非タスク指向対話システムにおける攻撃的ユーモア発話の生成

上垣 貴嗣　　藤倉 将平　　菊池 英明
早稲田大学 人間科学部
{t-gappy@fuji., 0spiral1@asagi., kikuchi@}waseda.jp

1 はじめに

近年, 自然言語による対話機能を持つ対話システムが社会に普及してきている. 日常的な雑談を行う非タスク指向対話システムには, 今後の発展と普及の中で, 人との多様な関係性が想定される. 人同士の会話における, 相手との関係性に応じた発話戦略の類型論としてポライトネス理論 [1] がある. 対話システムでも, 相手との関係性を考慮した発話を行うことで会話を円滑に進めることができ, 対話継続欲求に良い影響があると考えられる. そこで本研究では, 比較的親しい間柄で行われるポジティブ・ポライトネス戦略発話の中で, 特に親しい友人同士に見られるからかいやいじり, 皮肉などの攻撃的ユーモア発話 [2] に着目する. ユーザーの行動/状況報告に対して, 親しい友人関係を想定して攻撃的ユーモア応答を行うシステムを構築する.

2 提案手法

本研究の提案手法の流れを図 1 に示す.

攻撃的ユーモア応答が可能なモデルを構築するため, ルールベースの前処理を用いて Twitter から取得した行動/状況報告のツイート・リプライペア (Twitter データ) に対し, 半教師あり学習に類するラベリングを行う. Twitter データを一部分割し, 実際にあったツイートに対してお笑い芸人に攻撃的ユーモア応答を作成してもらう (芸人データ). 芸人データに対し, 攻撃的ユーモア度と暴言度について印象評定実験を行い, それぞれの得点で一定基準を満たすペアを「攻撃的ユーモア (humor)」, それ以外を「その他一般 (other)」としてラベリングする. 芸人データをもとに, ツイート・リプライペアに対する攻撃的ユーモア判別モデルを構築する. 攻撃的ユーモア判別モデルは, BERT[3] の日本語事前学習済みモデル [4] をファインチューニングする形で学習する. 芸人データ作成に利用しなかった Twitter データを攻撃的ユーモア判別モデルにかけてラベリングする (攻撃的ユーモア判別後データ). 攻撃的ユーモア判別後データと芸人データを合わせて対話システム構築用データとし, seq2seq モデルでの対話システム構築に利用する.

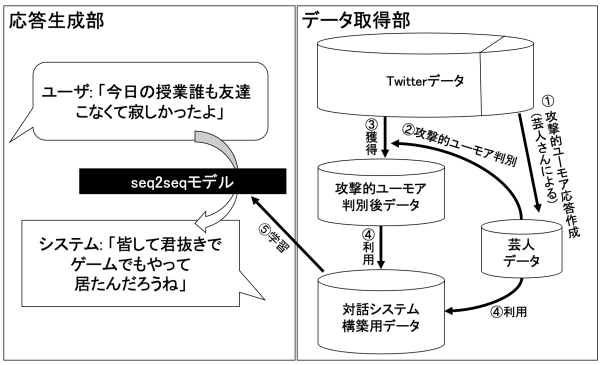

図 1: 提案手法の流れ

対話システムのモデルには, seq2seq の一種である Transformer[5] を使用する. seq2seq モデルの学習では, 推論時の条件分岐や言語モデルの効率的な学習を期待し, 特殊トークンを用いた学習を行う. 学習データのラベルに合わせて, 入力発話の文末に <humor> と <other> いずれかの特殊トークンを付与した上で, 同一のモデルを学習する.

3 評価実験

3.1 実験に用いたデータ

2019 年 6 月 28 日から 2019 年 12 月 19 日にかけて約 777 万件のツイート・リプライペアを Twitter から取得した. 提案手法の流れに沿って前処理と分割を行なった結果, 芸人データは 1346 件 (入力文ユニーク数 673 件), 残りの Twitter データは 78721 件となった.

芸人データの攻撃的ユーモア度と暴言度には, 8 名の被験者による 5 段階の印象評定得点の平均値を利用した. 本研究における攻撃的ユーモアを「攻撃的ユーモア度が平均値 (2.5) 以上かつ暴言度が 3.0 以下のもの」と定義し, humor ラベルが 546 件, other ラベルが 800 件となった.

3.2 攻撃的ユーモア判別モデルの評価

本研究で定義する攻撃的ユーモアに対する, ラベリングの妥当性を測るため, 攻撃的ユーモア判別モデルの評価を行った. 精度は F 値において 0.72 となった.

攻撃的ユーモア判別モデルを, 芸人データ作成に利用しなかった Twitter データへ適用し, 13287 件の humor ラベルペアと 65434 件の other ラベルペアを取得した.

3.3 攻撃的ユーモア応答生成の評価

システムの攻撃的ユーモア応答生成能力を測るため, 32 人の被験者による印象評定実験を行った. 同じ入力に対して, システム応答の humor 群と other 群, 人手作成の芸人データテスト群の 3 群を設けた. 刺激は 1 群あたり 68 件とし, 各群の発話・応答ペアに「攻撃的ユーモアであるか」の 5 段階評価を求めた. 印象評定得点をユーザーごとに平均 0, 分散 1 となるよう標準化し, 刺激ごとに平均値を算出した.

Wilcoxon signed-rank test の結果, 全群の間で $p < .01$ の有意差が見られた (図 2). 提案手法によって, 芸人データほどでは無いが, その他一般応答より攻撃的ユーモアらしい応答が可能であると示した.

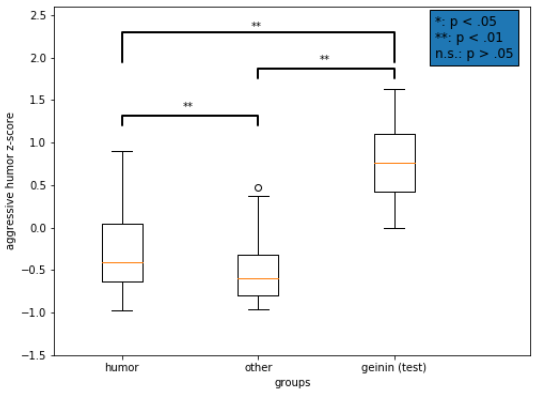

図 2: 攻撃的ユーモア印象評定の結果

3.4 対話継続欲求の評価

攻撃的ユーモア発話が対話継続欲求に与える影響を検討するため, 32 人の被験者による印象評定実験を行った. 同じ入力に対して, システム応答の humor 群と other 群の 2 群を設けた. 刺激は 1 群あたり 68 件とし, 各群の発話・応答ペアに「応答として有り得るか」「継続して対話したいか」の 5 段階評価を求めた. 加えて, 被験者特性として攻撃的ユーモア志向尺度 [6] と日常生活における笑いに関する自己評価 [7] の 13 項目への回答を求めた.

被験者特性合計点の平均値である 36.19 を境目として, 被験者を high_humor 群と low_humor 群の 2 群に分けた. 被験者群は各群に 16 人ずつ被験者が含まれることとなった. 各印象評定得点をユーザーごとに平均 0, 分散 1 となるよう標準化し, 刺激ごとに被験者群ずつ平均値を算出した. 応答の適切さによる対話継続欲求への影響を減らすため, humor 群と other 群でいずれも「応答として有り得るか」の標準化得点が 0 以下のものを削除し, 刺激は 1 群あたり 37 件となった.

Wilcoxon signed-rank test の結果, 「継続して対話したいか」について 2 群間で有意差は見られなかった. 一方で, 攻撃的ユーモア志向尺度 [6] と日常生活における笑いに関する自己評価 [7] の得点が高い被験者 (high_humor) において, humor 群の平均値に上昇が見られ, 攻撃的ユーモア発話が対話継続欲求に良い影響がある可能性を示した (表 1).

表 1: 被験者群/刺激群別の対話継続欲求の平均と分散

	high_humor		low_humor	
	humor	other	humor	other
平均値	**0.019**	**0.014**	0.005	0.109
分散	0.275	0.607	0.289	0.690

参考文献

[1] Brown. P., Levinson. S.C.: “Politeness: Some universals in language usage”, Cambridge University Press, 1987.

[2] 大津友美: "親しい友人同士の会話におけるポジティブ・ポライトネス -「遊び」としての対立行動に着目して-", 社会言語科学, Vol.6, No.2, pp.44-53, 2004.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding”, Proc. NAACL 2019, Vol.1, pp.4171-4186, 2019.

[4] Yohei Kikuta: "Bert pretrained model trained on japanese wikipedia articles", https://github.com/yoheikikuta/bert-japanese, 2019.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin: "Attention Is All You Need", Proc. 31st NIPS, pp.6000–6010, 2017.

[6] 上野行良: "ユーモアに対する態度と攻撃性及び愛他性との関係", The Japanese Journal of Psychology, Vol.64, No.4, pp.247-254, 1993.

[7] 伊藤理絵, 本多薫, 渡邊洋一: "攻撃的ユーモアを笑う", 山形大学人文学部研究年報, No.8, pp.215-227, 2011.