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

我藤勇樹　　秋葉友良
豊橋技術科学大学 情報・知能工学課程
gato.yuki.am@tut.jp　akiba@cs.tut.ac.jp

概要

ソーシャルメディアなどで真偽不明の情報が拡散されることが増加しているため、真偽不明の情報を検証することの必要性が高まっている。本研究では、NTCIR-16 QA Lab-Poliinfo-3のFact Verificationタスクに則り、議会議事録を対象としたファクトチェックを行う。パッセージ検索と含意関係認識を用いることで、NTCIR-16のFormal Runで参加チーム中トップの成績を達成した。

1 はじめに

近年、デマの拡散がオンライン上で加速していることが報告されている。真偽不明の情報が広がることにより、社会に深刻な影響を及ぼす可能性がある。このような状況に対して、NTCIR-16 QA Lab-PoliInfo-3[1]では、Fact Verificationタスクが開催された。このタスクでは、与えられた要約の事実性を議会議事録で実際に議論されているかで判定する。また、事実であると判定した場合、その根拠となる発言部分を議会議事録から特定する。要約を議会議事録の一部に関する主張と捉えると、議会議事録を対象としたファクトチェックと考えることができる。

本研究では、Fact Verificationタスクに則り、議会議事録を対象としたファクトチェックを行う。まず、検索手法により、議会議事録の中から主張に関連するパッセージを特定する。そして、主張と検索したパッセージを用いて、含意関係認識により主張の真偽を判定する。また、主張が真である場合、根拠文として主張に関連する箇所を議会議事録から抽出する。本手法を用いてFact Verificationタスクに結果を提出したところ、参加チーム中でトップの成績を達成した。

2 関連研究

現在、ファクトチェックの多くは信頼できるファクトチェック機関により人手で行われている。しかし、誤った情報や偽の情報は、正しい情報よりも拡散されやすいという性質がある。そのため、情報の真偽の検証には早さが求められている。

Hansenらは、フェイクニュースや都市伝説を対象とした自動ファクトチェックについて調査している[1]。真偽不明の主張に対して、Webを外部知識としてGoogle検索で根拠を取得し、古典的な機械学習モデルや事前学習モデルを用いてファクトチェックを行なっている。

また、Wikipediaをドメインとしたファクトチェックのワークショップ[2]も開催されている。このワークショップでは、Wikipediaの記事に関する主張について、Wikipedia記事を根拠にファクトチェックを行う。結果として、事前学習モデルを用いたファクトチェックモデルが優れた結果を達成している。

3 提案手法

本研究では、パッセージ検索と含意関係認識を用いてファクトチェックを行う。まず、パッセージ検索により特定のパッセージを検索する。そして、主張と検索したパッセージを用いて含意関係認識を行う。主張が真である場合、議会議事録から根拠を抽出する。

3.1 パッセージ検索

主張が真である場合、その主張は議事録の特定のトピックについて述べたものである。そのため、主張の真偽を判定するためには、そのトピックの箇所を議会議事録から特定する必要がある。

[1] https://poliinfo3.net

本研究では、検索するパッセージの単位をセグメント単位と文単位とする。セグメント単位では、まず、表1に示す正規表現を用いて、議会議事録の中でトピックが切り替わる箇所を特定する。具体的には、開始表現に該当する文と終了表現に該当する次の文をセグメントの開始文とする。その後、各セグメントに対して主張との類似度を計算し、最も類似度の高いセグメントをパッセージとする。この際、議会議事録上で連続するセグメントは結合したものもセグメントと呼び、セグメントの連続も検索対象とした。文単位では、議会議事録の各文に対して主張との類似度を計算し、類似度の上位n文をパッセージとする。

主張とパッセージの類似度検索における計算には、BM25+を用いる方法とsentence-BERTによるembeddingのコサイン類似度を使う方法の2手法を試した。

BM25+[3]は、クエリを含む長い文書より、クエリを含まない短い文書の方がスコアが高くなるというBM25の問題点を解決したランキング関数である。BM25+の式は以下のようになる。

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

$$IDF(q_i) = \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \quad (2)$$

sentence-BERT(以下、SBERT)は、事前学習済みBERTを文書ベクトル化用にファインチューニングし、pooling層を追加したモデルである。

表 1 セグメンテーションに用いる正規表現

	正規表現
開始表現	^まず|^最初に|^初めに|^次に|^次いで|^続いて|^最後に|^終わりに|^では
	|^[一二三四五六七八九十]+点目
	|^[^、]+についてで(す|あります|ございます)(が|けれど)
	|^終わり(ま|で)す。|^以上で|^ありがとうございま
	|他の質問に(ついて|つきまして)は|質問いたします。|^一方で
終了表現	伺い[^、]*ます。|お尋ね[^、]*します。|お答えください。
	(見解|所見|答弁)を求め[^、]*ます。
	(いかがで|どうで)(しょうか|すか)。|ありませんか。
	|.+質問を(終わります|終了します|いたします)。
	|(お答え|回答)(を?いたします|を?申し上げます)。

3.2 含意関係認識

一般的な含意関係認識では、2文の間に含意関係が成立するかどうかを判定する。本研究では、主張が真であるならば、その主張で検索したパッセージは主張を含意すると仮定する。これにより、含意関係認識を用いて主張の真偽を判定することができる。

図1に本研究で提案するファクトチェックモデルを示す。まず、主張文を用いて、パッセージ検索により議会議事録から関連するパッセージを抽出する。その後、主張とパッセージをセパレータで結合し、ファインチューニングした日本語学習済みモデルで含意関係認識を行う。

日本語学習済みモデルには、BERT、RoBERTa、BERT+biLSTMの3種類のモデルを用いる。BERT[5]は、Transformerをベースにしたエンコーダであり、事前学習としてMasked Language ModelとNext Sentence Predictionを学習している。RoBERTa[6]は、BERTの派生モデルで、Next Sentence Predictionを学習しないなど改良を加え、性能向上を図ったモデルである。

図2に、BERT+biLSTMモデルを示す。BERTモデルとRoBERTaモデルでは、主張文とパッセージ全文のペアを入力する。それに対し、BERT+biLSTMモデルは、主張文とパッセージの1文ずつをペアとし、全てのペアを1つずつ入力する。そして、全ての[CLS]の埋め込みをbiLSTM[7]に入力し真偽を判定する。

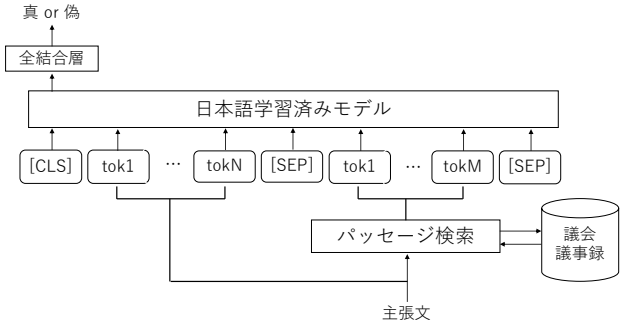

図 1 ファクトチェックモデル

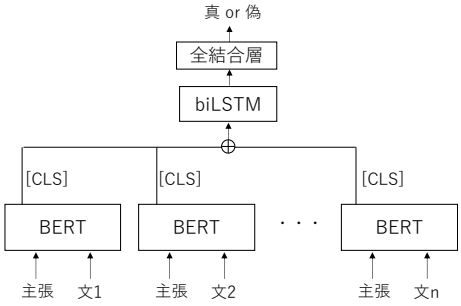

図 2 BERT+biLSTMモデル

4 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、NTCIR-16 QA Lab-PoliInfo-3 Fact Verificationタスクで配布されたDry Runの学習データを用いる。表2にデータセットのフォーマットと例を示す。このタスクでは、平成23年から平成27年の間に開催された東京都議会を対象としている。要約には、東京都議会により人手で作成された「都議会だより」を用いている。

データ数は1,024件であり、学習用、検証用、テスト用に7:1:2の割合で分割する。また、データの属性のうち、Date、Speaker、UtteranceSummaryのみ用いる。

表 2 データセットのサンプル

属性	データ例
ID	00002
Prefecture	東京都
Date	2023/9/28
Meeting	平成23年_第3回定例会
Speaker	知事
UtteranceSummary	首都の知事として強い危機感に立ち、 現場を踏まえて緊急になすべきことを建言した。
UtteranceType	answer
ContextSummary	放射能対策に一丸で取り組め。
ContextWord	新内閣への建言
RelatedUtteranceSummary	知事が込めた想いは。
StartingLine	8275
EndingLine	8283
DocumentEntailment	TRUE

4.2 共通設定

パッセージ検索における検索手法について、BM25+では、$k_1 = 1.2$、$b = 0.75$、$\delta = 1.0$とする。SBERTでは、Hugging Faceを用いて東北大学の乾研究室が公開しているBERT-base[2]を利用する。SBERTは、議会議事録から作成した115,750件の学習データでファインチューニングを行う。また、類似度の計算にはコサイン類似度を用いる。パッセージ単位について、文単位では、予備実験よりn=7とする。

含意関係認識モデルについて、BERT-baseモデルとBERT-largeモデルでは、東北大学の乾研究室が公開しているモデルを利用する。RoBERTa-base[3]モデルでは、rinnaが公開しているモデルを利用する。これらのモデルは、820件の学習データで含意関係認識用にファインチューニングを行う。

主張が真である場合、根拠文の抽出を行う。具体的には、表1の正規表現で議会議事録をセグメントに分割し、主張文をクエリとしてBM25+で最も類似度の高いセグメントを抽出する。

4.3 評価指標

評価指標は、recall、precesion、F値とする。以下に、主張が真の場合の評価指標の式を示す。主張文が偽の場合は、真偽判定の結果が正解ならrecallとprecisionを1、不正解なら0とする。

$$recall = \frac{\text{予測した根拠文のうち正解だった根拠文の数}}{\text{正解の根拠文の数}} \quad (3)$$

$$precision = \frac{\text{予測した根拠文のうち正解した根拠文の数}}{\text{予測した根拠文の数}} \quad (4)$$

4.4 パッセージ検索手法の比較実験

パッセージ検索手法において、各検索手法と各パッセージ単位の比較を行う。

4.4.1 比較手法

検索手法は、BM25+とSBERT、パッセージ単位は、主張文との類似度上位1セグメントと類似度上位7文とする。含意関係認識モデルには、BERT-baseを用いる。

4.4.2 結果

表3に実験結果を示す。結果より、検索手法はBM25+、パッセージは類似度上位7文を用いる手法が最も高いF値であることがわかった。

表 3 パッセージ検索の比較結果

パッセージ検索		含意関係認識モデル	recall	precision	F
手法	パッセージ				
BM25+	上位1セグメント	BERT-base	0.8846	0.9121	0.8869
	上位7文	BERT-base	**0.9048**	**0.9336**	**0.9080**
sBERT	上位1セグメント	BERT-base	0.7516	0.7810	0.7550
	上位7文	BERT-base	0.8742	0.9021	0.8770

[2] https://huggingface.co/cl-tohoku
[3] https://huggingface.co/rinna

4.5 含意関係認識モデルの比較実験

含意関係認識モデルについて比較を行う。

4.5.1 比較手法

検索手法はBM25+、パッセージは類似度上位7文とする。含意関係認識モデルは、パッセージ全文を入力するモデルであるBERT-base、BERT-large、R-oBERTa-baseとする。パッセージの1文ずつ入力するモデルでは、BERT-baseモデルを用いて、[CLS]埋め込みを全て連結するBERT+biLSTM(連結)モデル、[CLS]埋め込みの平均を用いるBERT+biLSTM(平均)モデル、biLSTMを取り除き全結合層で分類を行うBERT+LSTMなしモデルとする。

ベースラインモデルとして、主張文の全ての名詞が、話者の発言に存在するかどうかで判定するルールベース手法を用いる。

4.5.2 結果

表4に実験結果を示す。結果より、BERT-largeモデルが最もF値が高いことがわかった。ルールベース手法であるベースラインと比較して、F値で約+0.18を達成した。BERT-baseとRoBERTa-baseを比較すると、僅かにBERT-baseの方がF値が高い。これは、RoBERTaではNext Sentence Predictionを学習していないのに対して、BERTでは学習しているためだと考えられる。

図3に、ルールベース手法では正解できなかった負例を示す。これらの例は、BERT-largeでは全て正解できている。主張文の名詞が話者の発言に全て存在する場合でも、BERTでは意味を捉えて分類できている。

表 4 含意関係認識モデルの比較結果

パッセージ検索		含意関係認識	recall	precision	F
手法	パッセージ	モデル			
ベースライン			0.7432	0.7474	0.7400
BM25+	上位7文	BERT-base	0.9048	0.9336	0.9080
		BERT-large	**0.9146**	**0.9444**	**0.9183**
		RoBERTa-base	0.8949	0.9222	0.8972
		BERT + biLSTM(連結)	0.8656	0.8951	0.8692
		BERT + biLSTM(平均)	0.6843	0.7113	0.6878
		BERT + LSTMなし	0.8361	0.8648	0.8391

地域連携クリティカルパスの活用促進や医療機関相互の連携体制を確保。
研修会等で支援センター活用を働きかける。支援策充実を働きかける。
東京都では昭和五十六年以前の旧耐震基準で建てられたマンションが多い。 今回の会議で知事が検討する議論は、今後行われる東京緊急対策二〇一の一つにあたり、 東京都のマンション耐震化促進については否定するものと考えられる。

図 3 ルールベース手法で不正解だった負例

5 おわりに

本研究では、議会議事録を対象として、パッセージ検索と含意関係認識を用いたファクトチェックの研究を行なった。結果として、BM25+で類似度上位7文を検索し、BERT-largeモデルで含意関係認識を行うことで、ルールベース手法と比較してF値で約+0.18の結果を達成した。

今後の課題として、ソーシャルメディアやニュースなどのファクトチェックに適用できるかどうか調査していきたい。

謝辞

本研究は、JSPS科研費19K11980および18H01062の助成を受けた。

参考文献

[1] Casper Hansen, Christian Hansen, and Lucas Chaves Lima. 2021. Automatic Fake News Detection: Are Models Learning to Reason? Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 80-86.

[2] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 809–819.

[3] Yuanhua Lv and ChengXiang Zhai. Lower-bounding term frequency normalization. In Proceedings of CIKM'2011, pages 7-16.

[4] Nils Reimers and Iryna Gurevych. 2019. sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982-3992.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

[6] Yinhan Liu, Myle Ott, Naman Golyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

[7] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735-1780.