記事へのコメント生成によるフェイクニュースの早期検出

発表者: 情報学専攻 メディア情報学 プログラム 学籍番号 1930115 柳 裕太
指導教員: 田原 康之 准教授

1 はじめに

現代において、あらゆる情報の入手と拡散が簡単にできるソーシャルメディアは生活の重要な一部となった一方、悪意によって読者を騙して誤った風説を流布するために作られた情報であるフェイクニュースが問題となっている。

現在、フェイクニュース検出に向け有識者が事実関係を確認するファクトチェックが行われている。

図 1 はファクトチェックの一例である [1]。ファクトチェックは属人的な作業であることに加えて結果公表まで時間がかかるため、調査結果はフェイクニュースに比べ拡散されにくい。このため、機械学習によってフェイクニュースを自動で検出する研究が行われている。

検出性能の向上において、記事そのものがもつ情報に加えてソーシャルメディア上での反響を示すソーシャルコンテキスト (リツイート・いいね・リプライなど) を考慮することが有効であることが先行研究で示されている [2]。しかしながら、ソーシャルコンテキストはユーザの拡散によって生まれるため、この場合も早期の検出には向かない。

本研究では、記事と実際に寄せられたコメントから信憑性の学習を行い、記事と少量のコメントから別のコメントを生成してから真偽判断するモデルを提案する。このモデルは、フェイクニュースそのものを生成するモデル [3] を拡張する形で実装することでコメントの生成を実現する。

我々は提案モデルの検出性能を実際に投稿された情報をもつデータセットによって検証した。

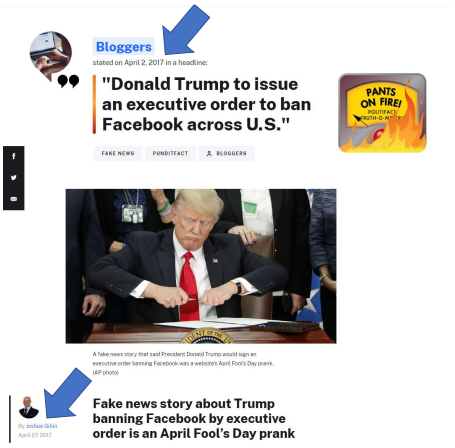

図 1 北米で行われたファクトチェックの一例。青矢印はフェイクニュース投稿日時とファクトチェック結果投稿日時を示し、両者には 25 日もの間が開いている。

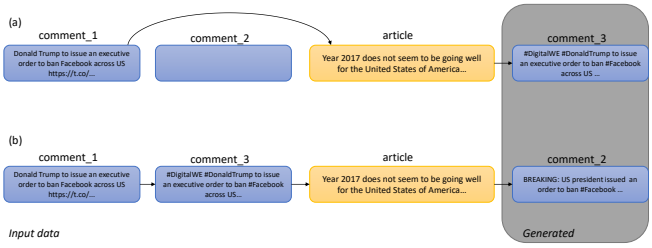

図 2 提案モデルのコメント生成例。(a) は記事と 1 件の実際に寄せられたコメントからコメントを生成している。(b) は (a) で生成したコメントを含めた状況で更にコメントを生成している。

2 手法

提案モデルによる文章生成の流れは図 2 の通りである。本研究ではベースとなった Grover モデル [3] に倣い、提案モデルは以下の同時分布として定義する。

$$p(\text{article}, \text{comment\_1}, \text{comment\_2}, \text{comment\_3}) \quad (1)$$

コメント生成学習時は、ベースとなった Grover モデルと同様に記事とコメントのセットを 2 つの集団に分け、無作為に要素を削除する。コメントの場合は 10%、記事本文の場合は 35% の確率で歯抜けにしてから一方の集団から学習を行い、もう一方での生成におけるクロスエントロピー誤差を最小化するように訓練される [3]。

記事とコメントのセットの末尾にはセットの終端を意味するトークンである [CLS] を追加し、またこのトークンが真偽を分類する際に使われる。これは Grover モデルがベースとしている GPT-2 がとる手法 [4] と同一である。図は実際の記事とコメントのセットを真偽分類するまでの流れである。まず、記事に寄せられたコメント群から実験に使用するために無作為に 3 件選出し、コメント生成学習する。真偽分類では、3 件の実際に投稿されたコメントから 1 件削除し生成されたコメントを追加し真偽分類する。

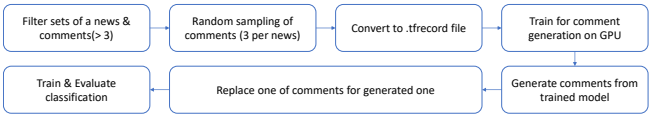

図 3　実験の流れ

表 1　分類成績

入力データ	適合率	再現率	F 値
記事本文のみ	0.647	0.615	0.631
+ 実在コメント 2 件	**0.682**	0.750	**0.714**
+ 生成コメント 1 件	0.590	**0.790**	0.675

3　実験

ニュース記事とコメントのセットは FakeNewsNet データセット [5] から取得した。このデータセットは北米で英文記事を対象にファクトチェックを行う団体である PolitiFact(政治ニュース中心) と GossipCop(芸能ニュース中心) の判断結果を元に正しいニュースとフェイクニュースのラベルが付けられている。我々はまず実験手法に合わせるために記事に対して最低 3 件以上コメントが寄せられているセットに対して無作為に 3 件選出した。生成コメントの有無によって真偽分類の結果に影響が出るか調べた。ベースラインとして 2 つの入力データを用意した。1 つは生成されたコメントを入力せず、記事と実際に投稿された 2 件のコメントから分類させた場合、もう 1 つは実際に投稿されたコメントも入力せず、記事のみから分類させた場合であった。この実験では、PolitiFact では十分な量の学習を行うにはセット数が少なかったため、GossipCop から真偽で各 2000 セットを用意して行った。

4　結果

実験結果は表 1 の通りである。提案モデルは再現率において全体ベストとなったものの、適合率においては生成モデルを使わない方が優秀であることが読み取れる。また、生成されたコメントには共通して文法面にさらなる改善の必要性が残された。

5　考察

表 1 より、提案モデルは再現率は優秀だったが適合率に大きな課題を残した。これはソーシャルコンテキストが制限されている状況でも、提案モデルにより多くのフェイクニュースの検出できることを意味する。

この傾向はファクトチェックが必要なニュースを探す際に役立つことを示唆している。ただし適合率が低く他のモデルより多く正しいニュースをフェイクニュースとして誤って検出するため、改善が求められている。今後は、より多くのデータセットを用いた場合に傾向が変化するか調べる必要がある。

6　結論

本研究では、フェイクニュースの早期発見における問題点の解決を試みた。我々は、ユーザのコメントはニュース記事を評価する際で重要な情報をもたらすものの、ニュース拡散の初期段階ではコメントが少ない点に着目する。そこで、Grover モデルを拡張したニューラルネットワークモデルを作成し、分類に有用なコメントを生成することを提案する。提案モデルのコメント生成による早期発見性能を評価するために、実際のニュースとそれに寄せられたコメントを対象に実験を行った。その結果、コメントを生成するプロセスが、ファクトチェックによって真偽を判定する際に役立つ可能性が示唆されている。

参考文献

[1] Joshua Gillin. Politifact - fake news story about trump banning facebook by executive order is an april fool's day prank, Apr 2017.

[2] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 943–951, New York, NY, USA, 2018. ACM.

[3] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 9054–9065. Curran Associates, Inc., 2019.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[5] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *ArXiv*, Vol. abs/1809.01286, , 2018.