QA Lab-PoliInfo-3 における Fact Verification

横手健一
日立製作所
kenichi.yokote.fb@hitachi.com

木村泰知
小樽商科大学
kimura@res.otaru-uc.ac.jp

渋木英潔
国立情報学研究所

秋葉友良
豊橋技術科学大学

小川泰弘
名古屋大学

石下円香
国立情報学研究所

1 はじめに

国立情報学研究所が主催する NTCIR プロジェクトの一つとして QA Lab-PoliInfo を開催してきた. QA Lab-PoliInfo は質問応答や自動要約などの自然言語処理のアプローチにより, 政治情報における信憑性問題を解決することを目指す. 本稿では,NTCIR-14 QA Lab-PoliInfo segmentation の後続タスクとなる Fact Verification タスクについて述べる.

2 Fact Verification の位置付け

本章では,Fact Verification タスク（以降「本タスクと呼ぶ」）について, 先行タスクとの差異, 本タスクの狙い, および関連研究を述べる。

2.1 先行タスク:NTCIR-14 QA Lab-PoliInfo segmentation

NTCIR-14 QA Lab-PoliInfo segmentation[1] では, 原文書とその中の一部を要約したテキストを入力として与え, テキストが原文書のどこを要約したかを文レベルで特定する. 原文書として東京都議会の発言記録を対象とし, 要約テキストとして, 都議会の活動を発信する広報紙である「都議会だより」を対象とする. いずれも, インターネット上に公開されている[2][3]. 発言を引用, 要約して生成した二次情報は, 発言の一部が欠落することにより, 発言者の本来の意図とは異なった印象を読者に与えてしまう危険性がある. 二次情報に対応する実際の発言を同時に提示することで, 誤った解釈で二次情報が流通することを防止する.

2.2 Fact Verification タスクの狙い

NTCIR-14 QA Lab-PoliInfo segmentation が対象とした広報紙「都議会だより」は, 東京都議会の発言記録を元に作成している. 従って, 都議会だよりの各記述に対応する実際の発言は必ず存在し, その必ず存在する発言情報（要約テキストに対応する原文書中の開始行と終了行）を特定する問題として扱うことができた. 一方で, 実際の発言が存在しない架空の二次情報を入力した時, 原文書中に対応する発言が存在しないことを提示することはできなかった. 本タスクは,NTCIR-14 QA Lab-PoliInfo segmentation が対象とした「東京都議会の発言記録」,「都議会だより」に加えて, 実際の発言が存在しない架空の「都議会だより」を作成し, 評価の対象とする. これによって, 誤った解釈で二次情報が流通することだけでなく, 根拠が存在しない偽の二次情報が流通することも防ぐ手段を検討する.

2.3 関連研究

本節では, フェイクニュース問題や Fact Checking を扱う関連タスク事例について述べる.

2.3.1 SemEval-2017 Task 8: RumourEval

SemEval-2017 Task 8: RumourEval タスク [4] は, ソーシャルメディアに投稿された噂 (Rumour) とそれに対する返信投稿の各やり取りに対して, 賛成反対などのスタンスに関する 4 種類のラベルに分類するサブタスク A と, その噂が正しいかどうかの信憑性を 0 から 1 の確信度で出力するサブタスク B を実施した. ラベル付与の正解はクラウドソーシングで作成し, 信憑性の正解はジャーナリストのような専門家によって作成した. サブタスク A の評価はラベル付与の正答率, サブタスク B の評価は確信度の平均平方二乗誤差 (RMS Error) を用いた.

2.3.2 The Fact Extraction and VERification (FEVER)

The Fact Extraction and VERification (FEVER) Shared Task[5] は, 人手で作成した主張 (Claim) に対して,

正しい (SUPPORTED), 誤り（REFUTED）, 情報不足 (NotEnoughInfo) の 3 種類のラベルに分類する判定と,SUPPORTED と REFUTED の場合はその根拠となるテキストを Wikipedia から抽出するタスクを実施した.Claim は Wikipedia のテキストから言い換えるなどの手法で人手で生成し, ラベルと根拠テキストの付与も人手で行った. 根拠テキストを漏れ無く人手で収集することは困難のため, タスク参加者のシステムを用いて収集できた根拠テキストも利用した. 評価は, ラベル分類の正答率と, 抽出した根拠テキストの文レベル F-measure によって実施した.

3 データセットの説明

本章では, データセットの構成について述べる. 表 1 は, 本タスクの原文書に相当する「東京都議会の発言記録」のカラム名称とサンプルデータである. 発言記録を行で分割したテキストを最小単位とし,「ID」は各分割テキストに一意な値である.「Line」は何行目の分割テキストであるかを示す.「Prefecture」は本タスクでは東京都を値に持つ.「Title」は発言がなされた会議の名称を示し,「Volume」と「Number」は会議名を識別するための情報を示す.「Year」「Month」「Day」は発言した日時を示し,「Speaker」は発言者,「Utterance」は発言の内容を示す. 表 2 は, 本タスクの要約テキストに相当する「都議会だより」のカラム名称とサンプルデータである.「架空の都議会だより」についても同様のデータ構造を持つ. 会議中の一つの質問とそれに対する回答の組を最小単位（以降,「答弁」と呼ぶ）とし,「ID」によって各答弁に一意な値を割り当てる.「Prefecture」は東京都を値に持ち,「Meeting」は答弁がなされた会議の名称,「Date」は答弁がなされた日時を示す.「MainTopic」と「SubTopic」は答弁を分類するためのカテゴリ情報を示し, 複数の答弁が一つの MainTopic,SubTopic に属する.「QuestionSpeaker」は質問の発言者,「AnswerSpeaker」は回答の発言者,「QuestionSummary」は質問の要約,「AnswerSummary」は回答の要約を示す. 「QuestionStartingLine」,「QuestionEndingLine」,「AnswerStartingLine」,「AnswerEndingLine」は, 要約テキストが「都議会だより」の場合, 要約の元となった「東京都議会の発言記録」データの Line の値を示す. NTCIR-14 QA Lab-PoliInfo segmentation タスクでは,Line の値が入力済みの「都議会だより」データを教師データとして配布し, 未入力の「都議会だより」データに対して Line を推定するタスクとして実施した. 要約テキストが「架空の都議会だより」の場合, 要約元が存在しないことを示す固有の表記を設けることを検討している.

表 1 東京都議会会議録のデータ構造

カラム名称	データサンプル
ID	130001_230617_99
Line	99
Prefecture	東京都
Volume	平成 23 年_第 2 回
Number	1
Year	23
Month	6
Day	17
Title	平成 23 年_第 2 回定例会 (第 7 号)
Speaker	石原慎太郎
Utterance	都庁舎にも義援物資を持った都民が 長蛇の列をなし、都に寄せられた 義援金は約六億円にも上っております。

表 2 都議会だよりのデータ構造

カラム名称	データサンプル
ID	Segmentation-2018-JA- FormalTestGS-00001
Prefecture	東京都
Date	23-9-28
Meeting	平成 23 年_第 3 回定例会
MainTopic	放射能対策に一丸で取り組め スポーツの力で復興の後押しを
SubTopic	新内閣への建言
QuestionSpeaker	鈴木あきまさ（自民党）
QuestionSummary	知事が込めた想いは。
AnswerSpeaker	知事
AnswerSummary	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。 日本再生に向けて 速やかに行動して、 都民・国民の不安を 振り払ってもらいたい。
QuestionStartingLine	7975
QuestionEndingLine	7989
AnswerStartingLine	8275
AnswerEndingLine	8283

4 評価方法

本章では, 本タスクの評価指標について述べる. 先行タスクである NTCIR-14 QA Lab-PoliInfo segmentation[1] で実施した Precision,Recall,F-measure に加えて, 要約テキストが「架空の都議会だより」である時に, 要約元が存在しないことを判定できたかどうかに関する指標を設ける. 例えば, 前記 SemEval-2017 Task 8: RumourEval タスクのように 0

から 1 の確信度を出力することを求め, それらの平均平方二乗誤差 (RMS Error) で評価する方法を検討している.

5 おわりに

本稿では,NTCIR-14 QA Lab-PoliInfo segmentation の後続タスクとなる Fact Verification タスクについて, タスクの狙い, 関連研究, データセットの構成, 評価指標を紹介した. 本タスクは予備テスト（Dry Run）期間と本テスト（Formal Run）期間を設け,Dry Run 期間中は仮のデータセットと評価指標で実施し, 本タスクの課題洗い出しを行う. その後に, 正式なデータセットと指標を決定し,Formal Run を実施する計画である.

参考文献

[1] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, et al. Overview of the ntcir-14 qa lab-poliinfo task. In *Proceedings of the 14th NTCIR Conference*, 2019.

[2] 会議録・速記録 | 東京都議会. https://www.gikai.metro.tokyo.jp/record/.

[3] 都議会だより. https://www.gikai.metro.tokyo.jp/newsletter/.

[4] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*, 2017.

[5] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pp. 1–9, Brussels, Belgium, November 2018. Association for Computational Linguistics.