Transformer-based Interpretable Conversational Artificial Intelligence

Zhuchen Cao

October 22, 2022

1 Introduction

Conversational artificial intelligence (AI) refers to technologies that can verbally interact with users, such as conversational agents or chatbots. In recent years, conversational AI has been implemented in essential applications in education (e.g. robot teaching), finance (e.g. insurance consulting services) and other fields. The most representative of these applications are conversational agents, which can respond meaningfully, naturally, and logically to the users' diverse input (text, images, and audio) based on natural language processing (NLP).

The transformer is a deep learning model based on a self-attention mechanism that allows the model to learn to assign corresponding weights to various elements and positions of one sequence to generate a representation. The recent successful development of NLP is inseparably linked to transformer models. As a critical branch of NLP, many transformer-based conversational agent technologies, such as Blender (announced by Facebook), have gradually been developed [1].

Although the transformer-based sequence-to-sequence (seq2seq) model has a high average performance, it is of concern in some high-stakes decision-making tasks. The vast majority of the current deep learning-based conversational agents work in a black-box manner; thus, their decision-making logic is uninterpretable. Such model results often lack explicit, readable intrinsic logical support (i.e. the models do not comply with the principle of interpretability) [2]. This research proposal will start with two typical drawbacks and explore a potential research direction regarding increasing model interpretability while leveraging state-of-the-art deep learning algorithms.

2 Background Significance

As mentioned, deep learning architecture has one thing in common: the model is designed to be complex enough to simulate the complexity of the real world. The

most traditional conversational agents are often constructed on preset scripts, which cannot fully meet the users' requirements concerning complexity and can only express very rigid logic [3]. However, today's deep learning techniques seem to be heading towards the other extreme, uninterpretable complexity.

Although the transformer model has a sophisticated architecture, this complexity often makes it difficult to understand the logic behind the model, even for the creator. Such a phenomenon results in two typical drawbacks of conversational agents based on deep learning: uncontrollable output and difficulty in tunability.

Uncontrollable output refers to the overly broad possible output space of deep learning-based conversational agents. A transformer model must be extensively pretrained on big data, which often comes from network materials, such as Twitter, Wikipedia, and others [4]; thus, it becomes challenging to remove noise from the training data effectively. Transformer-based language models indiscriminately learn from noisy data, and the process of generating conversations is similar to that of a black box. Such characteristics lead to two risks in the output content. First, the conversational agent could produce a failure in some specific knowledge blocks. Second, due to the model's limited knowledge of morality and the complex training data with noise, the output content could violate the constraints of ethics and morality.

In practical applications, transformer-based conversational agents lack good performance in highly specialised tasks, such as program coding teaching and medical teaching. Even on general tasks, the advantages of transformer models are more reflected in their superior average performance. In specific moments, the models can perform very poorly.

Ethical and moral hazards are a more severe problem. Conversational agents are widely used in the education and medical industries, and people generally expect AI to be knowledgeable and gentle. However, the facts do not always align with the researcher's imagination. Due to the uninterpretable program logic and the noisy training data, the output content of the conversational agent may involve violence, sexism, racial discrimination or other uncontrollable tendencies. Such ethical and moral risks can even have tragic results in the education of minors and psychological counselling.

Difficulty in tunability refers to the difficulty deep learning-based conversational agents have with fixing errors quickly (i.e. learning from previous mistakes). Taking Facebook's Blender model as an example, the final application model still has 175B parameters despite being streamlined. The massive number of parameters makes the model extremely complex, making it difficult to adjust effectively after training. For example, in a task where the model is considered seriously wrong on a particular question, it becomes incredibly challenging to adjust effectively for this specific error without affecting the model's

average performance. Although technical means related to transfer learning can be used, it requires time, computing power and well-designed datasets.

Although transformer-based conversational agents have impressive performance, this technique is still challenging to apply fully in industrial projects due to these two main drawbacks. Modern industry requires a controllable, reliable, high-performance conversational agent. Therefore, using the latest technology to architect an interpretable conversation AI is necessary.

3 Literature Review

Dating back to the last century, the realisation of speaking with devices began with the creation of the first fully documented conversational agent, ELIZA [5], by researchers at MIT in 1966. ELIZA responds logically by analysing the input utterance (keyword search). In the following decades, various conversational agents were invented one after another. However, they have shortcomings, such as difficulty maintaining dialogue and understanding the users' logic and emotion.

In 1995, Wallace developed the ALICE model based on the AI Markup Language (AIML) mechanism [6]. The AIML mechanism allows developers to add rules to conversational agents. In theory, the more rules there are, the more brilliant it is. However, ALICE's shortcomings are also evident. The artificially added rules cannot address the complexity of human language, so when encountering a case that has not been seen before, ALICE will reply that there are no relevant phrases [7]. Early conversational agents were based on visible finite rules, limited by their insufficient complexity and difficulty understanding language logic.

Beginning in 2010, technology companies began to focus on personal voice assistants, such as Siri [8], launched by Apple in 2010. At first, Siri was based on machine learning operations to realise the interaction between software and users, but around 2014, a revolution took place in the domain of NLP. Before 2014, mainstream NLP projects still used machine learning algorithms, such as the support vector machine or logistic regression. However, since 2014, many researchers have been experimenting with neural network (NN) techniques to implement NLP tasks.

A typical technique is the NN-based dependency parser proposed by Chen and Manning [9]. A dependency parser [10] is a technique for analysing the grammatical relationships between phrases in a sentence. An NN-based dependency parser is faster and more accurate than a traditional statistic-based one. In addition, in 2014, Apple announced using NN technology in Siri.

Along with developing deep learning techniques in the field of NLP, conver-

sational agents have been increasingly researched since 2016. In 2017, a seq2seq recurrent NN based on long short-term memory (LSTM) [11] was applied to the field of conversational agents by Yin et al. [12]. This technology, called DeepProbe, has achieved good performance. Its architecture is based on the LSTM attention decoder, which can also be considered a prototype of the attention mechanism application in conversational agents. In the same year, Vaswani et al. [13] published an attention-only transformer architecture dispensing with recurrence and convolutions. The birth of the transformer also symbolises a new era in conversational agent technology and even NLP.

Based on transformer's success in NLP, in 2020, Google and Facebook launched a new generation of conversational agents, Meena [14] and Blender [1]. Among them, the Blender model, based on the poly-encoder architecture, broke many records previously held by Meena. The poly-encoder [15] used by Blender combines the advantages of a bi-encoder and cross-encoder for efficient and fast global self-attention learning. Although the Blender team pays considerable attention to word security in downstream training, Blender's safety is concerning due to the substantial toxic language in the Reddit pretraining dataset [16]. Blender, one of the most successful conversational agents, still faces problems with uncontrollable output, which tends to use high-frequency words, sometimes repeating meaningless answers, and so on.

Many researchers have expressed concern about the research direction of the transformer model, which blindly pursues computing power and complexity. In 2020, Arrieta et al. proposed the concept of explainable AI (XAI) through algorithmic means, seeking clues to the model's decisions [17]. An example of XAI is the explainable method of image convolutional NNs (CNNs) proposed by Pope et al. [18]. In 2019, Murdoch et al. codified the concept of interpretable AI (IAI), which can readily demonstrate the intrinsic relationships that have been learned [2]. Compared with XAI, which focuses on seeking data (clues) from the model and using other tools for analysis, IAI emphasises that the model should exhibit reliable decision-making causality. In 2022, Wahdea and Virgolina proposed the DAISY model [19], a conversational agent model that completely abandons deep NNs (DNNs). The modular design concept embodied in DAISY is inspiring, and it has shown performance that is not inferior to mainstream DNN-based models on several tasks.

4 Aims and Research Directions

This research aims to create an interpretable conversational AI by combining the latest NN (transformer) with traditional statistics and machine learning techniques. Next, this research proposal will describe the general ideas through several research questions.

4.1 Increasing Model Interpretability

A viable research direction is constructing a modular conversational AI rather than having a deep learning model make all the decisions. Essentially, a conversation AI that meets industry requirements usually has the following process: analyse the input, query the knowledge base and output the answer. Current conversational AI leaves these three steps to deep learning models, which is the root of the problem. The transformer is crucial in the NLP domain because it efficiently models the relationships between elements within the same sequence. The strength of the self-attention architecture lies in parsing relationships between words, but does it necessarily have an advantage in generating text? The transformer is exciting for qualitative text analysis, such as classification and semantic search tasks. Using a transformer as an input analysis tool rather than a conversational AI decision-maker is a promising direction.

Inspired by the DAISY model [19] and some of the early conversational AI applications, such as ALICE [6], it is possible to experiment with a modular-based design. To provide a simplified concept: the input text can be embedded into tensors using a transformer model and paired with content from the knowledge base via a traditional clustering algorithm (e.g. K-means [20], Gaussian mixture [21], or even DNN) and finally input back. Although the actual architecture will be much more complicated than this example, the whole process is traceable, and the uncertainty of the DNN is limited to a minimal range. While the ideal architecture would still have some black-box work, the interpretability of the model would rise to a new level. Such a modular design would benefit the most if it could be aided by research in the direction of XAI (using tools to extract possible logical features from the transformer).

Lastly, this modular design will allow combining more statistics and machine learning algorithms. The DNN will improve when encountering mature NLP algorithms, such as the term frequency-inverse document frequency (TF-IDF) [22], best matching 25 (BM25) [23], and latent Dirichlet allocation (LDA) algorithms [24].

4.2 Increasing Output Controllability and Model Tunability

Controllability and tunability are two natural improvements that increase model interpretability. As mentioned, a modular design could consider assigning model output tasks to more traditional machine learning or statistical algorithms, whose output is more controllable and tunable than deep learning. Such a design does not negate the advantages of DNNs in terms of output, for example, in a knowledge base where relevant information is retrieved, and keywords allow the transformer model to present more natural human-like output. The output is based on a well-designed knowledge base; thus, the risk of the output is low. These are only preliminary ideas, and in practice, there will be many

problems, such as how to choose the form of the knowledge base, which must be further explored.

4.3 Dealing with Possible Performance Degradation

First, this research does not avoid the DNN model headed by the transformer, so the performance degradation is minimal. On the contrary, due to the modular design and improvement of interpretability, the industrial and commercial value of the model will rise steadily.

The most promising aspect of deep learning-based conversational AI is its natural human-like feedback, which fulfils the impression of robot intelligence. However, such an impression often overestimates the transformer model and blindly pursues its overly complex architecture.

5 Methodology

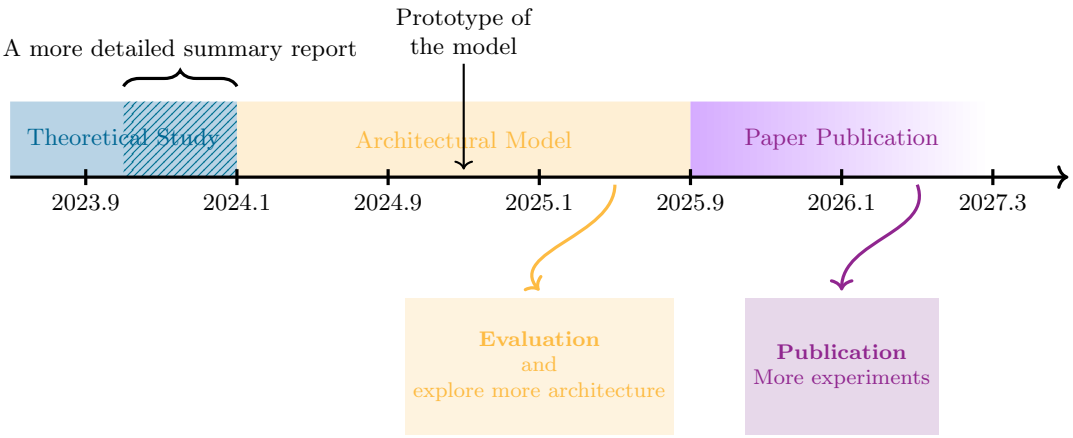
Currently, model research in AI is essentially algorithm research. The researcher will start with the mathematical theory and explore the mathematical nature of the self-attention mechanism. The researcher has rich experience in mathematics learning and paper reading; therefore, this starting point is feasible. Today's conversational AI can be understood at a higher level only after combining theoretical and experimental knowledge. Although the theoretical exploration is relatively broad and does not belong to the research method, it will be the starting point of no doubt. With the popularity of the open-source organisation and the mature research paper reading system, the in-depth study of the mathematics theory regarding AI is convenient and not subject to experimental conditions.

After a systematic study, the author plans to develop the initial model architecture on Python. The complete structural change of the transformer and the pretraining on big data are not considered for the time being; therefore, the requirements for computing power are not high. A working computer with a 30-series GPU can cover most task requirements, including transfer learning, fine-tuning and more. The author has experience using Amazon and Google cloud services, so it is feasible to use cloud computing services to realise the experiment and structure of conversational AI.

The system for evaluating conversational AI is not quite mature, but some standard metrics, such as coherence, can be used [25]. After completing the initial model architecture, the researcher can adjust it using related evaluation metrics. Moreover, to evaluate the model systematically, the researcher can compare the model output with the existing conversation dataset to evaluate its performance.

When the architecture of conversational AI is acceptable, the researcher will aim to refactor the model using C++, which has better performance than Python in terms of computing efficiency. Most Python deep learning libraries, such as Torch and TensorFlow, are based on C++ frameworks. Using C++ is worth attempting from the perspective of software performance. In general, this research is not based on computing power, but the core is exploring algorithms and model architecture.

6 Time Frame



7 Conclusion

The current AI research is entering a post-transformer era. Pursuing larger models, more data, and more complex architecture is desirable, but it is only suitable for leading information technology companies, such as Google's AI labs. Based on the shortcomings of DNN-based conversational AI, this research proposal proposes a pragmatic research direction, namely modular design, combining deep learning and traditional algorithms. If a modular design can be developed in conversational AI, this technology would be widely applicable in medical, financial, education and other fields, creating substantial social value. From an academic viewpoint, the direction of this research proposal is based on the latest technologies and theories and has creative value.

References

[1] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.

[2] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[3] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 373–383. Springer, 2020.

[4] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019.

[5] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[6] Maria das Graças Bruno Marietto, Rafael Varago de Aguiar, Gislene de Oliveira Barbosa, Wagner Tanaka Botelho, Edson Pimentel, Robson dos Santos França, and Vera Lúcia da Silva. Artificial intelligence markup language: a brief tutorial. *arXiv preprint arXiv:1307.3091*, 2013.

[7] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[8] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[9] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[10] Pasi Tapanainen and Timo Jarvinen. A non-projective dependency parser. In *Fifth Conference on Applied Natural Language Processing*, pages 64–71, 1997.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] Zi Yin, Keng-hao Chang, and Ruofei Zhang. Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2131–2139, 2017.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

[15] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*, 2019.

[16] Matej Gjurković and Jan Šnajder. Reddit: A gold mine for personality prediction. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 87–97, 2018.

[17] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[18] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10772–10781, 2019.

[19] Mattias Wahde and Marco Virgolin. Daisy: An implementation of five core principles for transparent and accountable conversational ai. *International Journal of Human–Computer Interaction*, pages 1–18, 2022.

[20] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[21] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[22] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[23] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[24] Hua Yu and Jie Yang. A direct lda algorithm for high-dimensional data—with application to face recognition. *Pattern recognition*, 34(10):2067–2070, 2001.

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