Research Statement

Yanchao Yu

Edinburgh Napier University
Edinburgh, Scotland, UK
y.yu@napier.ac.uk

1 Motivation

Research in an academic environment provides a unique opportunity to solve problems and envision the future of technology. However, most theories/approaches are usually introduced in an over well-defined environment, heavily influenced by dataset bias and lack robustness to uncommon configurations from the realistic world. Research, to me, is usually a bridge between an ambitious idea and a practical problem in the real world. For example, while bringing intelligent systems/robots from the laboratory to the physical world, they need to capable of natural everyday conversation with their human users about their physical surroundings. Among other competencies, this involves learning and adapting mappings between (words, phrases, and sentences) in Natural Language (NL) and perceptual aspects of the external environment – this is widely known as the grounding problem.

2 Previous Research

My principal research lies in approaching this grounding problem by building teachable systems/robots that can learn novel visual scenes (e.g. objects and features) in an individual user's language (e.g. "this is my red pen."), from the physical world through conversations with human users. There are many objects/actions/events occurring that a pre-trained system hardly addresses in a short period in the real world. Hence, in contrast with most research in this field, which relies on a large amount of pre-processed static training data, my research aims at learning new knowledge dynamically and proactively. I believe that an intelligent system that receives feedback from humans through conversations can more effectively learn novel knowledge than pre-trained systems.

In order to address such interactive learning problem on the vision and language task, I build a interactive teachable agent by:

- designing an interactive multimodal framework, consisting of vision and dialogue modules. The framework contributes to not only understand and produce natural, human-like conversations, but also learn the mapping between information from different modalities (e.g. vision and language) [1, 2]).
- releasing an open-domain human-human conversations on the task of learning visual attributes of different objects — BURCHAK[1] (see [3]). It offers a huge variation of task-oriented dialogue strategies and capabilities and refers to a wide range of natural incremental dialogue phenomena (e.g. overlapping, self-repair and repetition, fillers). I also investigate how these dialogue phenomena and strategies affect the learning performance during the learning process (see [4–6]).
- training an optimised dialogue policy using Reinforcement Learning with a hierarchical MDP by exploring a relative balance between performance (e.g. learning accuracy or task success) and human involvement. It learns to 1) process Natural human conversations, and 2) handle a form of active learning: the agent

[1]Other researchers have also used and cited this corpus in their research projects.

can only acquire helpful information about the visual scene through feedback from the tutor if it cannot be confident about its answers (see [7–9]). A teachable system was developed for robotics demonstration and planed to deploy for education and research.

The interactive multimodal framework that I proposed will be the foundation of my future research towards building the next generation personal intelligent assistants/robots that are more efficient to develop, understand and learn novel personal descriptions about the visual scene (i.e. people, events, visual objects or features), and more generalisable across multimodal tasks.

2.1 Interactively Teachable Framework

I proposed an interactive multimodal framework to build a teachable system that can learn visual attributes of different objects from the physical environment through natural language interaction with human partners. This framework contains three basic but core components: a vision module, a DS-TTR dialogue module and an adaptive learning policy to select actions to achieve a good learning performance (i.e. accuracy) but with less human involvement in a long-term learning task.

Vision Module It produces the high-dimensional visual feature representations for specific objects from the physical world. It considers each visual attribute (e.g. colour and shape) equally as a binary attribute classifier using Logistic Regression SVM classifier model with Stochastic Gradient Descent (SGD). The visual classifiers ground visual attribute words (such as 'red', 'green', 'square' etc.) that appear as parameters of the dialogue acts used in the system (see [1, 10]).

DS-TTR Dialogue Module I designed the dialogue module using the Dynamic Syntax (DS) and Type Theory with Records (TTR). DS is a word-by-word incremental semantic parser/generator, based around the Dynamic Syntax grammar framework [11] especially suited to the fragmentary and highly contextual nature of dialogue. In DS, dialogue is modelled as the interactive and incremental construction of contextual and semantic representations [12]. These contextual representations are of the fine-grained semantic content that is jointly negotiated/agreed by the speakers, as a result of processing variety of natural utterance, such as questions, answers, clarification requests, corrections etc. on the other hand, TTR, as an extension of standard type theory for semantic and dialogue modelling [13, 14], allows information from various modalities (incl. vision and language) into a common single semantic framework. For the grounding task here, it allows the system to encode partial knowledge about objects and their attributes to be extended, as and when this information becomes available.

Through interaction between various parts of the system, at any point in time, the system has access to an ontology of (object) types and attributes encoded as a set of TTR Record Types, whose individual atomic symbols, such as 'red' or 'square' are grounded in the set of classifiers trained so far. (more details see [6])

The interactive multimodal framework I proposed here is a generalised framework, in which any single components can be easily replaced by a more robust technique. The framework is able to build visually teachable intelligent systems that are more efficient to learn visual-grounded word meaning through natural conversations with real humans in long-term learning.

2.2 Human-Human Dialogue Corpus on Learning

Different to machines, humans may produce a more natural, spontaneous dialogue, which is incremental, and thus gives rise to dialogue phenomena such as self- and other-correction, continuation, unfinished sentences, etc. [3]. These phenomena are not simply noise for language understanding and are used by listeners to guide linguistic processing[15]; I have proposed the interactive multimodal framework to incrementally learn visual concepts through natural conversations with real humans. In order to help the framework cope with natural, human-like conversations, I collected a number of natural dialogues between humans for interactive learning of visually grounded word meanings through ostensive definition by a tutor to a learner — BURCHAK. To our knowledge, this is the first Human-Human dialogue corpus on the visual attribute learning task.

Task Design on Learning shape and colour I designed a novel visual attribute learning/tutoring task given to the participants. It involves a pair of participants who talk about visual attributes (e.g. colour and shape) through a series of visual objects. The overall goal of this task is for the learner to discover groundings between visual attribute words and aspects in the physical world through interaction. However, since humans have already known all groundings, such as "red" and "square", the task is assumed in a second-language learning scenario, where each visual attribute, instead of standard English words, is assigned to a new unknown word in a made-up language (see examples in Fig. 1). (see more details in [3])

Incremental Dialogue Phenomena The corpus refers to a wide range of incremental dialogue phenomena, such as overlapping, self -correction and -repetition, continuation, pause and filler. These phenomena are interactionally and semantically consequential, and contribute directly to how dialogue partners coordinate their actions and the emergent semantic content of their conversation. They also strongly mediate how a conversational agent might adapt to their partner over time. For example, self-interruption, and subsequent self- correction as well as hesitations/fillers aren't simply noise and are used by listeners to guide linguistic processing[15]; similarly, while simultaneous speech is the bane of dialogue system designers, interruptions and sub- sequent continuations are performed deliberately by speakers to demonstrate strong levels of understanding[16].

Dialogue Strategy and Capability As noted above, the dialogue collection is designed for the task of interactively learning visual attributes, i.e. colours and shapes, the corpus contains a list of task-oriented dialogue capabilities, such as question -asking and -answering, acknowledgement, correction, etc. On this simple learning task, each dialogue capability surprisingly involves into a large variety of expressions. On the other hand, these dialogues also offer interesting dialogue strategies, for instance, initiative, uncertainty as well as context-dependency. These dialogue strategies and capabilities, especially uncertainty, are likely to affect the overall learning performance — find out a relative balance between the performance (e.g. learning accuracy or task success) and human involvement.

A series of investigations [7] indicate that, in order to achieve a better performance, i.e. good learning accuracy but less dialogue effort, the system should take the initiative in the dialogue, takes into account the uncertainty as well as context-dependency. This result has been applied to improve the interactive multimodal framework I proposed above.

This human-human dialogue corpus was applied to train and evaluate the interactive multimodal framework I introduced above by building a user model that resembles human behaviours in the learning-oriented conversation. It contributes to an intelligent system that behaves like a real human who can performs and copes with natural, more complex and flexible conversations with others.

3 Learn How to Learn: Interactively Teachable System Implementation

My research also contributes to the field of human-robot collaboration in future research. Currently, some intelligent systems are able to complete simple tasks automatically by themselves following a series of pre-defined rules. But regarding more complicated situations with massive variations and uncertainties, the system always needs much more help from human users. As noted above, this research is attempting to explore a relative balance between the performance (e.g. learning accuracy or task success) and human involvement: One of the milestones and also a challenge of this research is that it expects to have an intelligent system which can determine whether and when it needs to request help from people, and during the rest of time, it is able to work on tasks by itself. In order to achieve this challenge, I applies an optimised learning agent using Reinforcement Learning, as follows:

- **Adaptive Policies for Lone-term Visual Attribute Learning.** My work [7–9] formulates this learning problem into two sub-tasks (i.e. when and how to learn), which trained using Reinforcement Learning with a hierarchical Markov Decision Process (MDP), consisting of two interdependent MDPs. It performs a form of active learning on visual concepts from a human partner. The former MDP aims at determining when to ask for information/feedback from human partners, which hinges on the previous learning performance (i.e. classification accuracy), and the latter one learns to choose among different dialogue strategies and

capabilities to acquire useful information but with less dialogue efforts by the human tutor. The learnt agent performs a form of active learning: the learner only asks a question about an attribute if it isn't confident enough already about that attribute.

- **Solution of Incremental Dialogue Phenomena** I introduced a mechanism of dialogue act inference that automatically predicts the most appropriate dialogue act based on certain completed semantic sub-trees following a set of pre-learned rules in the DS-TTR framework [17]. The experiment results illustrate that the new framework with DS-TTR model can keep a comparably good performance on the learning/grounding task without misunderstanding user's intent while interacting with a simulated tutor (trained on realistic data from the corpus, can randomly generate dialogue phenomena).

- **Deployment of Interactively Teachable System.** I deployed this spoken teachable system onto the Furhat Robot [18] that is able to learn to identify and describe simple colours and shapes of different objects from spoken conversations with real users. This system will be applied to conduct user studies to compare, evaluate and iteratively improve the proposed multimodal framework for different learning tasks. To support the system development, I designed and implemented a generic bridge framework, which can help easily deploy different interactive multimodal systems onto the IrisTK platform (which is provided by the Furhat Robot). The implemented teachable robot will be demonstrated on the year of Robotics 2017 in Edinburgh. The bridge framework will also be introduced for computer science education (applied in the tutorial and courses).

As an extension of the teachable system, I lead a a collaborative project with a PhD student from Heriot-Watt University and a visiting PhD student that deploy the same approach to learn to dynamically generate a semantic map with daily objects (e.g. pen, cup, bottle) instead of visual features (e.g. colour, shape and material). (see [19])

4 Future Directions

At the University of Glasgow, I am enthusiastic about establishing a research group specialising in multi-modal and multi-party conversations with AI. I look forward to continuing my principal research to improve the teachable interactive approach, proactively learning more complicate visual environment. for example:

- integrating the Distributional Semantics into the framework, make the use of the word embedding-based features or similarity scores from existing libraries (e.g. GLOVE and Word2Vector) as features in a continuous space MDP to learn novel attribute concepts

- jointly learning a sequence of decisions on dialogue acts and the perceptual features through dialogues using a multi-modal deep reinforcement learning

- a crowd-sourced learning capability, i.e. learning through novel knowledge online by interacting with many human tutors in real-time

Critically, due to the strong industrial links and applications, I believe my research fits well with the strategic areas of funders such as EPSRC. I will be making my 'New Investigator Award' application in the area of Trustworthy Autonomous Systems (TAS) in the first two years of my lecturer-ship.

On the other hand, I would like to research the application of interactive multi-modal approaches to other subjects in Artificial Intelligence, e.g. 1) improving the machine translation performance using visual-language representations, and 2) learning to dynamically generate scene graph representations through natural, daily human conversations.

I am also collaborating with several academic and industrial partners, including the Institute of Geology and Geophysics Chinese Academy of Sciences (IGGCAS) and Hebei Medical University. We had discussed the impact of Conversational AI (CA) on self-automated machines and data analysis, for example, automatic supervision conversations between a virtual interactive supervisor and self-automated unmanned aerial vehicle on, real-time Natural Language summary of visual data analysis (on images or videos) and interactively learning/monitoring cardiovascular disease from Ultrasound Cardiogram (UCG) in healthcare. I plan to pursue potential joint-funding

opportunities between EPSRC/Newton Fund and the National Natural Science Foundation of China (NSFC) when opportunities when they arise.

References

[1] Y. Yu, O. Lemon, and A. Eshghi, "Interactive learning through dialogue for multimodal language grounding," in *SemDial 2015, Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue, Gothenburg, Sweden, August 24-26 2015*, 2015, pp. 214–215.

[2] ——, "Comparing dialogue strategies for learning grounded language from human tutors," in *SemDial 2016, Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue, New Brunswick, NJ, USA,, July 16-18, 2016*, 2016, pp. 44–54.

[3] Y. Yu, A. Eshghi, G. Mills, and O. Lemon, "The burchak corpus: A challenge data set for interactive learning of visually grounded word meanings," in *Proceedings of the Sixth Workshop on Vision and Language*, Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–10.

[4] Y. Yu, O. Lemon, and A. Eshghi, "An incremental dialogue system for learning visually grounded language (demonstration system)," in *SemDial 2016, Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue, New Brunswick, NJ, USA,, July 16-18, 2016*, 2016, pp. 120–121.

[5] Y. Yu, A. Eshghi, and O. Lemon, "Incremental generation of visually grounded language in situated dialogue (demonstration system)," in *INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK*, 2016, pp. 109–110.

[6] ——, "Interactively learning visually grounded word meanings from a human tutor," in *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*, 2016.

[7] ——, "Training an adaptive dialogue policy for interactive learning of visually grounded word meanings," in *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, 2016, pp. 339–349.

[8] ——, "Learning how to learn: Grounding word meanings through conversation with humans," in *Proceedings of Machine Intelligence - Human-Like Computing (MI20-HLC)*, 2016.

[9] ——, "Learning how to learn: An adaptive dialogue agent for incrementally learning visually grounded word meanings," in *Proceedings of the First Workshop on Language Grounding for Robotics*, Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 10–19.

[10] ——, "Comparing attribute classifiers for interactive language grounding," in *Proceedings of the Fourth Workshop on Vision and Language*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 60–69.

[11] R. Cann, T. Kaplan, and R. Kempson, "Data at the grammar-pragmatics interface: The case of resumptive pronouns in English," *Lingua*, vol. 115, no. 11, R. Borsley, Ed., pp. 1475–1665, 2005, Special Issue: On the Nature of Linguistic Data.

[12] A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver, "Feedback in conversation as incremental semantic update," in *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK: Association for Computational Linguisitics, 2015.

[13] R. Cooper, "Records and record types in semantic theory," *Journal of Logic and Computation*, vol. 15, no. 2, pp. 99–112, 2005.

[14] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*. Oxford University Press, 2012.

[15] H. H. Clark and J. E. Fox Tree, "Using *uh* and *um* in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[16] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[17] Y. Yu, "Optimising strategies for learning visually-grounded word meanings through interaction," Ph.D. dissertation, Edinburgh, The UK, 2018.

[18] S. A. Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Systems - COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds., ser. Lecture Notes in Computer Science, vol. 7403, Springer, 2011, pp. 114–130. DOI: 10.1007/978-3-642-34584-5_9.

[19] A. Vanzo, J. L. Part, Y. Yu, D. Nardi, and O. Lemon, "Incrementally learning semantic attributes through dialogue interaction," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, E. André, S. Koenig, M. Dastani, and G. Sukthankar, Eds., International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018, pp. 865–873.