ARCIDUCA: Annotating Reference and Coreference In Dialogue Using Conversational Agents in games

Massimo Poesio
Queen Mary University
m.poesio@qmul.ac.uk

Richard Bartle
University of Essex
rabartle@essex.ac.uk

Jon Chamberlain
University of Essex
jchamb@essex.ac.uk

Julian Hough
Queen Mary University
j.hough@qmul.ac.uk

Chris Madge
Queen Mary University
c.j.madge@qmul.ac.uk

Diego Perez-Llebana
Queen Mary University
diego.perez@qmul.ac.uk

Matt Purver
Queen Mary University
m.purver@qmul.ac.uk

Juntao Yu
University of Essex
j.yu@essex.ac.uk

Abstract

The objective of ARCIDUCA is to address the twin challenge of developing conversational agents (CAs) able to deal with coreference and reference, and of creating datasets for training such agents, by having CAs generate through interaction the needed training data, which can then be used to improve those agents as well as train agents for other domains. A core hypothesis of the project is that the most effective way to motivate enough individuals to participate in such interactions is by embedding these interactions in online games-with-a-purpose.

1 Introduction

The development of architectures such as the encoder/decoder model (Sutskever et al., 2014) and the Transformer (Vaswani et al., 2017) has brought about an explosion of interest in neural architectures for conversational agents (CAs) (Vinyals and Le, 2015; Bordes et al., 2017; Zhang et al., 2018; Dinan et al., 2019b; Gao et al., 2019; Ram et al., 2018; Dinan et al., 2019a). CA research has since shifted towards CAs capable of engaging in more complex and task-oriented dialogue such as restaurant booking (Bordes et al., 2017) or question answering (Dhingra et al., 2017) . The results on these tasks show that CAs carrying out more complex tasks require the ability to carry out more in-depth interpretation (Quan et al., 2019; Roller et al., 2020). Achieving this requires, on the one hand, architectures capable of carrying out such aspects of interpretation, typically incorporating models of dialogue memory and representations of task-specific knowledge (Sukhbaatar et al., 2015; Dinan et al., 2019b). On the other end, training such models requires appropriate resources. Recently, a number of datasets have become available for end-to-end training of task-oriented CAs; these include the datasets available through ParlAi,[1] Amazon's MultiDOGO (Peskov et al., 2019) and Facebook's Dialogue Dodecathlon (Shuster et al., 2020). However, none of these datasets is also annotated with information about the semantic and discourse interpretation of utterances required to train modules for these tasks. The objective of ARCIDUCA is to develop conversational agents (CAs) able to deal with coreference and reference, and of creating datasets for training such agents, by having the CAs themselves generate through interaction the needed training data, which can then be used to improve those agents as well as train agents for other domains.

[1] https://parl.ai/docs/tasks.html

2 The approach

Datasets and Architectures for Coreference in Dialogue Coreference is prevalent even in the shortest conversations (Müller, 2008; Quan et al., 2019; Grobol, 2020). However, current neural architectures for conversational agents mostly do not resolve coreference. Such CAs can only react appropriately when generating the correct response does not require understanding coreference. Part of the problem is that despite impressive recent improvements (Lee et al., 2017; Joshi et al., 2019), coreference research is still mostly focused on written text. This research gap is largely due to a lack of resources. Training a coreference resolver on written text and domain-adapting it to dialogue has proven ineffective, as coreference in dialogue involves different phenomena and is more involved than coreference in text (Müller, 2008; Grobol, 2020). But the largest annotated corpus of coreference in dialogue, the TRAINS subset of our own ARRAU corpus (Uryupina et al., 2020), is too small to train a high performance coreference resolver for CAs. One objective of the project is to create more substantial datasets to study the problem. Also, there is a need for CA architectures including specific modules that enable them to interpret coreference. Some such architectures have recently appeared, such as GECOR (Quan et al., 2019), based on a copying architecture that solves coreference as an incomplete utterance restoration task. (Quan et al., 2019) showed that adding a coreference resolver to a task-oriented CA can substantially improve performance. In the project we will experiment with such architectures.

Games with a Purpose Games with a Purpose (GWAPs) (von Ahn, 2006) have emerged as an alternative to traditional micro-task crowdsourcing (Snow et al., 2008). GWAPs, particularly when run over large periods, can collect large amounts of annotations: e.g., our own *Phrase Detectives* (Poesio et al., 2013), designed to collect labels for coreference, accumulated over 5.7 million coreference judgments from more than 60,000 players over the last fifteen years; the third release of the corpus has now 400,000 markables, twice the number of ONTONOTES. But there is a fundamental difference between conversation and written text: the latter is designed to be read by third parties, whereas, e.g., (Clark and Schober, 1989) have shown that overhearers to a conversation only acquire a partial understanding of what was said.

Games and AI In recent years, games have become one of the most widely used platforms to test progress on machine learning-based AI agent theories (Silver et al., 2016). This progress became visible when DeepMind AlphaGo (Silver et al., 2016) mastered the GO game using a combination of Monte Carlo Tree Search and Deep Learning, but progress since has been accelerated through competitions such as General Video Game AI (Perez et al., 2019) and the development of platforms for rapid experimentation such as MALMO (Johnson et al., 2016) or Unity/ML (Juliani et al., 2018).

Collecting conversational data through conversational learning in games The dominant paradigm for CAs training discussed above (pre-training against an annotated corpus, followed by fine-tuning via reinforcement learning through interaction with other agents) is also the approach used in Game AI, which recently led to an exciting synergy between the two areas of AI, whereby Game AI platforms would be used to train conversational agents as well. One example of this synergy is the MALMO project at Microsoft (Johnson et al., 2016), a platform for training agents in Minecraft which was extended to allow training of conversational agents (Allison et al., 2018; Szlam et al., 2019). More recently, Hockenmaier's group developed an extension of MALMO to allow conversational agents to learn to interact, and used the extension to introduce the Minecraft Collaborative Game Task (Narayan-Chen et al., 2019). In parallel with this, Facebook launched project LIGHT (Urbanek et al., 2019)–an open platform for collecting conversations in a very rich textual fantasy game with extensive crowdsourced resources entirely described in natural language. In ARCIDUCA, we aim to train conversational agents able to interpret coreference and reference by embedding them in LIGHT and the Minecraft Collaborative Game.

Collecting judgments through clarification questions The obvious way to enable a CA to acquire information about interpretation is by making it able to ask **clarification questions** (CQs) as to that interpretation (Purver et al., 2003). As far as we know, this has not yet been attempted for coreference, or for CAs. The one proposal along these lines we are aware of (Thomason et al., 2019) was developed to learn grounded reference for robots. What we propose to do is to adopt a similar strategy for improving conversational agents in games' ability to interpret both references and coreference, but also recording these judgments in the form of an annotated corpus.

3 Progress so far

The project officially started in February 2022, but work started beginning of 2021 with the preparation of the CODI-CRAC 2021 shared task on anaphora resolution in dialogue (Khosla et al., 2021), a second edition of which is currently running. One of the outcomes of this work is the creation of the CODI-CRAC corpus of anaphoric reference in dialogue, covering four well-known domains including AMI, LIGHT, PERSUASION and SWITCHBOARD, and is currently the largest such dataset for English. A second outcome of the shared task has been the development of the Universal Anaphora scorer (Yu et al., 2022), currently being revised to make it more suitable to score coreference in dialogue, e.g., by allowing for discontinuous markables Next work was fine-tuning of a coreference resolver for the LIGHT domain and its incorporation in a conversational agent for the LIGHT domain based on the poly-encoder architecture from (Humeau et al., 2020).

Acknowledgements

ARCIDUCA is funded by EPSRC, EP/W001632/1 The creation of the CODI-CRAC corpus was funded in part by the DALI project (ERC project 695662), in part by HITS Heidelberg (Michael Strube), in part by funding from CMU (Carolyn Rose and Lori Levin).

References

Fraser Allison, Ewa Luger, and Katja Hofmann. 2018. How players speak to an intelligent game character using natural language messages. *TDGRA*, 4(2).

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of ICLR*.

Herbert H. Clark and Michael F. Schober. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21:211–232.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proc. of ACL*, pages 484–495. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, JasonWilliams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019a. The second conversational intelligence challenge (convai2). ArXiv preprint arXiv:1902.00098.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proc. of ICLR*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural approaches to Conversational AI*, volume 13 of *Foundations and Trends in Information Retrieval*. Now.

Loïc Grobol. 2020. *Coreference resolution for spoken French*. Ph.D. thesis, Université Sorbonne Nouvelle.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proc. of ICLR*.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The Malmo platform for artificial intelligence experimentation. In *Proc. of IJCAI*, pages 4246–4247.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Arthur Juliani et al. 2018. Unity: A General Platform for Intelligent Agents. *arXiv preprint arXiv:1809.02627*.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The codi-crac 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proc. of the CODI/CRAC Shared Task Workshop*.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proc. of EMNLP*.

M.-C. Müller. 2008. *Fully Automatic Resolution of It, This And That in Unrestricted Multy-Party Dialog*. Ph.D. thesis, Universität Tübingen.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proc. of the 57th Annual Meeting of the ACL*, pages 5405–5415.

Diego Perez, Simon M. Lucas, Raluca D. Gaina, Julian Togelius, Ahmed Khalifa, and Jialin Liu. 2019. *General Video Game AI*. Morgan Claypool.

Denis Peskov, Nancy Clarke, Jason Krone, Brigitta Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. In *Proc. of EMNLP*. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: ... *ACM Transactions on Intelligent Interactive Systems*, 3(1).

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse & Dialogue*, pages 235–255. Kluwer.

J. Quan, D. Xiong, B. Webber, and C. Hu. 2019. GECOR: An end-to-end generative ellipsis and coreference resolution model for ... In *Proc. of EMNLP*, Hong Kong. ACL.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigru. 2018. Conversational AI: The science behind the Alexa Prize. ArXiv abs/1801.03604.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. 2020. Open-domain conversational agents:... ArXiv preprint arXiv:2006.12442.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded cas. In *Proc. of the ACL*. Association for Computational Linguistics.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, and others. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Arthur Szlam, Jonathan Gray, Kavya Srinet, Yacine Jernite, Armand Joulin, Gabriel Synnaeve, Douwe Kiela, Haonan Yu, Zhuoyuan Chen, Siddharth Goyal, Demi Guo, Danielle Rothermel, C. Lawrence Zitnick, and Jason Weston. 2019. Why build an assistant in minecraft? ArXiv: 1907.09273.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. 2019. Improving grounded natural language understanding through human-robot dialog. In *Proc. of ICRA*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, Samuel Humeau, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. ArXiv preprint arXiv:1903.03094.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. ArXiv:1706.03762.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the Deep Learning Workshop at ICLR*, Lille.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Juntao Yu, Nafise Moosavi, Silviu Paun, Sopan Khosla, Sameer Pradhan, and Massimo Poesio. 2022. The universal anaphora scorer 1.0. In *Proc. of LREC*.

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 preprint arXiv:1801.07243*.