2R-07

感情を考慮した対話システム

石田将大　長名優子

東京工科大学 コンピュータサイエンス学部

1　はじめに

GPT (Generative Pre-trained Transformer)[1]-[3] は、OpenAI で開発された Transformer[4] のデコーダ部分をベースとした言語モデルである。名前にもある通り、事前学習は GPT の重要な要素の一つである。事前学習では、大規模なラベルなしデータを用いて本来のタスクではないタスクの学習を行う。ラベルなしのデータは、ラベルありデータに比べて楽に多くのデータを用意することができ、大規模なデータを用いた学習を行いやすいというメリットがある。

現在、この GPT のモデルを用いて、さらに対話に特化させたものが ChatGPT としてよく知られている。この ChatGPT をはじめ、コンピュータとの対話を行うシステムではテキストベースでの対話が主流となっている。それに対し、人同士の対話では会話の内容だけでなく、声のトーンや、態度などからも情報を得ながら会話している。また、同じ漢字でも読み方を間違えただけで、大きく印象が変わってしまうこともある。テキストベースの会話では適切に相手に伝わらず、すれ違いが起きることもある。そこで、本研究では、感情を考慮した対話システムを提案する。

2　Transformer

Transformer[4] は、もともと自然言語処理向けに考案されたモデルであるが、他のタスクでも高い性能を発揮できることがわかり、様々なタスクで使用されている。Transformer の最大の特徴は従来の手法で用いられていた再帰や畳み込みを用いずに Attention 機構のみを用いて構築されていることである。Attention のみを用いることで、学習時間を短縮しつつ、高精度な結果を出すことが可能となっている。構造は図 1 のようになっており、左側がエンコーダ、右側がデコーダである。入力された文章は Input Embedding 層でベクトルに変換され、Positional Encoding で位置情報が埋め込まれる。それから Multi-Head Attention 層で Attention を計算していくような流れになっている。

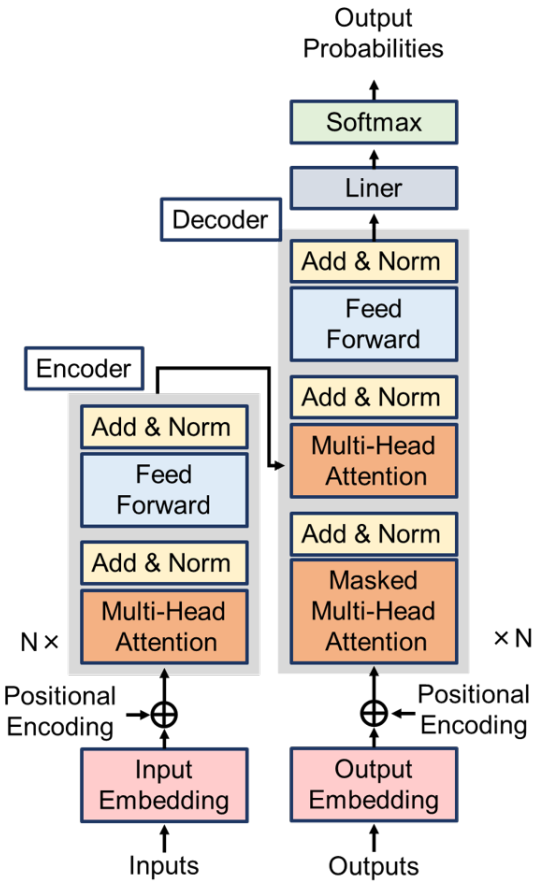

図 1: Transformer の構造

グレーの四角で囲まれている部分の処理を N 回繰り返し、多層化する。

3　感情を考慮した対話システム

提案する感情を考慮した対話システムの構造を図 2 に示す。提案システムは、GPT[1]-[3] を参考に基本的には Transformer[4] のデコーダ部分を使用したモデルである。図 2 の右の部分が Transformer のデコーダ部分をベースにしたものになっている。入力としては、音声信号を考えている。音声信号に対して、Wisper[5] を用いて音声認識を行ったものに対して Text Embedding を行う。また、同時に、Empath[6] を用いて音声感情認識を行い、ユーザの発話に含まれる感情の情報を抽出しする。この情報を Positional Encoding の情報を埋め込むのと同じタイミングで埋め込む。

Dialogue System considering Emotions
Masahiro Ishida and Osana Yuko(Tokyo University of Technology, osana@stf.teu.ac.jp)

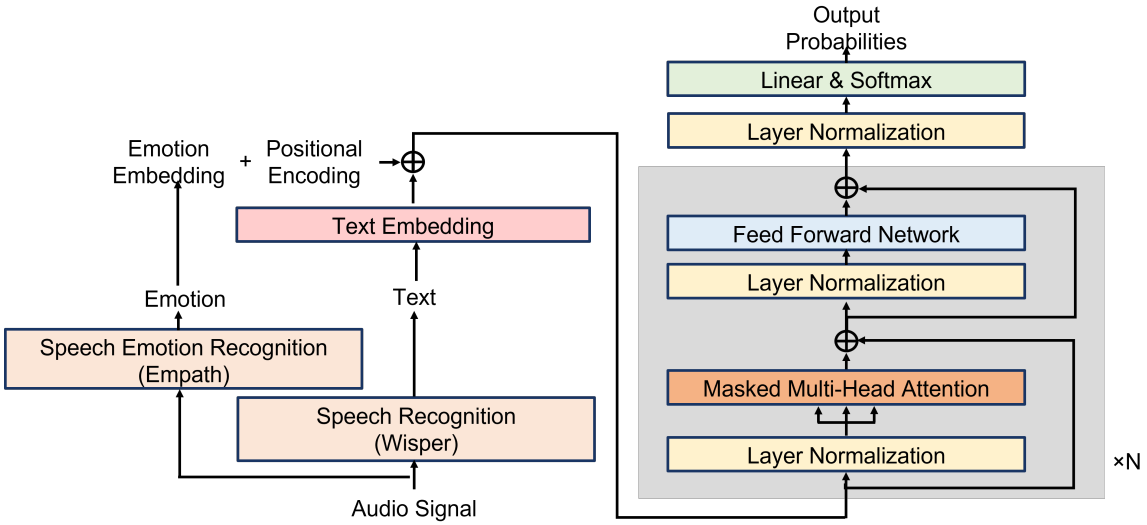

図 2: 感情を考慮した対話システム

3.1 学習

学習は、最初に感情埋め込みを行わないモデルで事前学習を行う。事前学習は、GPT を参考に Common Crawl[7] などの大規模データを用いて行う。その後、感情埋め込みを含め、ファインチューニングを行う。その際に用いるデータセットは音声対話コーパスに対し音声認識や音声感情認識を行える API を利用して生成する。

3.2 感情埋め込み

感情の埋め込みは、Empath を用いて音声感情認識して得られる感情ベクトルを利用して行う。Empath は、平常、怒り、喜び、悲しみ、元気度の 5 項目を 0~50 の数値で返すため、これを感情ベクトルとする。感情ベクトルは、単語単位ではなく発話単位で埋め込むため、1 つの発話に含まれる単語にはすべて同じベクトルが感情埋め込みに用いられることになる。

4 計算機実験

計算機実験を行い、提案システムにおいて埋め込みを行わない場合と比べて応答が変化することを確認した。

参考文献

[1] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever : “Improving language understanding by generative pre-training,” https://openai.com/research/language-unsupervised (2024/01/03 参照).

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever : “Language models are unsupervised multitask learners,” https://paperswithcode.com/paper/language-models-are-unsupervised-multitask(2024/01/03 参照).

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei : “Language models are few-shot learners,” https://openai.com/research/language-models-are-few-shot-learners (2024/01/03 参照).

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin : “Attention is all you need,” https://arxiv.org/pdf/1706.03762.pdf, (2024/01/03 参照).

[5] Wisper, https://platform.openai.com/docs/guides/speech-to-text/, (2024/01/03 参照).

[6] Empath, https://webempath.net/lp-jpn/, (2024/01/03 参照).

[7] Common Crawl, https://commoncrawl.org, (2024/01/03 参照).