31.51	35.33	63.92	43.54	47.15
DialogLED-large	—	44.14	30.43	32.36	54.46	48.84	47.23
DialogLED-large	window	46.29	32.22	34.05	53.48	46.98	45.86
DialogLED-large	BART	49.60	**37.88**	**38.90**	61.70	**52.18**	**51.57**

Table C.2: Complete concept-based evaluation results on the summarization test set.

LCHQA-Summ: Multi-perspective Summarization of Publicly Sourced Consumer Health Answers

Anonymous ACL submission

Abstract

Community question answering forums provide a convenient platform for people to source answers to their questions including those related to healthcare from the general public. The answers to user queries are generally long and contain multiple different perspectives, redundancy or irrelevant answers. This presents a novel challenge for domain-specific concise and correct multi-answer summarization which we propose in this paper.

1 Introduction

Community Question Answering (CQA) platforms like Yahoo!Answers, Stack Exchange, Reddit, Quora, etc., are vast repositories of question-answer pairs where common people ask questions as well as contribute answers across various domains. One such domain is healthcare. People not only seek answers from experts but also from the general public which is facilitated by these websites. The reasons for sourcing laymen contributed answers could be to avoid the use of medical jargon in the language used by the experts (Boyd et al., 2018), opportunity to freely express themselves (Park and Conway, 2018) and share their experiences (Alvaro et al., 2015). The posts also give a fair idea of public opinion on specific health issues (Odlum and Yoon, 2015). However, often these answers are long-winded and irrelevant. These challenges necessitate summarization of answers in CQA forums, especially for healthcare domain which directly impacts the well-being of people.

Majority of the existing works in answer summarization is in the general domain (Liu et al., 2008; Fabbri et al., 2019, 2021). There has been a limited study towards summarizing answer in the healthcare domain (Savery et al., 2020; Abacha et al., 2021; Demner-Fushman et al., 2020), which is confined to expert sourced answers. To the best of our knowledge healthcare related question-answers from CQA forums have not been harnessed yet. To bridge the gap, we bring forward an abstractive multi-document summarization approach for consumer health answer summarization. We also observe that these answers present several perspectives. For example, in Table 1, Answers 1 and 3 describe the the cause of hay fever symptoms while Answer 2 shares a personal experience and possible treatments. Answer 4 provides some suggestions that can potentially solve the problem. This motivates us to tackle the summarization problem while covering the different perspectives as done by (Fabbri et al., 2021).

Towards this, we frame our research objectives as follows: (i) Develop a novel gold standard Laymen-sourced Consumer Health Question Answer Summaries (LCHQA-Summ) dataset with summaries covering the breadth of perspectives across various healthcare topics. (ii) Propose an automated health answer summarization pipeline to generate perspective-specific answer summaries.

2 Proposed Plan of Research

2.1 Data Collection and Annotation

We begin by collecting dataset from popular CQA forum –Yahoo! Answers[1]. In particular, we plan to use Yahoo! L6 corpus that consist of 4.5 million questions across different topics, their answers and metadata such as question categories, number of answers, best answer, date, language etc. Since, our goal is focused on consumer healthcare domain, we selected the “Health" category which has 21 subcategories like Allergies, Diabetes, Heart Diseases and so on. It is also necessary to remove outliers in terms of number of answers which can range from as low as zero and as high as 2235 answers in response to a single query. We finally retain posts where number of answers range from $4 - 6$. The final data includes 77K question-answer pairs. To curate a gold dataset of manually written multi-

[1] http://answers.yahoo.com

Proceedings of the First Workshop on Natural Language Generation in Healthcare (NLG4Health), pages 23 - 26

Question	Why are my hay fever symptoms worse early in the morning and how I can stop suffering the first two hours after I wake up?
Context	Allergies
Answer 1	It's because the pollen counts are higher in the morning. Plants release their pollen earlier in the day, thus anyone with hayfever will find this part of the day more annoying.
Answer 2	I have similar problems. When I wake up I have a stuffy nose but then in like an hour or two and I'm fine. I take Zyrtec every morning but before I go to bed I take a Benadryl and that seems to help.
Answer 3	Because pollen is released early in the day, rises with the warm air and falls again in the evening.
Answer 4	It may help if you wash your hair in the evening to get rid of any pollen that might be left in there.
Summary Perspectives:	
Perspective 1	Plants release pollen early in the day.
Perspective 2	Pollen counts are higher in the morning.
Perspective 3	I have similar problems for an hour or two after I wake up.
Perspective 4	Taking Benadryl before bed and Zyrtec in the morning has helped me.
Perspective 5	Washing your hair at night can get rid of any left-over pollen.

Table 1: An example illustrating question, context and answers from Yahoo! L6 dataset. This is followed by an abstractive summary of the answers showcasing 5 different perspectives.

perspective abstractive summaries from the data we sample a subset of the data and put forward the following annotation strategy:
(1) Validate if a question is related to medical domain or not, that is if it pertains to diseases or conditions, drug or treatment, medical diagnosis or therapeutic procedure, any other related medical topic. This helps to weed out any irrelevant question, especially in more generic sub-categories like "Other-Health" and "Other-Health & Beauty".
(2) The next step is to generate abstractive multi-perspective summaries of answers to valid medical question. Based on our preliminary dataset analysis, we have identified 6 major perspectives—*information*, *cause*, *treatment*, *suggestion*, *experience* and *clarification* that describes most of the consumer answers. Example of such summary is shown in Table 1, where perspective 1 and 2 describe cause of the problem, 3 and 4 narrates experience as well as treatment and 5 suggests solution.

2.2 Automated Summarization Pipeline

For obtaining system generated multi-perspective summaries of consumer health answers, we devise a three-step pipeline described next.

Relevant Sentence Extraction: This step is to be applied at the sentence level with the goal of finding the answer sentences that are relevant to the question. As a baseline we would use BM25 (Robertson et al., 1994) to compute relevance of each answer sentence to a given question and retain those with score above a threshold as relevant. A similar approach is measuring semantic similarity between the embeddings of each answer sentence and question using cosine similarity or mutual information. For this we propose to use Sentence-BERT(Reimers and Gurevych, 2019) (SBERT) and UmlsBERT (Michalopoulos et al., 2021) representations.

Perspective Type Identification: Allocating perspective labels to a relevant answer sentence can be treated as a multi-label classification problem (For example, Perspective 4 in Table 1 can be both an *experience* and a *treatment*). Given the success of transfer learning along with zero-shot and few-shot approaches in text classification (Chalkidis et al., 2020; Zhang et al., 2019), we propose to adapt a Natural Language Inference (NLI) based transfer learning approach as done by (Yin et al., 2019) for assigning perspective labels to the sentences. Based on the performance of this method we would also experiment with more refined rules to improve performance across specific classes.

Summarization of answers: In the final stage of the pipeline, we aim to propose a perspective-guided multi-document answer summarization approach focusing on answer summary generation conditioned over the given perspective. We plan to infuse the perspective in terms of the external knowledge to the pre-trained encoder decoder models such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) which has shown state-of-the-art performance on the answer summarization task(Yadav et al., 2021; Mrini et al., 2021). Towards this, we will begin by inducing perspective information into the encoder as well as decoder to train the model which incorporates the underlying perspective while generating the summary.

3 Conclusion

Overall in this paper we present the novel problem of multi-perspective abstractive answer summarization from CQA forums focusing on the healthcare domain. We outline a data annotation process, followed by a three-step approach for automatic summary generation with a focus on the perspectives present in these answers.

References

Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.

Nestor Alvaro, Mike Conway, Son Doan, Christoph Lofi, John Overington, and Nigel Collier. 2015. Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use. *Journal of biomedical informatics*, 58.

Andrew D Boyd, Karen Dunn Lopez, Camillo Lugaresi, Tamara Macieira, Vanessa Sousa, Sabita Acharya, Abhinaya Balasubramanian, Khawllah Roussi, Gail M Keenan, Yves A Lussier, et al. 2018. Physician nurse care: A new use of umls to measure professional contribution: Are we talking about the same patient a new graph matching algorithm? *International journal of medical informatics*, 113:63–71.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online. Association for Computational Linguistics.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2):194–201.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R Fabbri, Xiaojian Wu, Srini Iyer, and Mona Diab. 2021. Multi-perspective abstractive answer summarization. *arXiv preprint arXiv:2104.08536*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, and Yong Yu. 2008. Understanding and summarizing answers in community-based question answering services. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 497–504, Manchester, UK. Coling 2008 Organizing Committee.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021. UCSD-adobe at MEDIQA 2021: Transfer learning and answer sentence selection for medical summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 257–262, Online. Association for Computational Linguistics.

Michelle Odlum and Sunmoo Yoon. 2015. What can we learn about the ebola outbreak from tweets? *American journal of infection control*, 43 6:563–71.

Albert Park and Mike Conway. 2018. Tracking health related discussions on reddit for public health applications. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:1362–1371.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):1–9.

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021. NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 291–301, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3914–3923.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1031–1040, Minneapolis, Minnesota. Association for Computational Linguistics.

Towards Development of an Automated Health Coach

Rommy Márquez-Hernández[1], **Leighanne Hsu**[1], **Kathy McCoy**[1], **Keith Decker**[1], **Ajith Vemuri**[1], **Greg Dominick**[2], **Megan Heintzelman**[2]
[1]Department of Computer and Information Sciences
[2]Department of Behavioral Health and Nutrition
University of Delaware
{marquez,lhsu,mccoy,decker,kumar,gdominic,mheintz}@udel.edu

Abstract

Human health coaching has been established as an effective intervention for improving clients' health, but it is restricted in scale due to the availability of coaches and finances of the clients. We aim to build a scalable, automated system for physical activity coaching that is similarly grounded in behavior change theories. In this paper, we present our initial steps toward building a flexible system that is capable of carrying out a natural dialogue for goal setting as a health coach would while also offering additional support through just-in-time adaptive interventions. We outline our modular system design and approach to gathering and analyzing data to incrementally implement such a system.

1 Introduction

It is well-known that eating a balanced diet and engaging in regular moderate-to-vigorous physical activity (MVPA), among other healthy behaviors, promotes better health and reduces the risk of cardiovascular disease and other chronic illnesses (Tsao et al., 2022). However, people may struggle to develop and integrate healthier behaviors on their own (Kivelä et al., 2014; Willard-Grace et al., 2015). Health coaching is a behavioral health intervention that is demonstrably effective in improving motivation and confidence and is strongly associated with behavior change (Dennis et al., 2013; Eakin et al., 2007; Mahon et al., 2018; Oddone et al., 2018). Health coaches utilize behavioral theories and evidence-based strategies in a client-centered approach to help clients set goals that are challenging yet achievable, supported by action plans and coping plans that include strategies to overcome barriers like lack of time or poor weather (Kivelä et al., 2014; Oddone et al., 2018). As such, goals and dialogue are highly specific and tailored to the client.

However, human health coaching is limited by coach availability, cost to potential clients, and the retrospective nature of the feedback (Hill et al., 2015). Most attempts to automate this process thus far have been mostly limited to theoretical studies or systems with pre-scripted, non-tailored dialogues, if there is any interactivity at all (op den Akker et al., 2014, 2015; Bickmore et al., 2011, 2013; Svetkey et al., 2015; Kramer et al., 2020). While some successfully demonstrate the acceptability of automated systems, even with scripted interaction, the feedback also identifies a user desire for increased tailoring with regards to timing and response to collected user data or context during coaching sessions (Bickmore et al., 2013; Mitchell et al., 2021).

Another type of intervention, Just-in-Time Adaptive Intervention (JITAI), leverages technology to monitor a user's state and deliver support at a time when it is most needed and the user is most receptive to act upon it in the moment (Nahum-Shani et al., 2018; Schembre et al., 2018; Spruijt-Metz et al., 2015). For instance, this may include nudges like "walk around the block an extra time" if the user is out for a walk.

To our knowledge, no system has yet combined automated, interactive coaching with real-time knowledge of user progress and JITAI. To this end, we aim to build an automated system capable of helping clients set achievable goals through interactive discussion and support them in achieving those goals. We focus on physical activity (PA) coaching, but this infrastructure is modular and extendable to other health coaching areas in which goals can be clearly defined or real-time user context leveraged, including stress management or adapting to a prescribed diet.

In this paper, we present our preliminary work in building a dialogue and messaging system for an application capable of coaching a user to set and achieve goals and provide useful just-in-time messaging. We will contextualize the messaging components within the greater system architecture

Proceedings of the First Workshop on Natural Language Generation in Healthcare (NLG4Health), pages 27 - 39

we are building upon in section 3 and then detail the approach we've taken to build our proposed dialogue and messaging component in sections 4 and 5. Finally, we will describe our plans for future experiments and evaluations.

2 Related Work

The rise in popularity and availability of wearable technology and biometric sensors offers the opportunity to create similarly theoretically-driven, evidence-based behavioral interventions (DiClemente et al., 2001; Fjeldsoe et al., 2009; O'Reilly and Spruijt-Metz, 2013; Bort-Roig et al., 2014; Danaher et al., 2015; Farmer and Tarassenko, 2015; Wang et al., 2015a; King et al., 2016; Lobelo et al., 2016). However, the resulting apps generally do not adhere to the American College of Sports Medicine (ACSM) recommendation of 150 minutes per week of moderate-intensity aerobic physical activity or 75 minutes per week of vigorous-intensity aerobic physical activity (Middelweerd et al., 2014; Guo et al., 2017; Modave et al., 2015). They lack guidance establishing realistic and appropriate behavioral goals, do not assist users in modifying goals over time, display messages that are not personalized, and do not account for contextual or situational barriers, such as weather and emotional states, that can significantly influence physical activity intentions and behavior (Düking et al., 2020; Rupp et al., 2018; op den Akker et al., 2014, 2015; Muntaner et al., 2016).

Active work on JITAI systems emphasizes their basis in behavioral theory, user relevance, and actionable feedback (Wang et al., 2015b; Hardeman et al., 2019). However, most do not truly account for context or barriers and instead use simple, canned messages delivered at preset moments (Klasnja et al., 2018; Lentferink et al., 2017; Saponaro et al., 2017; Mair et al., 2022). More recently, Saponaro et al. (2021) and Ismail et al. (2022) demonstrated that individualized, contextualized JITAI nudges are significantly better received than non-JITAI nudges.

Some automated coaching systems exist (op den Akker et al., 2014), but most are limited in interactivity, and efficacy varies. Many dialogue-driven health coaching systems are largely theoretical (Bickmore et al., 2011; op den Akker et al., 2015), although a few extend to more practical implementations with varying degrees of tailoring and interactivity (Svetkey et al., 2015; Bickmore et al., 2013). Several other embodied conversational agents were included in a recent survey (Kramer et al., 2020). Of these, most rely on scripted dialogue selection, and the others provide limited text interaction at best, lacking the flexibility to adequately tailor to users' unique goals and values. A detailed comparison between text-based coaching and human health coaching was performed in Mitchell et al. (2021), demonstrating the feasibility of an automated system with a wizard-of-oz setup. While participants appreciated the automated coaching system, they lamented the lack of tailoring and context sensitivity. Some analysis work has been done on counseling dialogues (Pérez-Rosas et al., 2017, 2018; Althoff et al., 2016) and, more recently, on coaching dialogues (Gupta et al., 2020a,b, 2021). This latter work is ongoing, but focuses primarily on post-dialogue SMART goal summarization and health coach assistance rather than interactivity. Thus, to our knowledge, no one has yet developed a system to conduct a dynamic, flexible, and interactive coaching dialogue.

Human coaching dialogue adheres to a specific procedural structure and language. Generation for dialogue in this context has added constraints that increase its complexity compared to other domains. While the intents and data content will be provided by the respective messaging policies, it is also important to incorporate personalization (Cawsey et al., 1997, 2000; Marco et al., 2006; Colineau and Paris, 2011), empathy (Prendinger and Ishizuka, 2005), and additional constraints for a health domain, as well as constrained generation (He and Li, 2021; Miao et al., 2019; Mou et al., 2015; Li and Sun, 2018) and style transfer (Jin et al., 2020; Toshevska and Gievska, 2021).

3 Automated Coaching System

The automated coach brings together ideas and solutions in human behavioral theory, physical activity monitoring, cooperative multi-agent architectures, and natural language processing to build an integrated approach to reducing sedentary behavior while increasing users' overall physical activity.

The automated coach is designed to ultimately take over certain basic tasks from a traditional human coach: it will be able to meet with users on a weekly basis to negotiate goals and talk about the user's progress. A number of schemas exist for creating a well-defined goal; we discuss these in section 4. The system may also leverage user

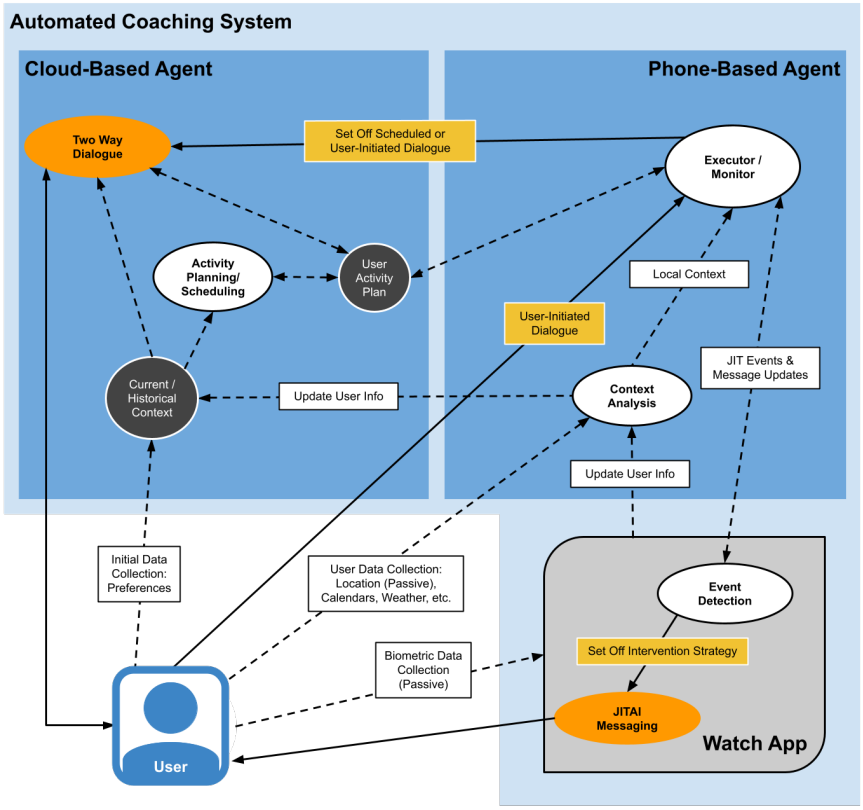

Figure 1: Interaction between the User and the Automated Coaching System: white/orange ovals represent major modules, gray circles represent stored information, rectangles represent actions and data transfer between modules.

qualities or other initially-provided information to quickly adapt to users' preferred intervention strategies. Unlike a traditional human health coach, it will be available for user-initiated coaching sessions at any time and will be able monitor the user's real-time goal progress. This will allow the automated coach to better support the user, as well as allow it to send personalized just-in-time messaging when needed.

Unlike typical one-size-fits-all solutions for just-in-time interventions such as "remind the user to walk at 10 min before the hour if they have not yet reached 250 steps," which do not work particularly well (Saponaro, 2020), the action plan and user preferences are compiled into a context- (state-) sensitive strategy enumerating possible conditions under which the coach should nudge with particular message content (remind, congratulate, suggest, or otherwise interact with) their user. Positive (or negative) reactions to these nudges are used to adaptively learn a better nudging strategy by taking into account both timing and message content.

The automated coaching system is built on a multi-agent architecture (Graham et al., 2003) for privacy and scalability, and further extended by our contributions toward coaching domains with an eye toward individualization, data integration, domain flexibility and (agent) behavior transferability (Vemuri et al., 2021). As can be seen in Figure 1, each user is allocated two personalized agents: one autonomous cloud-based agent running continuously on a server that is responsible for data collection, learning, dialogue understanding, and generation; and an app-based agent on the user's smartphone that handles local data collection from the user's smartwatch, user dialogue interface, and summary graphics. Although we have previously built-out versions of this using the FitBit Charge platform, true JITAI is not possible with that due to lack of real-time sensor access (Vemuri et al., 2021). Our current version uses the Apple Watch platform, which allows for an on-watch app that can be configured by the phone agent to detect specific, context-sensitive intervention events (prolonged inactivity, walking, stress) for JITAI. A centralized Dashboard agent (not pictured) is also used for running trials to give at-a-glance access to current participant status.

Agents are able to process data and communicate with each other concurrently. Processing can

be triggered by communications, state/sensed context/goal changes, and also pre-scheduled agent behaviors. Just-in-time notifications are triggered directly on the watch or phone (depending on the type), and do not require dialogue responses. Dialogues can be initiated by the phone agent at scheduled times, or by the user.

4 Approach

Bickmore et al. (2011) and Bickmore et al. (2013) described human health coaching as the gold standard for automated coaching systems to aim for. Such a system relies on a rich library of information representing user data, preferences, and coaching knowledge and principles. The dialogue and messaging system architecture is outlined in section 5. This system will interface with the agent architecture described in section 3 for timing and information for messages as can be seen in Figure 1. It will be capable of handling one-way JITAI messages (see section 5.3) and two-way, interactive health coaching dialogues.

We frame the coaching dialogue as a task-oriented dialogue. However, most task-oriented dialogues consist of rigidly defined simple tasks (e.g., booking a flight or negotiating a price) defined by a few parameters that the system needs to elicit from the user to complete the task. Dialogue policy, which determines each system intent (e.g., request information, offer a suggested value) and directs the dialogue, is similarly simplistic and limited: a task is complete when the parameters have been filled and an operation or query happens successfully (e.g., a flight is successfully booked). Parameters can be modified or updated until the task is completed. Additionally, there is no need for information to carry over from session to session; once a price is agreed upon, for instance, the task is complete, and there are no further exchanges on the subject.

Health coaching instead centers around a reflective discussion to achieve a more loosely-defined objective: setting a well-defined goal with realistic strategies for completing it. The dialogue is completed when the goal and strategies are not only fully defined by their parameters, but sufficiently motivated and supported to improve the user's success. The latter is accomplished not through filled parameters, but a series of reflective questions to ensure the user has thought their goal through thoroughly. This goal is then revisited at the subsequent coaching session, where a new goal may be set or a new coping plan may be created to assist the client in overcoming unforeseen barriers. Additionally, understanding barriers or support systems requires some representation of a health coach's world knowledge.

To ensure that the top-down approach aligns with practice and data, we also examined coaching dialogues. Due largely to patient privacy and protection, few publicly available datasets exist within the health coaching domain. There is one recently released dataset containing health coaching dialogues conducted via SMS text message (Gupta et al., 2020a). This dataset is tagged with a two-level labeling structure. One level covers stages and phases, breaking down the overall weekly dialogue into goal setting and goal implementation stages, which further break down into phases such as refining, anticipating barriers, negotiation, and follow up. Additionally, they identify SMART goal components, which break down a goal into Specificity, Measurability, Attainability, Realism, and Time-bound components.

This dataset released after development on our system had begun, and the coaching paradigm is different from our face-to-face data. Due to their curtailed nature, text messages often lack certain nuances, context, and cues compared to verbal interaction (Mitchell et al., 2021). Messages tend to be more curt and elaborate less, which affects the style of questions that need to be used to elicit the same information. Discussion of action plan, barriers, and coping strategies is thus unsurprisingly significantly more limited in this dataset, which focuses more on the goal parameterization. However, since our target coaching format is also text message-like, it will still be crucial for designing a text message coaching session, as well as for training the natural language understanding components described briefly below in section 5.1.

To ensure that our automated system is rooted in core health coaching concepts and behavior change theory, we examined coach training materials and guidelines provided by our health coaching team or publicly available online. These included outlines as well as coaching roleplay videos. Based on these materials, we developed a dialogue model described below in section 5.2. This model was further refined by examining data collected through a tangential, developmental study. We will refer to this data as dataset 1.

4.1 Data and Annotation

Dataset 1, currently being collected through BeSMART feasibility trial (Heintzelman et al., 2022), closely mimics the face-to-face coaching sessions that the coaching team regularly conducts with clients. Clients meet with their coach one-on-one initially for approximately an hour, and then subsequently for twenty or thirty minutes, generally with at least a week between meetings. For the short feasibility trial, our coaches did not receive information about goal progress between meetings, but we intend to correct this in subsequent studies (see section 7), as the proposed automated coaching system will have access to users' goal progress and other context.

Data collection for dataset 1 is ongoing, but the existing data is being broadly analyzed to further refine the dialogue model. The data was collected over video call, and the audio was automatically transcribed. The video was discarded for patient privacy, and the audio was kept only for quality control; the transcripts were manually cleaned for major mistranscriptions only. Verbal fillers, restarts, and exchanges consisting only of repeated acknowledgements will be automatically removed in a preprocessing step later, prior to data annotation.

Our coaches use a slightly different strategy to that of Gupta et al. (2020a). In addition to developing SMART goals, our health coaches utilize FITT (Frequency, Intensity, Time/duration, and Type of activity) and the W5 (What activity, Where, When, Who is supporting or accompanying, and Will any preparation be needed) to better assist clients in visualizing how their goal and action plan (details and strategy of how to achieve the goal) will fit into their daily schedule.

5 Dialogue and Messaging System

In this section, we detail the dialogue and messaging model and how the natural language interfaces will be built upon it. We will focus primarily on the dialogue model, as the one-way JITAI messaging is largely driven by the multi-agent architecture described previously and requires no interaction and much less tailoring of wording than the eventual dialogue system.

The overall dialogue system architecture is shown in Figure 2. We chose a traditional, modularly built dialogue system over an end-to-end neural network because the latter is unable to handle the level of complexity, control, and constraint that a health coaching system requires. This system is a modified dialogue state architecture. The dialogue schema and models are hierarchical. The modularity of this system allows for an evolving implementation. The current policy and generation modules are fully rule-based, which allows us to ground the overall structure of the dialogue in theory and coaching protocols. However, these will be incrementally swapped for dynamic, data-driven, learned implementations as the rest of the dialogue system develops to support them.

5.1 Natural Language Understanding & Dialogue State

During a dialogue exchange, for a given user input, the Natural Language Understanding (NLU) module identifies a number of different levels of slot and message labels, conditioned upon the system's prior request, if any. These labels update the dialogue state tracker, which keeps track of the information that has been provided by the user. It effectively captures the history and current knowledge state of the system based solely on the user's messages. This knowledge state representation is multilayered. At the top, the user directly or indirectly conveys an intent. In system-initiated, scheduled dialogues, the system is expected to direct the flow of conversation, determining when to move onto the next subdialogue. On the other hand, our system will eventually also accommodate user-initiated dialogues, which would start with a new intent without a prior message. A given input will also have one or more dialogue acts (e.g., whether the user is requesting information, setting parameters for their goal, or suggesting a possible coping plan to overcome a barrier). At a more fine-grained level, we will need a classifier to identify the task-specific labels (e.g., goal or action plan components, barriers, or support figures). Sentiment and uncertainty analyses will be added later to direct the policy and generation to produce clearer, more appropriate, or empathetic messaging.

While these understanding components will be based on the some of the same techniques used in summarization (Gupta et al., 2020b, 2021) or dialogue state tracking (Young et al., 2010), the policy and generation components allows an additional advantage of requesting confirmation to reduce mistakes in the summarization and dialogue state tracker. These labels will feed into the dialogue policy and eventually add to the knowledge

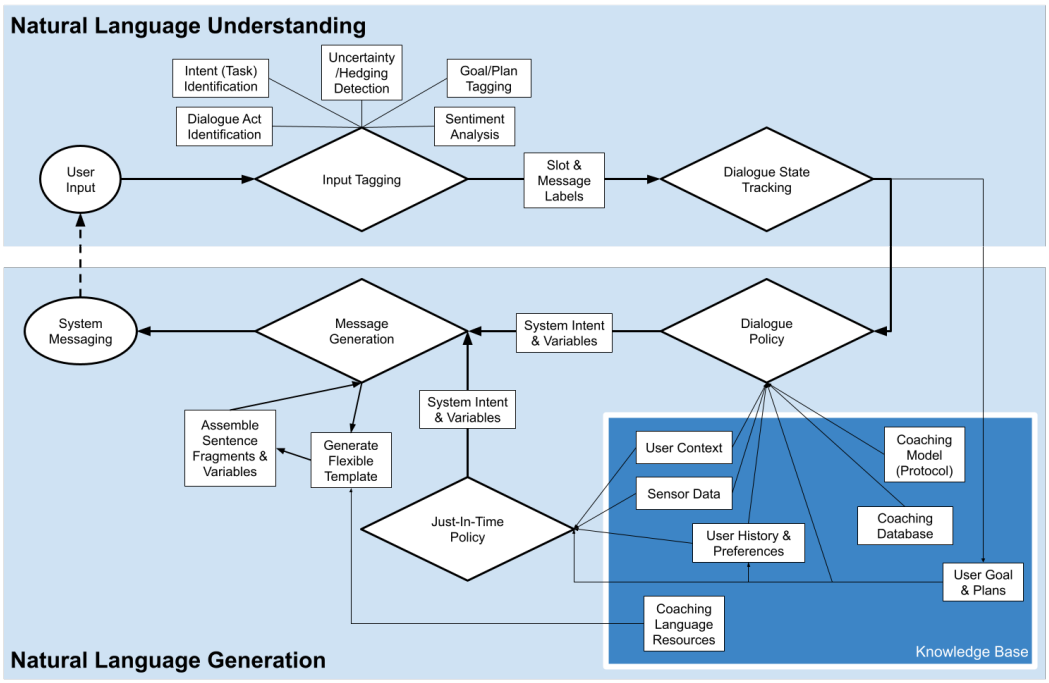

Figure 2: Dialogue State and Messaging Architecture: diamonds represent major modules, rectangles represent information transfer and functionality for those modules. The knowledge base is a logical representation of the context, history, and other information stored across the agents and apps in the automated coaching system.

base by updating the cloud-based agent. The NLU components can be built separately and combined to produce a multi-layered representation of the dialogue state. We will test a variety of features, including word embeddings and system intents of the previous turn, across a variety of machine learning classifiers. Gupta et al. (2020b) found that phase/stage classification (roughly equivalent to our intent/task classification) was more accurate when the SMART components were included as features, so we will test certain labels as potential features for other labels as well.

A particular challenge lies in understanding barriers. These are not necessarily unique to each user; barriers such as lack of time/space are common. We plan to build an expandable database, seeded initially with categorized examples from our own dataset, to represent barriers and potential solutions.

5.2 Dialogue Policy

Health coaching dialogues follow a particular pattern of subdialogues, which we refer to as the "backbone". Coaches establish rapport and get to know their clients before discussing anything goal-related, building up knowledge about their client that will help the coach guide the goal-setting subdialogue that follows. Coach and client establish a specific and realistic goal and an action plan to achieve it, discuss anticipated barriers and brainstorm resolutions and coping strategies, and discuss how the client's support network may help them in achieving this goal, either by reminding them or joining in the physical activity or holding them accountable for it. In follow-up sessions, emphasis is placed on exploring patient success and developing coping strategies for previously unanticipated barriers. Coaches thus guide clients in establishing a structure and pathway for success. This strategy guides our policy development.

The hierarchical dialogue policy is a key component in allowing us to direct the conversation in a sensible manner by supplying intents to the generation system. These intents dictate the kind of information the system wants to request from the user, such as slot values, clarifications, or confirmations. It draws from the knowledge base as needed, which contains user history and data and coaching knowledge. The high-level backbone, rooted in coaching theories and protocols, remains the same in all iterations. It will comprise multiple subdialogues for the goal-setting process (e.g., past goal progress and reflection, (re)negotiation of new goals, barrier resolution, etc.) and direct the flow of conversation sequentially through each of

these as the previous subdialogue completes. Each subdialogue will also have a policy, which will be rule-based initially. However, in time, we would like to expand the subdialogue policies to be more flexible. In freeform dialogue, users may provide more information in a response than was initially asked for. Humans naturally adjust their intents accordingly to avoid asking for the same information or to focus instead on discussing the extra information. We will conduct experiments to learn efficient, flexible strategies from the coaching transcripts to complete the subdialogue task. In later iterations, we will refine on collected dialogues and further incorporate the user's context, sensor data, and history to dynamically plan the ongoing dialogue for a more fluid and natural conversation.

5.3 Just in Time Adaptive Intervention (JITAI) Messages

JITAI messages are one-way messages that do not require a response and have their own policy. This policy is mostly driven by the agent architecture and will be based on the user's context, sensor data, and history, allowing us to implement logic for the timing of different types of JITAI messages. The logic will identify moments such as "the user has achieved a weekly goal" and "the user planned to exercise in the morning but it is supposed to rain" with their associated JITAI messages as well as identify whether it would be appropriate to send at that time. These topics have been explored before (Hardeman et al., 2019; Nahum-Shani et al., 2018; Mair et al., 2022; Ismail et al., 2022; Mutsuddi and Connelly, 2012; Muller et al., 2017).

While there is a consensus that the focus of JITAI messaging is to provide the user with the support they need at the time they need it, so that they can accomplish the goal of (in our case) increasing their PA, there is not much focus placed on categorizing the messages themselves other than to say that they are personalized/tailored messages that are motivational or encouraging (Nahum-Shani et al., 2018; Mair et al., 2022; Ismail et al., 2022). In order to preserve clarity, we have separated our JITAI messages into two main categories: anticipatory (which aim to to reduce barriers such as the weather, time of day or year, time management, planned meetings or events) and opportunistic (which provide encouragement and motivation at moments when there is perceived dwindling enthusiasm or when the user could take advantage of times they already unknowingly partake in activity). Anticipatory messaging can be planned in advance and delivered to the user at appropriate times that can be determined without complex sensing data (e.g., in the morning before leaving for work). The timing of opportunistic messaging is much more delicate as they must be delivered "in the moment" to be effective.

An example scenario for an anticipatory message would be to send the participant a message, while they are getting ready to leave their house to go to work, that lets them know that they are about to encounter one of their barriers and reminds them of the strategy they had already planned.

> *Remember to pack your umbrella! You planned to walk during your lunch break and there is a 50% chance of rain this afternoon.*

The purpose of this message is to anticipate a barrier that could cause the participant to fail at their goal for that day if not corrected in time.

Similarly, an example scenario for an opportunistic message would be to send the participant a message if their heart rate is going down and they only have 5 minutes left to finish their goal for the day.

> *Don't give up now! You only have 5 minutes left to go!*

The purpose of this message is to encourage the user at an opportunistic time to finish the goal they had set for themselves for that day.

5.4 Message Generation

Once either the dialogue or JITAI policy has determined the overall intent and data content of a given message, the next step is message generation. Due to the fact that we are implementing a complex task-oriented dialogue system, the fact that we have limited health coaching data and the importance of preserving the coaching language (e.g., the coaches must remain positive and encouraging through interactions, and there are guidelines for things that coaches should or should not say), the message generation will at first remain template-based. While Neural Natural Language Generation (NNLG) models have been improving greatly, they have many pitfalls when it comes to task-oriented dialogue systems. These include introducing hallucinated content (Reiter, 2018; Erdem et al., 2022), poor sentence planning and discourse operations (Reed et al., 2018), and not approximating human

generated text on complex problems (Wiseman et al., 2017; Erdem et al., 2022), especially in situations with a limited dataset.

While using a template-based method will help us avoid these pitfalls, they can be too structured and repetitive, which can hurt the user experience. Therefore, in an effort to introduce variety to our wordings over time, we plan to use what we call "flexible message templates." We first begin by identifying sentence fragment sections that we can put together to form our flexible templates. Each flexible template is made up of sentence fragment sections and any needed variables (e.g., proposed goal, dates, proposed strategies). Then, each sentence fragment section within the flexible template is replaced by one of multiple sentence fragment options that will together create a relatively unique message. We call them flexible templates both because each sentence fragment section could be used for multiple different templates and because in generating our templates in this way, we can create multiple different ways of saying the same message despite the overall generation being templated.

As we gather data during the collection of dataset 1 (as mentioned in section 4.1), we are looking to augment the number of flexible templates that cover the same purpose and content. However, since dataset 1 is speech-based while our system is text-based, we will need to handle the inherent differences between text and speech interactions and what that will mean for how our automated coach will need to differ from the human coach. As was encountered in Mitchell et al. (2021), during text-based interactions the health coaches felt like they could not have conversations that were as in-depth and nuanced because they were not just missing auditory input but also visual (e.g, body language, facial expressions, etc.). Additionally, they found that the health coaches found it hard to transition to a text platform because they had difficulties connecting to the user when they received short and ambiguous responses. As a result, we will make two assumptions: (1) the messages in text-based conversation need to be more direct and (2) the user is less likely to elaborate on little input.

Once we have augmented the flexible templates, instead of randomly selecting which flexible template to use in any given instance, we will explore ways to select the best template based on the conversation history and the users past reactions. We look to consider features such as message structure variability (e.g., if the last message had a prepositional phrase at the beginning, the next message should not), missing information (e.g., if we need to know three pieces of information, how many have already been given and what is remaining), and vocabulary variability (e.g., back-to-back messages should not use similar wording). We are taking inspiration from Razavi (2021), whose dialogue manager LISSA uses the user's last response to choose the best next response from multiple possible options.

In order to add more variety to the automated coach's speech, we aim to incorporate text style transfer techniques in order to affect the tone of the output by making adjustments in the emotion portrayed and politeness without needing to affect the content (Jin et al., 2020; Toshevska and Gievska, 2021). This requires user sentiment components for the NLU and dialogue policy and allows for the creation of a more empathetic, likable coach (Prendinger and Ishizuka, 2005).

Once we have more data following additional trials, we would also like to use information retrieval and constrained generation techniques to automate the generation of our flexible templates and sentence fragments. Recently, constrained generation research has put a focus on lexical constraints (He and Li, 2021; Miao et al., 2019; Mou et al., 2015; Li and Sun, 2018), which suits our needs in preserving the coaching language where we need to put soft and hard constraints on keywords or sentence formats that must be in the output and those that cannot appear in the output.

6 Formative Evaluation

The first iteration of our Dialogue and Messaging System will be both derived (as described above) and evaluated on dataset 1. We are aware that evaluating a system on the data that was used to derive it will bias it. However, since this will only be the rudimentary Dialogue and Messaging System, we do not believe the risk to be too great, since we will be further refining and evaluating the system with further trials. The evaluation needs to be separated into two parts: evaluating the NLU component and evaluating the message generation.

We mentioned above that basing our text-based system on the speech-based dataset 1 will affect the message generation by forcing us to make two assumptions: (1) the messages in text-based conversation need to be more direct and (2) the user is

less likely to elaborate on little input. These two assumptions will also affect the evaluation of our generated messages since we cannot evaluate on whether the two messages (one from dataset 1 and one generated by our system) are equivalent. Instead, we will need to evaluate on whether both messages ask the user for the same information given the same prompt.

Evaluating the NLU component could also be complicated due to the same assumptions. In this case, since our system is expecting more direct messages, the NLU component would expect that the user's response would be more straightforward. However, we can see *what* the NLU component can correctly recognize and this could be a worse-case situation. In addition, we can evaluate it on whether it reacts correctly to a message. Therefore, we will be evaluating the system on whether it correctly identifies the parameters it is expecting and on whether the policy correctly prompts the message generation.

7 Next Steps

Once we have a working Dialogue and Messaging System, we plan to lead two trials in order to evaluate and improve the system: Trial Alpha and Trial Beta.

Trial Alpha. In this trial, we will be generating a dataset we plan to release and evaluating our two-way dialogue. As with the collection of dataset 1, we will once again be collecting data from real user-human health coach interaction. This time, however, all interactions will be text-based and the human coach will have the same information as our automated coach will.

We expect the data labelling will function similarly as it did for dataset 1 (as described in the section 4.1). However, we hope that the data will be much cleaner and much more catered to the text domain. As previously mentioned, there was another dataset released in 2022 by Gupta et al. (2020a), but it does not cover barrier resolution and strategy negotiation. Therefore, we believe this dataset of labelled data will be very helpful in improving future health coaching research.

As far as evaluating our system goes, we will not need to evaluate our system around base assumptions like we will have to do for the Formative Evaluation. Therefore, the evaluation will be focused on four factors: (1) given two messages (one from dataset 2 and one generated by our system), is the content of both equivalent?, (2) is the language from the generated messages appropriate for a health coach? (3) are the parameters the NLU component is expecting reasonable?, and (4) does the NLU component correctly identify the parameters, and does the policy correctly prompt the message generation? To answer all these questions we will be using both standard metrics, such as BLEU-4 (Papineni et al., 2002), and human health coach manual evaluation.

Trial Beta. This trial is the first time that users will be using our system. It will serve to evaluate both our two-way dialogue and our JITAI messaging. By this point we hope to assess (1) how users respond JITAI messages and timing, (2) how users respond to our automated coach as opposed to the human coach in two way dialogue, (3) how successful the NLU component is at properly understanding the user, and (4) how successful the automated coach is when compared to the human coach. The goals and focus of this trial are subject to change based on the results of Trial Alpha.

8 Conclusion

Increasing engagement in MVPA and reducing sedentary behavior is a national priority for improving cardiovascular health. While wearable PA monitors show promise in initiating PA change, they do not assist the user in updating their PA goal, nor do they provide personalized messaging to assist the user in overcoming barriers to PA. Human coaching, following sound theoretical models of behavior change, has been demonstrated to be effective, but is hard to scale and misses the potential of "just-in-time" behavior suggestions and encouragement, as the coach is not always readily available.

Our Automated Coaching System is an integrated system to provide personalized, evidence-based, just-in-time feedback as well as interactive coaching including goal (re)negotiation, targeted at increasing PA and reducing risk for cardiovascular disease. Our system focuses on PA, but this infrastructure is modular and extendable to other health behaviors, including stress management and sleep hygiene.

Acknowledgements

This research was partially informed by data collected through National Institutes of Health award R21-AG056765-01.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Timothy W. Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient Educ Couns.*, 92(2):142–148.

Timothy W Bickmore, Daniel Schulman, and Candace L Sidner. 2011. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *Journal of biomedical informatics*, 44(2):183–197.

Judit Bort-Roig, Nicholas D. Gilson, Anna Puig-Ribera, Ruth S. Contreras, and Stewart G. Trost. 2014. Measuring and influencing physical activity with smartphone technology: A systematic review. *Sports Medicine*, 44(5):671–686.

Alison Cawsey, Ray B. Jones, and Janne Pearson. 2000. The evaluation of a personalised health information system for patients with cancer. *User Modeling and User-Adapted Interaction*, 10:47–72.

Alison Cawsey, Bonnie Webber, and Ray Jones. 1997. Natural language generation in health care. *Journal of the American Medical Informatics Association*, 4(6):473–482.

Nathalie Colineau and Cecile Paris. 2011. Motivating reflection about health within the family: the use of goal setting and tailored feedback. *User Modeling and User-Adapted Interaction*, 21:341–376.

Brian G. Danaher, Håvar Brendryen, John R. Seeley, Milagra S. Tyler, and Tim Woolley. 2015. From black box to toolbox: Outlining device functionality, engagement activities, and the pervasive information architecture of mHealth interventions. *Internet Interventions*, 2(1):91–101.

Sarah M Dennis, Mark Harris, Jane Lloyd, Gawaine Powell Davies, Nighat Faruqi, and Nicholas Zwar. 2013. Do people with existing chronic conditions benefit from telephone coaching? a rapid review. *Australian Health Review*, 37(3):381–388.

Carlo C. DiClemente, Angela S. Marinilli, Manu Singh, and Lori E. Bellino. 2001. The role of feedback in the process of health behavior change. *American Journal of Health Behavior*, 25(3):217–227.

Peter Düking, Marie Tafler, Birgit Wallmann-Sperlich, Billy Sperlich, and Sonja Kleih. 2020. Behavior change techniques in wrist-worn wearables to promote physical activity: Content analysis. *JMIR mHealth and uHealth*, 8(11):e20820.

Elizabeth G Eakin, Sheleigh P Lawler, Corneel Vandelanotte, and Neville Owen. 2007. Telephone interventions for physical activity and dietary behavior change: a systematic review. *American journal of preventive medicine*, 32(5):419–434.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Andrew Farmer and Lionel Tarassenko. 2015. Use of wearable monitoring devices to change health behavior. *JAMA*, 313(18):1864.

Brianna S. Fjeldsoe, Alison L. Marshall, and Yvette D. Miller. 2009. Behavior change interventions delivered by mobile telephone short-message service. *American Journal of Preventive Medicine*, 36(2):165–173.

J. Graham, K. Decker, and M. Mersic. 2003. Decaf: A flexible multi-agent system architecture. *Autonomous Agents and Multi-Agent Systems*, 7(1–2):7–27.

Yi Guo, Jiang Bian, Trevor Leavitt, Heather K Vincent, Lindsey Vander Zalm, Tyler L Teurlings, Megan D Smith, and François Modave. 2017. Assessing the quality of mobile exercise apps based on the american college of sports medicine guidelines: A reliable and valid scoring instrument. *Journal of Medical Internet Research*, 19(3):e67.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020a. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256.

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2020b. Goal summarization for human-human health coaching dialogues. In *FLAIRS Conference*, pages 317–322.

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2021. Summarizing behavioral change goals from sms exchanges to support health coaches. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–289.

Wendy Hardeman, Julie Houghton, Kathleen Lane, Andy Jones, and Felix Naughton. 2019. A systematic review of just-in-time adaptive interventions (jitais) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 16(1):31.

Xingwei He and Victor OK Li. 2021. Show me how to revise: Improving lexically constrained sentence generation with xlnet. In *Proceedings of AAAI*, pages 12989–12997.

Megan P Heintzelman, Gregory M Dominick, Ajith Vemuri, and Keith Decker. 2022. Development of the be smart feasibility trial to increase physical activity in midlife adults. In *ANNALS OF BEHAVIORAL MEDICINE*, volume 56, pages S253–S253. OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA.

Briony Hill, Ben Richardson, and Helen Skouteris. 2015. Do we know how to design effective health coaching interventions: a systematic review of the state of the literature. *American Journal of Health Promotion*, 29(5):e158–e168.

Tasnim Ismail, Dena Al Thani, et al. 2022. Design and evaluation of a just-in-time adaptive intervention (jitai) to reduce sedentary behavior at work: Experimental study. *JMIR Formative Research*, 6(1):e34309.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *arXiv preprint arXiv:2011.00416*.

AC King, EB Hekler, and LA Grieco. 2016. Effects of three motivationally targeted mobile device applications on initial physical activity and sedentary behavior change in midlife and older adults: A randomized trial. *PLOS ONE*, 11(6):e0156370.

Kirsi Kivelä, Satu Elo, Helvi Kyngäs, and Maria Kääriäinen. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling*, 97(2):147–157.

Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. 2018. Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of heartsteps. *Ann behave med*.

Lean L Kramer, Silke Ter Stal, Bob C Mulder, Emely de Vet, and Lex van Velsen. 2020. Developing embodied conversational agents for coaching people in a healthy lifestyle: Scoping review. *Journal of medical Internet research*, 22(2):e14058.

Aniek J Lentferink, Hilbrand KE Oldenhuis, Martijn de Groot, Louis Polstra, Hugo Velthuijsen, and Julia EWC van Gemert-Pijnen. 2017. Key components in ehealth interventions combining self-tracking and persuasive ecoaching to promote a healthier lifestyle: A scoping review. *J Med Internet Res.*, 19(8):e277.

Jingyuan Li and Xiao Sun. 2018. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. *arXiv preprint arXiv:1806.07000*.

Felipe Lobelo, Heval M. Kelli, Sheri Chernetsky Tejedor, Michael Pratt, Michael V. McConnell, Seth S. Martin, and Gregory J. Welk. 2016. The wild wild west: A framework to integrate mHealth software applications and wearables to support physical activity assessment, counseling and interventions for cardiovascular disease risk reduction. *Progress in Cardiovascular Diseases*, 58(6):584–594.

Susan Mahon, Rita Krishnamurthi, Alain Vandal, Emma Witt, Suzanne Barker-Collo, Priya Parmar, Alice Theadom, Alan Barber, Bruce Arroll, and Elaine Rush. 2018. Primary prevention of stroke and cardiovascular disease in the community (prevents): Methodology of a health wellness coaching intervention to reduce stroke and cardiovascular disease risk, a randomized clinical trial.

Jacqueline Louise Mair, Lawrence D Hayes, Amy K Campbell, Duncan S Buchan, Chris Easton, and Nicholas Sculthorpe. 2022. A personalized smartphone-delivered just-in-time adaptive intervention (jitabug) to increase physical activity in older adults: Mixed methods feasibility study. *JMIR formative research*, 6(4):e34662.

C. Di Marco, P. Bray, H.D. Covvey, D.D. Cowan, V. Di Ciccio, E. Hovy, Joan Lipa, and C. Yang. 2006. Authoring and generation of individualized patient education materials. *AMIA Annu Symp Proc*, pages 195–199.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.

Anouk Middelweerd, Julia S Mollee, C Natalie van der Wal, Johannes Brug, and Saskia J te Velde. 2014. Apps to promote physical activity among adults: a review and content analysis. *International Journal of Behavioral Nutrition and Physical Activity*, 11(1).

Elliot G. Mitchell, Rosa Maimone, Andrea Cassells, Jonathan N. Tobin, Patricia Davidson, Arlene M. Smaldone, and Lena Mamykina. 2021. Automated vs. human health coaching: Exploring participant and practitioner experiences. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).

François Modave, Jiang Bian, Trevor Leavitt, Jennifer Bromwell, Charles Harris III, and Heather Vincent. 2015. Low quality of free coaching apps with respect to the american college of sports medicine guidelines: A review of current mobile apps. *JMIR mHealth and uHealth*, 3(3):e77.

Lili Mou, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2015. Backward and forward language modeling for constrained sentence generation. *arXiv preprint arXiv:1512.06612*.

AM Muller, A Blandford, and L Yardley. 2017. The conceptualization of a just-in-time adaptive intervention (jitai) for the reduction of sedentary behavior in older adults. *mHealth*, 3:37.

Adrià Muntaner, Josep Vidal-Conti, and Pere Palou. 2016. Increasing physical activity through mobile device interventions: A systematic review. *Health Informatics Journal*, 22(3):451–469.

Adity U Mutsuddi and Kay Connelly. 2012. Text messages for encouraging physical activity are they effective after the novelty effect wears off? In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 33–40. IEEE.

Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462.

Eugene Z Oddone, Jennifer M Gierisch, Linda L Sanders, Angela Fagerlin, Jordan Sparks, Felicia McCant, Carrie May, Maren K Olsen, and Laura J Damschroder. 2018. A coaching by telephone intervention on engaging patients to address modifiable cardiovascular risk factors: a randomized controlled trial. *Journal of general internal medicine*, 33(9):1487–1494.

Harm op den Akker, Miriam Cabrita, Rieks op den Akker, Valerie M. Jones, and Hermie J. Hermens. 2015. Tailored motivational message generation: A model and practical framework for real-time physical activity coaching. *Journal of Biomedical Informatics*, 55:104–115.

Harm op den Akker, Valerie M. Jones, and Hermie J. Hermens. 2014. Tailoring real-time physical activity coaching systems: a literature survey and model. *User Modeling and User-Adapted Interaction*, 24(5):351–392.

Gillian A. O'Reilly and Donna Spruijt-Metz. 2013. Current mHealth technologies for physical activity assessment and promotion. *American Journal of Preventive Medicine*, 45(4):501–507.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137.

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, 19:267–285.

Seyedeh Zahra Razavi. 2021. *Dialogue management and turn-taking automation in a speech-based conversational agent*. Ph.D. thesis, DAI-A 83/4(E), Dissertation Abstracts International.

Lena Reed, Shereen Oraby, and Marilyn Walker. 2018. Can neural generators for dialogue learn sentence planning and discourse structuring? *arXiv preprint arXiv:1809.03015*.

Ehud Reiter. 2018. Hallucination in neural nlg.

Michael A Rupp, Jessica R Michaelis, Daniel S McConnell, and Janan A Smither. 2018. The role of individual differences on perceptions of wearable fitness device trust, usability, and motivational impact. *Applied ergonomics*, 70:77–87.

Matthew Saponaro. 2020. *Adaptive Real-time Coaching in Free-living Conditions*. Ph.D. thesis, University of Delaware.

Matthew Saponaro, Ajith Vemuri, Greg Dominick, and Keith Decker. 2021. Contextualization and individualization for just-in-time adaptive interventions to reduce sedentary behavior. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 246–256, New York, NY, USA. Association for Computing Machinery.

Matthew Saponaro, Haoran Wei, and Keith Decker. 2017. Towards learning efficient intervention policies for wearable devices. In *Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017 IEEE/ACM International Conference on*, pages 298–299. IEEE.

Susan M Schembre, Yue Liao, Michael C Robertson, Genevieve Fridlund Dunton, Jacqueline Kerr, Meghan E Haffey, Taylor Burnett, Karen Basen-Engquist, and Rachel S Hicklen. 2018. Just-in-time feedback in diet and physical activity interventions: systematic review and practical design framework. *Journal of medical Internet research*, 20(3):e106.

D Spruijt-Metz, C Wen, and G O'Reilly. 2015. Innovations in the use of interactive technology to support weight management. *Curr Obes Rep*, 4(4):510–519.

Laura P Svetkey, Bryan C Batch, Pao-Hwa Lin, Stephen S Intille, Leonor Corsino, Crystal C Tyson, Hayden B Bosworth, Steven C Grambow, Corrine Voils, and Catherine Loria. 2015. Cell phone intervention for you (city): a randomized, controlled trial of behavioral weight loss intervention for young adults using mobile technology. *Obesity*, 23(11):2133–2141.

Martina Toshevska and Sonja Gievska. 2021. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*.

Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Alvaro Alonso, Andrea Z Beaton, Marcio S Bittencourt, Amelia K Boehme, Alfred E Buxton, April P Carson, Yvonne Commodore-Mensah, et al. 2022. Heart disease and stroke statistics—2022 update: a report from the american heart association. *Circulation*, 145(8):e153–e639.

Ajith Vemuri, Keith Decker, Matthew Saponaro, and Greg Dominick. 2021. Multi agent architecture for automated health coaching. *Journal of Medical Systems*, 45(11).

Julie B. Wang, Lisa A. Cadmus-Bertram, Loki Natarajan, Martha M. White, Hala Madanat, Jeanne F. Nichols, Guadalupe X. Ayala, and John P. Pierce. 2015a. Wearable sensor/device (fitbit one) and SMS text-messaging prompts to increase physical activity in overweight and obese adults: A randomized controlled trial. *Telemedicine and e-Health*, 21(10):782–792.

Julie B Wang, Lisa A Cadmus-Bertram, Loki Natarajan, Martha M White, Hala Madanat, Jeanne F Nichols, Guadalupe X Ayala, and John P Pierce. 2015b. Wearable sensor/device (fitbit one) and sms text-messaging prompts to increase physical activity in overweight and obese adults: a randomized controlled trial. *Telemedicine and e-Health*, 21(10):782–792.

Rachel Willard-Grace, Ellen H. Chen, Danielle Hessler, Denise DeVore, Camille Prado, Thomas Bodenheimer, and David H. Thom. 2015. Health coaching by medical assistants to improve control of diabetes, hypertension, and hyperlipidemia in low-income patients: A randomized controlled trial. *The Annals of Family Medicine*, 13(2):130–138.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Comput. Speech Lang.*, 24(2):150–174.

Personalizing Weekly Diet Reports

Elena Monfroglio and **Luca Anselma** and **Alessandro Mazzei**
Dipartimento di Informatica, Università di Torino
elena.monfroglio@edu.unito.it, luca.anselma@unito.it, alessandro.mazzei@unito.it

Abstract

In this paper we present the main components of a weekly diet report generator (DRG) in natural language. The idea is to produce a text that contains information on the adherence of the dishes eaten during a week to the Mediterranean diet. The system is based on a user model, a database of the dishes eaten during the week and on the automatic computation of the Mediterranean Diet Score. All these sources of information are exploited to produce a highly personalized text. The system has two main goals, related to two different kinds of users: on the one hand, when used by dietitians, the main goal is to highlight the most salient medical information of the patient diet and, on the other hand, when used by end users, the main goal is to educate them toward a Mediterranean style of eating.

1 Introduction

The diet has a huge impact on the health of people, and a number of studies have tried to apply artificial intelligence techniques to this domain. On the one hand, there is a growing interest in applying computational techniques in all the aspects of food production, monitoring, consumption (Min et al., 2019). On the other hand, diet is one of the main factors influencing human health and it has been studied in the field of health informatics (Mazzei et al., 2020; Balloccu et al., 2020).

In the domain of health, it has been shown that one of the main risk factors in the onset of chronic diseases lies in the adoption of an unhealthy diet (Jayedi et al., 2020). Specifically, following a Mediterranean diet provides many health benefits (Godos et al., 2019; Schwingshackl et al., 2017; Galbete et al., 2018). However, following a diet is often difficult both for the specific complexity of the domain, and for the human tendency to transgress on eating behaviors (Anselma et al., 2017). A diet can be seen as a set of quantitative or qualitative rules and constraints, and technological tools could support users both in keeping track of the historical data and of the user progress, and obtaining motivation by means of their educational and persuasive roles. A virtual dietitian that reasons about eaten meals and that communicates through natural language suggesting corrective actions can be helpful in this task. The Multimedia Application for Diet Management (Anselma and Mazzei, 2015, 2018, 2020) (MADiMan[1]) was born in 2015 in order to build a virtual dietitian that is able to: (i) let the user choose the meal to eat through a mobile application, (ii) analyze the ingredients of the recipe and their quantity through the NLU module, (iii) evaluate the compatibility of the chosen meal with the principles of a diet through the Reasoner module, (iv) determine what the consequences of eating a particular dish are, (v) show these consequences to the user through natural language with messages for educational and informational purposes and motivating users to pursue their goals.

A recent development of MADiMan (Mazzei et al., 2020) concerns the integration of the *Mediterranean diet score* (*Med Score* henceforth) originally proposed in (Stefanadis, 2006). By using a food ontology, MADiMan is able to reason both (i) on macronutrient-based constraints typical of medical diets (e.g., *eat 0.8 g of proteins per kilogram of body weight per day*), and (ii) on food-based constraints typical of Mediterranean diet (e.g., *use daily olive oil in cooking*). The Med Score (0-55) is based on the specific scores (0-5) obtained over the consumption of 11 food categories during a week. Some categories prescribe to eat no more than a limit (e.g., no more than 2 portions of red meat per week), and others not less than a limit (e.g., not less than 5 portions of fish per week).

The MADiMan system includes modules that accompany the user in real-time in the contingent choices of individual meals, but a drawback is the

[1] http://di.unito.it/madiman

Proceedings of the First Workshop on Natural Language Generation in Healthcare (NLG4Health), pages 40 - 45

absence of a summary that allows the user to consolidate the results obtained at the end of the week. Furthermore, the implementation design lacks a proper personalization of the messages, since the NLG module does not take into account any personal data/preferences or the emotional state of the users.

This work has the intent to fulfill this limitation by producing a longer weekly report that educates the user. This automatic report is built with a higher degree of personalization, by formalizing different *user models*, in order to support different types of users who can access the platform and to implement the related communication strategies. Consequently, the information flow analyzed so far is enriched with a long report, which is sent to the user in the form of an e-mail on a weekly basis, in order to represent the habits held in the past week and to suggest which behaviors to encourage for the future and which ones to avoid.

The main research goal of this paper is to evaluate the impact of personalization on the quality of automatically generated weekly diet reports. With this aim, we first describe the main design choices in the DRG system and then we give the results of a preliminary evaluation of the system.

The paper is structured as follows: in Section 2 the concept of user model is introduced and its implementation is described. In Section 3, we describe DRG, a multilingual (Italian/English) generator that follows the typical modules of an automatic NLG system, in relation to the persuasiveness and the educational impact of the generated messages. In Section 4, we provide the results of an initial evaluation of the system and, finally, in Section 5 we conclude the paper describing some ongoing developments.

2 The User Models

In the domain of e-Health, personalization can play a role for achieving some form of engagement (DiMarco et al., 2007), and user models play a key role in personalizing automatically generated messages. For the diet domain, a user model contains both personal information about the health status (e.g. weight) as well as user's preferences on specific topics. In particular, DRG has been designed by considering two specific categories of users, that are the *patients* and the *dietitians*. The personalization of the messages is based on the different goals that these two kinds of users have. Note that in the first case the personalization needs to consider just the patient user model, but in the second case the personalization needs to consider both the dietitian (the message addressee) and the patient (the message topic).

On the one hand, the messages generated by DRG for *patients* have to be informative, motivational and educational. The final goal of the system is to educate the patients toward a better understanding of the Mediterranean diet principles using an emotional engaging language based on some psychological heuristics. On the other hand, the messages generated by DRG for *dietitians*, that are medical specialists on nutrition (in some cases physicians), should be as short as possible, should contain information just on bad behavior of the patient, and should use a technical lexicon without emotional content. Note that we decided to not communicate information on the good behavior to the dietitians since we think that in a support system for an expert is more important to produce a summary of the problems. However, if a dietitian prefers otherwise, it is possible to adopt a different policy by changing the DRG configuration.

On the basis of these differences, the patient user model contains: (1) numerical personal/medical information on the user, storing sex, age, weight, height, and BMI (Body Mass Index); (2) a 1-to-4 point scale for representing the stress level based on the DASS-21 questionnaire (Lovibond and Lovibond, 1995); (3) a Boolean variable representing the interest of the user for food sustainability, that is a sort of sensitivity to environmental issues. Using this source of information we can produce a specific personalization for the specific patient. In contrast, with the aim to produce a *technical* message, all the dietitians, for a specific patient, will read the same message.

3 The DRG Architecture

The DRG Architecture (Figure 1) follows the standard modular architecture of symbolic NLG (Reiter and Dale, 2000; Reiter, 2007). The generation flow starts from numerical data representing the weekly diet of a patient. The diet reasoner, a module of the MADiMan system, produces and stores in a relational database the information regarding the dishes eaten during a week, their recipes, their nutritional values and their Med Scores. Also the user model information of the various users are stored, in the same relational database.

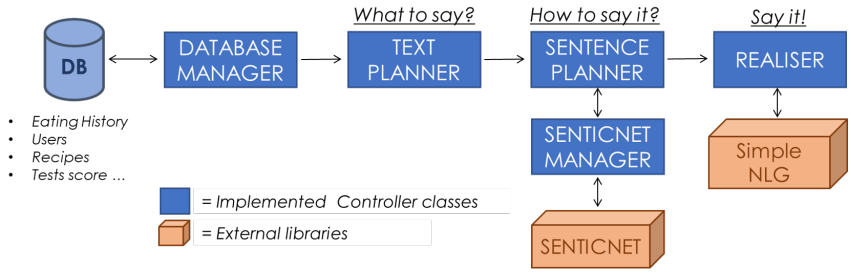

Figure 1: The DRG architecture.

Following (Reiter, 2007), we divided the generation process into three macro-phases implementing the specific generation tasks. The *text planning* phase implements the selection of information units to communicate (content determination) and the order in which they appear (text structuring). The creation of these information units follows the idea to aggregate together semantically equivalent information. The different categories of food are aggregated based on their scores over four different possible values: *very good*, *good*, *bad* and *very bad*. Text structuring decides in which order the information should be presented: following the so-called *sandwich technique*, we communicate the units following the *very good, bad, good, very bad* order, that is alternating a positive and a negative communication. In Figure 2 we report an example of text plan. Note that the text plan contains the Med Score value, computed by the reasoner, rather than the frequency of consumption of each food category.


```
{"language":"English",
 "user name":"Giulia",
 "user age":44,
 "user gender":"F",
 "user stress":0,
 "domain knowledge":0,
 "Med score":31,
 "last Med score":26,
 "very good":[{"name":"cer","score":5},
              {"name":"veg","score":5},
              {"name":"fish","score":5},
              {"name":"oil","score":5}],
      "good":[{"name":"pot","score":3},
              {"name":"leg","score":3},
              {"name":"poul","score":3}],
      "bad":[{"name":"fru","score":2}],
 "very bad":[{"name":"rmeat","score":0},
             {"name":"dairy","score":0}],
 "best dish":206,
 "worst dish":288,
 "total environment score":1000.82,
 "bad category environment":"rmeat"}
```


Figure 2: An example of text plan.

The *sentence planning* phase is responsible for building the syntactic structures of the messages. So, starting from the sequence of information unit produced in the text planning, a rule-based sentence planner decides both the syntax and the lexical items of the sentences. We defined a fixed schema based on a sequence of ten elements: (a) greetings, (b) Med Score, (c) encouragement, (d) very good score, (e) bad score, (f) good score, (g) very bad score, (h) best and worst dish of the week, (i) environmental impact, (j) educational notion on the Mediterranean diet. For each element (a-j), the sentence planner will use a specific quasi-tree, that is a sort of unordered and unlexicalized dependency tree (Anselma and Mazzei, 2020). The quasi-tree will be instantiated, producing a complete structure ready for realization, considering both the text plan and the user model. For instance, greetings (a) depend on the age, whilst the best/worst dishes (h), as well as the environmental impact (i), are not provided for dietitians. Moreover, for patients with a high level of stress the system does not provide information on the "very bad" category in order to not exacerbate their stress. For instance, in Figure 3 a sentence plan generated for dietitians is presented.

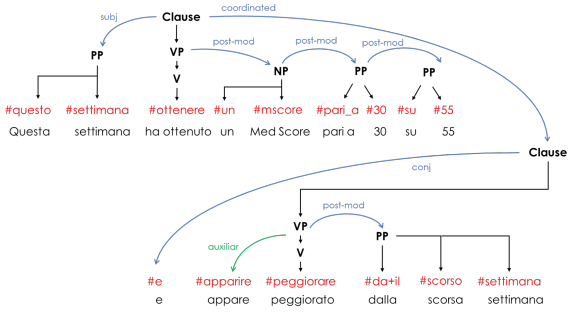

Figure 3: A sentence plan for the sentence "This week he has got a Med Score of 30 out of 55 and it seems to have gotten worse since last week." (translation from Italian).

The final process of the pipeline is the realization phase that accounts for function words insertion and inflection. Following the previous implementation of MADiMAn and, in order to build a bilingual Italian/English generator, for this phase we used the Italian porting (Mazzei et al., 2016) of the SimpleNLG realizer (Gatt and Reiter, 2009).

In Section 3.1, we give some details on the lexicalization that personalizes the sentences on the basis of some emotions.

3.1 Using SenticNet for lexicalization

In the field of NLG a number of works consider the use of affective strategies for the realization of an emotionally engaging text (de Rosis and Grasso, 2000; Mahamood and Reiter, 2011).

To give different emotional nuances to the final messages, we decided to use an emotional lexicon. SenticNet is a multilingual knowledge base designed for text sentiment analysis and provides a list of 150 000 lemmata, each one with different types of information, including primary and secondary emotions. Crucially, we used SenticNet for associating the stress level contained in the user model with emotion types. In SenticNet the lemmata are associated with emotions via type and polarity as described in the *Hourglass of Emotions* (Susanto et al., 2020) model. Specifically, the emotions are classified in four categories (introspection, temper, attitude and sensitivity), each one with six different polarity levels. Our idea is to associate with each stress level a specific type of emotion to mitigate the stress. We stipulate that, in correspondence to the stress levels, the types of emotions will be selected in this specific ascending order: sensitivity, introspection, attitude and temper. In the case of more lemmata with a same type of emotion, the *SenticNetManager* algorithm will prefer the lemma with highest polarity. In this way, we constrain DRG to select the least negative term. Thus, we built a DRG emotional lexicon for English and Italian by intersecting the original SimpleNLG lexicon with the SenticNet lexicon. Moreover, for each leaf of the quasi-trees, we defined a specific *synset* of words belonging to the DRG emotional lexicon. In this way, the *SenticNetManager* will choose among the words in the synset the best one in correspondence to a specific user stress level. For instance, let us suppose that a synset of a quasi-tree contains three lemmas: *choice*, *idea* or *decision*. On the one hand, in SenticNet *choice* and *idea* correspond both to the same emotion type, that is temper, that will be selected by *SenticNetManager* in the case of high stress; since *choice* has a higher polarity value will be preferred over *idea*. On the other hand, *decision* is related to introspection and it will be selected in case the stress is medium-low.

4 Initial Evaluation of DRG

We are aware that message personalization does not always correspond to an effective improvement for the end user (Reiter et al., 2003). So, in order to evaluate DRG, we performed two different evaluations. A first preliminary evaluation consisted in submitting a number of Italian and English texts generated by DRG to an adjunct professor of dietistic (henceforth, the *expert*). The evaluation was set up by generating ten different reports simulating the diet of ten patients. These simulations consist in randomly selected dishes from a database of recipes recovered from well-known web sites (e.g. BBC Food). For eight simulations, DRG generated a report personalized for the patient, and for two simulations DRG generated a report personalized for dietitians. By considering the specific user for which the text is generated, the expert had to evaluate a report in terms of: (i) readability, that consists in the linguistic quality of the report, (ii) accuracy or content quality and (iii) usefulness, that is the effective educational support that the system could provide to the patient. The general feedback of this first evaluation was positive, with a good level for all the three measurements. However, the expert suggested to improve the system in three directions: (i) to integrate the Med Score with information about macro/micronutrients (i.e. cholesterol, proteins, etc.); (ii) to provide more details about the ingredients; (iii) to enhance the personalization considering the patient's BMI.

A second preliminary and still ongoing evaluation was performed only for Italian language to have the patients' feedback. Similarly to the first evaluation, DRG generated four texts for four different patients, on the basis of a simulation consisting of randomly selected dishes. Moreover, we built a baseline text by simply listing all the information contained in the text plan (cf. Figure 2). In Table 1 we report an example of text generated by DRG and the corresponding baseline text.

The evaluation was set up in the form of an online form with the testing hypothesis that the users would prefer highly personalized report over the

	Italian version	English version
DRG	Ciao, Davide. Questa settimana hai ottenuto un Med Score pari a 23 su 50 e, inoltre, sei peggiorato dalla scorsa settimana. Non mollare! La quantità di patate, pesce ed olio era quasi eccellente. E inoltre hai fatto un lavoro fantastico con cereali e verdura. La merenda di venerdì era una scelta eccellente perché il piatto King Ranch Chicken Casserole ha una buona quantità di cereali e verdura. Gli esperti sconsiglierebbero il piatto Creamy Au Gratin Potatoes che hai mangiato il lunedì scorso a colazione perché la quantità di latte e derivati non è buona. Ricorda: una pessima dieta uccide più del fumo.	Hi Davide. This week you got a Med Score of 23 out of 50 and, furthermore, you have not improved since last week. Do not give up! The amount of potatoes, fish and oil was almost excellent. Furthermore, you've done a fantastic job with cereal and vegetables. Friday's snack was an excellent choice because the King Ranch Chicken Casserole dish has a good amount of grains and vegetables. Experts would advise against the Creamy Au Gratin Potatoes dish you ate for breakfast last Monday because the amount of milk and derivatives is not good. Remember: a bad diet kills more than smoking.
Baseline	Questa settimana hai ottenuto i seguenti punteggi: - Med Score: 23 su 50 - Carne rossa, latticini e carne bianca: 0 su 5 - Legumi e frutta: 1 su 5 - Pesce: 3 su 5 - Olio e patate: 4 su 5 - Cereali e verdura: 5 su 5 - Migliore piatto: King Ranch Chicken Casserole - Peggior piatto: Creamy Au Gratin Potatoes	This week you obtained the following scores: - Med Score: 23 out of 50 - Red meat, diary and poultry: 0 out of 5 - Legumes e fruit: 1 out of 5 - Fish: 3 out of 5 - Oil and potatoes: 4 out of 5 - Cereal and vegetables: 5 out of 5 - Best dish: King Ranch Chicken Casserole - Worst dish: Creamy Au Gratin Potatoes

Table 1: The text generated by DRG and the corresponding baseline text used for the evaluation. The Italian version is on the left and the English version, not used for evaluation, is on the right.

baseline. The form presents a user description, the baseline text and the DRG text (using a Latin square arrangement), and asks to evaluate the readability, the accuracy and the usefulness of each text by means of a 7-point Likert scale. Four pairs of reports along with a user description are shown. The four different cases were constructed by varying both the weekly dishes (randomly extracted) and the type of user for whom the report is generated. Currently, only five testers participated to the second evaluation, as reported in Table 2. We are aware that the small number of testers cannot guarantee statistically significant results. However, we can speculate that the readability score confirms the appealing of personalization in the linguistic quality of the text.

5 Conclusions and Ongoing Work

In this short paper we presented the main properties of DRG, that is a symbolic natural language generation system for building weekly report on Mediterranean diet. We are still evaluating our system by using the procedures described. Moreover, in order to have a more realistic and significant feedback on DRG, we are going to involve some students in dietistic in a form-based evaluation.

	Readability	*Accuracy*	*Usefulness*
DRG	5.65	5.3	5.45
Baseline	5.45	5.75	5.55

Table 2: Preliminary evaluation results (average over 7-point Likert scale).

As a future work, we intend to design a more complete comparative evaluation based on an ablation strategy. We intend to generate different versions of the report by excluding/exploiting the various components of DRG. In particular, we want to evaluate the contribution of the emotional lexicon in a A/B test.

References

Luca Anselma and Alessandro Mazzei. 2015. Towards diet management with automatic reasoning and persuasive natural language generation. In *Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings*, number 9273 in Lecture Notes in Computer Science, pages 79–90. Springer. ISBN 978-3-319-23484-7.

Luca Anselma and Alessandro Mazzei. 2018. Designing and testing the messages produced by a virtual dietitian. In *Proc. of 11th International Conference on Natural Language Generation (INLG 2018)*, pages 244–253. ACL.

Luca Anselma and Alessandro Mazzei. 2020. Building a Persuasive Virtual Dietitian. *INFORMATICS*, 7(3):1–27.

Luca Anselma, Alessandro Mazzei, and Franco De Michieli. 2017. An artificial intelligence framework for compensating transgressions and its application to diet management. *Journal of Biomedical Informatics*, 68:58–70.

Simone Balloccu, Steffen Pauws, and Ehud Reiter. 2020. A NLG framework for user tailoring and profiling in healthcare. In *Proceedings of the First Workshop on Smart Personal Health Interfaces co-located with 25th International Conference on Intelligent User Interfaces (IUI 2020)*, volume 2596, pages 13 – 32. CEUR-WS.

Fiorella de Rosis and Floriana Grasso. 2000. *Affective Natural Language Generation*, pages 204–218. Springer Berlin Heidelberg, Berlin, Heidelberg.

Chrysanne DiMarco, H. D. Covvey, Peter Bray, Donald Cowan, Vic DiCiccio, Eduard Hovy, Joan Lipa, and Doug Mulholland. 2007. The development of a natural language generation system for personalized e-health information. In *Proceedings of the 12th World Congress on Health (Medical) Informatics*, volume 129, pages 2339–2340. IOS Press.

Cecilia Galbete, Lukas Schwingshackl, Carolina Schwedhelm, Heiner Boeing, and Matthias B Schulze. 2018. Evaluating Mediterranean diet and risk of chronic disease in cohort studies: an umbrella review of meta-analyses. *European journal of epidemiology*, 33(10):909–931.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90–93, Stroudsburg, PA, USA. Association for Computational Linguistics.

Justyna Godos, Raffaele Ferri, Filippo Caraci, Filomena Irene Ilaria Cosentino, Sabrina Castellano, Fabio Galvano, and Giuseppe Grosso. 2019. Adherence to the Mediterranean diet is associated with better sleep quality in italian adults. *Nutrients*, 11(5):976.

Ahmad Jayedi, Sepideh Soltani, Anna Abdolshahi, and Sakineh Shab-Bidar. 2020. Healthy and unhealthy dietary patterns and the risk of chronic disease: an umbrella review of meta-analyses of prospective cohort studies. *British Journal of Nutrition*, 124(11):1133–1144.

P.F. Lovibond and S.H. Lovibond. 1995. The structure of negative emotional states: Comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. *Behaviour Research and Therapy*, 33(3):335–343.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, Edinburgh, UK. Association for Computational Linguistics.

Alessandro Mazzei, Antonio Lieto, Mirko Di Lascio, and Luca Anselma. 2020. Adopting the mediterranean diet score in a diet management system. In *13th International Joint Conference on Biomedical Engineering Systems and Technologies-HEALTHINF*, pages 670–676. SCITEPress.

Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *ACM Comput. Surv.*, 52(5).

E. Reiter, R. Robertson, and L. Osman. 2003. Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence*, 144:41–58.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proc. of the 11th European Workshop on Natural Language Generation*, ENLG '07, pages 97–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Lukas Schwingshackl, Carolina Schwedhelm, Cecilia Galbete, and Georg Hoffmann. 2017. Adherence to mediterranean diet and risk of cancer: an updated systematic review and meta-analysis. *Nutrients*, 9(10):1063.

Demosthenes B. Panagiotakos, Christos Pitsavos, Christodoulos Stefanadis. 2006. Dietary patterns: A Mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk. *Nutrition, Metabolism and Cardiovascular Diseases*, 16(8):559–568.

Yosephine Susanto, Andrew G Livingstone, Bee Chin Ng, and Erik Cambria. 2020. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

Author Index