7N-06

災害情報採集のための会話 window 導入の評価

藤田俊之 †　　小林亜樹 ‡

† 工学院大学大学院工学研究科 電気・電子工学専攻 ‡ 工学院大学情報学部情報通信工学科

1 はじめに

リアルタイムに災害情報を得るために，Social Networking Service (SNS) の一種である Twitter が注目されている．Twitter を情報源として自然言語処理を行う際，1 投稿単位である tweet を 1 処理単位である文書とする方法が考えられるが，短文性などの問題が指摘されている [1]．この問題に対し，筆者らは reply-chain によりもたらされる reply-tree を会話木と定義し，1 会話木を 1 文書とすることで話題の連続性と自然言語処理の問題解決を提案してきた．しかし，1 会話木中にも話題の遷移が見られることがあり，採集誤りを引き起こすことがわかった [2]．そこで，本稿では細かい粒度での災害情報採集のために，会話木の部分を指す会話 window の導入を提案する．この会話 window を分類時に活用することで，より細かい粒度で会話を分析し，災害言及 tweet の採集を目指す．また，準リアルタイムでの採集を行うため，リアルタイム性を損なわない範囲の tweet 集合から教師なし分類器を用いた教師データの作成も行う．これらの内容を元にした 3 章における試作システムを用いて，災害例として 2019 年に発生した台風 Hagibis における tweet 群を対象とした 4 章の通りに評価実験を行い，会話 window による効果を示す．

2 提案手法

本章では，まず，2.1 節で会話木や会話 window について定義を行い，提案手法の内容について紹介する．提案手法は 2.2 節の教師データ作成部と 2.3 節の言及度算出部の 2 つに分かれる．教師データ作成部では，言及度算出部で用いる教師あり分類器の教師データを機械的に作成する．言及度算出部では，採集対象とする tweet が属する会話木に対して災害に言及している度合である言及度を求める．

2.1 会話 window

一連の reply のやりとりのうち，reply を行った tweet を reply 元 tweet, それに対して reply をされた tweet を reply 先 tweet とする．Reply 元 tweet と reply 先 tweet が連鎖している一連の tweet 群を一つの会話と見做す．また，tweet をノードとし，reply 元 → reply 先の reply 関係を有向辺として持つグラフをモデル化すると，根付き有向木として見做せる．これを会話木と呼ぶ．会話木の条件は，会話木 C の頂点（tweet）集合 $T = V(C)$ とするとき，その要素数を用いて，$|V(C)| \geq 2$ とし，reply 関係を持たない tweet は会話木として見做さないこととする．

会話 window では，会話木におけるエッジ（reply 関係）を距離 1 とし，各 tweet から k-近傍を 1 文書として見做し，以後 k をパラメータとする．

2.2 教師データ作成部

ここでは，言及度算出時に用いる教師あり分類器の教師データを機械的に作成する方法について説明する．採集対象とする tweet よりも過去に投稿された tweet 群を 2 種の教師なし分類器の入力とし，どちらの分類器においても判定結果が同一であった文書だけを用いて，これを教師データとする．本稿で用いる教師なし分類器は，それぞれ LDA, Twitter-LDA により文書ベクトルを得て，k-means によってクラスタリング後，災害を代表するような代表語の割合を参照して災害言及文書，災害非言及文書に分類する．

2.3 言及度算出部

言及度算出部では，採集対象とする会話木を入力として，会話木単位で災害に言及している度合である言及度を推定する．言及度の推定時には，2.2 節によって得られた教師データを元に学習した教師あり分類器を用いる．

言及度の算出（推定）は，会話 window 単位の文書集合 $D = \{d_1, \ldots, d_N\}$ を教師あり分類器への入力とし，文書 $d_i (d_i \in D)$ が災害言及判定の場合はスコア $s_i = 1$，災害非言及判定の場合はスコア $s_i = 0$ とする．このとき，会話 c の言及度 m_c は各文書 d_i のスコア s_i の平均値とし，次の式で求められる．

$$m_c = \frac{\sum_i^N \mathrm{s}_i}{N} \tag{1}$$

また，教師あり分類器として，ランダムフォレスト，ロジスティック回帰，ナイーブベイズを 3 種を弱分類器としたスタッキングによるアンサンブル学習器を用いる．

3 試作システム

提案手法に基づく採集システムを試作した．会話を含む tweet 集合の取得の為，Twiter 社が提供する StreamingAPI, LookupAPI を用いる．こうして取得した tweet 集合を時間ごとに区切り，時系列順に $\mathrm{slot}_1, \mathrm{slot}_2, \ldots$ とする．ここで，採集対象の tweet 群を $\mathrm{slot}_i (i > 1)$ としたとき，教師データ作成時に用いる tweet 群は slot_{i-1} とする．

Evaluation of Window on a Conversation Tree for Disaster Information Collecting

†Toshiyuki Fujita　‡Aki Kobayashi

†Electrical Engineering and Electronics, Kogakuin University Graduate School

‡Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

4 評価

会話 window におけるパラメータ k による言及度に対する効果を確認する．試作システム上で会話 window におけるパラメータ $k = 1$ の場合を提案手法とし，比較手法はパラメータ $k = 0$，つまり従来の 1 tweet 1 文書分類と等価である場合とし，1 文書とする tweet の範囲を reply 関係を辿ることで拡張することによる言及度への影響を確認する．

4.1 実験条件

2019 年 10 月に発生した台風 Hagibis が伊豆半島に上陸した付近の時刻 19:00:00 (JST) を slot_1 の始点時刻とする．

4.2 評価指標

人手による判定を用いて得られた各会話木の正解言及度 y と試作システムによって得られた推定言及度 $\hat{y}$ との RMSE 値を次のように求め，評価指標として用いる．このとき，評価対象とする slot から無作為に抽出した 100 個の会話木を対象として評価するため，n =100 である．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2}$$

4.3 結果

実験結果として，slot_1 における試作システムの正解言及度 y と推定言及度 $\hat{y}$ の分布を図 1, 2 にグラフを示す．横，縦軸はそれぞれ正解言及度 y, 推定言及度 $\hat{y}$ の階級を示しており，マス目内に示す値は度数を表している．各マスは度数の値とカラーバーに沿って色付けされている．このときの RMSE 値を表 1 に示す．

まず，表 1 から，$k = 0$，つまり 1 tweet 1 文書分類とした場合と比較し，$k = 1$ とした場合が RMSE の値からより正解の言及度に近いと言える結果となった．

次に，図 1, 2 から，各会話木の言及度 $\hat{y}$ が $k = 0$ のときから $k = 1$ でどのように変化したのか見ていくと，$k = 0$ における言及度 $\hat{y}$ が $0.2 \leq \hat{y} < 0.4$ かつ正解の言及度 y が $0.9 \leq y \leq 1.0$ に含まれる会話木の計 6 個は $k = 1$ では 4 つの会話木が $0.6 \leq \hat{y} < 0.7$ に，残りの 2 つの会話木が $0.0 \leq \hat{y} < 0.1$ となった．また，同じく正解の言及度 y が $0.9 \leq y \leq 1.0$ で $k = 0$ における言及度 $\hat{y}$ が $0.5 \leq \hat{y}$ の会話木は全て $k = 1$ ではグラフ右側にシフトし正解の言及度により近くなっていた．これは 1 文書となる tweet を広げることで同じような話題かつ 1 文書ごとの語が増えたことにより，1 tweet 1 文書時では会話木のうち災害言及として掬えなかった部分を災害言及として判定できるようになった場合があったことが要因の可能性として考えられる．また，いくつかの文書では $k = 0$ から $k = 1$ で正解の言及度から離れてしまった場合も見られた．これは，1 tweet 1 文書では災害言及として判定できていた部分を $k = 1$ とした場合では災害非言及として判定してしまったためである．

これらのことから，パラメータ k によって 1 文書とする tweet の範囲を広げる事で，全体として RMSE 値の低下に寄与することが見られたが，一方で単体で見ると RMSE 値が悪くなった会話木があることが分かった．これらの要因についての分析は今後の課題である．

表 1: Slot_1 における提案手法と比較手法の RMSE 値．

	提案手法（k=1）	比較手法（k=0）
slot_1	0.26	0.29

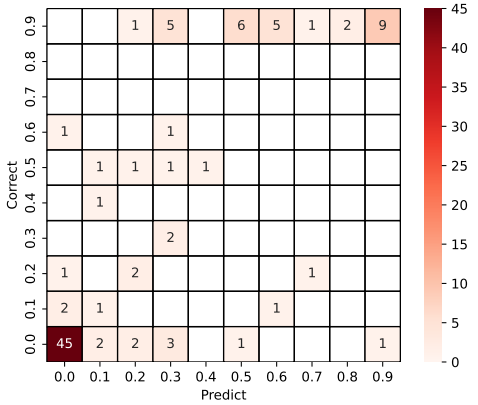

図 1: 比較手法における正解言及度 y と推定言及度 $\hat{y}$ における二次元ヒストグラム（パラメータ $k = 0$）．

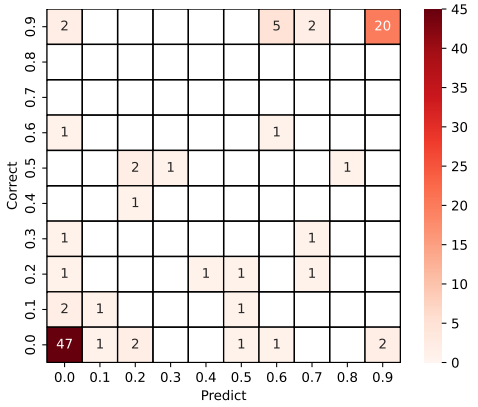

図 2: 提案手法における正解言及度 y と推定言及度 $\hat{y}$ における二次元ヒストグラム（パラメータ $k = 1$）．

5 おわりに

会話 window を導入した採集手法を提案し，試作システム上で会話 window のパラメータ k による効果として，1 tweet 1 文書と見做す場合と比較し分類精度の向上に寄与する部分があることを確認した．今後は，適切なパラメータ k の調査や会話木上の枝の分岐などを考慮することからより話題の転換を捉えることがことが考えられる．

参考文献

[1] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li, “Compareing twitter and traditional media using topic models,” Proc. of ECIR 2011, 2011.

[2] T. Fujita and A. Kobayashi, “Tweet Classification Using Conversational Relationships,” The 11th International Workshop on Advances in Networking and Computing (WANC’20), Nov.2020.