He Can Talk! Oh My Word, He can Talk!

Anonymous AACL-IJCNLP submission

Abstract

Voice interactions with computers (e.g. voice assistants) currently suffer from a number of problems which prevent human quality, voice conversation.

These problems can be divided into:

1. Open domain conversation
2. Prosody understanding and generation that operates in real-time
3. Computing resources efficient.

In this work we show how all of these problems can be addressed by creating an end-to-end voice in, voice out neural network. We show that appropriate learning can improve voice interactions and bring them closer to human levels.

We demonstrate that by using large language, graph and voice models, a natural language interaction is achieved.

[Ivan to add a sentence to explain what is new]

1 Introduction

Currently, neural generation techniques have powered many inspiring applications, e.g., poem generation (Yang et al., 2018), neural machine translation (NMT) (Bahdanau et al., 2015) and chatbot (Zhao et al., 2017). Conditional (also known as controllable) text generation is an important task of text generation, aiming to generate realistic text that carries a specific attribute (e.g., positive or negative sentiment).

A common solution is to encode the condition into a vector representation and then integrate it with the text generation process (Kingma ...).

These existing neural models have achieved encouraging results. However, when a new condition is added (e.g., a new topic for categorical generation), they require a full retraining or finetuning. This process is both time-consuming and computationally inefficient (Houlsby et al., 2019).

Both fine-tuning and retraining are not desirable in real-world applications since the delivery (e.g., transmitting updated weights through the Internet) and client-side re-deployment (e.g., distribute updated weights to users) of large-scale weights are often difficult.

Inspired by the recent success of Variational Auto-Encoder (VAE) (Kingma and Welling, 2014) based post-hoc conditional image generation strategy (Engel et al., 2018), we provide a new perspective for flexible conditional text generation.

2 Preliminaries

2.1 General framework

3 Methology

block diagram comes here..

3.1 Language Understanding And Generation

3.1.1 Large Attention Language Model

3.1.2 Autonomous Language Generation Control

3.1.3 Knowledge Graph

3.2 Speech Modality

3.2.1 Inputs

The baseline requirement from the user's speech is for KAMI to know what the user said. For this we employ Google's

¡ORI¿

In addition to knowing what the user said, the system can benefit by knowing how the user said it – what emotional state affected the user and caused the stress and intonation that are evident in the speech data recognized by KAMI.

Such emotion analysis on the user's voice characteristics is called "sentiment extraction".

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

3.2.2 Outputs

The baseline requirement from KAMI's speech is to convey the content of KAMI's responses, the spoken words, to the user.

For this we employ the ¡WHATEVER – Ori to insert¿

text-to-speech system.

In addition to conveying the words with which KAMI is responding to the user, the system imposes on the spoken words an appropriate prosody for conveying any required emphasis as well as KAMI's emotional state.

4 Experiments

4.0.1 Setup

4.0.2 Hyperparameters

4.0.3 Evaluation

5 Results

6 Discussion

7 Related work

¡models papers here¿

8 Conclusion

In this paper, we present a novel speech generation framework for flexible conditional voice interaction. The extensive experiments demonstrate the superiority of the proposed framework against the existing alternatives on conditionality and diversity while allowing a new type of social oriented conversation. Further more, we achievied new state of the art results.

¡actual numbers¿

9 Acknowledgments

Thank yo Gadi for being what you are...

10 References

A Appendix

You may include other additional sections here.

B Credits