議事録の質疑応答の対応関係に基づくクラスタリングによる議題の抽出

大杉 了斗
豊橋技術科学大学
ohsugi.ryoto.dv@tut.jp

秋葉友良
豊橋技術科学大学
akiba@cs.tut.jp

増山繁
東京理科大学
masuyama@rs.tus.ac.jp

1 はじめに

新型コロナウイルス感染症の流行に伴い, 政府や地方議会の対策をすばやく市民に伝える重要性が高まっている. しかし, 会議内容を要約して市民に伝えるためには, 長い時間や手間を要する. この問題に対処するため, 2020 年開催のワークショップ NTCIR15 における QA Lab-PoliInfo タスクのサブタスクとして, 新たに Topic Detection タスクが設立された. このタスクの目的は, 東京都議会の議事録から議題の一覧を提示することである. [1]

本稿では, このタスクに取り組んだ手法を提案する. 提案手法は会議の質疑応答の関係に着目し, この関係を利用しない手法に比べ性能が向上していることを示す.

2 タスク概要

Topic Detection タスクの目的は, 議事録 (2.1 節) から「議題（トピック）の一覧」を議員ごとに提示することである. どのような粒度のトピックを, どのように提示するかについてはタスクで定められておらず, 出力するトピックの定義は参加者によって異なる. そこで本研究では, 都議会だより (2.2 節) のサブトピックに着目した. 都議会だよりとは, 議会の活動を知らせる広報紙のことで, 新聞折り込みや都の施設, 公共機関窓口で配布される. そのため, 都議会だよりのトピックの粒度や提示方法は, タスクの背景である市民に伝えるということに適していると考えられる. また, 都議会だよりは議会職員により人手で作成されているため, 信頼性も高い. よって本研究では, 都議会だよりのサブトピックにできるだけ類似するトピックのリストを出力するシステムの構築を目的とする.


```
{
  "Date": "2020/6/3",
  "Prefecture": "東京都",
  "ProceedingTitle": "令和二年東京都議会会議録第十一号〔速報版〕",
  "URL":"https://www.gikai.metro.tokyo.jp/record/proceedings/2020-2/03.
  html",
    "Proceeding": [ {
      "Speaker": "議長（石川良一君）",
      "Utterance": "これより本日の会議を開きます。¥n"
    },{
      "Speaker": "議長（石川良一君）",
      "Utterance": "昨日に引き続き質問を行います。¥n 四十七番たきぐち学君。
      〔四十七番たきぐち学君登壇〕¥n"
    },{
      "Speaker": "四十七番（たきぐち学君）",
      "Utterance": "質問に先立ち、新型コロナウイルス感染症でお亡くなりになられた方々に対して心からの哀悼の意を表します。¥n　初めに、新型コロナウイルス感染症の迅速、適切なデータの把握と都民への周知について伺います。¥n　…（省略）…¥n　また、機能別団員制度の導入によって、基本団員との役割分担と連携強化を図り、総合的な活動力向上に結びつけるべきと考えますが、見解を伺い、私の質問を終わります。（拍手）¥n〔知事小池百合子君登壇〕
    },{
      "Speaker": "知事（小池百合子君）",
      "Utterance": "たきぐち学議員の一般質問にお答えいたします。¥n　リスクコミュニケーションについてのご質問がございました。¥n　都民が誤った情報に惑わされることなく、感染症を正しく恐れ、予防に向けた適切な行動をとるためには、収集した情報に専門家の知見もいただきながら、わかりやすいメッセージを発信することが重要です。…
```


図 1 Topic Detection タスクにおける議事録の例

2.1 議事録

Topic Detection タスクにおける議事録[1)] の例を図 1 に示す. 議事録は会議の発言を書き起こしたもので, 会議の日付や会議名に加え, 発話者名と発話内容のペアのリストが含まれる. 会議は質問者がまとめて質問をし, 知事や委員会長が各質問に回答するという一括質問一括答弁方式がとられている.

2.2 都議会だより

都議会だより[2)] の例を図 2 に示す. 都議会だよりは議事録を要約したもので, 内容はメイントピック, 議員名, サブトピックと要約の一覧から構成される. サブトピックは図 2 の「新型コロナ対策」と「災害対策」であり, 短いフレーズで表現される.

1） https://www.gikai.metro.tokyo.jp/record/proceedings/
2） https://www.gikai.metro.tokyo.jp/newsletter/

介護事業者のICT導入支援を
災害時の情報連絡体制の強化を

たきぐち学（都ファースト）

新型コロナ対策

〔1〕都独自で感染拡大のフェーズに応じた公表基準を定めるべき。
〔2〕ウィズコロナの介護のあり方を見据えタブレット等介護事業者のICT導入の更なる強化を。

福祉保健局長　〔1〕第二波に備え都の公表基準を整理。〔2〕補助対象を全サービス種別に拡大し補助基準額引き上げ等ICT機器導入促進。

災害対策

発災時、防災業務に精通したリエゾンの自治体派遣等、情報連絡体制強化を。

知事　2年から全ての区市町村に職員派遣体制を整え初動対応に備える。今後実践的な研修で災害対応力を向上し連携した訓練を実施。

図 2　都議会だよりの例

3　提案手法

提案手法の処理の流れを図 3 に示す. まずは候補となるトピックを抽出する (3.1 節). この候補トピックは大量に抽出されるため, 似た意味を持つ候補トピックをまとめるためクラスタリングを行う (3.2 節). クラスタリングは質問からの候補と回答からの候補を分けて 2 回行う. その後, 互いのクラスタを結合する (3.3 節). 最後に結合したクラスタから最終的に出力するトピックを選択する (3.4 節).

3.1　候補トピック抽出

議事録の質問と回答部分から, 候補トピックを抽出する. 図 3 における Step 1 の処理である. トピックは「〜について」という語句の前方に存在する場合が多いため, 正規表現を使ってこの部分を抽出する. 本研究では, 以下の正規表現を使用した.

（名詞｜接続助詞以外の助詞｜接頭詞｜自立動詞）＋について

正規表現の前半部分は, 品詞にマッチする. 品詞の特定には MeCab[2] を用いた. 品詞部分を抽出し, このフレーズを候補トピックと定義する. また, 候補を区別するため, 質問側から抽出された候補は質問候補, 回答側から抽出された候補は回答候補と定義する. さらに抽出精度を高めるため, 以下のような処理を行った.

- 抽出位置が文頭以外の場合は不適切なトピックである可能性が高いため, この候補トピックは除外する.
- 単語の区切りとして「, 」が使われる場合があるため, 「, 」の前方の単語が名詞の場合は「, 」を「・」に変換して抽出する.

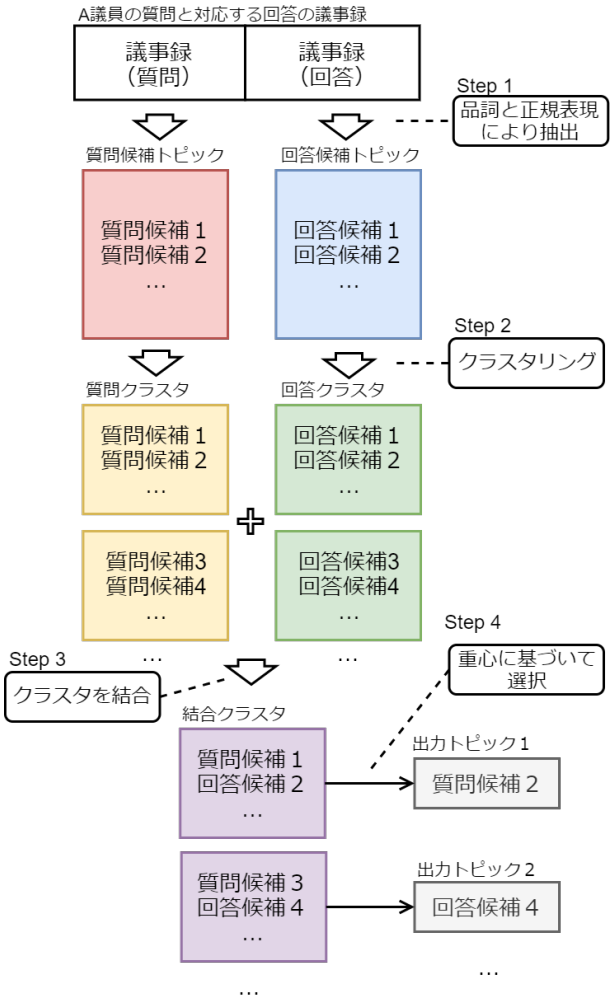

図 3　提案手法の概略

3.2　クラスタリング

候補トピックをクラスタリングする. 図 3 における Step 2 の処理である. クラスタリングの際に候補トピック同士の比較を行う必要があるため, 前処理として候補トピックを事前訓練済み Sentence BERT(SBERT)[3] の埋め込みに変換する. SBERT は, Siamese Network を組み込んだ BERT モデルで, 意味的に近いデータは埋め込み空間での距離が近くなるように学習されている. そのため, 通常の BERT よりも高い精度で埋め込み間の距離を求めることができる. 本研究では, 事前訓練された日本語用の Sentence BERT モデル [4] を用いた. クラスタリングは質問候

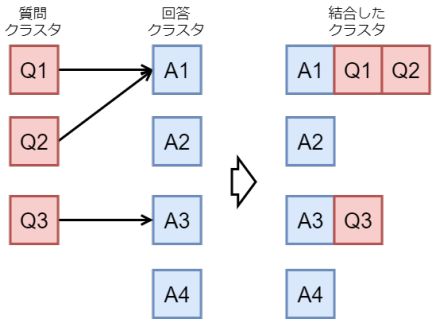

図 4　クラスタの結合例

補と回答候補を分けて行う. クラスタリングアルゴリズムには scikit-learn の階層型クラスタリング [5] を用いた. 結合方法はウォード法, メトリックにはコサイン距離を用いた. クラスタ数は距離のしきい値により決定される. 本研究では, 過去の都議会だよりのサブトピック数を参考にして, しきい値を 28.0 に決定した. クラスタ群を区別するため, 質問候補をクラスタリングしたものを質問クラスタ, 回答候補を回答クラスタと定義する.

3.3　クラスタ結合

質問クラスタと回答クラスタを結合する. 図 3 における Step 3 の処理である. 結合の例を図 4 に示す. まず, クラスタ数が少ない側（例では質問側）から, 最も類似度が高い反対側のクラスタを参照する. 類似度はクラスタの重心のコサイン類似度を用いる. 次に, クラスタ数が多い側（例では回答側）のクラスタと, 参照されている反対側のクラスタを結合する. この結合したクラスタは, 結合クラスタと定義する.

3.4　候補トピック選択

それぞれの結合クラスタから, 出力トピックを 1 つ選ぶ. 図 3 における Step 4 の処理である. 候補トピックの埋め込みとクラスタの重心とのコサイン類似度を計算し, 類似度が最も高い候補トピックを選択する. 選択した候補トピックは最終的な出力となり, これを出力トピックと定義する.

4　実験方法

4.1　評価方法

Topic Detection タスクはオープンタスクであり, 評価は行われない. しかし, 入力データに使用した議事録の都議会だよりは既に発行されているため, 人手により評価を行った. 各出力トピックに対して, その

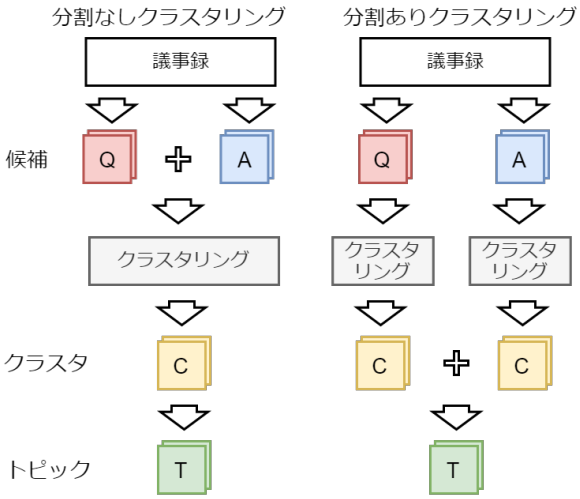

図 5　手法の比較

トピックが都議会だよりに含まれていれば正解, 含まれていなければ不正解を割り当てる. 正解数をカウントし, Precision, Recall, F 値を計算する. ここで正解数をカウントする際, 同じ内容の出力トピックがあった場合（具体的には, 複数の出力トピックが同じ都議会だよりのサブトピックと関連している場合）は, 1 つとしてカウントする. これはクラスタリングの目的が重複を除去することであり, 重複を正解としてカウントすることは適切でないためである.

4.2　分割なしクラスタリング手法

評価結果の比較のため, 質問, 回答を区別せずにクラスタリングする手法も実装した. 提案手法との違いを図 5 に示す. 提案手法では, 候補トピックを質問候補と回答候補に分割して 2 回クラスタリングを実行する. 対してこの手法では, 質問候補と回答候補で分割せずまとめてからクラスタリングを実行する.

5　実験結果

出力例として出力トピックと, それに対応する都議会だよりのサブトピックを表 1 に示す. 出力トピックは都議会だよりとの比較のため, 並び替えている.

評価結果を表 2 に示す. 「分割なし」は 4.2 節の手法で, 「分割あり」は 3 節の提案手法を表している. 質問と回答を分けてクラスタリングした提案手法では, 分けずにクラスタリングした手法と比べて F 値が 0.44 向上した.

表 1　出力結果の例

出力トピック	都議会だより
新型コロナウイルス感染症への対策	新型コロナウイルス対策
ベビーシッターの活用	子育て支援
シニア予備軍向けの読本	認知症対策
受動喫煙対策	犯罪被害者支援
民生児童委員への活動支援	民生委員・児童委員
首都高の大規模更新とまちづくりとの連携	学校の ICT 環境整備
テナントビルなどの耐震改修	中小企業支援
経団連など産業界とも未来の東京	商店街の活性化
練習会場となる公立施設	災害時の電力確保
都立公園における他競技の利用と調和する形でのラグビーができる場の整備	地域の防犯カメラ
保存・管理	東京大会
	文書管理
	大会経費の剰余金

表 2　評価結果

手法	出力数	正解数	Precision	Recall	F 値
分割なし	239	94	0.393	0.537	0.454
分割あり	237	101	0.426	0.577	0.490

6　考察

提案手法による性能向上の理由として考えられることは, クラスタの質が向上することが挙げられる. 分割なしクラスタリングでは, 一度に入力する候補数が増え, 関連性が低い候補トピックまで同一のクラスタに含まれてしまうことがある. このような候補トピックは重心の値をずらしてしまい, 出力トピックの選択に悪影響を与えてしまう. そこで分割してクラスタリングすることで, このような問題を避けられていると考えられる.

また, どちらの手法でも F 値が低くなってしまった. Recall に対して Precision が低い値であることから, Precision が F 値を低下させているとわかる. Precision 低下の原因として, 全てのトピックが都議会だよりに掲載されるわけではないということが挙げられる. これにより正解トピック数が減ることで, 正解数も減り, Precision が下がってしまう（Recall は正解数を分母に持つため影響を受けない）. 都議会だよりのトピックに欠けがある点については, Ogawa ら [6] の論文でも言及されている. Precision の低下は本研究の問題設定によるものであり, これを解決するためにはネットリポート[3)]などの別の要約データを用いる事が必要である.

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

7　関連研究

トピックに関する研究は多く, 例えばトピックモデルが挙げられる. トピックモデルの実装として LDA[7] が広く使われている. LDA は文書中に存在する潜在的なトピックを, 単語の確率分布として表現するモデルである. このトピックの表現方法が本研究との大きな違いである. LDA におけるトピックは, 具体的なテキスト表現を持たない. 一方で本研究のトピックは, 文書中に出現する具体的なテキストを表現している. このテキストを抽出することが本研究の目的である.

次に, Topic Detection タスクにおける他参加者の研究を紹介する. Ogawa ら [6] は, 本研究と同様に, 都議会だよりのサブトピックのような短いフレーズを抽出している. 本研究との違いは問題設定で, 議論の対象となったトピックを全て抽出するという問題設定になっており, 自動的なクラスタリングは行っていない. Hirai ら [8] は, 単語の共起構造と LDA を用いた手法により, トピックの抽出を行っている. トピックの表現は, 本研究と同様に短いフレーズである. しかし, 都議会だよりのような具体的なターゲットは想定していない. Ibrk チームは, doc2vec による分散表現と k-means によるクラスタリングにより, トピックを抽出している. 本研究とは, 問題設定とトピックの定義が大きく異なる. 問題設定は, あらかじめキーワード（「新型コロナウイルス」など）を設定しておき, それに関する 20 つの単語集合を求めるというものである. また, トピックの表現はフレーズではなく単語集合である.

8　おわりに

本研究では, Topic Detection タスクに取り組んだ手法を提案した. 議事録の関係に着目した手法を提案し, この関係を利用しない手法に比べ性能が向上していることも確認した. 今後の課題は, さらに議事録の特徴に着目し, 議事録上での抽出位置を利用した手法等を検討したい.

謝辞

本研究は JSPS 科研費 19K11980 の助成を受けた.

参考文献

[1]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, KenjiAraki, Satoshi Sekine, and Noriko Kando. Overview of thentcir-15 qa lab-poliinfo-2 task. In *Proceedings of The 15th NTCIRConference*, 12 2020.

[2]Taku Kudo, Nippon Telegraph, and Telephone. Mecab, 2001. https://taku910.github.io/mecab/.

[3]Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[4]Sentence embeddings with bert & xlnet, 2019. https://github.com/sonoisa/sentence-transformers.

[5]Clustering — scikit-learn, 2007. https://scikit-learn.org/stable/modules/clustering.html.

[6]Yasuhiro Ogawa, Yuta Ikari, Takahiro Komamizu, and Katsuhiko Toyama. Nukl at the ntcir-15 qa lab-poliinfo-2 task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, 2020.

[7]David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research 3*, 2003.

[8]Yuya HIRAI, Yo AMANO, and Kazuhiro TAKEUCHI. Ntcir-15 qa lab-poliinfo2 dialog topic detection based on discussion structure graph. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, 2020.