平成 30 年度 筑波大学情報学群情報科学類 卒業研究中間報告

モジュール分割された日本語対話システムの作成

知能情報メディア主専攻　201611350　江畑 拓哉

指導教員　Claus Aranha（コンピュータサイエンス専攻）
櫻井鉄也（コンピュータサイエンス専攻）

提出日　2010 年 9 月 30 日

1　序論

現在世界では様々な目的を持った対話システムが研究されている。主にそれは、ユーザを何らかの目標に導くためのタスク指向システム、対話そのものを目的とした非タスク指向システムの二種類に分かれている。またこの二つを組み合わせたシステムも開発されており、 Apple 社の Siri がこれに該当する。

本研究では上記の二つを満たし、かつより低コストなシステムを構築するため、様々な目的を持ったモジュールを作成し組み合わせる手法を提案する。尚ここで言う低コストとは、学習時間やモデルの更新の容易さなどの意味で用いている。

2　研究概要

2.1　研究目標

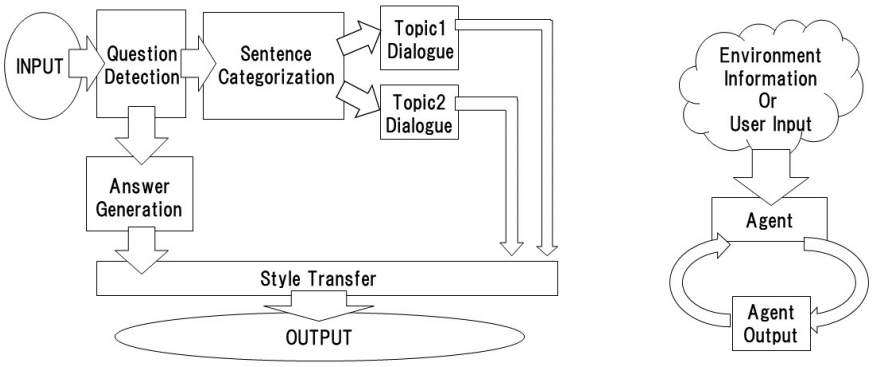

図 1　システム概図

本研究は 図 1 のシステムを構築することを目的とする。

例えばそのエージェントの名前に関する質問が来た場合は、 Question Detection を通り Answer Generation で自分の名前に関する文章を生成し、 Style Transfer を経由して回答を出力する。ゲームに関する入力が来た場合は、 Question Detection を通過し Sentence Categorization で ゲームに関する Topic Dialogue にアクセスし出力する。最終的にそれらの情報を Style Transfer を通して出力する。Style Transfer とは口調変換であり、エージェントに様々なペルソナを持たせるために用いられる。そのペルソナはエージェントの外見などと同期することでより対話感を出すことが出来ると考えている。またエージェントは天候などの環境データ、今までの対話データからそのステータスが変化することを想定しており、例えばあるものに対する好感度はそれまでの対話から形成されるようにする。

2.2　利用するデータ

このシステムにおいて必要とするデータは主に二種類である。

- Style Transfer を行うためのデータ
  　このデータには、元の文と付与された文のペアを集めたものと、スタイルが付与されていない文の群と付与された文の群を用いるものがある。現在の進捗としてはデータ生成の問題から前者を用いている。
- 対話を行うためのデータ
  　入力文と出力文のペアを集めたものである。収集手法として、1. 実際に手作りする 2. アニメや演劇などから集める 3. SNS から集める を行い、特に後者 2 つについては有効なデータをフィルタリングする手法を研究している。

2.3 それぞれのモジュールに適用した手法と今後の予定

- Question Detection
  入力文に対して LSTM を適用し質問番号を出力する手法によって殆どの分類が行えることがわかった。
- Answer Generation
  あらかじめ用意したテンプレートにエージェントのステータスから取り出した値を当てはめることで生成する。
- Sentence Categorization
  SCDV [4] のような手法や、sentiment analysis で使われる手法 [3] を検討している。また単語埋め込みとして、 fasttext [2] や Word2Vec を採用したいと考えている。いずれも十分なサイズのデータを収集できておらず実験が出来ていない。代替としてアニメの台本から入出力のペアを切り出した対話として成立しているデータ、一般的な対話を集めたデータを用いて RNN, CNN を用いた二値分類実験を行い、後者で 8 割 5 分ほどの検出をすることが出来た。
- Topic Dialogue
  日本語でも Sequence to Sequence [6] のような手法を用いた様々な取り組みが行われているが、芳しい結果を得られたものは少ない。しかし公開されているデータセットを確認すると、英語のデータセットと比較して、入力と出力のペアで意味が通らないものがあるなど、有効なデータセットとは言えなかった。自作の日常会話を集めたデータセットで Sequence to Sequence を用いた実験をしたところ、データセットの近傍の入力に対し意味のある出力を得た。ところがデータ量が少ないため (1k 対話程度) これをいかに増やしていくかが今後の課題となる。
- Style Transfer
  スタイルを付与されていないものと、付与されたもののペアをデータセットとして、 Sequence to better Sequence [5] と、DAE [7] から着想を得た提案手法と CopyNet [7] を用いて実験を行った。結果として前者 2 つはチューニングが難しい代わりに自然な文を生成することに成功し、後者は比較的高速に文を生成できる他にある程度未知語に対応可能であった。今後 Sequence to better Sequence に CopyNet を組み合わせたシステムを提案したいと考えている。
- データ収集について
  手作りでデータを作成する手法に関しては誤字脱字や意味の通らない文を含まない理想的なデータになるように作成している。アニメや演劇からの収集については、長すぎる文は必要な部分のみにし、専門的過ぎる文は削除している。こちらは著作権などの問題があるため、この対処を調査中である。SNS からのデータの収集については Twitter を用いているが、ノイズの多いデータが混じっているため、そのデータクリーニングについて既存の手法 [8] を参考にしながら研究している。
- ステータス更新を行う手法
  文章に関しては Word2Vec [1] や fasttext [2] を用いた極性判定を用いて実験する予定だ。

References

[1] Yoav Goldberg and Omer Levy. *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. 2014. arXiv: 1402.3722.

[2] Armand Joulin et al. *Bag of Tricks for Efficient Text Classification*. 2016. arXiv: 1607.01759.

[3] Yoon Kim. *Convolutional Neural Networks for Sentence Classification*. 2014. arXiv: 1408.5882.

[4] Dheeraj Mekala et al. *SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations*. 2016. arXiv: 1612.06778.

[5] Jonas Mueller, David Gifford, and Tommi Jaakkola. "Sequence to Better Sequence: Continuous Revision of Combinatorial Structures". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 2536–2544. URL: http://proceedings.mlr.press/v70/mueller17a.html.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215.

[7] P. Vincent et al. "Extracting and composing robust features with denoising autoencoders." In: (2008).

[8] 稲葉 通将, 神園 彩香, and 高橋 健一. "Twitter を用いた非タスク指向型対話システムのための発話候補文獲得". In: 人工知能学会論文誌 29.1 (2014), pp. 21–31. DOI: 10.1527/tjsai.29.21.