# 自然言語処理によるフェイクニュース判断の自動化

発表者: 総合情報学科 メディア情報学 コース 学籍番号 1510151 柳 裕太
指導教員: 大須賀 昭彦 教授, 田原 康之 准教授, 清 雄一 准教授

## 1 背景

インターネット環境の浸透により、フェイクニュースの急速な拡散が大きな社会問題として近年取り沙汰されている。主に人々の扇動を目的に作られたフェイクニュースは、受け手に専門知識がない場合判断が難しい。そこで有識者が発信された情報の信頼性を判断・公表するファクトチェックが主に米国で盛んになっている状況である。しかしながら新たに発信される情報とファクトチェックには時間差があり、その間に拡散されてしまう状況となっている。そこで過去のファクトチェック結果による機械学習モデルを構築することで、新たに発信されるフェイクニュースに対応することを考えた。

今回の実験には、発言引用と真偽の要素を併せ持つデータセットが必要である。使用したのは、ファクトチェック機関 politifact.com のファクトチェック結果からなる LIAR dataset を使用した [1]。LIAR dataset 掲載の真偽データは politifact.com の評価方法”Truth-O-Meter”に準拠 [2] しており、真偽の度合いによって”True”, ”Mostly True”, ”Half True”, ”Mostly False”, ”False”, そして不正確かつばかばかしい主張を示す”Pants-on-Fire”までの 7 段階評価となっている [1]。

今回の実験では、自然言語処理によってどれだけ正確に信頼性を自動で判定することができるか調べたものとなっている。

## 2 課題

課題として挙げられるのは、以下の 2 点である。

- ファクトチェック機関によって蓄積された結果と、新型機械学習手法との親和性が不明である点
- 新たなニュースとファクトチェック結果が出るまで時間差がある点

特に後者では、ファクトチェック機関の調査が終わるまで少々の時間差があるため、ファクトチェック結果が公表されても浸透しないというケースが起こりやすい。そこで、その時間差を信頼性を自動算出・公表することで、フェイクニュースの拡散に足止めを行うことを考えた。

## 3 アプローチ

今回扱う課題を解決するためには、以下のアプローチが必要である。

- 最新の自然言語処理技術によってファクトチェックを自動化する
- 自動ファクトチェックシステムを SNS 上へ適用する

## 4 現在の進捗

前者のアプローチに関して、実際に予備実験を行った。全体の流れとしては、以下の図 1 の通りである。

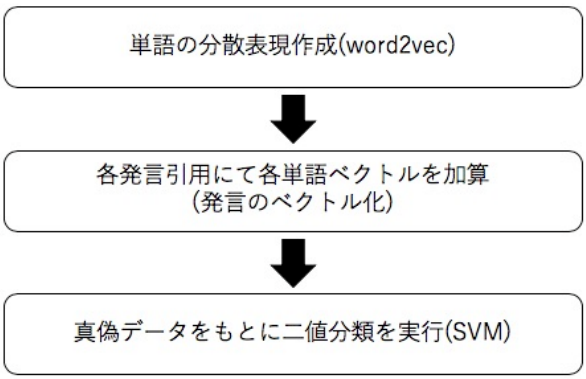


図 1 予備実験の流れ図

まずは全体データセットの発言引用を元に word2vec による単語の分散表現を作成する。その次に各発言内にて、発言がもつ各単語ベクトルを加算・正規化することで、発言のベクトル化を行う。最後に、各発言ベクトルと真偽を元に SVM で教師あり学習・検証を実行する。

LIAR データセットの真偽値に関しては、”Pants-on-Fire”〜”Mostly-False”を False、”Half-True”〜”True”を True として扱った。なお、その場合の全体件数は True は XXXX 件, False は XXXX 件だった。

5 分割交差検定を行った結果が以下の表 1 の通りである。Precision(適合率) は全体でのモデルの正答率、Recall(再

表 1　予備実験結果

| データ名 | 数値 |
|---|---|
| Precision | 0.723439 |
| Recall | 1.000000 |
| F1 score | 0.839530 |

現率) は True 全体内での True 看破率を表し、F1 score(F 値) は上記 2 値の調和平均を表す。

この結果より、word2vec+SVM の手法は良好な分類成績を示すことが判明した。

## 5　今後の展望

今後は、このフェイクニュース検出システムを実際の SNS 上で運用する場合を想定したモデル作りが必要である。そのために必要なものは、対象となる情報を選定することと、検証方法を考案する必要がある。

また、今回扱ったデータセットは全て英語・米国における情報だった。もしもこれを日本語で運用することを考えた場合、日本語で LIAR dataset とほぼ同じ情報形式を持つデータセットが必要である。そのためには日本における第三者機関によるファクトチェックの活発化が求められるが、残念ながらそのような運動があまりみられないのが現状である。

また，引用文献はキチンと入れましょう [**?**]．引用は，先人に対するリスペクトなので，よほど独立性が高い研究でない限り必要となります．

## 参考文献


[1] Wang, William Yang. *"liar, liar pants on fire": A new benchmark dataset for fake news detection*. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)(2017):422-426.

[2] Angie Drobnic Holan. *The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking*. https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/. (Sep. 18, 2018).