NTCIR-15 QA Lab-PoliInfo-2 のデータセット構築

木村　泰知† 　渋木　英潔†† 　高丸　圭一††† 　秋葉　友良†††† 　石下　円香††

内田　ゆず††††† 　小川　泰弘†††††† 　乙武　北斗††††††† 　佐々木　稔†††††††

三田村照子†††††††††

横手　健一* 　吉岡　真治** 　神門　典子††

† 小樽商科大学　〒 047-8501 北海道小樽市緑 3 丁目 5-21

†† 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2

††† 宇都宮共和大学　〒 320-0811 栃木県宇都宮市大通り 1 丁目 3 番 18 号

†††† 豊橋技術科学大学　〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

††††† 北海学園大学　〒 062-8605 北海道札幌市豊平区旭町 4 丁目 1-40

†††††† 名古屋大学　〒 464-0814 愛知県名古屋市千種区不老町

††††††† 福岡大学　〒 814-0180 福岡県福岡市城南区七隈 8 丁目 19-1

†††††††† 茨城大学　〒 316-8511 茨城県日立市中成沢町 4 丁目 12 番 1 号

††††††††† カーネギーメロン大学　5000 Forbes Ave, Pittsburgh, PA 15213 アメリカ合衆国

* 日立製作所　〒 185-8601 東京都国分寺市東恋ヶ窪一丁目 280 番地

** 北海道大学　〒 060-8628 北海道札幌市北区北 13 条西 8 丁目

E-mail: †kimura@res.otaru-uc.ac.jp

あらまし　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における信ぴょう性の問題を解決するために Shared task である NTCIR-15 QA Lab-PoliInfo-2 を開催している．このタスクでは，議会会議録を対象として 3 つのサブタスク (Stance Classification, Dialog Summarization, Entity Linking) を設計した．Stance Classification は議会における議員の発言から議案への賛否を推定するサブタスクである．Dialog Summarization は議論の構造を考慮しつつ，議会における質疑と答弁を要約するサブタスクである．Entity Linking は，発言から法律名を抽出し，表記の揺れや曖昧性を解消し，知識ベースと結びつけるサブタスクである．本稿では，これら 3 つのタスクの概要と，それぞれのサブタスクのために構築したデータセットについて述べる．

キーワード　QA Lab-PoliInfo-2, 政治情報, 議会会議録

1 はじめに

近年，政治にまつわるフェイクニュースが社会的な問題になりつつある．信ぴょう性の低い情報がソーシャルメディアを介して拡散され，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ぴょう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といえる．

このため，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 や UK Hansard コーパスなどが，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 [4] [5] [6] [7]．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，発言文字列を外部の言語資源とを Linking する機構が必要不可欠である．そこで，筆者ら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いる [1]．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）と言った研究を進めることを検討している．

これらの問題を解決することを目指して，評価型ワークショップ NTCIR において Shared task QA Lab-PoliInfo を行っている．2019 年後半から 2020 年にかけて実施している NTCIR-15 QA Lab-PoliInfo-2 の概要を図 1 に示す．QA Lab-PoliInfo-2 は 3 つのサブタスクで構成される．Stance Classification サブタスクでは，議会における発言に基づいて会派の立場（議案へ

1 : http://local-politics.jp/

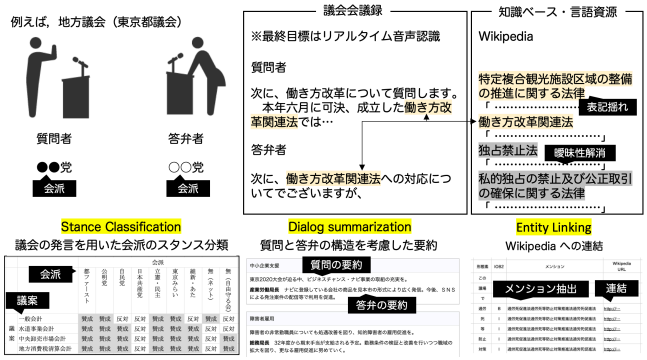

図 1　PoliInfo2 の 3 つのタスクの関係図

の賛否）を抽出することを目指す．Dialog Summarization サブタスクでは，本会議で行われる一般質問および代表質問における質疑の自動要約を目指す．Entity Linking サブタスクでは，発言内で言及されている法律名を抽出し，知識ベース・言語資源の該当項目と結びつけることを目指す．

これらのタスクはいずれも，フェイクニュース検出やファクトチェックのために必要な基礎的な技術であるといえる．なお，図 1 に示すような実時間のフェイクニュース検出やファクトチェックを実現することを考えると，議会での発言は音声認識され，逐一 Stance Classification や Entity Linking の結果が出力され，議論終了時には Dialog Summarization の結果が出力されるべきである．本タスクの範囲では，議会の書き起こしテキストである会議録を音声認識の出力の代替として用いる．

本稿では，まず 2 章で NTCIR における QA Lab タスクの背景について述べた上で，3 章で現在実施している QA Lab-PoliInfo-2 タスクにおける 3 つのサブタスクの概要と，タスクのためのデータセットの構築について述べる．

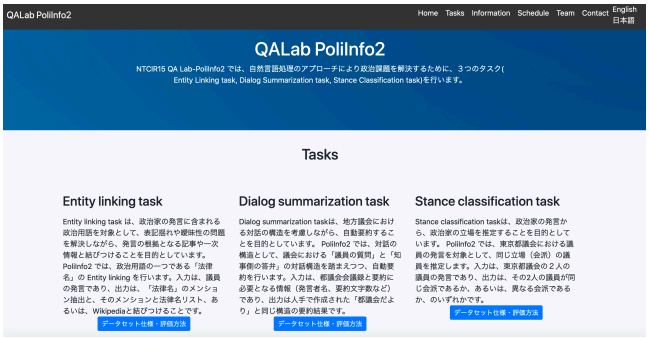

図 2　PoliInfo2 のウェブサイト

2　NTCIR における QA Lab タスクの背景

QA Lab タスクは評価型ワークショップ NTCIR[2]において，現実世界における質問応答システムの実現を目指して開催されている．NTCIR-11 の QA Lab [8] を第 1 回として NTCIR-13 の QA Lab-3 [9] まで，世界史の大学入試問題を対象として実施してきた．

4 回目以降の QA Lab-PoliInfo[3]では政治情報を対象としたサブタスクを質問応答の枠組みで捉えている．地方議会会議録コーパス [10] を対象に，議会発言における議員の意見やその根拠，立場などを抽出し，議論の構造を明確にして，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最終的な目標としている．

第 4 回となる NTCIR-14 QA Lab-PoliInfo では，Segmentation サブタスク，Summarization サブタスク，Classification サブタスクの 3 つのサブタスクを行った．

Segmentation は，要約されて引用された議会中の発言に対して，会議録の引用元の範囲を特定することを目指したサブタスクである．Summarization は，会議録中の特定の範囲から，発言者の意図が誤解されないように指定された文字数の範囲内で要約文を生成するサブタスクである．Classification は，議会における発言に，政治課題に対する意見が含まれるか，また意見に対する事実検証可能な根拠が示されているかを分類するサブタスクである．QA Lab-PoliInfo タスクには，15 組のチームが参加し，2018 年 11 月から 12 月にかけて Formal Run が行われた [11] [12]．

本稿で述べる NTCIR-15 QA Lab-PoliInfo-2 は，QA Lab としては第 5 回であり，上述の NTCIR14 QA Lab-PoliInfo の成果と問題点を踏まえて，図 1 のイメージを実現するために，下記の 3 つのサブタスクを提案している．

（1）　Stance Classification
（2）　Dialog Summarization
（3）　Entity Linking

本タスクのウェブサイト [4]を図 2 に示す．このウェブサイトを通じて，参加者に共通のデータセットを配布するとともに，現時点の最も良い手法を表示するリーダーボードを設置することで，政治情報を対象とした自然言語処理研究の促進を目指す．

次節に，各サブタスクの概要とタスクのために構築したデータセットの設計について述べる．

3　サブタスクの概要とデータセットの構築

本節では，QA Lab-PoliInfo-2 の 3 つのサブタスク（Stance Classification，Dialog Summarization，Entity Linking）の概要とデータセット構築について述べる．

3.1　Stance Classification

3.1.1　タスクの概要

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，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 サブタスクでは政治家の発言か

2：http://research.nii.ac.jp/ntcir/index-ja.html

3：https://poliinfo.github.io/
4：https://poliinfo2.net/

ら，所属する会派の立場，すなわちそれぞれの議案に対する賛否を推定することを目標とする．対象とする議会は東京都議会である．

本サブタスクの入力，出力，評価は以下の通りである．

入力	東京都議会会議録 (定例会/委員会)，および，出題ファイル
出力	各議案に対する会派の「賛成 or 反対」
評価	議案ごとの正解率の総和

入力は，議会会議録全文（本会議および委員会），審議に付されたされた議案のリスト，会派のリスト，発言者した議員の所属会派名である．出力は，議案に対する各会派の立場（賛成または反対）である．評価は，議案ごとの賛否の正解率の総和である．

$$Score = \sum_{i=1}^{\text{議案数}} \frac{\text{議案 } i \text{ における賛否の正解数}}{\text{会派数}} \quad (1)$$

3.1.2 データセットの構築

Stance Classification の正解データは，審議に付された議案に対する各会派の賛否である．これはすべて「都議会だより」に掲載されている．たとえば，図 3 の右側を見ると，平成 31 年度第 1 回定例会で審議された知事提出議案「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ことが分かる．議案に対する各会派の賛否が，最終的には図 4 のようなクロス表の形で得られることを目的とする．

Stance Classification サブタスクでは，東京都議会会議録（定例会，および，委員会）に加えて，機械処理が容易な json 形式で，賛否推定に必要な情報を含んだ出題ファイルを提供する．

東京都議会会議録のデータ構造を表 1 に，json 形式の例を図 5 にそれぞれ示す．Json ファイルには，日付 (Date)，都道府県名 (Prefecture)，議会のタイトル (ProceedingTitle)，URL (東京都議会の URL)，会議録 (Proceeding) が含まれる．会議録 (Proceeding) は，発言者 (Speaker) と発言 (Utterance) を辞書型として一つの要素としたリスト型である．

表 1　Stance Classification 東京都議会会議録のデータ構造

Field name	Description
Date	日付
Prefecture	都道府県
ProceedingTitle	議会のタイトル
Proceeding	辞書型 Speaker, Utterance を要素としたリスト
URL	会議録の URL

出題ファイルのデータ構造を表 2 に，json 形式の例を図 6 にそれぞれ示す．出題ファイルでは，識別番号 (ID)，都道府県名 (Prefecture)，会議名（例えば「平成 31 年度第 1 回定例会」）(Meeting)，議会の開催日 (Date)，議案のカテゴリ (BillClass)，議案のサブカテゴリ (BillSubClass)，議案名 (Bill)，議案番号 (BillNumber)，当該会議中の発言議員名と所属会派のリスト (SpeakerList) が情報として与えられる．これに対して参加者は会議録を利用して議案に対する各会派の賛否を推定し，ProsConsPartyListBinary に解答を出力する．ProsConsPartyListBinary は，各議案に対する会派の立場として，「賛成」「反対」の 2 値のいずれかを埋めることになる．正解となる会派の賛否は，議会事務局の職員により作成された「議会だより」に記載されている結果を用いることとした．Leaderboard では，各会派の賛否が正しく推定できたのかを 2 値の結果で評価を行う．しかしながら，会派の賛否は，知事や議員の発言が記録されている会議録から判断できない場合がある．つまり，議会事務局では，会議録に賛否の言及がない場合でも，「賛成」「反対」の結果を記述している場合がある．そこで，本データセットには，会議録に記述がないために賛否の判定ができない「言及なし」という選択肢も含めた 3 値で答えるProsConsPartyListTernary も用意した．「言及なし」を含めた意図としては，記載されている内容から正しい情報を抽出するために，「記述されていないため，わからない」ということを明確にする必要があると考えているためである．

東京都議会のホームページにおいて，議会会議録および正解データとなる議会だよりの両方が公開されている会議は 20 年分存在する．これらに基づいて，上述のデータセットを作成し，サブタスクを実施を進める．

表 2　Stance Classification 出題ファイルのデータ構造

Field name	Description
ID	識別番号
Prefecture	都道府県
Meeting	会議名
MeetingStartDate	会議開始日 (Date 型)
MeetingEndDate	会議終了日 (Date 型)
Proponent	提案者
BillClass	大カテゴリ
BillSubClass	小カテゴリ
Bill	議案名
BillNumber	議案番号
SpeakerList	議員と会派　※辞書型
ProsConsPartyListBinary	会派と賛否 (2 値) ※辞書型 賛成、反対
ProsConsPartyListTernary	会派と賛否 (3 値) ※辞書型 賛成、反対、言及なし

3.2 Dialog Summarization

3.2.1 タスクの概要

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，東京都議会をはじめとする多くの地方議会では，一問一答方式ではなく，一括質問一括答弁方式がとられており，質問と答弁が離れた位置に存在する．また，質問に対して，知事が答弁する場合と，総務部長や教育

図 3　Stance Classification のデータセット構築方法

		会派								
		都ファースト	公明党	自民党	日本共産党	立憲・民主	東京みらい	維新・あた	無（ネット）	無（自由守る会）
	一般会計	賛成	賛成	反対	反対	賛成	賛成	反対	賛成	反対
議	水道事業会計	賛成	賛成	賛成	反対	賛成	賛成	賛成	反対	反対
案	中央卸売市場会計	賛成	賛成	賛成	反対	賛成	賛成	賛成	反対	賛成
	地方消費税清算会計	賛成	賛成	賛成	反対	賛成	賛成	賛成	反対	賛成

図 4　Stance Classification のイメージ


```
{
"Date": "2001/8/8",
"Prefecture": "東京都",
"ProceedingTitle": "平成十三年第一回臨時会会議録",
"URL": "https://www.gikai.metro.tokyo.jp/record/extraordina
"Proceeding": [
{
"Speaker": "null",
"Utterance": "　出席議員（百二十六名）\n一番谷村　孝彦君\n二番東村
},
{
"Speaker": "議会局長(細渕清君)",
"Utterance": "議会局長の細渕でございます。\n　本日は改選後初の議会で
}
}
```


図 5　Stance Classification 東京都議会会議録のフォーマット (Json 形式)


```
{
"ID":"PoliInfo2-StanceClassification-JA-Dry-
"Prefecture":"東京都",
"Meeting":"平成31年第1回定例会、第1回臨時会",
"MeetingStartDate":"2019/2/20",
"MeetingEndDate":"2019/3/28",
"Proponent":"知事提出議案",
"BillClass":"予算",
"BillSubClass":"31年度予算",
"Bill":"一般会計",
"BillNumber":"第一号議案",
"SpeakerList":{
 "増子ひろき":"都ファースト",
 "吉原修":"自民党",
 "東村邦浩":"公明党"
},
"ProsConsPartyListBinary":{
 "都ファースト":"賛成",
 "公明党":"賛成",
 "自民党":"反対",
},
"ProsConsPartyListTernary":{
 "都ファースト":null,
 "公明党":null,
 "自民党":null,
}
},
```


図 6　Stance Classification 出題ファイルのフォーマット (Json 形式)

長のような知事以外の出席者が答弁する場合がある．さらには，知事による答弁を補足する形で複数の出席者が答弁することもある．従って，質疑（質問と答弁の組）を要約するためには，議論の構造を考慮することが求められる．したがって，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．

本サブタスクの入力，出力，評価は以下の通りである．評価について，複数の手法を並行して用いる予定であるが，公式サイトの Leader Board では，ROUGE-1 Recall を用いて順位を決める．

3.2.2　データセットの構築

議会における一般質問および代表質問の概要は都議会だより

入力	東京都議会会議録，および，出題ファイル
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

に掲載されている（図 9 右）．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，質問項目ごとに質問者と答弁者が示されており，質問とそれに対応する答弁が簡潔にまとめられている．そこで，都議会だよりに記された質疑の要約を Dialog Summarization の正解として用いることとする．図 3 の例で

は，左側が要約前の議会会議録であり，右側に黄色いで示した範囲の文字列が正解データとなる．

Dialog Summarization サブタスクでは，東京都議会会議録（定例会）に加えて，機械処理が容易な json 形式で，要約に必要な情報を含んだ出題ファイルを提供する．

Dialog Summarization の東京都議会会議録は，NTCIR14 Summarization と同じデータ構造であり， Stance Classification のデータ構造と異なる．Dialog Summarization の東京都議会会議録のデータ構造を表 3 に，json 形式の例を図 7 にそれぞれ示す．

表 3　Dialog Summarization 東京都議会会議録のデータ構造

Field name	Description
ID	識別子 (市町村コード $_{年}$月日$_{行}$数)
Line	行番号
Prefecture	都道府県名
Volume	回、第一回定例会
Number	号、何日目
Year	年
Month	月
Day	日
Title	表題
Speaker	発言者名
Utterance	発言


```
{
    "ID":"130001_230617_2",
    "Line":2,"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":"和田宗春",
    "Utterance":"ただいまから平成二十三年第二回東京都議会定例会を開会いたします。"
},
```


図 7　Dialog Summarization 東京都議会会議録のフォーマット (Json 形式)

出題ファイルのデータ構造を表 4 に，json 形式の例を図 6 にそれぞれ示す．出題ファイルでは，識別番号 (ID)，議会の開催日 (Date)，都道府県名 (Prefecture)，会議名（例えば「平成 31 年度第 1 回定例会」）(Meeting)，議会だよりに書かれている 1 人の質問者の全質問を端的に表す単文（図 9 の例では「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」）(MainTopic)，質問項目（図 9 の例では「産業振興」や「ダイバーシティ・東京」）(SubTopic)，質問者名および答弁者名 (QuestionSpeaker，AnswerSpeaker)，質問および答弁の開始行･終了行 (QuestionStartingLine，QuestionEndingLine，AnswerStartingLine，AnswerEndingLine)，生成する質問･答弁の要約の字数制限 (QuestionLength，AnswerLength) が与えられる．これに対して参加者は会議録に基づいて，質問と答弁の要約を生成し，QuestionSummary，AnswerSummary にそれぞれの解答を出力する．

表 4　Dialog Summarization のデータ構造

Field name	Description
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
AnswerStartingLine	答弁の開始行　※リスト型
AnswerEndingLine	答弁の終了行　※リスト型


```
{
"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 8　Dialog Summarization のフォーマット (Json 形式)

3.3　Entity Linking

3.3.1　タスクの概要

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を明らかにする必要がある．一次情報は，過去の会議録，法令集，文書等に記載されている可能性があり，これを現在の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と考えられる．Entity Linking サブタスクでは，参照すべき一次情報が，会議録外の知識ベース・言語資源に集約されていることを想定して，議会での発言と wikipedia を結びつけることを目指す．

議会会議録に含まれる政治家の発言のうち，法律名を対象とする．法律名は正式名称の文字数が長いことが多く，話しことばにおいては，揺れや曖昧性が生じる．

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．

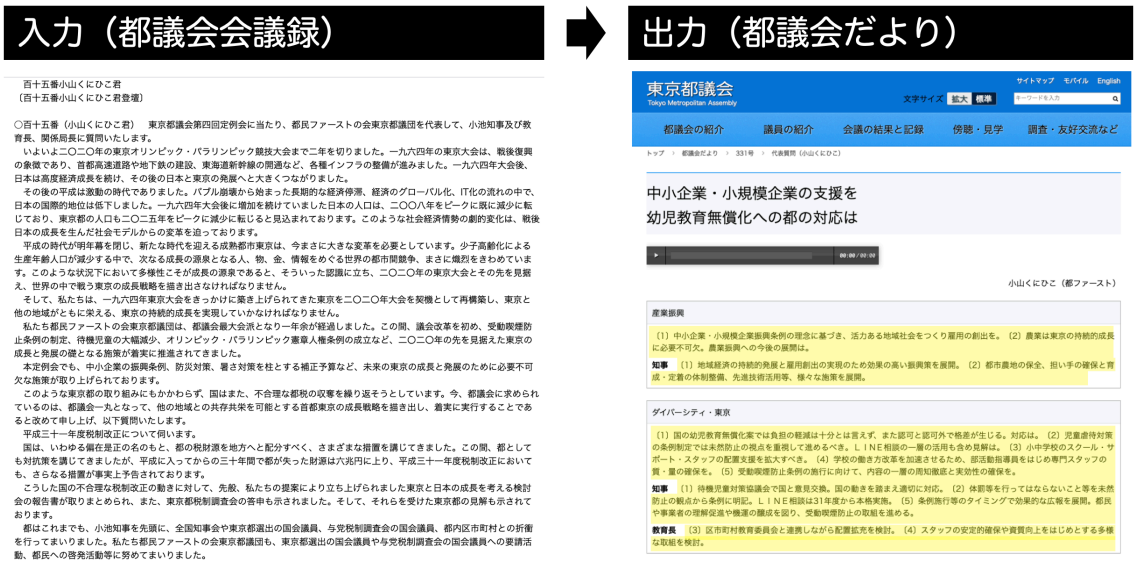

図 9　Dialog Summarization のデータセット構築方法

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。

他の発言においては「ＩＲ推進法」という表記が見られるが，これは異なる法律を指すものである．また,「ＩＲ法」という曖昧な表記で記述されることがある．本サブタスクでは，まず，会議録から法律名のメンション抽出を行い，次に，表記の揺れや曖昧性を解消して，Wikipedia（知識ベース）への結びつけを行う．

本サブタスクの入力，出力，評価は以下の通りである．

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	抽出された法律名（メンション）
	メンションに対応する Wikipedia URL
評価	抽出：形態素単位の抽出精度
	連結：連結した URL の正解率

3.3.2　データセットの構築

本サブタスクは法律名を対象とするため，東京都議会会議録に加えて，国会会議録を対象とする．入出力の形式は AIDA CoNLL-YAGO Dataset format [5] [13]．に基づいた 4 つのカラムから構成される．このデータ構造を表 5 に示す．

以下の会議録中の記述に基づいて Entity Liniking の具体例を述べる．

表 5　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	Wikipedia タイトル
column 5	Wikipedia URL

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。

出題の時点では，第 1 カラムに形態素に分割された会議録が入力されている．形態素解析は，UniDic 辞書を用いて形態素解析ツール MeCab により行う．

参加者はメンション抽出および結びつけを行い，その結果を，図 10 に示すように，第 2 カラムに IOB2 タグ（具体的には B または I），第 3 カラムにフルメンション，第 4 カラムに Wikipedia への URL を記入する．

4　おわりに

本稿では，NTCIR-15 QA Lab-PoliInfo-2 における 3 つのタスク (Stance Classification, Dialog Summarization, Entity Linking) の概要と各サブタスクのために構築したデータセットについて述べた．

NTCIR-15 QA Lab-PoliInfo-2 は，2020 年 2 月 1 日より 2020 年 4 月 30 日まで参加を受け付けている．本稿で述べたデータセットを用い，参加者相互で議論を行いながら，2020 年 5〜6 月に予備テスト（Dry Run）を，2020 年 7 月に本テスト（Formal Run）を実施する予定である．

5：https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/aida/downloads/

形態素	IOB2	メンション	Wikipedia URL
この			
議場			
で			
過労	B	過労死促進法過労死等防止対策推進法過労死促進法	http://…
死	I	過労死促進法過労死等防止対策推進法過労死促進法	http://…
等	I	過労死促進法過労死等防止対策推進法過労死促進法	http://…
防止	I	過労死促進法過労死等防止対策推進法過労死促進法	http://…
対策	I	過労死促進法過労死等防止対策推進法過労死促進法	http://…
推進	I	過労死促進法過労死等防止対策推進法過労死促進法	http://…
法	I	過労死促進法過労死等防止対策推進法過労死促進法	http://…
が			

図 10　Entity Linkig のフォーマット（TSV 形式）

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．

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 Josiane Mothe Fionn Murtagh Jian Yun Nie Laure Soulier Eric Sanjuan Linda Cappellato Nicola Ferro Patrice Bellot, Chiraz Trabelsi, editor, *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, Avignon, France, September 2018. 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 Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France, September 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 Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France, September 2018. CEUR-WS.org.

[4] Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

[5] 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, 1 2017.

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

[7] F. Nanni, S. Menini, S. Tonelli, and S. P. Ponzetto. Semantifying the uk hansard (1918-2018). In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 412–413, June 2019.

[8] Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. Overview of the ntcir-11 qa-lab task. Proceedings of the 11th NTCIR Conference, 2014.

[9] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, and Noriko Kando. Overview of the ntcir-13 qa lab-3 task. Proceedings of the 13th NTCIR Conference, 2017.

[10] 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. The COLING 2016 Organizing Committee, 2016.

[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. Overview of the ntcir-14 qa lab-poliinfo task. In *Proceedings of the 14th NTCIR Conference*, Tokyo, Japan, June 2019.

[12] 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 Testbeds and Community for Information Access Research - 14th International Conference, NTCIR 2019, Tokyo, Japan, June 10-13, 2019, Revised Selected Papers*, pp. 122–135, 2019.

[13] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pp. 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.