検索エンジンによる上位検索ページを情報源とするフェイクニュース自動検出のためのデータセット作成*

尾崎 諒介† 前田 竜冶† 宇津呂 武仁‡ 村瀬 一之†
福井大学 工学部 知能システム工学科† 筑波大学 システム情報系 知能機能工学域‡

1 はじめに

Twitter や Facebook といった SNS の普及により大量の情報を簡単に入手できるようになった．その反面，その情報には不正確な記述や信憑性の薄い意見が混在しているため，事実かどうかを判断する必要性は高くなった．人を欺くために作為的に書かれたニュースをフェイクニュースと呼ぶ．近年，このフェイクニュースの拡散が社会問題となっている．実例の 1 つとして，2016 年の米大統領選挙において，Facebook 上でのフェイクニュース拡散がドナルド・トランプ氏の当選を助けたという非難が Facebook に殺到したという件があげられる．フェイクニュースが拡散する背景としては，インターネットのおかげで不特定多数のユーザーが任意の内容を簡単に世界に発信することができるようになり，真実と虚偽の境界が曖昧になってきていることが挙げられる．

この動向に伴って，機械学習を用いたフェイクニュース自動検出の研究が盛んになりつつある．しかし，未だフェイクニュース検出タスクのための実用的なデータセットが豊富にあるとは言いがたい．そこで本稿では，有用なデータセットの作成を目的として，人間が実際に行うフェイクニュース検出の過程を模倣して，検索エンジンを利用して収集した外部ソースからの情報を用いてフェイクニュース検出を行う方式により，フェイクニュース検出タスク用のデータセットを作成する手法を提案し，その手法を用いて実際に小規模データセットを作成した．

2 関連研究

2.1 含意関係認識

含意関係認識とは 2 つの文 Text，Hypothesis が与えられた時，Text から Hypothesis の意味を推論できるかどうかを判定するタスクである．このタスクは，機械翻訳，文書分類をはじめとする自然言語処理のさまざまな研究に広く共通している．

含意関係認識のためのデータセットとして，Stanford Natural Language Inference (SNLI) corpus [1] や，SNLI を基にさらに豊富な語彙や含意関係を加えた Multi-genre NLI (MultiNLI) [7] が挙げられる．SNLI データセットは前提文と仮説文の関係を人の手によって，含意，中立，矛盾の 3 種類のラベル付けがなされた約 57 万のデータからなる．このデータセットは，英含意関係の研究のベンチマークデータセットとして活用されており，自然言語処理技術の発展に貢献している．しかし，データ内の語彙に大きな偏りがありデータ作成時にバイアスを含んでしまっているため，前提文を無視し仮説文のみを用いたラベル推定が 66%の精度で可能である，という報告 [5] もされている．SNLI を用いた Chen らによるモデル [2] では，BiLSTM ベースのモデルに外部からの語彙知識を与えた学習を行うことで，精度 89.1%と人によるテスト精度 87.7% [3] を上回る結果を残している．

2.2 フェイクニュース検出

フェイクニュース検出タスクも，事実 (前提) とニュース記事 (仮説) の関係が正しいかどうかを判定するという意味で本質的には，含意関係認識のタスクと言える．ニュース記事に事実との矛盾があった場合，その記事はフェイクニュースとみなされる．しかし，ある事柄が事実かどうかを調べることは容易ではない．[4] ではファクトチェックの方法の 1 つとして，記事の主要となる主張について外部ソースと比較し真偽を判定する方法を挙げている．本研究においてはこの手法に基づいてのデータセット作成を行う．

既存のフェイクニュース検出のためのデータセットとしては，Fake News Challenge (FNC) データセット，LIAR データセット (4.1 節) [6] などがある．FNC データは FakeNewsChallenge[1] というフェイクニュー

*Developing a Data Set for Fake News Detection based on Search Results Web Pages

†Ryosuke Ozaki, Tatsuya Maeda, Kazuyuki Murase, School of Engineering, University of Fukui

‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba

[1] http://www.fakenewschallenge.org/

ス自動判別のための競技会で公開され使用されたデータである．このデータセットは，実際のニュース記事の見出しと本文間の関係を agree，disagree，discuss，unrelated の 4 値に分類されたデータで構成される．しかし，このデータセットは様々な記事の本文と見出しを組み合わせることによってデータを作成しているため，必然的に無関係ラベルの数が多くなっており，無関係ラベルデータが全体の約 7 割を占めるといったデータの不均衡問題がある．

この競技会での優勝チーム[2] は，1 次元の畳み込みニューラルネットワーク (CNN) と勾配ブースティング決定木 (GBDT) によるアンサンブル学習手法を取ることで，テスト精度 88.6%を誇っている．しかし，実際は無関係ラベルの正解率が 98%あるのに対し，agree ラベルは 44%，disagree ラベルは 7%とデータセットの偏りによる影響が見られる．このような点からも，フェイクニュース検出のためのより有用なデータセット作成が必要であると考えられる．

3 検索エンジンによる上位検索ページを情報源とするフェイクニュース自動検出の枠組み

本節では，フェイクニュース自動検出の枠組み (図 1) について述べる．人間がある事柄の真偽を判定するとき，インターネットの検索エンジンを用いてその事柄について検索し，検索結果のそれぞれの主張を読み理解した上で最終的にその事柄がフェイクニュースなのかそうでないのかの判断をする．このように人間は検索エンジンを用いた検索結果を情報源として事柄の真偽を判定する．そこで，本論文では，フェイクニュース自動検出の枠組みとして，人間によるフェイクニュース判定の手順を模倣する方式を採用する．また，人間がインターネットの検索エンジンを用いてある事柄の真偽判断する場合，検索クエリはある事柄そのままの文にすることが多い．そのことを前提に，本論文では，真偽判断の対象文と検索クエリは同一文としてそれをステートメントと呼ぶ．本論文におけるフェイクニュース自動検出の枠組みの概要を以下に示す．

1. ステートメントを検索エンジンで検索する．
2. 検索結果ページ一つ一つの本文の主張とステートメント間の含意関係を True，False，関連するが判定不可，無関係の 4 値に自動分類する．
3. それぞれの分類をもとに最終的にステートメントがフェイクニュースかそうでないか判断をする．

[2] https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html

FNC データでは記事の見出しと本文間の関係を agree，disagree，discuss，無関係の 4 値に分類したことに対して，本論文では，ステートメントと検索結果ページ間の関係は，以下の True，False，関連するが判定不可，無関係の 4 値に分類する．

- True — 賛成または同じ主張
- False — 反対または偽だと主張
- 関連するが判定不可 — 関連記事であるが真偽判断は不可
- 無関係 — 無関係な記事

4 フェイクニュース自動検出システムのためのデータセット作成

本節ではまず，データセット作成において必要となった PolitiFact サイトについて述べ，次にフェイクニュース自動検出システムに用いることを前提としたデータセット作成手順について述べる．

4.1 PolitiFact

PolitiFact[3] は主にアメリカの政治にまつわる発言や事柄 (ステートメント) についての信憑性の事実確認を行うサイトである．検証対象となる発言を転記し，ジャーナリストによって True，Mostly True，Half True，Barely True，False，Pants on Fire の 6 段階スコアによってその内容の評価を行っている．PolitiFact を基にフェイクニュース自動検出のために作成されたデータが LIAR データセット (表 1) [6] である．LIAR データは PolitiFact のステートメントとそのラベルで構成されている．その他に Speaker，Context の情報が付与されている．Label は PolitiFact に基づく 6 値である．

本研究では，このデータセットから True と Pants on Fire のラベルがついたステートメントをいくつか取り出し，外部ソースからの判断情報を付与することで新たなデータセットに用いる．

4.2 データセット作成手順

図 1 のフェイクニュース自動検出システムの枠組みに沿って，以下の手順でデータセットの作成を行う．(1) LIAR データのステートメントを検索クエリとして Google 検索機でそのまま検索する，(2) 検索結果の上位 20 個のサイトの URL を Web スクレイピングをすることにより自動収集する，(3) 取得した URL を用いてその記事に書いてあるタイトルと本文を自動抽出する，(4) ステートメントと記事から抽出した内容つまりステートメント・検索結果ページ間の含意関係のラベルを決定する．

[3] http://www.PolitiFact.com/

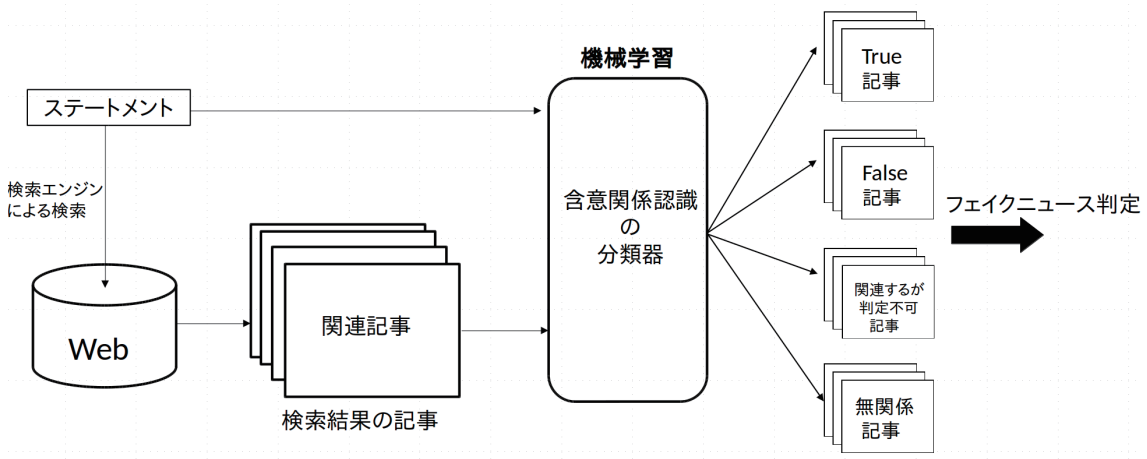

図 1: フェイクニュース自動検出システム

表 1: LIAR データとラベル説明

Statement	Speaker	Content	Label	ラベル説明
During his tenure as mayor, he saw Houwtons crime rates drop to the lowest levels in more than 25 years.	Bill White	the biography on his compaign Web site	True	正確な情報である
If you got rid of the income tax today you'd have about as much revenue as we had 10 years ago.	Ron Paul	interview	Mostly True	正確ではあるが，追加情報が必要
There are mechanism in place to shut down the Internet privately	Andrew Napolitano	comments on the 'Glenn Beck' show	Half True	部分的には正確であるが重大な情報が不足している
The American people support defunding Obamacare and oppose shutting down the government.	Marco Rubio	a press release	Barely False	若干正確ではあるが，重大な情報を無視している
74 percent of Republicants support the Affordable Care Act.	donna Edwards	a round table discussion on CNN's "State of the Union"	False	不正確な情報である
Photo of "tea party" protests shows crowd sprawling from Capitol to Washington Monument.	Blog Posting	consercative blogs	Pants on Fire	不正確なだけでなく，馬鹿げている

5 ステートメントと検索結果ページ間の含意関係人手付与及び含意関係に基づくフェイクニュース判定

本節では 4.2 節のデータセット作成手順「(4) ステートメント・検索結果ページ間の含意関係のラベルの決定方法」について説明し，次にそのラベルに基づいたフェイクニュース判定の説明をする．

5.1 手順

ステートメントと検索結果ページ間の含意関係付与 (ラベル付け) の決定は人手によって行う．ラベル付け方法は検索された URL の記事にリンクしその内容を理解して，ステートメントとの含意関係を True，False，関連するが判定不可，無関係の 4 値に分類する．この作業を 1 ステートメントにつき 20 記事行い，全部で 40 ステートメントつまり 800 記事分を行う． 1 人だけの判断ではバイアスがかかるので，英語を得意とする計 4 人に同様の作業をしてもらい 4 人の分類判断の多数決でステートメント・検索結果ページ間のラベルを決定する．4 人の分類判断が分かれてしまった場合はラベルを [同数] とする．また Wikipedia，PDF，Google ブックス，動画サイトなどの記事で無いものはすべて無関係に分類する． このようにしてステートメントと検索結果ページ間のラベルを多数決で 800 個決定し，それを作業者間多数決ラベルとする．

また，1 つのステートメントがフェイクニュースであるかどうかの判定は集められた検索結果上位 20 個の作業者間多数決ラベルの True と False の多数決によって決定する．1 つのステートメントに対しての 20 個の作業者間多数決ラベルの内，True か False の多い方をそのステートメント自体のラベルとし，検索結果ページ間多数決ラベルと呼ぶ．同数の場合は検索結果

表 2: PolitiFact ラベル・検索結果ページ間多数決ラベルの組ごとの作業者間多数決ラベルの統計

PolitiFact ラベル	検索結果ページ間多数決ラベル	ステートメント数	作業者間多数決ラベル				
			True	False	関連するが判定不可	無関係	同数
True	True	16	178	18	20	83	21
False	False	19	25	169	26	137	23
True	False	1	2	9	2	4	3
False	True	1	8	6	2	0	4
False	True	1	4	2	1	11	2
False	同数	1	3	3	2	11	1
False	同数	1	5	5	2	7	1
合計		40	225	212	55	253	55

ページ間多数決ラベルを [同数] とする．検索結果ページ間多数決ラベルが True の場合ステートメントは事実，False の場合はフェイクニュースとする．このようにして 40 ステートメントすべての検索結果ページ間多数決ラベルを決定する．

5.2 考察および分析

ステートメントと検索結果上位 20 ページを元に人手によってラベル付けされた検索結果ページ間多数決ラベルの精度は PolitiFact ラベル (True か Pants on fire) との一致具合で評価する．40 ステートメントの検索結果ページ間多数決ラベルの内，2 ステートメントが [同数] なので，それらを差し引いた 38 ステートメントの検索結果ページ間多数決ラベルの内，35 ステートメントが PolitiFact ラベルと一致した．検索エンジンによる上位検索ページを情報源とステートメントの真偽判定をすることは高い精度で PolitiFact ラベルの実現ができることがわかった．表 2 は PolitiFact ラベルと検索結果ページ間多数決ラベルが一致の場合と不一致の場合に分けそれぞれのステートメント数と作業者多数決ラベルの内訳を示している．PolitiFact ラベルの Pants on Fire は検索結果ページ間多数決ラベルの False と同じ意味とする．

PolitiFact ラベルと検索結果ページ間多数決ラベルが一致しなかったステートメントは 3 個，検索結果ページ間多数決ラベルが [同数] であったステートメントが 2 個で，計 5 個のステートメントが PolitiFact ラベルの再現ができなかった．これらのステートメントが PolitiFact ラベルを再現できなかった理由は大きく分けて 2 つある．1 つ目は検索された結果ページにフェイクニュースの記事のほうが多く存在してしまう場合で，実際に

> Tens of thousands of fraudulent Clinton votes found in Ohio warehouse.

のステートメントはフェイクニュースであるが，検索結果においては，クリントンを陥れる意図のある記事や，フェイクニュースを信じた人によって書かれた記事が多く存在した．　2 つ目は，2 つの立場から真偽を判定でき，1 つの立場では真であり，もうひとつの立場では偽である場合で，

> America has already taken in one-fourth of Mexico's entire population.

ではアメリカに移民したメキシコ人の人数は 4 分の 1 に満たないが，現在アメリカに住んでいるメキシコ系アメリカ人を含めるとメキシコの人口の 4 分の 1 に匹敵するという 2 つの立場があるので作業者によって分類が分かれる．これらの理由から PolitiFact ラベルの実現ができなかった．

6 おわりに

本研究では人間が実際に行うフェイクニュース検出の過程を模倣した手法によるフェイクニュース自動検出タスク用のデータセットを作成する手法を提案し，その手法を用いて実際に小規模データセットの作成及びその評価を行った．データセットの作成方法において検索エンジンによる上位検索ページを情報源としてステートメントの真偽判定をすることにより，高い精度で PolitiFact ラベルの実現ができることがわかった．今後は，この手順のもとでフェイクニュース自動検出のための大規模なデータセット作成を行う．

参考文献

[1] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proc. 20th EMNLP*, pp. 632–642, 2015.

[2] Q. Chen, X. Zhu, Z. H. Ling, D. Inkpen, and S. Wei. Natural language inference with external knowledge. Vol. abs/1711.04289, 2017.

[3] Y. Gong, H. Luo, and J. Zhang. Natural language inference over interaction space. Vol. abs/1709.04348, 2017.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Exploration Newsletter*, Vol. 19, No. 1, pp. 22–36, 2017.

[5] 土屋雅稔. 含意関係認識コーパスの偏りによる性能評価への影響. Vol. 2017–NL–233, No. 5, pp. 1–6, 2017.

[6] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proc. 55th ACL*, pp. 422–426, 2017.

[7] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. Vol. abs/1704.05426, 2017.