ADELE: Evaluating and Benchmarking an Artificial Conversational Care Agent

Brendan Spillane
ADAPT Centre,
Trinity College Dublin
brendan.spillane@adaptcentre.ie

Benjamin Cowan
ADAPT Centre,
University College Dublin
benjamin.cowan@ucd.ie

Christian Saam
ADAPT Centre,
Trinity College Dublin
christian.saam@adaptcentre.ie

Vincent Wade
ADAPT Centre,
Trinity College Dublin
vincent.wade@adaptcentre.ie

Emer Gilmartin
ADAPT Centre,
Trinity College Dublin
gilmare@tcd.ie

Abstract
This paper provides an overview of a planned experiment to evaluate the latest iteration of ADELE, an artificial conversational agent to aid in the care of the elderly. It is being evaluated against an earlier iteration of itself and against the leading systems from the 2018 ConvAI2 Challenge at NeurIPS (formally NIPS) to measure its progress on producing improved natural social dialogue.

Author Keywords
Artificial Conversational Agent; Conversational Agent for the Elderly; Social Dialogue Agent; Evaluating; Benchmarking.

CCS Concepts
•**Human-centered computing** → **Natural language interfaces;** Empirical studies in HCI; •**Computing methodologies** → *Discourse, dialogue and pragmatics;*

Introduction
ADELE is being developed as an artificial conversational agent (ACA) to aid in the care of the elderly by providing them with health and well-being related advice and monitoring through social dialogue. Users will be able to converse with ADELE naturally through informal yet informed social dialogue on a variety of topics. Introduced in [23] and expounded upon in [22], ADELE was entered into the ConvAI2 challenge at NeurIPS in 2018 where it came second

in the automatic evaluation and sixth in the overall competition (Note: the competition entry was called "Adapt Centre") [5]. The 2018 engine was based on a Recurrent Neural Network Sequence-to-Sequence Model. The 2020 engine has evolved into a Transformer Decoder Model taking advantage of Transfer Learning from generic large-scale Language Models trained by Generative Pretraining. ADELE currently supports text-based interaction.

The purpose of this experiment is to evaluate and benchmark the dialogue produced by the 2020 version of ADELE to the 2018 iteration, and at least two other systems from the leading 2018 ConvAI2 entries. This will be undertaken using the DialCrowd Framework and Toolkit [14].

Background

The project is comprised of a team of machine learning, HCI, and personalisation experts who are working on the design and evaluation of ACAs to aid in the care of the elderly. ADELE is being designed to deliver health and well-being personal care advise and monitoring through social and informal dialogue in a similar fashion to a district care worker visiting the elderly in their own homes.

As such, it poses a number of research challenges to overcome. These include trying to better understand the nature of *good conversation* [1], how topic transitions [21] can be used to extend social talk [9, 8], and concerns about trust in ACAs for the care of the elderly [24]. The project is currently focused on evaluating ADELE's ability to produce high quality natural or human like social dialogue. In future, it will be evaluated as a means of delivering health and well-being personal care advise through social dialogue. Longer term, it will be evaluated in realistic care settings with elderly users.

Motivation

Population ageing has become an increasingly acute challenge in Europe [10], Japan [15], North America [4], and China [28]. Other countries are also beginning to experience the same problem. According to the UN [18], for the first time there are now more people in the world over 65 than under 5, and by the year 2050, 1 in 6 people in the world will be over 65, up from 1 in 11 in 2019. In China, many young people are now caring for the needs of up to four grandparents on their own [30].

Many countries have responded with large scale investment into research and technology to aid in the care of the elderly [11], though others have pointed out that this has been slow to result in business success [26, 12]. One area receiving particularr attention is ACAs which are capable of unconstrained natural language input. Laranjo et al. provide a detailed review of fourteen such natural language conversational agents [13]. ADELE furthers the state of the art by combining unconstrained natural language input with personalisation technologies to create a more natural social dialogue which can be used to deliver health and well-being advice and monitoring.

Evaluation Experiment

Hypothesis

This experiment will evaluate ADELE and benchmark its performance against an earlier iteration (2018) of itself and other leading ACAs. We aim to identify how ADELE performs on a number of key established metrics from the literature. We hypothesize that:

H_A: ADELE (2020) will produce more engaging, coherent and natural dialogue than the other dialogue agents.

Design

The number of ACAs being evaluated in this experiment is dependent on how many of the 2018 ConvAI2 Challenge entries are integrated with the DialCrowd experiment framework and toolkit [14]. Currently, 4 ACAs are integrated: ADELE 2018, ADELE 2020, and two other ACAs. If more of the ConvAI2 entries are integrated with DialCrowd, our experiment design will be adjusted accordingly. Currently, the experiment is being set up as a 4X1 between-subjects design. Participants will be randomly assigned to one of four groups to interact with one of the four ACAs without knowing which ACA they are interacting with.

Scenario and Dataset

The experiment scenario will be the Persona-Chat task which was used in the ConvAI2 Challenge. The purpose of the task is to model normal conversation when two interlocutors meet for the first time and try to get to know each other. Their purpose is to be engaging while learning about the other's interests and discussing their own interests to find common ground. The task involves asking and answering questions while maintaining a persistent persona which is provided in the dataset. Each of the systems will be trained on the Persona-Chat dataset [29]. It contains 10,907 dialogues with 162,064 utterances. The dataset also contains 1155 unique personas with at least 5 sentences with revised descriptions.

Metrics and Scales

The main concerns (in no particular order) guiding the selection of metrics for this evaluation are: 1) The dialogue is text based. 2) ADELE is designed to produce non task based social dialogue. 3) The Persona-Chat dataset was evaluated with four human metrics (fluency, engagingnes, consistency, and profile detection) [29]. 4) The metrics must suit the star rating scales (1-5 stars) DialCrowd supports. 5) A limited number of metrics is preferred to prevent task fatigue and to prevent too much overlap. 6) In this experiment, ADELE is being evaluated and benchmarked as an ACA and not on its ability to deliver health and well-being related advice, however it would be optimal if some of the metrics would also carry over into future evaluations. 7) The metrics should already be established in the literature.

Engagement is an obvious choice based on 1, 2, 3, 4, 6 and 7 [29, 20]. **Coherence** was selected based on 1, 4, 6, and 7 [27]. **Naturalness** was selected based on 1, 2, 4 and 7 [17, 20]. These will be measured using star rating scales (1-5) which produce ordinal type data. The mean score for each of the metrics will be reported for each of the ACAs. The mean of the three metrics will be used to satisfy the experiment hypothesis. In future, when ADELE ability to diliver healthcare and well-being related advice and monitering are being evaluated, additional metrics such as likeability, felt support, caring, and warmth etc. may also be used [17]. Participants will also be asked to provide qualitative feedback on the system via questions with free form text responses.

Participant Recruitment

Participants will be crowdsourced online via Amazon Mechanical Turk (AMT) and randomly assigned to one of the four groups. Participants will be over 18 and will be required to have native English language proficiency.

Sample Size

Estimating sample size a priori is difficult as there is currently no way to compute this for a Kruskal-Wallis H Test. However, Mahoney and Magel [16] maintain that an f-test can be used to approximate power, and that the Kruskal-Wallis Test may be more robust than a similar f-test. G*Power [7, 6] was used to approximate that at least 436 participants (4 groups of 109) will be sufficient to have a >0.95 power

of detecting an effect size of 0.2 which is considered small. This was chosen based on the work of Cohen [3], Coe [2], and Nelson [19]. Should more systems become available on DialCrowd the sample size will be adjusted accordingly. Post hoc, we will report the achieved effect size in the form of eta squared based on [25].

Statistical Analysis
Kruskal-Wallis H tests will be used to compare the means of the individual metric scores, and the overall combined mean, for each of the four systems. The results of combined mean will be used satisfy the hypothesis.

Importance of Evaluation and Benchmarking

The 2018 version of ADELE performed very well in the ConvAI2 Challenge coming 2nd on the automatic evaluation and 6th on the final human evaluation. Since then, ADELE has undergone significant redevelopment to integrate the latest technologies and improve the quality of the dialogue it produces. This experiment was conceived to focus on the human evaluation. By using the same Persona-Chat dataset and experiment scenario, it will provide a fair evaluation of both the ADELE 2018, and 2020 systems, and the two other leading systems. By adopting new but established human evaluation metrics, it will also be possible to see the strengths and weaknesses of each system.

This evaluation and benchmarking of ADELE as a ACA is important as this needs to be undertaken before it can be evaluated as a means of delivering health and well-being related advice through social dialogue. Longer term, completing these two evaluations is important so that ADELE can be evaluated in a staged care setting and subsequently in a real care setting with elderly users.

Constraints and Limitations of the Evaluation

Due to the nature of the ConvAI2 challenge we cannot exactly replicate all of the experiment conditions, particularly around the automatic evaluation. However, this has likely proved to be a boon, as in doing so would have severely limited innovation and control. Our solution, facilitated by DialCrowd, is to utilize the same Persona-Chat task and dataset, and to utilize suitable and established human evaluation metrics taken from the literature. This will allow others to easily replicate our experiment with their own systems to measure and benchmark their progress. The usefulness of DialCrowd and the release of the large Persona-Chat dataset must be acknowledged for facilitating this.

One criticism of DialCrowd and other experiment frameworks such as ParlAI is that currently they can also be quite limiting in terms of the experiment designs they support and individual experiment options. One example of the limitations is the single type of measurement scale (star rating scale) that is currently supported on DialCrowd. Star rating scales produce ordinal type data which limits the analysis to non-parametric statistics. Consequently, the experiment is forced to use a Kruskal-Wallis H test rather than an ANOVA.

Conclusion

This paper provided an overview of the planned evaluation and benchmarking of ADELE, a personalised ACA designed to aid in the care of the elderly. Our approach is to utilize the DialCrowd experiment framework and Persona-Chat dataset which will allow us to evaluate and benchmark the latest iteration, ADELE 2020 against ADELE 2018, and other leading systems from the domain. This allows others to replicate our approach and benchmark their ACAs against ADELE 2020. The experiment has received ethics approval and it will be executed in the coming weeks with the results being submitted for publication in the future.

Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant Number 13/RC/2106.

REFERENCES

[1] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, and et al. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 475–486. DOI: http://dx.doi.org/10.1145/3290605.3300705 event-place: Glasgow, Scotland Uk.

[2] Robert Coe. 2002. It's the effect size, stupid: What effect size is and why it is important. In *Proceedings of the British Educational Research Association Annual Conference*. 18. http://www.leeds.ac.uk/educol/documents/00002182.htm 00640.

[3] Jacob Cohen. 1969. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press. 00773 Google-Books-ID: kx0ajwEACAAJ.

[4] Eileen M. Crimmins, Hiram Beltrán-Sánchez, Lauren Brown, and Yongjie Yon. 2017. *Ageing in North America: Canada and the United States*. Oxford University Press. Google-Books-ID: fAJCDwAAQBAJ.

[5] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, and et al. 2019. The Second Conversational Intelligence Challenge (ConvAI2). *arXiv:1902.00098 [cs]* (Jan 2019). http://arxiv.org/abs/1902.00098 arXiv: 1902.00098.

[6] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (Nov 2009), 1149–1160. DOI: http://dx.doi.org/10.3758/BRM.41.4.1149 00000.

[7] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. DOI: http://dx.doi.org/10.3758/BF03193146 00000.

[8] Emer Gilmartin, Marine Collery, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R. Cowan, and Nick Campbell. 2017. Social Talk: Making Conversation with People and Machine. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents (ISIAA 2017)*. ACM, 31–32. DOI: http://dx.doi.org/10.1145/3139491.3139494

[9] Emer Gilmartin, Brendan Spillane, Christian Saam, Carl Vogel, Nick Campbell, and Vincent Wade. 2019. Stitching Together the Conversation—Considerations in the Design of Extended Social Talk. In *9th International Workshop on Spoken Dialogue System Technology (Lecture Notes in Electrical Engineering)*, Luis Fernando D'Haro, Rafael E. Banchs, and HaizhouEditors Li (Eds.). Springer, 267–273. DOI: http://dx.doi.org/10.1007/978-981-13-9443-0_23

[10] Emily M. Grundy and Michael Murphy. 2017. *Population ageing in Europe*. Oxford University Press, 11–18. https://global.oup.com

[11] Florian Kohlbacher and Benjamin Rabe. 2015. Leading the way into the future: the development of a (lead) market for care robotics in Japan. *International Journal of Technology, Policy and Management* 15, 1 (2015), 21. DOI: http://dx.doi.org/10.1504/IJTPM.2015.067797

[12] Marinka Lanne, Outi Tuisku, Helinä Melkas, and Marketta Niemelä. 2020. My business or not? The perspective of technology companies on shifting towards care robotics. *European Planning Studies* 28, 2 (Feb 2020), 296–318. DOI: http://dx.doi.org/10.1080/09654313.2019.1652249

[13] Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y. S. Lau, and et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (Sep 2018), 1248–1258. DOI: http://dx.doi.org/10.1093/jamia/ocy072

[14] Kyusong Lee, Tiancheng Zhao, Alan W. Black, and Maxine Eskenazi. 2018. DialCrowd: A toolkit for easy dialog system assessment. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 245–248. DOI:http://dx.doi.org/10.18653/v1/W18-5028

[15] Landis MacKellar and David Horlacher. 2000. Population Ageing in Japan: A Brief Survey. *Innovation: The European Journal of Social Science Research* 13, 4 (Dec 2000), 413–430. DOI: http://dx.doi.org/10.1080/13511610020017381

[16] Michelle Mahoney and Rhonda Magel. 1996. Estimation of the Power of the Kruskal-Wallis Test. *Biometrical Journal* 38, 5 (1996), 613–630. DOI: http://dx.doi.org/10.1002/bimj.4710380510

[17] M O Meira and A M P Canuto. 2015. Evaluation of emotional agents architectures: an approach based on quality metrics and the influence of emotions on users. (2015), 8.

[18] United Nations, Department of Economic, Social Affairs, and Population Division. 2020. *World population ageing, 2019 highlights.*

[19] Michael J. Nelson. 2013. Statistical Power and Effect Size in Informative Retrieval Experiments. *Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI* 0, 0 (Oct 2013). https://www.cais-acsi.ca/ojs/index.php/cais/article/view/523 00004.

[20] Zsofia Ruttkay, Claire Dormann, and Han Noot. 2004. Embodied Conversational Agents on a Common Ground: A Framework for Design and Evaluation. *From Brows to Trust: Evaluating Embodied Conversational Agents* (2004), 27–66. DOI: http://dx.doi.org/10.1007/1-4020-2730-3_2

[21] Brendan Spillane, Emer Gilmartin, Christian Saam, Leigh Clark, and Benjamin R. Cowan. 2018a. Identifying Topic Shift and Topic Shading in Switchboard. In *UK Speech 2018 Abstract Book*. UK Speech, 44. http://ukspeech.inf.ed.ac.uk/wp-content/uploads/2019/03/uk-speech-2018-abstract-book.pdf

[22] Brendan Spillane, Emer Gilmartin, Christian Saam, Benjamin R. Cowan, and Vincent Wade. 2018b. ADELE: Care and Companionship for Independent Aging.. In *Proceedings of the AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications co-located with the Federated AI Meeting (FAIM 2018), ICAHGCA@AAMAS 2018*, Vol. 2338. CEUR-WS.org, 18–24. http://ceur-ws.org/Vol-2338/

[23] Brendan Spillane, Emer Gilmartin, Christian Saam, Ketong Su, Benjamin R. Cowan, Séamus Lawless, and Vincent Wade. 2017. Introducing ADELE: A Personalized Intelligent Companion. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents (ISIAA 2017)*. ACM, 43–44. DOI: http://dx.doi.org/10.1145/3139491.3139492

[24] Brendan Spillane, Emer Gilmartin, Christian Saam, and Vincent Wade. 2019. Issues Relating to Trust in Care Agents for the Elderly. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19)*. ACM, 20:1–20:3. DOI: http://dx.doi.org/10.1145/3342775.3342808 event-place: Dublin, Ireland.

[25] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences* 21, 1 (2014), 7.

[26] Silvia Tulli, Diego Agustin Ambrossio, Amro Najjar, and Javier Rodrıguez Lera. 2019. Great Expectations Aborted Business Initiatives: The Paradox of Social Robot Between Research and Industry. *BNAIC/BENELEARN 2019 - Proceedings of the 31st Benelux Conference on Artificial Intelligence (BNAIC 2019) and the 28th Belgian Dutch Conference on Machine Learning (Benelearn 2019), Brussels, Belgium, November 6-8, 2019.* 2491 (2019), 10.

[27] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, and Angeliki Metallinou. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625* 4 (2018), 60–68.

[28] Xue-Qiang Wang and Pei-Jie Chen. 2014. Population ageing challenges health care in China. *The Lancet* 383, 9920 (Mar 2014), 870. DOI: http://dx.doi.org/10.1016/S0140-6736(14)60443-8

[29] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv:1801.07243 [cs]* (Sep 2018). http://arxiv.org/abs/1801.07243 arXiv: 1801.07243.

[30] Yuanting Zhang and Franklin W. Goza. 2006. Who will care for the elderly in China?: A review of the problems caused by China's one-child policy and their potential solutions. *Journal of Aging Studies* 20, 2 (Apr 2006), 151–164. DOI: http://dx.doi.org/10.1016/j.jaging.2005.07.002