

NTCIR-15 QA Lab-PoliInfo2 のタスク設計

木村 泰知¹ 渋谷 英潔² 高丸 圭一³ 秋葉 友良⁴
 石下 円香² 内田 ゆず⁵ 小川 泰弘⁶ 乙武 北斗⁷ 佐々木 稔⁸
 三田村 照子⁹ 横手 健一¹⁰ 吉岡 真治¹¹ 神門 典子^{2,12}

¹ 小樽商科大学 ² 国立情報学研究所 ³ 宇都宮共和大学 ⁴ 豊橋技術科学大学
⁵ 北海学園大学 ⁶ 名古屋大学 ⁷ 福岡大学 ⁸ 茨城大学
⁹ カーネギーメロン大学 ¹⁰ 日立製作所 ¹¹ 北海道大学 ¹² 総合研究大学院大学

kimura@res.otaru-uc.jp

1 はじめに

フェイクニュースなどが大きな社会問題となり、政治情報における信憑性判断技術が世界的に注目されている [1]. 我々は、議会における議員発言といった一次情報の提示こそがフェイクニュース対策の基本であるという考えから、国際評価型ワークショップ NTCIR-14¹において shared task である QA Lab-PoliInfo²を 2018 年 1 月から 2019 年 6 月にかけて開催した. QA Lab-PoliInfo では、地方議会会議録コーパス [3] を用いて、議員の発言に含まれる意見やその根拠や条件などを抽出し、関係性などを理解しやすいように整理して提示することを目標としている. そのために、会議録中の発言が引用として与えられた場合にその引用箇所該当する会議録の範囲を特定する Segmentation タスク、発言者の意図が誤解されないように要約する Summarization タスク、発言中の政治課題に対する意見と事実検証可能な根拠を分類する Classification タスクの 3 つのタスクを設計した.

我々は、QA Lab-PoliInfo での成果と問題点を踏まえて、引き続き NTCIR-15 において QA Lab-PoliInfo-2 を開催し、以下の 3 つのタスクを提案している.

1. Entity Linking
2. Dialog Summarization
3. Stance Classification

図 1 に QA Lab-PoliInfo-2 で目標とする課題の全体像と各タスクの位置づけを示す. Entity Linking は、NTCIR-14 の引用箇所をみつける Segmentation タスクと関連があり、表記揺れや曖昧性を解消しながら、会議録に含まれる法律名を対象として、知識ベース (Wikipedia) との連結を行うタスクである. Dialog Summarization は、NTCIR-14 の Summarization タスクを発展させたタスクであり、質問と答弁の対話構造を考慮しつつ、自動要約を行うタスクである. Stance Classification は、議会で議論されている複数の議題について、会派の賛成・反対を分類することで、会派の立場を明らかにするタスクである.

NTCIR-15 QA Lab-PoliInfo-2 では、参加者に共通のデータセットを配布するとともに、公式サイトにおいて、現時点の最も良い手法を表示する Leader Board を設置することで、政治情報を対象とした自然言語処理の研究を促進させる.

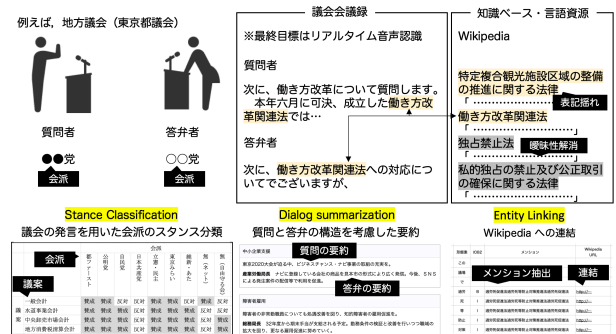


図 1: PoliInfo-2 の 3 つのタスクの関係図

表 1: Entity Linking のデータ構造

column 1	形態素
column 2	B (メンションの開始) I (メンションの続き)
column 3	メンション
column 4	正式名称
column 5	Wikipedia URL

本稿では、NTCIR-15 QA Lab-PoliInfo-2 における 3 つのタスク (Entity Linking task, Dialog Summarization task, Stance Classification task) の意義とテストセットとなるデータ構築について述べる.

2 Entity Linking task

政治家の発言の信憑性を判断するためには、発言の根拠となる一次情報が存在するのかが、明らかにする必要がある. 一次情報は、議会会議録、法律、新聞、Wikipedia などに記載されている可能性があり、発言と結びつけることで、フェイクニュース検出やファクトチェックに役に立つ. そこで、Entity Linking では、議会会議録に含まれる政治家の発言を対象として、表記揺れ、曖昧性、根拠の有無を明らかにすることに焦点を絞り、発言内容の根拠を見つけることを目標とする. 具体的には、法律名を対象として、表記揺れや曖

¹ <http://research.nii.ac.jp/ntcir/index-ja.html>

² <https://poliinfo.github.io/>

形態素	IOB2	メンション	正式名称	Wikipedia URL
この				
議場				
で				
過労	B	過労死等防止対策推進法	過労死等防止対策推進法	http://...
死	I	過労死等防止対策推進法	過労死等防止対策推進法	http://...
等	I	過労死等防止対策推進法	過労死等防止対策推進法	http://...
防止	I	過労死等防止対策推進法	過労死等防止対策推進法	http://...
対策	I	過労死等防止対策推進法	過労死等防止対策推進法	http://...
推進	I	過労死等防止対策推進法	過労死等防止対策推進法	http://...
法	I	過労死等防止対策推進法	過労死等防止対策推進法	http://...
が				

図 2: Entity Linkig のデータフォーマット (TSV 形式)

曖昧性を解決しつつ、一次情報（異なる言語資源）と結びつける。

例えば、下記の発言には「特定複合観光施設区域整備法案」「IR整備法案」「カジノ法案」のように異なる表記で法律名が記述されている。

発言に含まれる異なる表記の法律名の例

特定複合観光施設区域整備法案、いわゆる IR 整備法案について、最近の世論調査では、カジノ法案の成立は不要としている国民の方々七六％、自民党の支持の方々でも六四％に及びます。

また、他の発言では「IR推進法」という異なる法律名についての記述や「IR法」という曖昧な表記で記述されることがある。

Entity Linking タスクでは、法律名を対象として、下記の 3 つの課題を解決しながら、Wikipedia との連結を行う。

1. 表記ゆれの解決
2. 曖昧性の解決
3. リンク先が存在するのか、存在しないか、明確にする

Entity Linking は、会議録に含まれる法律名を抜き出し、知識ベースへ結びつけるタスクである。そのため、タスクは、メンション抽出と Wikipedia (知識ベース) への連結に分けることができる。下記に入力、出力、評価について述べる。

入力	1. 地方議会会議録, および, 国会会議録 2. Wikipedia dump (2019-12-01)
出力	法律名のメンション抽出 メンションに対応する Wikipedia URL への連結
評価	抽出: 形態素単位の抽出精度 連結: 連結の精度

入力と出力の形式は AIDA CoNLL-YAGO Dataset format を用いる³[1]. 表 1 に Entity Linking のデータ構造を示す。

³[https://www.mpi-inf.mpg.de/departments/databases-](https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/aida/downloads/)

Entity Linking のデータ構造について、下記の会議録に含まれる発言を例として説明する。

会議録に含まれる発言の例

この議場で過労死等防止対策推進法が全会一致で可決、成立し、翌年には過労死等の防止のための対策に関する大綱が閣議決定されました。

Entity Linking のデータフォーマットは、図 2 に示したように、形態素、IOB2、メンション、正式名称、Wikipedia への URL がタブで区切られている。形態素は、UniDic 辞書を用いて形態素解析ツール MeCab により分割した。

表 2: Dialog Summarization のデータ構造

ID	識別番号
Date	日付
Prefecture	都道府県
Meeting	会議名
MainTopic	メイントピック
QuestionSpeaker	質問者
SubTopic	サブトピック
QuestionSummary	質問の要約
QuestionLength	質問の字数制限
QuestionStartingLine	質問の開始行
QuestionEndingLine	質問の終了行
AnswerSpeaker	答弁者 ※リスト型
AnswerSummary	答弁の要約 ※リスト型
AnswerLength	答弁の字数制限 ※リスト型
AnswerStartingLine	答弁の開始行 ※リスト型
AnswerEndingLine	答弁の終了行 ※リスト型

3 Dialog Summarization

政治家の発言の信憑性を判断するためには、政治課題に関する議論がどのように行われているのか、知る必要があり、議論をしている相手の発言や文脈を考慮しなければならない。政治課題に関する議論は、議会において行われており、議会会議録として質問や答弁が残されている。しかしながら、議会会議録は、発言を書き起こした文書であり、まとめられておらず、読みづらいという問題がある。特に、東京都議会は、一問一答方式ではなく、一括質問一括答弁方式をとっており、質問と答弁の発言が離れた箇所にある。また、1人の質問者に対して、知事側の総務部長や教育長のように複数の答弁者が存在する場合があります。複数の発言をまとめて解釈する必要がある。このような特徴がある議会会議録は、数多くの発言者がいることに加えて、同じトピックでも離れた箇所に記述されていることから、議論の構造を考慮しつつ要約することが求められる。そこで、Dialog Summarization は地方議会における「議員の質問」と「知事側の答弁」という対話構造を考慮しながら要約することを目標としている。

下記に入力、出力、評価について記述する。

公式サイト⁴の Leader Board では、ROUGE-1 Recall を用いて順位を決める。

図 3 を用いて、Dialog Summarization のデータセット構築方法について説明する。Dialog Summarization

⁴[and-information-systems/research/ambiverse-nlu/aida/downloads/](https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/aida/downloads/)

入力（都議会会議録）

百十五番 小山くひこ君
(百十五番 小山くひこ君登録)

○百十五番（小山くひこ君） 東京都議会第四回定例会に当たり、都民ファーストの会東京都団を代表して、小池知事及び教育長、関係局長に質問いたします。

いよいよ二〇二〇年の東京オリンピック・パラリンピック競技大会まで二年を切りました。一九六四年の東京大会は、戦後復興の象徴であり、首都高速道路や地下鉄の建設、東海道新幹線の開通など、各種インフラの整備が進みました。一九六四年大会後、日本は高度経済成長を遂げ、その後の日本と東洋の発展へと大きくつながりました。

その後の平成は激動の時代でありました。バブル崩壊から始まった長期的な経済停滞、経済のグローバル化、IT化の流れの中で、日本の国際的地位は低下しました。一九六四年大会後に増加を続けていた日本の人口は、二〇〇八年をピークに減少に転じており、東京都の人口も二〇二〇年をピークに減少に転じると見込まれております。このような社会経済情勢の劇的な変化は、戦後日本の成長を生んだ社会モデルからの変革を迫っております。

平成の時代が閉幕を閉じ、新たな時代を迎える成熟都市東京は、今まさに大きな変革を必要としています。少子高齢化による生産年齢人口が減少する中で、次なる成長の源泉となる人、物、金、情報をめぐる世界の都市間競争、まさに激戦りをきわめています。このような状況下において多様な成長の源泉を創出し、そういった認識に立ち、二〇二〇年の先を見据えた東京の成長と発展の礎となる施策が着実に推進されていきます。

世界の中で競う東京の成長戦略を打ち出さなければなりません。

そして、私たちは、一九六四年東京大会をきっかけに築き上げられてきた東京を二〇二〇年大会を契機として再構築し、東京と他の地域がともに栄える、東京の持続的成長を実現していかなければなりません。

私たち都民ファーストの会東京都団は、都議会最大会派となり一年余りが経過しました。この間、議案改革を初め、受動喫煙防止条例の制定、待機児童の大規模減少、オリンピック・パラリンピック籌募人権条例の成立など、二〇二〇年の先を見据えた東京の成長と発展の礎となる施策が着実に推進されてきました。

本定例会でも、中小企業への振興策、防災対策、働き方改革を軸とする補正予算など、未来の東京の成長と発展のために必要不可欠な施策が取り上げられております。

このような東京の取り組みにもかかわらず、国はまた、不合理な都税の収奪を繰り返そうとしています。今、都議会に求められているのは、都議会一丸となって、他の地域との共存共栄を可能とする首都東京の成長戦略を打ち出し、着実に実行することであると改めて申し上げます。

平成三十一年度税制改正について伺います。

国は、いわゆる備前改正の名のもと、都の税財源を地方へ配分すべく、さまざまな措置を講じてきました。この間、都としても対抗措置を講じてきましたが、平成に入ってから三十四年間で都が失った財源は六兆円に上り、平成三十一年度税制改正においても、さらなる措置が事実上予告されております。

こうした国の不合理な税制改正の動きに対して、先般、私たちが提案により立ち上げられました東京と日本の成長を考える検討会の報告書が取りまとめられ、また、東京都税制調査会の答申も示されました。そして、それらを受けた東京都の覚悟も示されております。

都はこれまで、小池知事を先頭に、全国知事会や東京都選出の国会議員、与党税制調査会の国会議員、都内区市町村との折衝を行ってまいりました。私たち都民ファーストの会東京都団も、東京都選出の国会議員や与党税制調査会の国会議員への要請活動、都民への啓発活動等に努めてまいりました。

出力（都議会だより）

東京都議会
Tokyo Metropolitan Assembly

都議会の紹介 議員の紹介 会議の結果と記録 傍聴・見学 調査・友好交流など

中小企業・小規模企業の支援を
幼児教育無償化への都の対応は

小山くひこ（都ファースト）

産業振興

(1) 中小企業・小規模企業振興条例の理念に基づき、活力ある地域社会をつくり雇用の創出を。(2) 農業は東京の持続的成長に必要不可欠。農業振興への今後の展開は。

知事 (1) 地域経済の持続的発展と雇用創出の実現のため効果の高い振興策を展開。(2) 都市農地の保全、担い手の確保と育成・定義の体制整備、先進技術活用等、様々な施策を展開。

ダイバーシティ・東京

(1) 国の幼児教育無償化案では東京の都民は十分とは言えず、また認可と認可外で格差が生じる。対応は。(2) 児童虐待対策の条例制定では未然防止の視点を重視して進めるべき。L I N E 相談の一層の活用も含め見解は。(3) 小中学校のスクール・サポート・スタッフの配置支援を拡大すべし。(4) 学校の働き方改革を加速させるため、都活動員等へのタイムリングで効果的な広報を展開。都民や事業者の理解促進や推進の徹底を図り、受動喫煙防止の取組を進める。

知事 (1) 待機児童対策協議会で国と意見交換。国の動きを踏まえ適切に対応。(2) 体罰等を行ってはならないことを未然防止の観点から条例に明記。L I N E 相談は31年度から本格実施。(5) 条例施行等のタイムリングで効果的な広報を展開。都民や事業者の理解促進や推進の徹底を図り、受動喫煙防止の取組を進める。

教育長 (3) 区市町村教育委員会と連携しながら配置拡充を検討。(4) スタッフの定型的確保や賃金向上をはじめとする多様な取組を検討。

図 3: Dialog Summarization のデータセット構築方法

```

1 {
2   "AnswerEndingLine": [22522],
3   "AnswerLength": [150],
4   "AnswerSpeaker": ["知事"],
5   "AnswerStartingLine": [22499],
6   "AnswerSummary": [
7     "[1] 都は全国の先頭に立ち被災地復興を強力に後押ししていく。
8   ],
9   "Date": "24-2-28",
10  "ID": "Summarization-2020-Training-00001",
11  "MainTopic": "日本の未来のため東京が先頭に<br>帰宅困難者対策",
12  "Meeting": "平成24年第1 回定例会",
13  "Prefecture": "東京都",
14  "QuestionEndingLine": 22252,
15  "QuestionSpeaker": "宮崎章（自民党）",
16  "QuestionStartingLine": 22233,
17  "QuestionSummary": "[1] 被災地そして日本の未来のため東京は。
18  "SubTopic": "都政運営の基本姿勢"
19 },

```

図 4: Dialog Summarization の Json 形式

入力	東京都議会の会議録
出力	都議会だよりの要約に必要な情報
評価	都議会だよりの要約結果 ROUGE, および, 人手による評価

の正解は、都議会だよりを利用している。都議会だよりは、議会で記載された内容が議会事務局の職員により作られていることから、人手により作成された「正解の要約」とみなすことができる。また、都議会だよりでは、質問内容ごとに質問と答弁が対応しており、質問者と答弁者がわかりやすくまとめられている。この都議会会議録の対話構造を用いて、図5の右側にある黄色いエリアで囲まれたテキストを出力するデータセットを構築した。表2および図4は、データ構造とその構造に対応したJson形式である。

表 3: Stance Classification のデータ構造

ID	識別番号
Prefecture	都道府県
Meeting	会議名
Proponent	知事提出議案 or 議員提出議案
BillClass	議案の大分類
BillSubClass	議案の小分類
Bill	議案
BillNumber	議案番号
SpeakerList	議員と会派※辞書型
ProsConsPartyList	会派と賛否※辞書型

4 Stance Classification

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、知ることが必要である。政治家の立場を理解するためには、一つの政治課題に対する賛成・反対を明らかにするだけではなく、マネフェストのように複数の政治課題に対する賛成・反対を総合して判断しなければならない。地方議会では複数の政治課題に対して同じ立場の人が集まり「会派」をつくっている。Stance Classification taskでは「会派」を用いて、政治家の発言から、会派の立場を推定することを目標とする。具体的には、東京都議会における議員の発言を対象として、会派の各議案に対する賛成・反対の立場を推定する。

下記に入力、出力、評価について記述する。入力は、東京都議会会議録、議案、都議会の会派であり、出力は、各議案に対する会派の立場（賛成、あるいは、反対）となる。評価は、各会派の議案数ごとに、賛成・反対の立場があることから、それらの総数を分母として、正解率を計算する。

図5を用いて、Stance Classification データセット構築方法について説明する。Stance Classificationにおいても、都議会だよりを正解として利用することとした。都議会だよりには、議会で議論された議案ごとの採決結果が、賛成、反対に分けて、記述されている。



図 5: Stance Classification のデータセット構築方法

```

1  {
2    "ID": "PoliInfo2-StanceClassification-JA-Dry-Training-",
3    "Prefecture": "東京都",
4    "Meeting": "平成29年第3回定例会、第2回臨時会",
5    "Proponent": "知事提出議案",
6    "BillClass": "予算",
7    "BillSubClass": "29年度補正予算",
8    "Bill": "一般会計 (第1号)",
9    "BillNumber": "第一号議案",
10   "SpeakerList": {
11     "増子ひろき": "都ファースト",
12     "谷村孝彦": "公明党",
13     "秋田一郎": "自民党",
14     "大山とも子": "日本共産党",
15     "中村ひろし": "民進党",
16   },
17   "ProsConsPartyList": {
18     "都ファースト": "賛成",
19     "公明党": "賛成",
20     "自民党": "賛成",
21     "日本共産党": "賛成",
22     "民進党": "賛成",
23   }
24 },

```

図 6: Stance Classification の Json 形式

入力	東京都議会会議録 (定例会, および, 委員会)
出力	各議題に対する会派の「賛成 or 反対」
評価	各会派の議案数を分母した正解率

図 5 の右側では, 知事提出議案である「一般会計」についての「賛成」「反対」が会派単位で記述されている. そこで, 我々は議会会議録における各議員の発言を入力として, 議会だよりに記述されている「賛成」「反対」を出力するデータセットを構築した. 表 3 および図 6 はデータ構造とその構造に対応した Json 形式である.

5 おわりに

本稿では, NTCIR-15 QA Lab-PoliInfo-2 における 3 つのタスク (Entity Linking task, Dialog Summarization task, Stance Classification task) の意義とテストセットとなるデータ構築について述べた.

QA Lab-PoliInfo-2 では, 発言の根拠となる一次情報をみつけるタスク (Entity Linking task), 議論の構造をとらえて要約するタスク (Dialog Summarization task), 会派の立場を明らかにするタスク (Stance Classification task) のデータセットを公開する.

謝辞

本研究は JSPS 科研費 JP16H02912, JP16H01756, および, セコム財団の助成を受けたものです.

参考文献

- [1] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouni, P. Atanasova, S. Kyuchukov, and G. Da San Martino. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, 2018.
- [2] Y. Kimura, H. Shibuki, H. Otake, Y. Uchida, K. Takamaru, K. Sakamoto, M. Ishioroshi, T. Mitamura, N. Kando, T. Mori, H. Yuasa, S. Sekine and K. Inui, "Overview of the NTCIR-14 QA Lab-PoliInfo Task," Proceedings of the 14th NTCIR Conference, 2019.
- [3] Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu Uchida, Hokuto Otake, Shigeru Masuyama. Creating Japanese Political Corpus from Local Assembly Minutes of 47 Prefectures. Coling 2016 workshop, The 12th Workshop on Asian Language Resources, pp.78–85, 2016.