D-022

一日のトピック示す五七五:
ソーシャルビッグデータを用いた一日の出来事を表す五七五の生成

5-7-5 syllable structure indicating the topic of a day:
Generation of 5-7-5 syllable structure a day's topic using social big data

金子 卓矢* 横山 昌平*†‡
Takuya Kaneko Shohei Yokoyama

1 はじめに

新型コロナウイルス感染症（COVID-19）は 2019 年 12 月に中国の武漢市で感染者の第一例目が報告された後，世界的に流行しパンデミックとなった．そして，緊急事態宣言や 2020 年東京オリンピックの開催延期，ワクチン接種など多くの出来事が話題となった．コロナウイルスに関する出来事でも，日によって話題となるものは異なる．

そこで，ある一日の重要なトピックを SNS から見つけ出し短くまとめることにより，一目で一日の出来事を知ることが出来るようにする．そして，ただ短い文にまとめるのではなく，五七五とする事で印象に残りやすくなる．

Rafal らは Web の文書を資源とし，オリジナルのテンプレートに単語を当てはめることによって俳句の生成を行った [1]．Yael らは俳句コーパスを言語資源とし，連想語ネットワークを用いて俳句の単語を決定し，品詞 N-gram によって俳句の生成を行った [2].

本研究では Twitter の全量データ[1]を使用し，ツイートから単語をトピック分類し，コロナ禍における出来事を表す単語を見つけ出し，トピックごとに一日の特徴を示す五七五を生成した．

2 提案手法

五七五の生成はツイートから取得した単語を使用する．本手法の流れを図 1 に示す．

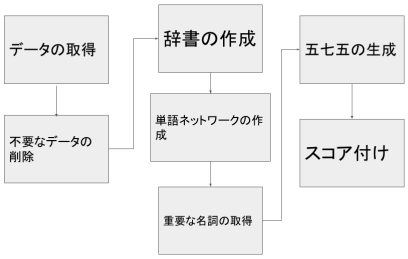

図 1 処理の流れ

2.1 単語ネットワークの作成

ツイートから単語がどのようにつながっているかというネットワークを作成する．ネットワーク作成の際，リツイートは同じツイートのデータのため除外する．また，アフィリエイトのツイートは同じような投稿をたくさんしており，出来事を表すツイートではないため除外する．本研究では，URL が含まれているツイートを取り除くことでアフィリエイトのツイートを除外している．

次に複数の名詞で一つの名詞を表す単語を扱うための辞書の作成を行う．ツイートの中から五七五となっている部分を見つけ，その中から名詞が続いている単語を辞書に追加する．

そして，形態素解析を行い，名詞と名詞の間にどのような単語が入るかというネットワークを作成する．形態素解析器に janome を使用し，作成した辞書を用いる．

2.2 特徴を表す名詞の抽出

一日の特徴を表す単語の抽出を行う．はじめに，トピックモデル [3] を用いて単語をトピック分類する．トピック分類は LDA(Latent Dirichlet Allocation)[4] を用いた．そして，term-score[5] を用いて，単語の特徴量を計算する．

term-score とは TF-IDF(term frequency inverse document frequency) の考え方をトピックに応用した指標である．あるトピックに偏って出現する単語がそのトピックで高い確率を占めるとき，高い指標となる．語彙 v のトピック k における term-score は式 (1) のようになる．

$$term - score_{v_k} = p(v_k) \cdot \log\left(\frac{p(v_k)}{(\Pi_{k=1}^{K} p(v_k))^{\frac{1}{K}}}\right) \quad (1)$$

ただし，$p(v_k)$ はトピック k における語彙 v の出現確率であり，K はトピックの総数である．

2.3 名詞の間に入る単語

それぞれのトピックの中から上位の名詞を取り出す．取り出した名詞の中から出現回数の多い組み合わせを取得する．

取得した名詞の間にどのような単語が入るか，単語のネットワークから出現数の多いものから当てはめ，五拍または七拍となる場合は五七五の候補とする．不自然な文を除外するため，累積確率を設定し，しきい値を超えない範囲で生成する．

2.4 スコア付け

生成した五七五を MLM(Masked Language Model)[6] を用いてスコア付けし，文として成り立っているかをランキングする．MLM のファインチューニングに使うデータは wikipedia の文から，五七五となっている部分を抜き出す．

抜き出したデータに対して transformers[2] を用いてファインチューニングを行う．ファインチューニングしたモデルを用いてスコア付けする．

3 結果と考察

日本時間の 2020 年 3 月 1 日から 2020 年 5 月 31 日のツイートの全量データから「コロナ，リモートワーク，

* 東京都立大学 Tokyo Metropolitan University
† 国立情報学研究所 National Institute of Informatics
‡ 東京大学 The University of Tokyo

1 株式会社 NTT データ「なずきのおと」を用いている

2 https://github.com/huggingface/transformers/blob/main/examples/tensorflow/language-modeling/run_mlm.py

オンライン授業，オリンピック」を含むツイートを取得し，いくつかの日付で五七五の生成を行った．

3.1 2020 年 3 月 30 日の五七五

2020 年 3 月 30 日のツイートを使用し，トピック数を 2 として五七五を生成した．

表 1 2020 年 3 月 30 日の五七五 (topic1)

五七五	スコア
病院は感染者全世界的	-3.960248498199244
病院で感染者全世界的	-4.553393180468407
国民のマスク着用医療器具	-5.100236789082183
病院も新型コロナウイルスか	-5.552575767387028
ウイルスの感染者数世界的	-5.724082908598636
病院の状況コロナ感染者	-5.90290106210689
感染者新型コロナウイルスか	-5.946444167325857
病院でコロナウィルスが世界的	-5.963350164021825
病院の新型コロナウイルスか	-6.020416926108325
国民の新型コロナウイルスか	-6.047607169111329

表 2 2020 年 3 月 30 日の五七五 (topic2)

五七五	スコア
コロナよりテレビ番組志村さん	-2.7522681977016923
最後までテレビ速報志村さん	-2.806518538914691
志村けんさんの追悼ニュースでし	-2.992971960078785
志村けんさんの悲しいニュースでし	-3.1642821091103457
最後までテレビ番組志村さん	-3.18147464518471
志村けんテレビ番組最後まで	-3.4920597938124196
学校はコロナ陽性最後まで	-3.5717031784263917
学校にテレビ番組志村さん	-3.7284460827547643
最後までコロナ肺炎志村さん	-3.8555224176331153
志村けんコロナ最新ニュースみる	-3.861168229458274

topic1 について，コロナウイルスの感染が拡大しているということについては，実際にこの時期は感染拡大をしているため適切である．しかし，病院での感染者が世界的というのは，実際の出来事と意味が異なってしまっている．

topic2 について，前日の 2020 年 3 月 29 日に志村けんさんがコロナウイルスにより死亡したため，2020 年 3 月 30 日の topic2 は適切であるといえる．しかし，七拍の部分が「さん」で始まっており，五七五の文としては不自然である．

3.2 2020 年 4 月 7 日の五七五

2020 年 4 月 7 日のツイートを使用し，トピック数を 2 として五七五を生成した．

表 3 2020 年 4 月 7 日の五七五 (topic1)

五七五	スコア
オンライン学校延期コロナ消え	-3.972329683144056
オンライン大学受験コロナ出る	-4.005633590300482
大学のコロナ大変気持ち落ち	-4.073017741843872
オンラインコロナ原因バイトクビ	-4.0773052441290405
学校でコロナ収まる休み方	-4.16387779233226
大学でコロナおさまる休みなし	-4.269922835031607
学校でコロナ収まる休みなし	-4.291267753723427
大学でコロナ収まる休み暇	-4.350272772048177
大学はコロナおさまる休みなし	-4.393809266327372
学校でコロナ収まる休み暇	-4.418937876958628

表 4 2020 年 4 月 7 日の五七五 (topic2)

五七五	スコア
病院にリモートワーク可能性	-1.5516351423859074
会社からリモートワーク可能性	-2.554544410133972
職場から感染者出る可能性	-3.7339571615555585
国民はリモートワーク可能性	-4.006212004283598
政府より感染者出る可能性	-4.225985340699472
病院に感染者出る可能性	-4.347327478312066
病院もリモートワーク可能性	-5.025027932960588
病院は感染者出る可能性	-5.058930461165782
感染者世界各国政府主義	-5.1101146618387485
国民がオリンピックの可能性	-5.251982808396921

topic1 は，1 回目の緊急事態宣言により，大学の授業などの開始が延期になったり，リモートワークが話題となったため適切な五七五であると考えられる．しかし，病院にリモートワークというのは話題となったわけではないため，一日の特徴を表すものとしては適切ではないと考えられる．

topic2 について，この日は感染が拡大しているが，生成された五七五は「コロナが収まる」や「コロナ消え」と実際の出来事とは逆の意味となっている．実際のツイートを確認すると，「コロナ収まるまで〜しない」や「コロナ消えてほしい」という内容のツイートがある．しかし，形態素解析を行うと「収まる」や「消え」で一つの単語として扱われる．そして，これらの単語を当てはめることによって五七五となるため，実際とは意味が異なる文となってしまう．

4 おわりに

本研究では，一日の特徴を表す五七五の生成を行い，新型コロナウイルスが流行し始めた頃についての特徴を調べた．ツイートから名詞を取得し，トピックを分類して五七五に使う単語を決定した．単語のネットワークから，名詞の間に入る単語を見つけることで，五七五の生成を行った．その結果，有名人がコロナに感染したということや，リモートワーク・オンライン授業が行われたことが生成された五七五で確認できた．

今後の課題として，現在は一日の特徴を表す五七五の候補として複数生成しているが，その中から一つを決定する．また，形態素解析を行い単語を当てはめると，実際の出来事とは異なる意味となってしまうことがあるため，意味が合う五七五を生成する．

参考文献

[1] Rzepka, Rafal, and Kenji Araki. "Haiku generator that reads blogs and illustrates them with sounds and images." Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), IJCAI, 2015.

[2] Yael Netzer et al. "Gaiku: Generating Haiku with Word Associations Norms". Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity, pp. 32 − 39, 2009.

[3] Hofmann, Thomas. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999.

[4] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[5] D.M.Blei, and J. D. Lafferty, "TOPIC MODELS" In A.Srivastava and M. Sahami, editors, Text Mining: Theoryand Applications. Taylor and Francis, 2009.

[6] Julian Salazar, Davis Liang, Toan Q. Nguyen, Katrin Kirchhoff. "Masked Language Model Scoring", ACL 2020.