NTCIR-17 QA Lab-PoliInfo-4 のタスク設計

小川 泰弘 [1]　木村 泰知 [2,3]　渋木 英潔 [4]　乙武 北斗 [5]　内田 ゆず [6]
高丸 圭一 [7]　門脇 一真 [8]　秋葉 友良 [9]　佐々木 稔 [10]　小林 暁雄 [11]

[1] 名古屋大学　[2] 小樽商科大学　[3] 理化学研究所　[4]BESNA 研究所
[5] 福岡大学　[6] 北海学園大学　[7] 宇都宮共和大学　[8] 株式会社日本総合研究所
[9] 豊橋技術科学大学　[10] 茨城大学　[11] 農業・食品産業技術総合研究機構
[1]yasuhiro@is.nagoya-u.ac.jp　[2]kimura@res.otaru-uc.ac.jp

概要

我々は，QA 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の信憑性の問題を解決するためのシェアードタスクである NTCIR-17 QA Lab-PoliInfo-4（以下，PoliInfo-4）を開催している．PoliInfo-4 では，議会会議録などを対象として四つのタスク，Question Answering-2，Answer Verification，Stance Classification-2，Minutes-to-Budget Linking を設計した．本稿では，本タスクの概要について述べる．

1　はじめに

近年，政治に関するフェイクニュースが社会的な問題になっている．特に 2016 年のアメリカ大統領選挙においては，信憑性の低い情報がソーシャルメディアを介して拡散され，選挙の結果に影響を与えたと言われている．現在において，そうした情報が民意の形成に偏りを生じさせることが懸念されている．また，政治家の発言自体も信憑性や根拠が曖昧な場合が多く，近年，政治家の発言に対するファクトチェック [1, 2, 3] の必要性も高まっている．

そこで我々は，自然言語処理技術を用いて政治に関するフェイクニュースの検出やファクトチェックに関する問題を解決することを目指している．

近年，EuroParl Corpus [4] や UK Hansard corpus [1)] などが，政治に関わる自然言語処理研究のための言語資源として利用されている [5, 6, 7, 8]．しかしながら現在のところ，日本語を対象とした研究データが少ないことに加えて，議会における議員の発言を対象とした研究や，議会における政策形成のための議論を対象とした研究は進んでいない．

例えば，ソーシャルメディアで拡散されるテキストと政治家の実際の発言が同一であるか検証するフェイクニュース検出技術や，政治家の発言に適切な根拠があるかを検証するファクトチェックを実現するためには，議会における議員の発言から必要な箇所を抜き出して要約したり，発言文字列と外部の言語資源とを関連付ける機構が必要不可欠である．そこで，我々は地方政治に関わる言語資源として地方議会会議録コーパスの整備に取り組んできた [9]．さらに，そうして収集・整備した議会会議録コーパスを活用して，議論の要約や，発言内容と根拠となる一次情報との結びつけ，発言者の態度（賛否）といった研究を進めるために，評価型ワークショップ NTCIR においてシェアードタスク QA Lab-PoliInfo を行っている．2018 年から 2019 年前半にかけて実施された NTCIR-14 QA Lab-PoliInfo [10, 11] においては，Segmentation，Summarization，Classification の三つのタスクを実施した．その後，NTCIR-15 において QA Lab-PoliInfo-2 [12] を，NTCIR-16 において QA Lab-PoliInfo-3 [13] を実施し，現在は NTCIR-17 において，QA Lab-PoliInfo-4 を実施している．それらで実施されたタスク間の関係を図 1 に示す．

PoliInfo-4 では，四つのタスク，Question Answering-

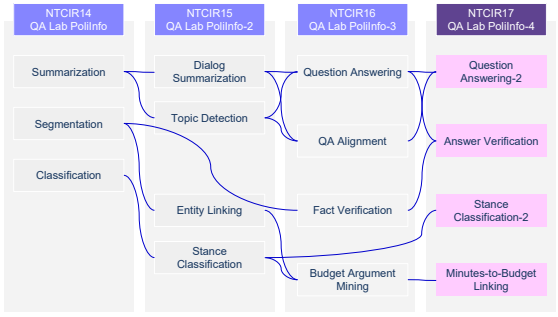

図 1　QA Lab-PoliInfo におけるタスクの推移

1）https://www.english-corpora.org/hansard/

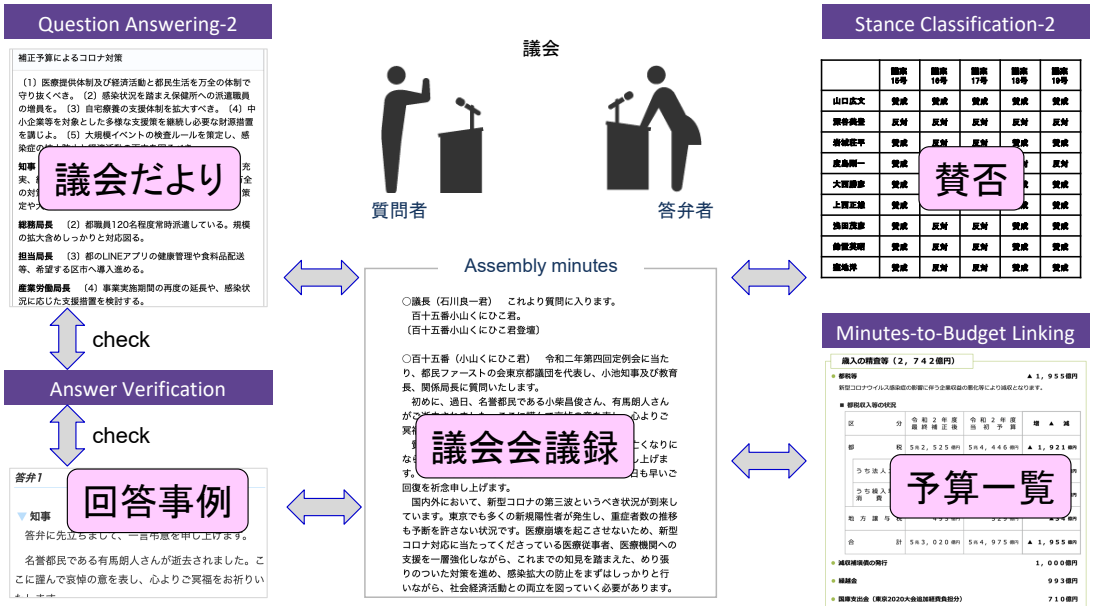

図 2　QA Lab-PoliInfo-4 の各タスクの関係

2, Answer Verification, Stance Classification-2, Minutes-to-Budget Linking を設計した．図 2 に各タスク間の関係を示す．本稿では，以下の各節において，それぞれのタスクの概要について述べていく．

2　Question Answering-2

Question Answering-2 タスク（以下 QA）は NTCIR-16 QA Lab-PoliInfo-3 における Question Answering タスクと同じタスクであり，議会会議録に対する質問に簡潔な回答を返すことを目的とする．

議会における質問と答弁は会議録として記録され，ウェブ上で公開されている．今回は東京都都議会本会議の会議録を対象とする．

通常の質問応答においては，一問一答，すなわち質問を 1 件したらそれに対する答弁があり，引き続き次の質問をする形式が想定される．しかし，都議会本会議においては，一括質問一括答弁という形式がとられている．これは最初に質問者が複数の質問をまとめてした後に，答弁者がまとめて答弁するという形式である．また，その場合に複数の担当者が答弁することもあり，質問の順番と答弁の順番が一致しないことが多い．また，質問の背景や答弁の根拠など述べられているため，質問や答弁自体も長い．そのため，会議録中からユーザが求める質問の回答を探すのは容易ではない．

東京都議会が発行する『都議会だより』においては，そうした質問と答弁の要約が対応付けられた形で掲載されており，都民への分かりやすい情報提供に努めている．『都議会だより』は，議会事務局の職員が作成していることから，人手による「正解の要約」とみなすことができる．しかし，『都議会だより』に掲載されるのは質問の一部にすぎない．

そこで本タスクでは，簡潔な質問を入力した場合，それに対する答弁の要約を返すシステムの構築を目指す．具体的には，まずシステムにはあらかじめ都議会会議録のデータが与えられている．そこに，『都議会だより』に掲載された質問の要約を入力とする．その際には，質問の発言者やそれに対する答弁者などの情報も与える．それを元に，システムは会議録中から質問の答弁となる箇所を発見し，それを要約したものを返す．評価においては，『都議会だより』に掲載された答弁の要約をゴールドデータとして ROUGE の値を計算するとともに，人手による判定も実施する．

このタスクは Watson [14] などの質問応答システムとは以下の 2 点で異なる．1 点目は，回答を探索する際に外部知識にアクセスする必要はなく，あくまで会議録における答弁が正解である点が異なる．そのため，会議が異なれば同じ質問に対して異なる答弁が正解となる可能性もある．

2 点目は，特殊なノンファクトイド型の質問応答になる点である．議会では質問の形式をした要望も多くあり，その場合の答弁も，その要望をどう応えるかという内容になっている．

数値の誤り（特に年号）

質問	緑地整備を進めるべき。
ゴールドデータ	2 年度早期に整備方針を改定、区市町との連携を更に深めながら整備を加速させ、ゆとりと潤いのある東京の実現を図る。
出力	27 年度早期に整備方針改定し、区市町との連携を更に深めながら整備を加速。

要約元の選定失敗

質問	コロナで経済的格差が鮮明に。国と連携し生活底上げを。
ゴールドデータ	生活資金の無利子貸し付け等を講じている、支援を国の取組含め検索できるサイトを立ち上げ、情報が届く仕組みも整えている。
出力	区市町村と連携し、各学校が現状に即した指導計画への再構築を行う。

図 3　QA における誤り例

3　Answer Verification

前節で述べた QA の結果は完璧ではない．前回の PoliInfo-3 においては，四つの観点による人手評価の結果，800 点満点で 499 点（62%）が最高であった．図 3 に誤りの例を示す．なお，ゴールドデータである『都議会だより』を同様に人手で評価した場合が 598 点（75%）であり，10 ポイント以上の差がある．

本研究の目的はフェイクニュースへの対抗であるが，現状では QA の出力結果が新たなフェイクニュースを生み出す原因となっている．そこで我々は，QA の出力がファクトかフェイクかを判定する Answer Verification タスク（以下，AV）を設計した．

AV においては，与えられた質問と答弁のペアが会議録の内容に合致しているか否かを判定するが，ここで問題となるのは学習データである．正例となるデータは，『都議会だより』の質問・答弁の要約を利用することが可能である．一方，負例については，PoliInfo-3 の人手評価の結果，不適切と判定されたものが利用可能である．PoliInfo-3 の QA のフォーマルランでは，100 個の入力に対してそれぞれ 5 個のシステムの出力結果を 4 人で評価した．合計 500 個のうち，半数以上の評価者が不適切と判定したものは 154 個にすぎない．そのため，学習データとして用いる負例の数が不足している．

この問題に対処するため，AV を二つのステージに分割する．一つは，正例・負例のラベル付きデータを作成する Fake-Answer Generation であり，もう一つは，質問・答弁のペアの適切性を判定する分類器を構築する Fact-Checking Classification である．

3.1　Fake-Answer Generation

Fake-Answer Generation においては，『都議会だより』中の質問の要約と会議録を利用し，以下のいずれかの答弁を作成する．

1. 本当はフェイクだが，分類器がファクトと判定
2. 本当はファクトだが，分類器がフェイクと判定
3. 本当はファクトで，分類器がファクトと判定
4. 本当はフェイクで，分類器がフェイクと判定

もともとの動機は，1. や 2. のような負例の不足であったが，3. と 4. のような正例も作成する．

作成方法は自動でも人手でも構わない．自動的に作成する場合，QA のシステム出力を人手で判定する方法が考えられる．また，一部の単語を別の単語に置換するルールなどを適用することも考えられる．図 3 の上部の例のように，前回の QA では数値に関して誤っているものがあった．数値の誤りはフェイクになりうるため，数値を自動的に置換するルールにより，負例を作成することが可能である．そうした誤りはニューラルモデルに基づくシステムではしばしば出現する．また，会議録では「今年度」となっているが『都議会だより』では「2 年度」となっている例があり，正誤の判定は容易ではない．

また，人手で作成する場合は，多くの参加者を募るために，ゲーミフィケーションを利用した手法も考えている [15]．その際には，参加者の敵役となる分類器を準備し，その分類器を騙せるかどうかをリアルタイムで試しながら答弁を作成する枠組も用意する．後述するように，参加者が敵役を騙せるような答弁をある程度作成・投稿した段階で，それを元に学習し直した分類器を新たな敵役として交代させることにより，参加者が継続して答弁を作成するように動機付けを行う．

3.2　Fact-Checking Classification

Fact-Checking Classification では，参加者は答弁がファクトかフェイクかを判定する分類器を構築する．この判定は，会議録の内容から答弁を導くことが可能か，という見方もできるため，自然言語推論の一種ともいえる．

学習データはタスクオーガナイザから提供する．その際に，Fake-Answer Generation で作成されたデータも提供する．実際の運用においては，PoliInfo-4 の実施期間中に何回かのサイクルを用意し，あるサイ

クルの Fake-Answer Generation で作成されたデータを，次のサイクルの Fact-Checking Classification の学習用に利用できるようにする．同様に，あるサイクルの Fact-Checking Classification で作成された分類モデルを，次のサイクルでの敵役として利用することも計画している．

4 Stance Classification-2

政治家の発言の信憑性を判断するためには，政治家がどのような立場で発言しているのかを知ることが必要である．そのためには，複数の議案に対する賛成・反対を総合して判断する必要がある．Stance Classification タスク（以下，SC）は，政治家の発言を元に，議案に賛成か反対かを判定するタスクである．

同様のタスクは，PoliInfo-2 でも実施した．その際には，議案・会議録・議員の所属政党の情報を与えた．しかし，実際の会議録を見ると，「議案第○号×××について、賛成の立場で討論いたします。」といった賛否を表明する発言が冒頭にあることが多く，この賛否表明を発見するだけで高い精度が得られてしまった．

そこで今回の SC-2 では，そうした賛否表明中の「賛成」もしくは「反対」の部分をマスクした発言を入力とし，その状態で賛否を判定するタスクとした．また，前節までの QA および AV とは異なり，一括質問一括答弁形式ではない会議録を対象とする．そのため，東京都議会以外の様々な地方自治体の会議録を収集し，そちらから出題する．

5 Minutes-to-Budget Linking

予算編成は，どのような施策をどの程度重視するかを具体的に表すものといえ，議会での議論おいて重要な位置を占める．PoliInfo-3 においては，Budget Argument Mining として，議会の会議録と予算を結び付けるタスクを実施した．今回実施する Minutes-to-Budget Linking タスク（以下 MBLink）は，Budget Argument Mining の後継となるタスクである．具体的には，会議録と予算表が与えられたとき，会議録のテキストと予算表の項目を結び付け，議論の根拠を抽出することを目的とする．

図 4 に MBLink の概要を示す．今回の MBLink では，入力となる会議録の HTML ファイルには，文ごと ID を埋め込み，同様に予算表の HTML ファイルにもセルごとに ID を割り当てておく．これらが与

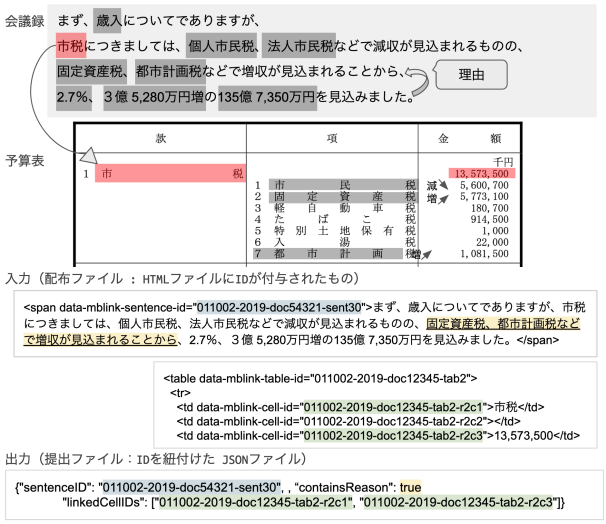

図 4 MBLink の概要

えられたとき，ある文が予算表の特定のセルの金額に関連するものであれば，そのセルの ID を付与し，さらのその文に予算の主張に関する理由が含まれているか否かの判定結果も付加する．

6 おわりに

本稿では，NTCIR-17 QA Lab-PoliInfo-4 における四つのタスク，Question Answering-2，Answer Verification，Stance Classification-2，Minutes-to-Budget Linking について述べた．今回の大会では，それぞれのタスクについて個別の発表 [16, 17, 18] も実施しているので，そちらも参照していただきたい．

現在，PoliInfo-4 では広く参加者を募っている．興味を引くタスクがあった場合は，是非とも PoliInfo-4 の Web サイト[2)] を参照いただくか，連絡をいただきたい．

今後は参加者とともに本タスクを実施していく．現在は予備テスト (Dry Run) を実施中であり，2023 年 6 月頃に本テスト (Formal Run) を実施予定である．予備テストの間も新規参加可能である．2023 年 12 月の NTCIR-17 カンファレンスにおいて各参加者に研究成果を発表していただき，PoliInfo-4 は終了となる．

2) https://sites.google.com/view/poliinfo4/home

謝辞

本研究は JSPS 科研費 21H03769，22H03901，22K12740，20K00576 の助成を受けたものである．

参考文献

[1] Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. Overview of the CLEF-2018 CheckThat! lab on Automatic Identification and Verification of Political Claims. In *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, Avignon, France, 2022. Springer.

[2] Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the CLEF-2018 CheckThat! lab on Automatic Identification and Verification of Political Claims, Task 1: Check-Worthiness. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France, 2018. CEUR-WS.org.

[3] Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the CLEF-2018 CheckThat! lab on Automatic Identification and Verification of Political Claims, Task 2: Factuality. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France, 2018. CEUR-WS.org.

[4] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pp. 79–86, 2005.

[5] Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, 2011.

[6] Astrid Van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. The debates of the European Parliament as linked open data. *Semantic Web*, Vol. 8, No. 2, pp. 271–281, 2017.

[7] Benjamin E Lauderdale and Alexander Herzog. Measuring political positions from legislative speech. *Political Analysis*, Vol. 24, No. 3, pp. 374–394, 2016.

[8] Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. Semantifying the UK Hansard (1918-2018). In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 412–413. IEEE, 2019.

[9] Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu Uchida, Hokuto Ototake, and Shigeru Masuyama. Creating Japanese political corpus from local assembly minutes of 47 prefectures. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 78–85, Osaka, Japan, 2016. The COLING 2016 Organizing Committee.

[10] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. Overview of the NTCIR-14 QA Lab-PoliInfo task. In *Proceedings of the 14th NTCIR Conference*, 2019.

[11] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. Final report of the NTCIR-14 QA Lab-PoliInfo task. In *NII Conference on Testbeds and Community for Information Access Research*, pp. 122–135. Springer, 2019.

[12] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Noriko Kando. Overview of the NTCIR-15 QA Lab-PoliInfo-2 task. In *Proceedings of The 15th NTCIR Conference*, 2020.

[13] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. Overview of the NTCIR-16 QA Lab-PoliInfo-3 task. In *Proceedings of The 16th NTCIR Conference*, 2022.

[14] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building Watson: An overview of the DeepQA project. *AI magazine*, Vol. 31, No. 3, pp. 59–79, 2010.

[15] 渋木英潔, 内田ゆず, 小川泰弘, 門脇一真, 木村泰知. ゲーミフィケーションに基づく QA データセット拡充手法の提案: QA Lab-PoliInfo-4 Answer Verification タスクに向けて. ARG 第 18 回 Web インテリジェンスとインタラクション研究会予稿集, pp. 9–12, 2022.

[16] 渋木英潔, 内田ゆず, 小川泰弘, 門脇一真, 木村泰知. NTCIR-17 QA Lab-PoliInfo-4 Answer Verification における GDADC の利用に向けての考察. 言語処理学会第 29 回年次大会, 2023.

[17] 高丸圭一, 内田ゆず, 木村泰知, 秋葉友良. 地方議会における議案への賛否に関する発言の分析―NTCIR17 QA Lab-PoliInfo4 Stance Classification-2 タスクに向けて―. 言語処理学会第 29 回年次大会, 2023.

[18] 木村泰知, 梶縁, 乙武北斗, 門脇一真, 佐々木稔, 小林暁雄. 議会会議録と予算表を紐づける Minutes-to-Budget Linking タスクの提案. 言語処理学会第 29 回年次大会, 2023.