6Q-08

状況に即した返信文候補提示手法における単語被覆率の適用

渡辺　紀文 † 松原　雅文 ‡ Goutam Chakraborty‡ 馬淵　浩司 ‡

岩手県立大学大学院ソフトウェア情報学研究科 † 岩手県立大学ソフトウェア情報学部 ‡

1. はじめに

携帯端末において SNS を利用し，ユーザ同士でコミュニケーションをとる機会は多い．SNS において返信文を作成する際には，返信文を推敲し，文字入力を行う必要があり，ユーザにとって負担となっている．

そこで，我々はユーザの負担を軽減させるために，返信文候補を状況ごとに複数提示する手法を提案している [1]．提示する返信文候補は受信文の内容に即しながら，それぞれ違う状況に対応することで，ユーザにとって様々な場面に適応可能である．なお，返信文候補は，ツイートを収集しクラスタリングを行うことで出力している．また，短文であるほうが返信文として有効であると仮定し，文字数に着目した評価を行っている．

本稿では，より短文であるほうが返信文として有効であるという仮定に基づいたうえで，評価指標に単語被覆率を加え，短文かつ端的に状況を示す返信文候補の提示する実験を行い，その結果から単語被覆率を適用する有効性について述べる．

2. 既存手法

単語被覆率を用いて代表ツイートを選択し，ユーザに提示するシステムとして，坂本ら [2] が提案した手法がある．Twitter は，特定のトピックに対してさまざまな情報が膨大に取得できるという特徴があるため，ユーザが情報を収集しやすいという利点がある．しかし，膨大なツイートが情報源となるために，ユーザがすべての情報に目を通すことは困難で，情報の取捨選択を行わなければならず，ユーザにとって大きな負担となる．

この手法では，特定のトピックを表すツイート群に対して，単語被覆率を用いて代表ツイートを抽出しユーザに提示することで，特定のトピックの要約を実現し，ユーザの負担を軽減させている．

本手法では，この単語被覆率を適用し，状況に即した返信文候補として代表ツイートの選択を行う．単語被覆率を用いることで，特定の状況を端的に示すツイートを返信文候補として提示することを目指す．

3. 提案手法

本手法の流れを 3.1.〜 3.3. に示す．

3.1. ツイートの収集

返信文候補になり得るデータを収集するために，受信文内に含まれる名詞でツイートの検索を行う．受信文内の名詞で検索を行うことで，受信文の話題に近いツイートを収集する．

3.2. クラスタリング

収集したツイート群に対してクラスタリングを行う．クラスタリングには一般的な手法である K-means 法を用いる．クラスタリングの素性には，状況説明の役割を果たしている用言の出現回数を用いる．用言を素性とすることで，それぞれ違う状況を示すツイートごとにクラス分類する．

また，各クラス内で最も出現回数の高い素性を，そのクラスの状況を示す代表素性とする．しかし，代表素性の出現回数がクラスに所属しているツイートの件数を超えていない場合は，特定の状況に依存していないクラスとして，代表素性を出力しない．

3.3. 単語被覆率の適用

クラスタリングによって形成された各クラス内のツイートに対して，単語被覆率を算出する．各クラス内において，単語被覆率が最上位のツイートを返信文候補としてユーザに提示する．クラス c に属するツイート t における単語被覆率 WC_t を式 (1) のように定義する．

$$WC_t = \frac{W_t \cap I_c}{W_t \cup I_c} \tag{1}$$

W_t はツイート t に含まれる用言と名詞の集合である．I_c はツイート t が属するクラス c における重要単語の集合である．重要単語とはクラス内において tf 値が上位 5 件の自立している用言と名詞を指す．つまり，多くの重要単語を含みながら，その他の単語を含まないツイートの値が高くなる．

このような処理により，収集したツイート群から返信文候補として有効なツイートをユーザに提示する．

Appropriate Word Coverage based on Situation for Presenting Candidate Words in Reply

Norifumi WATANABE†, Masafumi MATSUHARA‡, Goutam CHAKRABORTY‡, Hiroshi MABUCHI‡

†Graduate School of Software and Information Science, Iwate Prefectural University, ‡Faculty of Software and Information Science, Iwate Prefectural University

表 1: 最短文ツイートを出力した返信文候補 3 件

代表素性	返信文候補
良い	全国高校駅伝、我が母校は順位記録を更新！良かった！
すごい	高校駅伝、福島 3 位か。すごい。
ない	高校駅伝見てたら大幅遅刻なう

表 2: 単語被覆率を適用して出力した返信文候補 3 件

代表素性	返信文候補
良い	高校駅伝も留学生居るんだなぁ～まぁ、お前が良ければなって世界だな。
すごい	神村学園女子全国高校駅伝優勝すごいな
ない	なんもすることないからぼーっと高校駅伝見てる

4. 実験

4.1. 実験条件

単語被覆率の有効性を示すため，最も短文のツイートを出力した返信文候補と，単語被覆率を適用して出力した返信文候補との比較を行った．

実験では，受信文に含む名詞を入力することで，返信文候補を出力する．実験に用いた名詞は，2018 年 12 月 23 日 Google トレンド急上昇ワードに含まれた「有馬記念」「高校駅伝」「下町ロケット」「エスパー伊東」とした．

それぞれの名詞を検索クエリとして，2018 年 12 月 23 日 23 時 50 分時点の最新ツイート 1,000 件を収集する．また，クラスタリングのクラス数は 4 とし，形態素解析には MeCab，クラスタリングの素性には形容詞を用いた．

4.2. 実験結果と考察

検索クエリ「高校駅伝」を用い，最短文のツイートを出力した時の返信文候補を表 1 に，単語被覆率を適用して出力した時の返信文候補を表 2 に，それぞれ示す．いずれも 1 クラスは代表素性が出力されなかったため，提示される返信文候補は 3 つとなった．

表 1 から分かるように，単純に最短のツイートを返信文候補として出力したことで,「我が母校」や「遅刻」といった極めて限定的な表現が含まれる返信文候補や，今現在を表す「なう」が誤った形態素解析により形容詞として含まれた返信文候補が提示された．

これに対して，表 2 から分かるように，単語被覆率を適用して返信文候補を出力したことで,「留学生」に対する意見を含んだ返信文候補が提示された．これは,「留学生」についてツイートを行うユーザが多くいたことから，重要単語に「留学生」が選択され，提示される返信文候補が変化したことによるものであると考えられる．

また，検索クエリ「有馬記念」では，重要単語に，勝利した馬名が選択されたことから，単語被覆率を適用した時には馬名を含めた返信文候補の提示を行うことができた．同様に「下町ロケット」では「最終回」,「エスパー伊東」では「引退」などが重要単語に選択されたことから，重要単語は受信文の話題に沿った単語となっていることが確認でき，それらを用いた単語被覆率を算出することで有効な返信文候補が出力できていると考えられる．

しかし，表 2 で提示された「なんもすることないから」のような，重要単語が含まれていない冗長な表現の返信文候補が出力されてしまうこともあった．これは，単語被覆率を算出する際に，ツイート内の自立している用言と名詞のみに着目しているため，助詞や非自立の名詞がツイートに含まれても単語被覆率が低くならないためだと考えられる．

5. おわりに

本稿では，返信文候補提示手法に対して，評価指標に単語被覆率を適用し，実験結果から単語被覆率を適用することで，重要単語を含んだ有効な返信文候補の出力を可能となることが確認された．

今後は出力される返信文候補の冗長性を解消するために，助詞や非自立の名詞にも着目した単語被覆率の算出方法を検討する予定である．また，冗長な部分には重要単語が含まれていないという点から，ツイート全文を返信文候補とせず，重要単語が含まれている部分のみを抽出して返信文候補とすることも検討する．

謝辞

本研究の一部は JSPS 科研費 18K11358 の助成を受けたものである．

参考文献

1）渡辺 紀文，松原 雅文，Goutam Chakraborty，馬淵 浩司，“How 型質問文に対する返信文候補提示手法の有効性について”，情報処理学会第 80 回全国大会，7Q-86，March 2018.

2）坂本 翼，横山 昌平，福田 直樹，石川 博，“マイクロブログを対象としたリアルタイムな要約生成システムの試作”，第 3 回データ工学と情報マネジメントに関するフォーラム，F5-5，February 2011.