# NTCIR-15 QA Lab-PoliInfo2 のタスク設計

木村 泰知[1] 渋木 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.ac.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-14[1] において shared task である QA Lab-PoliInfo[2][2] 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に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 Segementation 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．

[1]http://research.nii.ac.jp/ntcir/index-ja.html
[2]https://poliinfo.github.io/

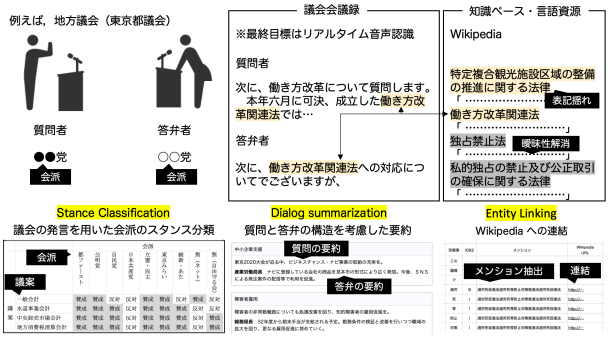


図 1: PoliInfo-2 の 3 つのタスクの関係図

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)<br>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，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，発言内容の根拠をみ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


Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

| 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）

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<br>2. Wikipedia dump（2019-12-01） |
| 出力 | 法律名のメンション抽出<br>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<br>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]ttps://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 Linkig 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，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


Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

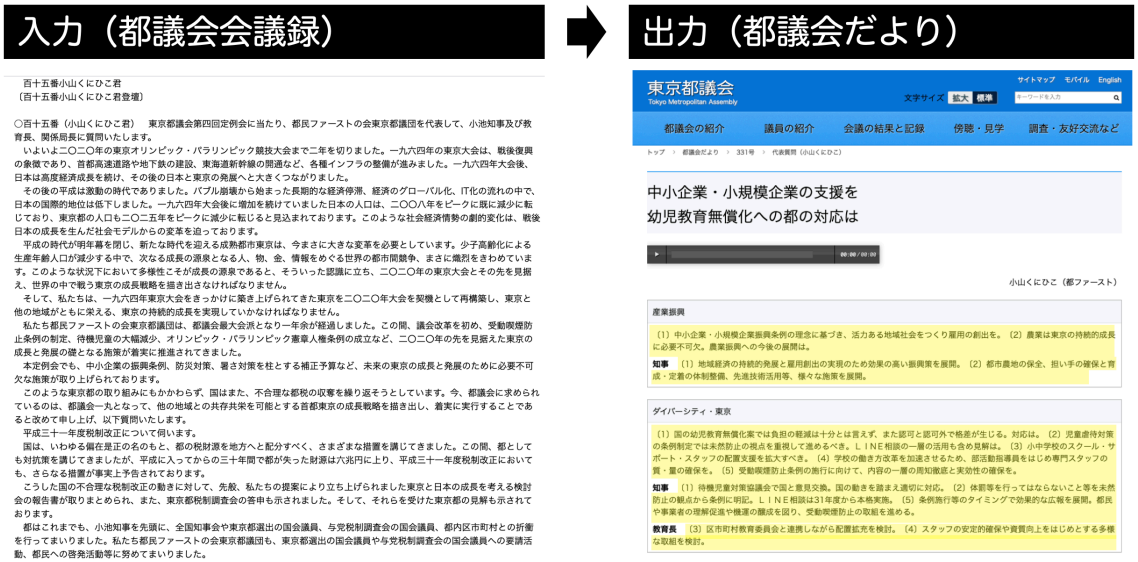


図 3: Dialog Summarization のデータセット構築方法


```
{
"AnswerEndingLine": [22522],
"AnswerLength": [150],
"AnswerSpeaker": ["知事"],
"AnswerStartingLine": [22499],
"AnswerSummary": [
"　〔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。
],
"Date": "24-2-28",
"ID": "Summarization-2020-Training-00001",
"MainTopic": "日本の未来のため東京が先頭に<br>帰宅困難者対策
"Meeting": "平成24年第１回定例会",
"Prefecture": "東京都",
"QuestionEndingLine": 22252,
"QuestionSpeaker": "宮崎章（自民党）",
"QuestionStartingLine": 22233,
"QuestionSummary": "〔1〕被災地そして日本の未来のため東京は
"SubTopic": "都政運営の基本姿勢"
},
```


図 4: Dialog Summarization の Json 形式

| | |
|---|---|
| 入力 | 東京都議会の会議録 |
| | 都議会だよりの要約に必要な情報 |
| 出力 | 都議会だよりの要約結果 |
| 評価 | 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，マニ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の採決結果が，賛成，反対に分けて，記述されている．


Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.


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


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


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. Zaghouani, 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. Ototake, 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 Ototake, 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.



Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.