事実検証モデルのための ラウンドトリップ翻訳を利用した疑似フェイクデータ生成

小林龍斗　　秋葉友良
豊橋技術科学大学 情報・知能工学課程
kobayashi.ryuto.jz@tut.jp　akiba@cs.tut.ac.jp

概要

フェイクニュースの拡散は，誰もが情報を共有できるソーシャルメディアが普及している昨今において重大な問題となっている．この問題に取り組むため，NTCIR プロジェクトにおける議会議事録を対象とした評価タスク QA Lab-Poliinfo-3[1]では，Fact Verification タスクが実施された．しかし我々は，同タスクで提供された学習データを使用して構築した分類器では，人手で作成されたフェイクデータを上手く検出できないことを実験により確認した．本研究では，この一つの要因であるフェイク学習データの不足に着目し，Round-Trip 翻訳を利用した疑似フェイクデータ拡張手法を提案する．評価実験により，提案手法で自動生成されたフェイクデータを学習に利用することで，人手フェイクデータに対する検出精度を向上させられることを確認した．

1 はじめに

近年，フェイクニュースやデマの拡散が社会問題になっている．ソーシャルメディア等を通じて拡散された真偽不明の情報は人々に誤解を与え，混乱を招く恐れがある．このような問題に取り組むため，NTCIR-16 QA Lab-Poliinfo-3 では，Fact Verification タスクが実施された．Fact Verification タスクでは，与えられた討論の要約が実際に議会で話し合われている内容と即しているかどうかを検証するモデルの精度が競われた．ここで我藤ら[1] は，パッセージ検索と含意関係認識モデルを組み合わせた手法を提案し，参加チーム中トップの成績を達成した．

しかし我々は，我藤らのモデルが人手で作成された巧妙なフェイクデータを上手く検出できないことを実験によって確認した．本研究では，この一つの要因と考えられるフェイクの学習データの不足に着目し，ラウンドトリップ翻訳を利用した疑似フェイクデータ生成手法を提案する．実験の結果，提案手法で生成した疑似フェイクデータをモデルの学習に利用することで，人手フェイクデータの検出精度を向上させられることが確認できた．

2 関連研究

事実検証の方法には様々な手法が提案されている．

我藤ら[1]は，主張文中の語句をクエリとして情報源から要約に関連する文を検索し，主張文と検索文との間に含意関係が成立するかどうかを判定することで，フェイクの検出を試みた．含意関係とは，ある主張 A が成立するとき，ある主張 B もまた成立する関係を示し，このとき主張 A は主張 B を含意していると言う．我藤らは，主張文が正しければ，それを含意する検索文との間に含意関係が成立すると仮定し，含意関係が成立するかどうかを判定するモデルを事前学習言語モデル BERT[2]を利用して構築した．本研究は我藤らの研究に基づいている．

また，Jawahar ら[3]は知識ベースを利用した事実検証方法を提案した．知識ベースとは，知識をコンピュータで扱えるよう形式化した特殊なデータベースを指し，Jawahar らの研究では，単語と単語が関係で接続された三連結のデータを集めた YAGO[4]を利用している．Jawahar らは知識ベースに含まれる知識を基に有向グラフを構築し，それをグラフ畳み込みニューラルネットワークで符号化することで，事実検証モデルに知識ベースを組み込んだ．また，ここではエンティティ操作によるフェイクデータ生成手法も提案されている．通常のエンティティ操作は，エンティティを無作為に置き換えることで達成されるが，ここではテキスト生成モデル GPT-2[5]を利用

[1] https://poliinfo3.github.io/

した手法が提案されており，置き換え元の語と関連性の高い語を予測して置換することで，検出難易度の高いフェイクを作成している．

3 提案手法

提案手法では，正しい主張文に特定の操作を自動的に加えることにより，フェイクデータを作成する．

3.1 文の操作

提案手法では，主張文に対して以下の操作を加えることによって，フェイクデータの生成を試みる．

- 否定の挿入・削除
  「A は B が好きだ」→「A は B が好きでない」
- 対義語への変換
  「A は B が好きだ」→「A は B が嫌いだ」
- 主語－目的語の交換
  「A は B が好きだ」→「B は A が好きだ」

3.2 Direct Manipulation

一つ目の手法として，主張文に対して何らかの操作を直接的に加える Direct Manipulation（以降，DM）を提案する．操作の模様を図 1 に示す．尚，DM では 3.1 節に示した 3 つの操作の内，否定の挿入・削除のみを実装する．

この手法は単純であるが，主張文を強引に書き換えることで不自然な文が生成される恐れがある．

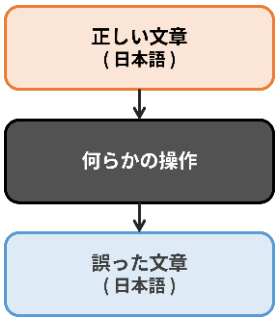

図 1　Direct Manipulation

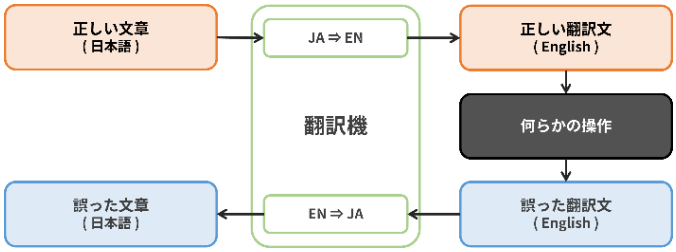

図 2　Round-Trip Manipulation

3.3 Round-Trip Manipulation

DM に対して，主張文を直接操作するのではなく，Round-Trip 翻訳を介して文を操作する手法を提案する．本稿では，Round-Trip Manipulation（以降，RTM）と呼ぶ．操作の模様を図 2 に示す．尚，RTM では 3.1 節に示した 3 つの操作を全て実装する．

RTM では，はじめに日本語で与えられる主張文を英語などの言語に翻訳し，機械翻訳によって得られた英語の翻訳文に対して何らかの操作を加え，それを再び翻訳することによって日本語のフェイクデータを生成する．翻訳機を介することにより，違和感のある文章の改変が緩和され，自然な文章が得られることを期待している．

4 評価実験

提案手法の有効性を検証するため，以下の評価実験を行う．

4.1 データセット

本実験では，昨年開催された NTCIR-16 QA Lab-PoliInfo-3 Fact Verification タスクにて配布された Formal Run の学習データ，テストデータを使用してデータセットを作成する．配布データのフォーマットを表 1 に示す．

データ数は 1,433 件あり，本実験ではここから Utterance Type が回答，Document Entailment が True のデータ 411 件を抽出して利用する．さらに，テストデータ作成に 60 件のデータを確保し，残りの 351 件を学習データ作成に利用する．

表 1　配布データのフォーマット

属性	データの内容
ID	データを一意に識別する番号
Prefecture	会議の開催地
Date	会議の開催年月日
Meeting	会議の識別情報
Speaker	発言者
UtteranceType	発言の種類（質問／回答のどちらか）
UtteranceSummary	発言の要約
ContextSummary	発言前後の対話全体の要約
ContextWord	発言が対象としている主題
RelatedUtteranceSummary	発言に関連する発言の要約
StartingLine	要約開始行
EndingLine	要約終了行
DocumentEntailment	要約の真偽

4.1.1 テストデータ

テストデータは人手で作成する．我々は，4 名の被験者の協力で人手フェイクデータを収集した．作成にはテストデータ作成用に確保した 60 件のデータを用い，要約文を何らかの形で書き換えることでフェイクデータを作成するよう被験者らに依頼した．結果，60 件の True データに対して 60 件の人手作成 False データを収集することができた．テストデータの例を付録の表　に添付する．テストデータは人手作成フェイクデータ 60 件，その参考となったテストデータ作成用のデータ 60 件を合わせた 120 件とする．

4.1.2 学習データ

学習データは提案手法によって自動的に作成する．作成には学習データ作成用に確保した 351 件のデータを利用する．

DM では，否定の挿入・削除操作（NEG）によってフェイクデータを生成する．実装には形態素解析ツール MeCab[2]を利用し，解析された動詞と活用の種類に応じて否定の挿入・削除を行う．

RTM では，否定の挿入・削除（NEG），対義語への変換（ANT），主語－目的語の交換操作（SOE）によってフェイクデータを生成する．実装には自然言語処理ツール Stanza[3]を利用し，翻訳機には DeepL API[4]を利用する．また，単純な Round-Trip 翻訳によって True の擬似データも同時に作成する．これは，フェイクデータのみを拡張することによる偏りを抑えることを目的としている．

以上の手法によって作成されたデータ例と件数を付録の表 4 に添付する．

4.2 実験ベースライン

本実験のベースラインとなるモデルを先行研究[1]に従って構築する．先行研究のモデルはパッセージ検索と事前学習言語モデルを利用した含意関係認識によって構築されている．

パッセージ検索では，文書ランク付け手法である BM25+[5]を利用する．クエリQに対する文書Dの BM25+スコアは次の計算式で表現される．パラメータはそれぞれ$k_1 = 1.2, b = 0.75, \sigma = 1.0$ とする．

$$\sum_{i=1}^{n} IDF(q_i) \cdot \left[\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} + \delta \right] \quad (1)$$

$$IDF(q_i) = \log \frac{N}{n(q_i)} \quad (2)$$

含意関係認識モデルには，東北大学の乾研究室が公開している BERT-base[5]モデルを利用する．このモデルに対して NTCIR-16 QA Lab-PoliInfo-3 Fact Verification タスクにて配布された Formal Run の学習データ 1023 件を利用し，含意関係認識モデルのファインチューニングを行う．先行研究のモデルの模式図を付録の図 3 に添付する．

実験ベースラインには，上記ベースラインモデル（以降，Baseline）に加え，120 件のテストデータで交差検証を行ったモデル（以降，Supervised）を設定する．テストデータ数が 120 件と少ないため，交差検証では，テストデータとして 1 件を残し，残り 119 件を学習データとして学習と推論を 120 回繰り返す leave-one-out 法を採用する．

4.3 評価指標

評価指標はテストデータ 120 件に対する分類精度である Accuracy，フェイク検出の Precision, Recall, F1 とする．

4.4 実験方法

実験では，4.2 節に従って作成される実験ベースラインモデルに対して，提案手法で自動生成された学習データを再学習させることで，テストデータ 120 件に対する分類精度が向上するかどうかを確かめる．

4.5 実験結果

次頁の表 2 に実験結果を示す．

4.5.1 ベースラインとの比較

Baseline の Recall を見ると，ベースラインのモデルが人手で作成されたフェイクデータを殆ど見分けられていないことが分かる．これに対し，他の殆どのモデルでは Recall が向上しており，適切な学習データを利用することは人手作成フェイクデータの判定に効果的だと言える．また，人手作成データを再

[2] https://taku910.github.io/mecab/
[3] https://stanfordnlp.github.io/stanza/
[4] https://www.deepl.com/pro-api?cta=header-pro-api
[5] https://huggingface.co/cl-tohoku

表 2　評価実験の結果

Training Data		Pre.	Rec.	F1	Acc.
Baseline		0.800	0.067	0.123	0.535
Supervised		0.793	0.767	0.780	0.783
DM	ORG ／ NEG	1.000	0.600	0.750	0.800
RTM	ORG ／ NEG	0.927	0.850	**0.887**	**0.892**
	RTT ／ NEG	0.957	0.633	0.762	0.808
	ORG ／ NEG + ANT	0.976	0.683	0.804	0.867
	ORG ／ NEG + SOE	1.000	0.383	0.554	0.808
	ORG + RTT ／ NEG + ANT	0.961	0.817	**0.883**	**0.892**
	ORG + RTT ／ NEG + SOE	0.867	0.767	0.820	0.825
	ORG + RTT ／ NEG + SOE + ANT	0.960	0.800	0.873	**0.883**

学習に用いた Supervised モデルと提案手法を採用したモデルでは，提案手法を採用したモデルの方が高い精度で分類することができた．Supervised モデルに再学習させられたフェイクデータは 60 件と数が少なかったのに対し，提案手法のモデルではデータ拡張によってフェイクデータを 300 件前後学習させることができた．この結果から，擬似データであっても，多量のデータを自動生成することは効果的だと言える．

4.5.2 提案手法の比較

二つの提案手法 RTM，DM を比較すると，擬似フェイクデータの作成手法としては Round-Trip 翻訳を利用した方が効果的だと言える．DM であまり精度が出なかった要因として，操作前と操作後の文の表現が殆ど変わらないことから，再学習の段階でモデルを混乱させているためだと考察する．

4.5.1 学習データの比較

RTM を利用したモデルの中では，F1 の観点において，ORG/NEG と ORG+RTT/NEG+ANT を学習させたモデルの精度が高かった．表全体を見ると，NEG, ANT を学習させたモデルの精度が高い傾向にあり，逆に，SOE は今回のテストデータに対してはあまり効果が無かったことが読み取れる．これはテストデータに主語ー目的語を交換することによって生成されたデータが殆どなかったためだと考えられる（表 3）．これに対し，否定の挿入・削除および対義語への変換を行ったフェイクは多く，これは NEG および ANT 学習させたモデルの精度が向上した一つの原因と考えられる．また，ORG/NEG+ANT，ORG/NEG+SOE を学習させたモデルと，これに RTT を加えた ORG+RTT/NEG+ANT，ORG+RTT/NEG+ANT を学習させたモデルを比較すると，RTT を加えたモデルの方が高い性能を達成した．この結果から，フェイクデータのみを増強するのではなく，True のデータも増強していく方が効果的だと考えられる．

表 3　テストデータの分類

分類	件数
否定の挿入・削除	25
対義語への変換	18
主語－目的語交換	2
数値の変更	4
単語の変更	8
その他	3

5 おわりに

本研究では，先行研究[2]の事実検証モデルの問題点に着目し，それを解決するための擬似フェイクデータ拡張アプローチを提案した．評価実験の結果，提案手法を利用して自動生成したフェイクデータをモデルの学習に利用することで，人手で作成されたフェイクデータに対する検出精度を大きく向上させることができた．本研究では，英語をピボット言語として日本語のデータを生成したが，提案手法は他言語のデータ拡張にも適用できる、

今後の課題として，提案手法は True データが存在することを前提としているため，True データに依存しないフェイク作成手法も検討したい．

謝辞

本研究の遂行に際して，4 名の被験者に協力いただいた．

参考文献

[1] 我藤勇樹，秋葉友良，パッセージ検索と含意関係認識による議会議事録を対象としたファクトチェック，言語処理学会 第 28 回年次大会 発表論文集，2022 年 3 月，pages 768-772,
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171-4186.
[3] Ganesh Jawahar, Muhammad Abdul-Mageed, Laks V. S. Lakshmanan, Automatic Detection of Entity-Manipulated Text Using Factual Knowledge, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, May 2022, pages 86-93.
[4] Thomas Pellissier Tanon, Gerhard Weikum, and Fabian M. Suchanek, YAGO 4: A reasonable knowledge base. In The Semantic Web, volume 12123 of Lecture Notes in Computer Science, May 2020, pages 583–596.
[5] Yuanhua Lv, ChengXiang Zhai, Lower-bounding term frequency normalization, In Proceedings of CIKM’2011, pages 7-16.

A 付録

表 4　データの作成件数と作成データの例

生成手法		件数	データ例	True/False
原文（ORG）		351	不安定な就労は本人にも社会にも不幸な事態。国は有効な対策打ち出していない。	True
RTT		351	不安定な雇用は、個人にとっても社会にとっても不幸な状況である。政府は有効な対策を打ち出せていない。	True
DM	NEG	310	不安定な就労は本人にも社会にも不幸な事態。国は有効な対策打ち出して**いる**。	False
RTM	NEG	337	不安定な雇用は、個人にとっても社会にとっても不幸な状況では**ない**。政府は効果的な対策を打ち出して**います**。	False
	ANT	252	不安定な雇用は、個人にとっても社会にとっても**幸運な**状況である。政府は**効果のない**対策は打ち出していない。	False
	SOE	205	不安定な状況は、個人と社会の両方にとって不幸な雇用である。その**対策**は、効果的な**政府**には至っていない。	False

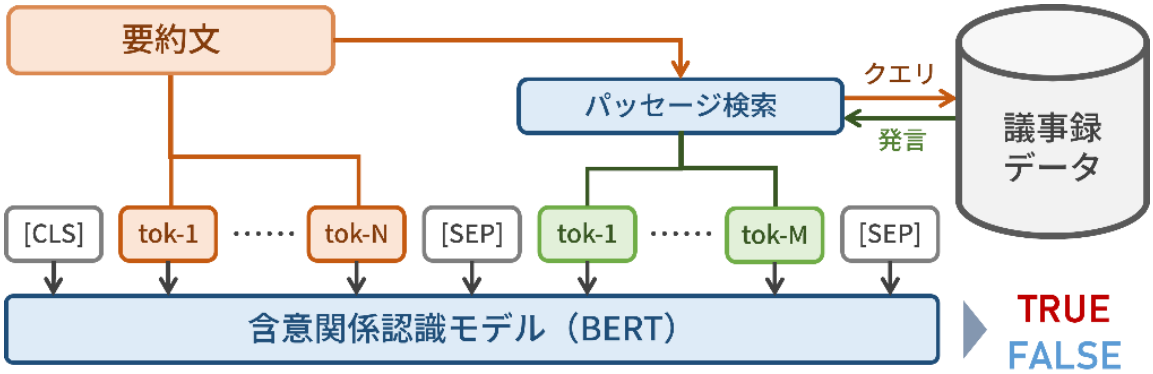

図 3　我藤らが提案した事実検証モデル