

QA Lab-PoliInfo-3 における質問と答弁の対応付け

小川 泰弘
名古屋大学
yasuhiro@is.nagoya-u.ac.jp

木村 泰知
小樽商科大学
kimura@res.otaru-uc.ac.jp

渋谷 英潔
国立情報学研究所

秋葉 友良
豊橋技術科学大学

横手 健一
日立製作所

石下 円香
国立情報学研究所

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とされてきている。2016年のアメリカ大統領選挙の頃から、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、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に欠ける場合があり、それらの発言に対するファクトチェックも求められている。

また日本においても、新型コロナウイルスへの対策について、国会や地方議会で検討され、その内容が報道されるが、取り上げられる内容は限られている。そのため、「〇〇について議論していない」のような事実と反する批判が起きたりする。

こうした状況を踏まえ、政治に関わる自然言語処理研究が注目されてきており、EuroParl Corpus[1]やUK Hansard Corpus¹⁾などが、言語資源として利用されている。しかしながら、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。

これらの問題を解決することを目指して、我々は評価型ワークショップ NTCIR²⁾において QA Lab-PoliInfo を実施してきた。2018年から2019年には、NTCIR-14 QA Lab-PoliInfo を、2019年から2020年にかけては、NTCIR-15 QA Lab-PoliInfo-2 を開催した。2021年から始まる NTCIR-16 においても、引き続き QA Lab-PoliInfo-3 を提案している。

QA Lab-PoliInfo-3 においては、Question Answering, Budget Argument Mining, Fact Verification, QA Alignment の四つのサブタスクをを提案しているが、本稿においては、会議録中の質問と答弁を対応付ける

QA Alignment について紹介する。

本稿では、まず第2節で NTCIR における QA Lab タスクの背景について述べた上で、第3節で QA Alignment の提案理由を述べる。第4節以降では、QA Alignment のデータフォーマット、参加者のタスク、正解データの作成方法、評価方法について、それぞれ述べる。第8節は、本稿のまとめである。

2 QA Lab-PoliInfo タスク

QA Lab-PoliInfo 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たものとして、NTCIR-14 において開催した。地方議会会議録コーパス [2] 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。

NTCIR-14 においては、Segmentation, Summarization, Classification の三つのサブタスクを実施した。Segmentation は、要約し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、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た [3][4]。

NTCIR14 QA Lab-PoliInfo の成果と問題点を踏まえて、NTCIR-15 においては、QA Lab-PoliInfo-2 として、当初は Stance Classification, Dialog Summarization, Entity Linking の三つのサブタスクを実施し、追加で Topic Detection のサブタスクを実施した。Stance Classification は、各議案に対する議員の発言から、

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

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

その所属会派の議案への賛否を推定するサブタスクである。Dialog Summarization は、一括質問一括答弁の形式の会議録の内容を、質問と答弁の対応を考慮して要約するサブタスクである。Entity Linking は、会議録中の発言から、表記揺れや曖昧性の問題を解決しながら、法律名を抽出するサブタスクである。Topic Detection は、会議録から質疑応答のトピックを抽出するものである。QA Lab-PoliInfo-2 の本テストには、14 組のチームが参加した [5]。

我々は、これまでの成果と問題点を踏まえ、NTCIR-16 において QA Lab-PoliInfo-3 を提案している。これは、Question Answering, Budget Argument Mining, Fact Verification, QA Alignment の四つのサブタスクを含んでいる。Question Answering は、会議録の質問の要約が与えられたとき、対応する答弁の要約を生成するサブタスクである。Budget Argument Mining は、予算項目や金額が与えられたときに、会議録中からそれについて議論している部分を抽出するとともに、その内容の議論ラベル（決定事項、例示、事実など）を判定するサブタスクである。Fact Verification は、要約と会議録が与えられたとき、その要約の内容が本当に会議録中に存在するかを判定し、存在する場合はその範囲を特定するサブタスクである。QA Alignment は一括質問一括答弁の形式における質問と答弁を対応付けるものであり、本稿において説明する。

3 QA Alignment の提案理由

本節では、QA Alignment の前提となる Dialog Summarization と Topic Detection について述べた後、QA Alignment の提案理由を述べる。

3.1 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、東京都議会をはじめとする多くの地方議会では、一問一答形式ではなく、一括質問一括答弁形式がとられており、質問と答弁が離れた位置に存在する。また、質問に対して、知事だけでなく、総務部長や教育長の

ような他の出席者が答弁する場合がある。

そうした問題点を解消し、「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た、Dialog Summarization のサブタスクを QA Lab-PoliInfo-2 では実施した。

3.2 Topic Detection

議会においては様々なテーマについて議論されるが、そのすべてが報道される訳ではない。そのため、実際には議会で扱われているのにもかかわらず、議論されていないといった批判が起きることがある。例えば去年は、「野党は新型コロナについて議論せず、云々」といったデマが流れたが、これも一種のフェイクニュースと言える。

そうした問題を解決するために、議会においてどのようなトピックが扱われたのかを検出する Topic Detection のサブタスクを、QA Lab-PoliInfo-2 では追加で実施した。このタスクはパイロットタスクという位置付けであり、トピックの定義についてはタスクオーガナイザ側で定義せず、参加者それぞれが考えるものとした。特にトピックの粒度は大きな課題である。このタスクでは、令和 2 年度東京都議会会議録からのトピック抽出を実施したが、ほとんどの質疑応答は、「新型コロナ」に関連するものである。そのため、トピックを単に「新型コロナ」とするだけでは不十分であり、その中に含まれるサブトピックの検出も必要になる。その場合、最も基本となるサブトピックの粒度は何かという点が問題になる。それに対して、ここでは一つの質疑応答には一つの基本トピックが含まれると考える。一つの質疑応答に複数のトピックが含まれる場合は例外とする。

その際に問題となるのは、何をもって一つの質疑応答と定義するかである。ここで注目するのは、会議録の答弁の冒頭において頻出する「〇点のご質問にお答えいたします。」という定型表現である。東京都議会会議録において、知事以外の答弁者が複数の質問に回答する場合、冒頭でこの発言をすることが多い。すなわち、答弁者が質問の個数を認識した上で回答していると言える。

3.3 QA Alignment

以上のように、Dialog Summarization および Topic Detection のサブタスクにおいては、一括質問一括答弁形式の構造を解析し、質問と答弁を対応付ける前処理が必要になると考えられる。しかし、実際に

表 1 会議録の形式 (JSON 形式のキー)

キー名	説明	キー名	説明
ID	識別番号	Year	年 (和暦)
Line	文番号	Month	月
Prefecture	都道府県	Day	日
Volume	会議名	Speaker	発言者
Number	会議番号	Utterance	発言内容
Title	会議録名		

は、タスクの参加者は質問と答弁を対応付けをすることなく、各タスクに取り組んだ。

その理由としては、end-to-end な手法が現在の主流であること、対応付けをしなくてもそれなりの性能が得られたこと、対応付けの正解が与えられていないことが考えられる。しかし最後に関しては、タスク参加者からの指摘で、東京都議会の「本会議ネットリポート³⁾」の存在が判明した。これは都議会会議録とは別に公開されており、その中で質問と答弁が対応付けられている。そこで今回は、このネットリポートを正解とし、質問と答弁を対応付けを自動的に実行するタスクを提案する。

4 データ形式

表 1 は、過去の QA Lab-PoliInfo および QA Lab-PoliInfo-2 において、都議会会議録のデータとして提供した JSON 形式のデータの形式である。今回の正解データは、表 1 のデータに、以下の三つのキーを追加したものである。

QuestionerID: 質問および答弁が、どの質問者に関するものかを示す。一括質問一括答弁形式のため、ある議員が質問した場合、その議員の発言とそれに対するすべての答弁で同じ値になる。

QorA: 値は Q, A, O であり、発言者が質問者か、答弁者か、それ以外 (議長など) かをそれぞれ示す。発言内容は問わない。

QAID: 値は非負整数。対応する質問と答弁に同じ値を付与する。なお、0 は特別な値であり、質問でも答弁でもない発言を示す。

QAID が 0 になるのは、以下のような場合である。

まず、答弁者の冒頭にある「〇点のご質問にお答えいたします。」などの発言。これは、特定の質問に対応する答弁ではなく、対応する質問がないため、QAID の値を 0 とする。なお、こうした発言はネットリポートには掲載されていない。

3) <https://www.gikai.metro.tokyo.jp/netreport/>

次に、質問者の発言に、質問でないものが含まれる場合。例えば、以下のような発言がある。

なお、売電を目的とする大規模な天然ガス発電所の新規建設に当たっては、… 都は土地の提供程度にとどめておくべきであると、この場で申し上げておきます。

これは意見を述べているだけであり、対応する答弁は存在しない。なお、ネットリポートでは、こうした発言は質問のタグが付かずに掲載されている。

5 参加者のタスク

テストデータにおいては、上記のキーのうち、QAID 以外はすべてタスクオーガナイザ側から提供する。これは、今回新規に追加した QuestionerID および QorA の値については、発言者の役職などの情報から容易に特定でき、本タスクの本質ではないためである。よって、参加者のタスクは一括質問一括答弁を適切に分割し、対応付けをすることである。

タスクとしては、質問数が少なく、答弁者も一人であれば比較的容易であるが、質問数が多くなり、さらに答弁者の人数が多くなると難易度が上がる。東京都議会では、代表質問と一般質問があり、代表質問は一般質問よりも多くの質問がなされる。

実際の対応付けの例を表 2 に示す。これは、平成 23 年第 2 回定例会での山下太郎議員の代表質問への対応付けの例である。山下議員の質問は 42 個あり、それに対応する答弁者は知事を含め全部で 12 人であった。QAID は、各答弁者がどの質問に回答したのかを示している。表 2 から分かる通り、答弁は必ずしも連続する質問に対応する訳ではない。特に知事の答弁は飛び飛びの質問に対応する傾向がある。

6 正解データの作成

正解データは、タスクオーガナイザ側で、都議会会議録のデータとネットリポートのデータを突合することで作成した。基本的には、ネットリポートの内容は会議録の発言を並び替えただけであるため、突合は容易であった。ただし、会議録の答弁の冒頭に「次に、〇〇についてでございますが」という発言がある場合、ネットリポートでは「次に、」の部分が削除されていた。また、一部に全角半角の相違がみられた。そのため、単純な完全一致で対応する発言が見つからない場合、ルーズなマッチにより対

表2 質問と答弁の対応付け

答弁者	答弁数	QAID
知事	7	1, 3, 4, 13, 29, 34, 42
教育長	8	17, 18, 22, 23, 24, 25, 26, 27
東京都技監	3	7, 8, 20
総務局長	4	2, 5, 10, 11
消防総監	1	6
建設局長	1	9
港湾局長	1	12
環境局長	4	14, 15, 19, 39
財務局長	1	16
産業労働局長	6	21, 30, 35, 36, 37, 38
福祉保健局長	4	28, 31, 32, 33
中央卸売市場長	2	40, 41

表3 対応付けの正解と出力結果の例

Line	QorA	QAID		Line	QorA	QAID	
		正解	出力			正解	出力
1	Q	1	1	11	A	0	0
2	Q	1	1	12	A	1	1
3	Q	1	2	13	A	1	1
4	Q	0	0	14	A	1	2
5	Q	2	3	15	A	0	0
6	Q	2	3	16	A	2	3
7	Q	2	3	17	A	3	3
8	Q	3	4	18	A	3	4
9	Q	3	4	19	A	4	5
10	Q	4	5				

応付けを実施した。なお、第4節で述べたように、「〇点のご質問にお答えいたします。」といったネットリポートで省かれている発言や、対応する答弁がない質問については、QAIDを0とした。

7 評価方法

タスクの評価には文献[6]などで提案される単語アライメントの評価を応用する。以下に具体例を示して説明する。

表3は19文からなる質疑応答がどのように対応しているかを示した例であり、QAIDの値で対応付けが示されている。ここで、正解データと評価対象となるシステム出力では、異なる対応付けを行っている。正解データでは質問は4個であり、例えば文1, 2, 3で一つの質問を形成し、それに対応する答弁が文12, 13, 14であることを示している。一方、システム出力では、質問は5個であり、正解データ

表4 表3をalignmentに変換した結果

正解		出力	
1-12	1-12	5-16	5-16
1-13	1-13	6-16	6-16
1-14		7-16	7-16
2-12	2-12		5-17
2-13	2-13		6-17
2-14			7-17
3-12		8-17	
3-13		9-17	
3-14	3-14	8-18	8-18
		9-18	9-18
		10-19	10-19

の最初の質問が二つに分割されている。システム出力を正解データと比較するためにalignmentを以下のように考える。例えば、システム出力の文1, 2の2文と文12, 13の2文が対応しているが、この場合、alignmentとしては、1-12, 2-12, 1-13, 2-13で、合計4個(2×2)あると考える。正解データの最初の質問は、3文と3文が対応しているので、計9個のalignmentができる。表3の対応付けから変換できるalignmentをすべて示したものが表4である。alignmentに変換することにより、正解データとシステム出力で異なるQAIDが付与された場合にも対処可能となる。実際、表3の例では、正解データの文10, 19のQAIDが4であるのに対して、システム出力では5であるが、表4に示すalignmentでは同じ10-19として検出されている。

このalignmentを元に適合率、再現率、F値を計算し、評価に用いる。この例では、適合率78.5%(11/14)、再現率64.7%(11/17)、F値71.0%となる。

8 おわりに

本稿では、NTCIR-16において提案しているQA Lab-PoliInfo-3におけるQA Alignmentタスクの概要及び、データセット、評価方法について述べた。

QA Lab-PoliInfo-3が採用された場合は、2021年7月にデータのリリースを行い、予備テスト(Dry Run)、本テスト(Formal Run)を実施後、2022年6月に本会議を開催する予定である。公式サイトでは、リーダーボードを用いて参加者の結果を随時掲載するだけでなく、TOチームによる実行プログラムなどを公開していく予定である。

参考文献

- [1] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5, pp. 79–86. Citeseer, 2005.
- [2] Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu Uchida, Hokuto Ootake, 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, 2016.
- [3] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ootake, 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.
- [4] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ootake, 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.
- [5] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ootake, 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. *Proceedings of The 15th NTCIR Conference*, 2020.
- [6] Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pp. 1–10, 2003.