4D-05

マイクロブログからの対話コーパスの自動構築

関田 崇宏* 白井 清昭†
北陸先端科学技術大学院大学 先端科学技術研究科

1 はじめに

対話コーパスは対話システムの研究に欠かせない言語資源である．しかしながら，対話の収録や書き起こしなどの作業を要するため，作成コストが高く，大規模な対話コーパスを構築することが難しいという問題がある．この問題に対し，Twitter からツイートとそれに対するリプライを抽出し，それを発話と応答の組とみなすことで擬似対話コーパスを自動構築した上で，これを対話システムの実現やモデル化に活用する研究が行われている．Ritter らは，Twitter からリプライを 3 回以上辿って得られる長さ 3 以上の対話を収集し，これを用いて発話行為を推定する教師なし機械学習手法を提案している [3]．東中らは，リプライを辿って得られる対話の多くは長さ 2 であることから，長さ 2 の対話を連結して長い対話を生成し，対話モデルを構築する手法を提案している [2].

しかしながら，Twitter から得られる擬似対話の中には，対話として不適切なものも多く含まれる．先行研究では，リプライの連鎖が真に対話として成立しているかについては考慮していなかった．大量の対話を獲得することで不自然な対話の影響を軽減するという考え方もあるが，自然な対話を選別して対話コーパスを構築する方が望ましい．本論文は，Twitter から取得されたリプライの連鎖が対話として適切であるかを判定し，不自然な対話を除外することで，質の高い対話コーパスを自動構築する手法を提案する [4].

2 提案手法

2.1 対話収集

Twitter から対話を収集する手続きを以下に述べる．

1. キーワードでツイートを検索する．
2. 得られたツイート t_1 が別のツイート t_2 のリプライであれば，t_2 を取得する．
3. 2. の操作をツイート t_i が別のツイートのリプライでなくなるまで繰り返す．
4. 得られたツイートを逆順に ($t_n t_{n-1} \cdots t_1$ の順に) 並べ，対話として保存する．
5. 1.〜4. の操作を繰り返す．

ツイートの検索や収集は Twitter API (Tweepy[1]) を用いる．本研究では長さ 3 以上の対話のみを収集する．

上記の手順 1. で用いる検索キーワードとして，単語親密度 [1] の高い単語を使用する．単語親密度の高い単語は，使用頻度が高いために多くの対話を収集するのに適しており，また親密度が高い単語を含む対話は自然な対話になりやすいと考えたためである．本研究では「全員」「再会」「恋人」など 1,686 個のキーワードを使用する．

2.2 不適切な対話の除去

収集した対話に対し，それが不適切な対話であるかを判定し，適切な対話のみ対話コーパスに収録する．本研究では以下の 4 つのルールによって不適切な対話を検出する．

画像・URL を含む対話を除去するルール ($\mathbf{R_{image}}$)

対話の中に画像を含むツイートが存在するとき，対話として成立しないときがある．以下の対話例 D1 では，画像 (*IMAGE*) を参照しなければ対話の内容を理解することができない．本研究では，対話コーパスを収録する際にはテキストのみを保存することを仮定しているため，このような対話は除去する．

D1
u_1: これ，どう思う? *IMAGE*
u_2: 究極の変顔だね

ただし，D2 のように画像を含んでいても対話として成立する場合もあるので，注意を要する．

D2
u_1: お祭り楽しい *IMAGE*(祭の画像)
u_2: いいなー

$\mathbf{R_{image}}$ は以下のいずれかの条件を満たす対話を除外するルールである．

1. URL と指示語 (「これ」など) を含むツイートが存在する．
2. URL のみ，もしくは URL とハッシュタグのみのツイートが存在する．
3. u_i が URL を含み，かつ u_{i+1} に「それ」などの指示詞を含む．

不特定多数のユーザへの呼びかけを除去するルール ($\mathbf{R_{invite}}$)

対話の起点となるツイートが不特定多数への呼びかけであるとき，そのツイートとそれへのリプライは対

Construction of Dialog Corpus from Microblog
*Takahiro Sekita , Japan Advanced Institute of Science and Technology, †Kiyoaki Shirai , Japan Advanced Institute of Science and Technology

[1] https://www.tweepy.org/

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.

話として成立しないことが多い．例えば，大喜利のお題を発信するツイートとそれに対するリプライは対話として適切ではない．

このような対話を検出する方法を検討したところ，対話の最初の発話 u_1 に対するリプライ数が多いとき，それが不特定多数のユーザへの呼びかけである場合が多いことがわかった．しかしながら，技術的な問題[2] から，リプライ数を基準に不適切な対話を検出するシステムを実装することは断念し，代わりに簡便な手法を採用する．

$\mathbf{R_{invite}}$ は，対話の起点となるツイート u_1 のユーザが大喜利を運営しているユーザであれば，その対話を除去するルールである．大喜利ユーザのリストは,「大喜利」でキーワード検索するなどして，人手で用意した．ユーザ数は 61 である．

セリフを含む対話を除去するルール ($\mathbf{R_{line}}$)

ツイートの中に他者のセリフがあるとき，対話として成立しないことが多い．例えば，以下の対話 D3 では，ひとつの発話に複数人のセリフがあり，対話としては不自然である．

D3 | u_1: 客「何？ここ禁煙なの？」ワシ「禁煙にさせて頂いてますすみません〜」客「チッ」

$\mathbf{R_{line}}$ は，以下の全ての条件を満たすツイート u_i を含む対話を除外するルールである．

1. u_i は 2 つ以上の括弧の組 (‘「’ と ‘」’) を含む．
2. 括弧の中の文字数が 6 文字以上である．
3. 括弧の次の単語が助詞ではない．

2. と 3. は，括弧が (セリフではなく) 単語を強調するために使われていないことを確認するための条件である．

短い発話を含む対話を除去するルール ($\mathbf{R_{short}}$)

あまりに短いツイートを含む対話は対話として成立しないことが多い．D4 はその例である．

D4 | u_1: 少し寝ようと思って寝たらこの時間つんだ
u_2: を

$\mathbf{R_{short}}$ は以下のいずれかの条件を満たすツイートを含む対話を除外するルールである．

- 1 文字のひらがなのみのツイート．ただし，間投詞となる「あ」「え」「お」を除く．
- 全角スペース，句点，読点のみのツイート．
- 絵文字のみのツイート．

3 評価実験

Twitter からの対話の収集は，2019 年 6 月 8 日から 2019 年 12 月 25 日にかけて実施した．収録した対話数は 92,207，平均対話長は 9.50 であった．

次に，ランダムに対話を 100 件選択し，本研究で提案するルールを用いて不適切な対話かどうかを判定した．提案手法は取得した対話のうち不適切なものを検出するが，本研究の目的は適切な対話を残して対話コーパスを構築することなので，適切な対話，不適切な対話それぞれの検出の精度，再現率，F 値を評価指標とする．評価結果ならびに対応表を表 1 に示す．NG は不適切な対話，OK は適切な対話を表わす．

表 1: 不適切な対話の検出手法の評価

	NG	OK
精度	0.75	0.70
再現率	0.32	0.94
F 値	0.45	0.80

		(正解) NG	(正解) OK
(判定)	NG	12	4
	OK	25	59

False Negative (不適切な対話を検出できなかった誤り) が多く，不適切な対話検出の再現率が低いが，その多くは画像を含む不適切な対話を検出できていなかった．ルール $\mathbf{R_{image}}$ をさらに洗練する必要がある．ただし，適切な対話を検出するという観点から見れば，精度，再現率は高く，概ね良好な結果が得られている．

不適切な対話を除外するルールを個別に評価する[3]．各ルールで不適切と判定された対話の中からランダムに 50 個を取得し，それが真に不適切であるかを人手で評価した．結果を表 2 に示す．

表 2: 各ルールによる不適切な対話の検出精度

$\mathbf{R_{image}}$	0.78	$\mathbf{R_{line}}$	0.74	$\mathbf{R_{short}}$	0.52*

$\mathbf{R_{image}}$ と $\mathbf{R_{line}}$ の精度は 75%程度であった．$\mathbf{R_{short}}$ については検出精度が低いが，2.2 項で述べた手法の実装が完了していないため，本稿では 3 文字以下の発話を含む対話を除外する簡便なルールを評価した．$\mathbf{R_{short}}$ の完全な実装と評価は今後の課題である．

近日中に，収集した全対話に対して不適切な対話を除去する処理を行い，対話コーパスを構築する予定である．

参考文献

[1] 浅原正幸. クラウドソーシングによる単語親密度の推定. 言語処理学会第 25 回年次大会, pp. 45–48, 2019.

[2] 東中竜一郎, 川前徳章, 貞光九月, 南泰浩, 目黒豊美, 堂坂浩二, 稲垣博人. 2 ツイートを用いた対話モデルの構築. 言語処理学会第 18 回年次大会, pp. 567–570, 2012.

[3] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *HLT-NAACL*, pp. 172–180, 2010.

[4] 関田崇宏. マイクロブログからの対話コーパスの自動構築. 修士論文, 北陸先端科学技術大学院大学, 3 2020.

[2] Twitter API では，ツイートに対するリプライの数は取得できない．ウェブスクレーピングでは取得できるが，Twitter 社はこれを許可していない．

[3] $\mathbf{R_{invite}}$ は人手で作成した大喜利ユーザのリストを用いたルールであるため，ここでは評価しない．

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.