Tanvi Dinkar

Heriot-Watt University
Edinburgh,
Scotland,
EH14 4AS

t.dinkar@hw.ac.uk
animatas.eu/network/esr/tanvi/

1 Research interests

Recently there has been an explosion of chatbot-style systems that utilise Large Language Models (LLMs) deployed in the real world. However, with this large scale deployment, the safety of these systems is critical (Bommasani et al., 2021; Bender et al., 2021; Weidinger et al., 2021; Bergman et al., 2022; Dinan et al., 2022a). While the NLP community has traditionally explored the ethical issues of text-based models (such as hate speech detection, inherent biases of the system etc), real-world conversations and dialogues differ *significantly* from structured, written text documents, and this brings with it its own unique set of safety challenges.

Firstly, a central theme of generative linguistics going back to von Humboldt, is that language is 'an infinite use of finite means', i.e there exists many ways to say the same thing. However, current research fails to account for this inherent variability of language, which results in a **lack of robustness** of these systems to: real-world use cases, noisy perturbations to the input, or even adversarial attacks (Jin et al., 2019; Moradi and Samwald, 2021; Wu et al., 2021).

Additionally, in real-world interactions, words alone don't sufficiently communicate intended meaning; listeners often arrive at meaning inferring several other speaker cues, such as prosody or even context. However, these unique *human-like* ways to communicate may be co-opted by designers of these systems to drive up user engagement, encouraging humans to relate to such systems in human-like ways – i.e. these systems are **anthropomorphised** or personified. Assigning human characteristics to dialogue systems can have consequences that could be on one hand, harmless, e.g. referring to automated systems by gender, but on the other, disastrous e.g., people following the advice or instructions of a system to do harm[1]. Based on these themes, I will present the research interests in my PostDoc (§1.1 and §1.2) on **safety and robustness** specific to **conversational AI**, including the relevant overlap from my PhD.

[1]A person recently has committed suicide, allegedly as a consequence of the harmful outputs generated from such a system (Xiang, 2023).

1.1 Robustness in Conversational AI: How do models perform in real-world conditions?

The real-world performance of text based models first interested me in my PhD, where I focused on how robust such models are to input transcripts arising from speech, given that they are pre-trained on massive amounts of written text. With this in mind, we investigated the representations of spontaneous speech phenomena present in speech transcripts – in particular fillers ('uh', 'um') – using deep contextualised word embeddings. A finding of the work was that Bi-directional Encoder Representations (BERT) (Devlin et al., 2019) already has existing representations of fillers, and their inclusion in the input decreased the uncertainty of the language model (Dinkar et al., 2020), despite research to suggest that other spontaneous speech phenomena increase uncertainty (Sen, 2020). Thus (somewhat surprisingly), LLMs may be robust to certain kinds of spontaneous speech phenomena.

In my post-doc I shifted focus to safety-critical contexts, deliberating on whether there are *scenarios where models **must be** robust to variability*. If so, what steps can be taken to ensure such guarantees? For the former question, it may be required legally for a chatbot to *always* disclose identity, such as California legislation stating *'[...] unlawful for a bot to mislead people about its artificial identity [...]'* (Legislature, 2018). Similar legislation could be widespread in the future (Montgomery, 2023). Another scenario is that a system may give a user false impressions of its 'expertise' and generate harmful advice in response to medically related user queries (Abercrombie and Rieser, 2022; Dinan et al., 2022b). In practice it may be desirable for the system recognise medical queries and avoid answering them. Thus the question remains, on how to create and ensure such guarantees for the output, given the inherent variability of language?

I collaborated with researchers to analyse the feasibility of applying formal verification methods to the NLP domain (work under review). These methods *ensure* that for every possible input, the output generated by a neural network satisfies the desired properties (such as consistently disclosing non-human identity). The work proposed semantically informed verification filters, which

Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 5–8
September 11–12, 2023.

essentially creates a geometric shape around a certain embedded input in a pre-trained LLM (such as a query 'are you a chatbot'), and *guarantees* that for every data point surrounding that input within that shape, the output of the network will generate the desired class (i.e. confirming non-human identity). We evaluated the work on the R-U-A-Robot dataset (Gros et al., 2021), a dataset containing multiple adversarial ways to ask 'are you a robot' and a medical safety dataset (Abercrombie and Rieser, 2022), a dataset comprised of medical queries annotated by expert practitioners. We found that the semantically informed filters capture not only the input, but also a large set of perturbations and adversarial attacks, allowing for robust representation in safety critical contexts. In the future we plan to focus on how to apply such methods to consider the sequentiality of dialogue, as initially asking the query 'are you a robot', may not have guarantees on subsequent followup query (i.e. 'no seriously?').

1.2 Anthropomorphism: What is the balance between naturalness and safety?

While a common goal of AI is to work towards more human-like (anthropomorphic) agents, research should also explore the trade-off between the naturalness of a system and safety of its deployment. Consider Google Duplex (Leviathan and Matias, 2018); a Text-to-Speech (TTS) system for accomplishing real world tasks over the phone. The *inclusion of spontaneous speech phenomena* (such as hesitations) led to highly natural sounding generated responses. However, these responses convinced the human recipients that they were conversing with another human, and also recieved widespread criticism (Lieu, 2018).

This illusion of agency can have harmful consequences when considering safety in conversational AI. NLP researchers have begun to investigate factors that induce personification and develop resources to mitigate such effects. However these efforts are fragmented, and many aspects of anthropomorphism are yet to be considered. Thus in recent work (Abercrombie et al., 2023), we discussed the linguistic factors that contribute to the anthropomorphism of dialogue systems (in Dinkar et al. (2023) with a focus on spontaneous speech phenomena), the harms that can arise, and the recommendations that designers should consider for the development, release, and descriptions of dialogue systems.

2 Spoken dialogue system (SDS) research

With chatbot style systems being widely deployed, there needs to be emergent research on safety and robustness, but focusing on real world contexts and the nature of dialogues, rather than (brittle) performance on carefully curated datasets. Ethically, more research needs to be done on the core set of communicative competencies truly required for different kinds of tasks in a dialogue system, to avoid users unnecessarily personifying and relying on the system.

3 Suggested topics for discussion

- Ethics of AI, e.g. (unnecessary) anthropomorphism in chatbots and LLMs
- Privacy concerns and data protection, e.g. when adding an LLM to an embodied robot, it not only involves collecting speech/text based inputs, but potentially using video surveillance to analyse input.
- Governance of AI, e.g. how can we create standards that publicly deployed chatbots need to meet (such as, via unit testing)?

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems.

Gavin Abercrombie and Verena Rieser. 2022. Risk-graded safety for handling medical queries in conversational ai. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*. pages 234–243.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. pages 610–623.

A Stevie Bergman, Gavin Abercrombie, Shannon L Spruit, Dirk Hovy, Emily Dinan, Y-Lan Boureau, and Verena Rieser. 2022. Guiding the release of safer e2e conversational ai through value sensitive design. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 39–52.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto,

Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR* abs/2108.07258. https://arxiv.org/abs/2108.07258.

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)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022a. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, pages 4113–4133. https://doi.org/10.18653/v1/2022.acl-long.284.

Emily Dinan, Gavin Abercrombie, A Bergman, Shannon L Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022b. Safetykit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 4113–4133.

Tanvi Dinkar, Chloé Clavel, and Ioana Vasilescu. 2023. Fillers in spoken language understanding: Computational and psycholinguistic perspectives. *Traitement Automatique des Langues* 63(3).

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 7985–7993. https://doi.org/10.18653/v1/2020.emnlp-main.641.

David Gros, Yu Li, and Zhou Yu. 2021. The R-U-A-Robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932* .

California State Legislature. 2018. California senate bill no. 1001.

Yaniv Leviathan and Yossi Matias. 2018. Google duplex: An AI system for accomplishing real world tasks over the phone. *Google AI Blog* .

Johnny Lieu. 2018. Google's creepy AI phone call feature will disclose it's a robot, after backlash. `https://mashable.com/2018/05/11/google-duplex-disclosures-robot`. Mashable. Accessed 2023-03-16.

Christina Montgomery. 2023. Hearing on "Oversight of AI: Rules for Artificial Intelligence". `https://www.ibm.com/policy/wp-content/uploads/2023/05/Christina-Montgomery-Senate-Judiciary-Testimony-5-16-23.pdf`. Accessed: 2023-06-01.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 1558–1570. https://doi.org/10.18653/v1/2021.emnlp-main.117.

Priyanka Sen. 2020. Speech disfluencies occur at higher perplexities. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*. Association for Computational Linguistics, Online, pages 92–97. https://aclanthology.org/2020.cogalex-1.11.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288* .

Chloe Xiang. 2023. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. `https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says`. VICE. Accessed: 2023-06-12.

Biographical sketch

Tanvi Dinkar is a Research Associate at Heriot Watt University, working on Safety in Conversational AI with Prof. Oliver Lemon. She completed her PhD at Télécom Paris, supervised by Prof. Chloé Clavel, Prof. Catherine Pelachaud and Prof. Ioana Vasilescu. Her PhD studied the representations of disfluencies for SLU, and how they can be informative signals of communication, rather than simply removed as noise. During her PhD, she was a Marie Curie Early Stage Researcher at ANIMATAS. Her research interests include safety and robustness in conversational AI, spoken language understanding, how NLP models are brittle compared to real-world dialogues, communicative strategies and pragmatics. Prior to this, she was a dialogue engineer at Nuance (now Microsoft), coding dialogue systems for the automotive industry. She decided to pursue research when she saw from customer tickets that the task oriented dialogue systems are not robust to people speaking naturally. She has two masters from the University of Edinburgh, one in Linguistics and one in Speech and Language Processing. Once upon a time, she completed an undergraduate degree in Journalism and Literature.