Vol.2021-NL-249 No.10
2021/7/28

予算項目に関連する議論を対応づける Budget Argument Mining のデータセット構築

木村 泰知[1,a)]　永渕 景祐[1]　乙武 北斗[2]　佐々木 稔[3]

概要：
本稿では，NTCIR16 QA Lab-PoliInfo-3 のサブタスク「Budget Argument Mining」のデータセット構築について述べる．Budget Argument Mining は，国，あるいは，自治体の予算審議の事項と議会における議論を対応づけることを目的としている．従来の Argument Mining との違いは，単一文書内の議論構造ではなく，複数文書にまたがる議論構造を予算という観点から分析する点にある．特に，構造化されている予算審議の情報と構造化されていない議会会議録の発言文を対象として，予算項目を軸に議論を対応づけることは，新たな取り組みといえる．本タスクでは，予算審議の情報（予算項目，金額，管轄省庁・部局名など）が与えられたときに，議会会議録に含まれる政治家の発言（金額表現を含む発言）と対応づけ，３つの議論ラベル「Claim（主張）」「Premise（根拠）」「その他」を付与する．本稿では，データセット構築に向けた，データ形式の設計，アノテーションの方法，および，結果について述べる．

1. はじめに

政治には，収入と支出を考慮し，お金の使い道を決める予算作成の役割がある[*1]．国の予算は，内閣で予算案が作成され，その予算案をもとに国会で議論された後に，正式な予算となる．また，地方自治体の予算は，知事や市長により予算案が作成され，議会で審議された後に成立する．このような過程を経て成立する予算は，どのような背景に基づいて予算案が作成され，どのような議論を経て成立しているのかを把握しづらい．

従来から，政治学や経済学の分野において，国や地方自治体の予算に関する研究が行われている．予算過程では，(1) 財政収支計画の作成，(2) 審議，(3) 執行，(4) 決算, の順番に進み，租税の配分や経費の配分について意思決定が行われる [1]．予算を含む審議の分析は，国会，あるいは，地方議会の会議録を対象として，TFIDF を用いた分析，あるいは，共起ネットワークを用いた分析を行っている [2][3]．地方会議録の議論構造を考慮した研究として，東京都議会会議録を対象とした Shared task の NTCIR-14 QA Lab-PoliInfo[*2]，NTCIR-15 QA Lab-PoliInfo-2[*3]がある [4]．しかしながら，これらのタスクは，質問と答弁の構造に着目したものであり，予算審議の議論構造を対象としていない．

自然言語処理の分野では，Argument Mining が注目されている[*4]．Argument Mining は，議論構造を解析するタスクであり，小論文などの論述文を対象として，文や節の談話単位のラベル（主張，根拠）を付与する論述構造解析のタスクが有名である [5][6]．論述構造解析では，論述文を入力として，談話単位間の論述関係（Support, Attack），談話単位の機能 (Claim, Premise) を出力としている．これらの研究は，論述文を対象とした研究が中心であり，政治における政治家の議論構造に対して研究の余地がある．

そこで，我々は，国や自治体の予算成立までの議論に着目し，どのような議論に基づいて成立したのかを簡単に把握できるシステムを開発することを最終目標としている．本研究では，国および地方自治体が公開している予算審議に関する「予算項目」と対応する議論を議会会議録からみつけ，結びつける Budget Argument Mining タスクの設計を進めている．予算項目に着目することで，全ての発言の

[1] 小樽商科大学
[2] 福岡大学
[3] 茨城大学
[a)] kimura@res.otaru-uc.ac.jp

[*1] https://www.sangiin.go.jp/japanese/kids/html/shikumi/ichinen.html
[*2] https://poliinfo.github.io/
[*3] https://poliinfo2.github.io/
[*4] http://acl2016tutorial.arg.tech/

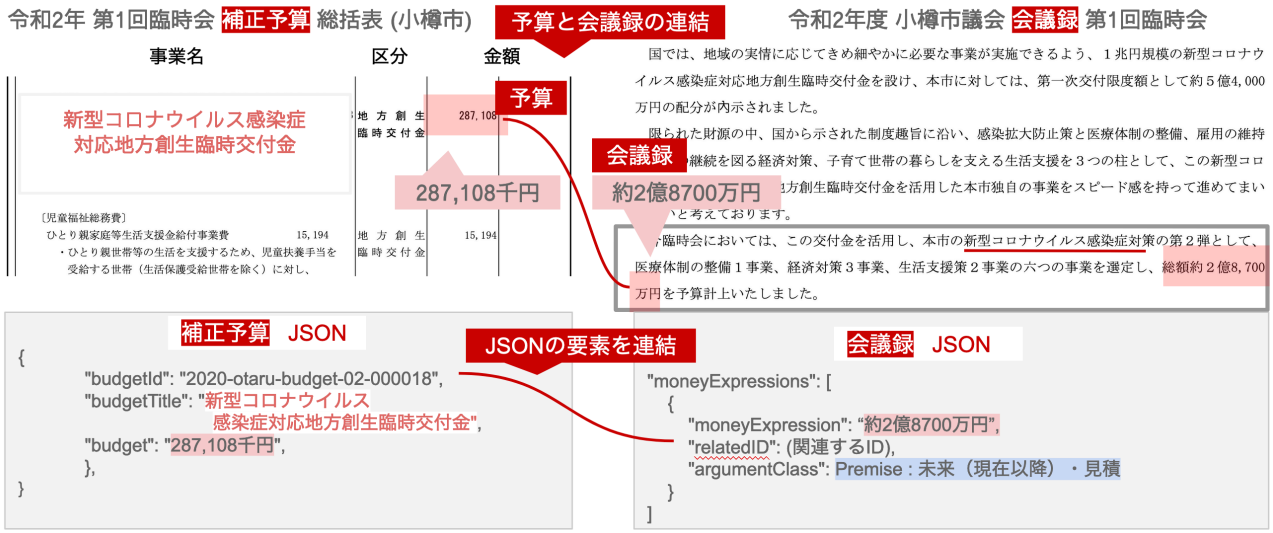

図 1 予算総括表と会議録を用いて予算項目と議論を結びつける Budget Argument Mining

議論構造を解析する必要はなく，効率よく，予算に関する議論を把握できるようになる．本稿では，データセット構築に向けた，データ形式の設計，アノテーションの方法，および，結果について述べる．

本研究の貢献は，下記の４つである．

- 文書内の議論構造ではなく，複数文書にまたがる議論構造を**予算項目**，および，**金額表現**という観点から結びつけている
- 構造化された予算情報[*5]と非構造の文書を連結を試みている
- Web アプリによるアノテーションツールにより効率的なアノテーションを行っている
- 異なる自治体（小樽市，茨城県，福岡市）に対応可能なフォーマットを提案し，データセットを構築している

2. 関連研究

金額表現の抽出

本研究では，予算審議の議論における金額表現に焦点を当てる．金額表現は，固有表現抽出における「金額（MONEY）」を自動で抽出する [7][8][9][10]．固有表現抽出における金額表現は，新しい表現が増え続ける人名や組織名と異なり，限られたパターンであることから，高い精度で抽出できる．金融に関する研究では，金額表現を対象としていることが多い．例えば，NTCIR15 の FinNum タスクでは，金融関連のツイートに含まれる数値表現が対象となる項目に対して，関連しているのか，関連していないのか，２値分類を行っている[*6]．

[*5] 現時点では構造化された情報と呼べないが，4. の予算書情報・決算書情報検索に記載した XML がより扱いやすい形式になれば，構造化データといえる．

[*6] https://sites.google.com/nlg.csie.ntu.edu.tw/finnum2020/finnum-2

会議録および予算

国会会議録は，検索 API が公開されていることから，議員の発言を対象としたテキスト分析が行われている [11]．地方会議録を対象とした研究には，NTCIR-14 QA Lab-PoliInfo，NTCIR-15 QA Lab-PoliInfo-2 がある [4][12]．NTCIR-15 QA Lab-PoliInfo-2 では，東京都議会を対象として，議論構造を考慮した自動要約，議案に対する各会派の賛否分類，法律名の表記揺れおよび曖昧性を解決する Entity Linking などのタスクがある．しかしながら，予算審議を対象とした研究は，行われていない．

Argument Mining

Argument Mining に関する研究は，論理学に基づいて，自然言語処理のアプローチにより議論構造をとらえる研究として注目されている [13][14][15]．論述構造解析は，議論構造を解析するタスクであり，文や節の談話単位のラベル（主張，根拠）を付与する Argument Mining の代表的なタスクである [5][6]．Argument Mining 分析に共通している処理は，議論構成要素の識別，節の属性識別，節間の関係識別である [15]．本研究では，Argument Mining の分析手順を参考にしつつ，予算の観点から，議会における議論構造を解析するタスクを設計する．

3. Budget Argument Mining とは

Budget Argument Mining は，国会，および，地方議会における「予算審議に関する項目」と「議会会議録に含まれる関連した議論」を結びつけるタスクである．

図 1 は，小樽市の「予算総括表」と「議会会議録」を例として，予算項目と議論を結びつける **Budget Argument Mining** タスクのイメージである．本タスクは，予算情報の予算項目を起点として，予算審議の議論，つまり，会議

録に含まれる関連した議員の発言に結びつける．

次の２つのステップにより，予算項目と関連する議論を結びつける．

(1) **節の属性識別 (Identifying Clausal Properties)**：会議録に含まれる議論の構成要素に対して，議論ラベルを付与する．

(2) **予算項目と議論構成要素の連結 (ID Linking)**：会議録の発言と予算表に含まれる対応する予算項目を ID を結びつける．

図 1 の予算総括表に「新型コロナウイルス感染症対応地方創生臨時交付金」の事業名があり，その金額は 287,108 千円と記載されている．ここで記載されている 287,108 千円の議論は，小樽市議会会議録の第 1 回臨時会で議論されており，「総額約 2 億 8,700 万円を予算計上いたしました.」という記述がある．節の属性識別では，議会会議録に含まれている金額表現「総額約 2 億 8,700 万円」を含む節に，根拠ラベルとして「Premise」の議論ラベルを付与する．予算項目と議論の連結は，議論ラベルが付与された「2 億 8,700 万円」を含む節に対して，予算項目の ID を対応づける．本タスクの入力，出力，評価方法を下記に示す．

入力	予算情報 (予算項目,ID, 金額, 部局名, 説明...)
	議会会議録（議会名, 発言者, 発言, 金額表現...）
出力	会議録の発言に対する議論ラベル推定
	会議録の発言と予算項目 ID の対応付け
評価	議論ラベル
	正解率 = 正解議論ラベル ÷ 議論ラベル数
	予算表への連結 (F 値)
	再現率　=出力に含まれる正解数 ÷ 正解数
	適合率　=出力に含まれる正解数 ÷ 出力数

次節以降で，予算情報，会議録，議論ラベルの詳細を述べる．

4. 予算情報とは

国，あるいは，地方自治体では，予算に関する書類が複数存在する．本節では，どのように，対象とする予算情報を選択したのかについて述べる．

Budget Argument Mining では，下記の条件を優先して，予算に関する情報を選択することとした．

- 省庁，年度，金額の記載がある
- 予算の項目（大・中・小）がある
- 会議録と結びつけるための情報が含まれている
- 形式が統一されている
- 省庁，あるいは自治体による違いが少ない

ここでは，厚生労働省を例として，予算に関する資料の 5 つの候補の利点と欠点について述べる．

予算の主要事項 [*7]

- 利点：省庁，年度，金額，予算項目の説明がある．会議録と結びつけるための情報が多い．
- 欠点：PDF ファイルであり，省庁により形式が異なる．PDF からテキストへ自動で変換しづらい．

予算概要[*8]

- 利点：省庁，年度，金額，予算項目の説明があり，会議録と結びつけるための情報が多い．
- 欠点：PDF ファイルであり，省庁により形式が異なる．

総括表 [*9]

- 利点：ほぼ全省庁で統一された形式である．
- 欠点：項目のみで詳細なし・存在しない省庁がある．PDF ファイルである．

個別表 [*10]

- 利点：総括表よりも，細かい情報が記述されいる．省庁によって，総括表が存在しない場合もあるが，個別表は存在する．
- 欠点：PDF ファイルである．

予算書情報・決算書情報検索 [*11]

- 利点：PDF，Excel，XML のファイルがある．
- 欠点：予算の項目に政策体系という項目が存在しない（説明という項目はある）．閲覧できるブラウザは InternetExplorer のみであり，文字コードは Shift JIS である．

予算情報は「日付」「予算項目」「前年度予算」「今年度予算」のように，共通の項目があり，数値データであることから，「予算書情報・決算書情報検索」の Excel や XML のように構造化データとして公開可能といえる．我々は，構造化データである XML を用いることを前提に進めていたが，公開されている XML の属性名が予算項目と対応しておらず，簡単に変換できないことから，利用を断念した．本研究では，PDF ファイルであるが，低コストで，データ整形が可能な「**予算概要**」を利用することとした．他にも，「予算概要」を選んだ理由としては，国，都道府県，市区町村においても，ほぼ同じ内容のファイルが存在するためである．最終的には，全国の 1,788 の自治体を対象とすることを目指しているため，予算情報の形式も，国，都道府県，市区町村の自治体で差がでないように進めている．

4.1 予算のデータ形式

本研究では，異なる自治体の予算情報でも，同じデータ形式としている．ポイントは，最も詳細に記述されている予算タイトルを「**予算項目**」として，**categories** により，

[*7] https://www.mhlw.go.jp/wp/yosan/yosan/21syokan/dl/01-01.pdf
[*8] https://www.mhlw.go.jp/wp/yosan/yosan/20hosei/02index.html
[*9] https://www.mext.go.jp/a_menu/kaikei/zaimu/1234699.htm
[*10] https://www.mhlw.go.jp/wp/yosan/other/r02/index.html
[*11] https://www.bb.mof.go.jp/hdocs/bxss010br2.html

予算項目の上位階層の異なるタイトル数に対応している点である．例えば，小樽市は上位階層に 2 つ，茨城県は上位階層に 2〜3 つ，福岡市は上位階層なしとなる．

表 1 予算のデータ形式

フィールド名	説明
budgetId	予算の識別子
budgetTitle	予算タイトル
typesOfAccount	会計種別（一般会計，特別会計）
department	管轄省庁・部局名
url	元データの URL
budgetItem	予算項目
categories	上位階層，カテゴリ (詳細→概要の順)
budget	今年度予算
budgetLastYear	前年度予算
description	説明部分
budgetDifference	比較増減額

5. 会議録とは

国会および地方議会の会議録は，いつ，どこで，だれが，なにを発言したのかを，そのまま書き起こして記録していることから，**一次情報**として利用できる貴重な言語資源である．しかしながら，各自治体でウェブ上に公開されている会議録は，話し言葉であるため，読みづらく，自治体によって公開形式が異なるため，利用しづらいという問題がある．

例えば，47 都道府県の議会会議録は，それぞれの自治体のウェブサイトに会議録の全文検索システムを公開しており，全文検索システムを提供する主要 4 社が 89%（=42/47）のシェアを持つため，ある程度，統一されているようにみえる．しかしながら，自治体ごとに検索システムや表示部分をカスタマイズしているため，同じ検索システムでも，自治体ごとに人手による確認が必要となる [16]．また，主要 4 社以外のシステムによって会議録を公開している 5 県については，それぞれ対応する必要がある．さらに，市町村の議会会議録の場合には，検索サービスを利用していない自治体も多く，PDF のみで公開していることから，そのままテキスト処理をすることができないことも多い．本研究では，自治体ごとに異なる形式を，少ない作業量で，同じデータ形式に整形しつつ，Budget Argument Mining タスクに利用する．

5.1 会議録のデータ形式

本節では，国会会議録を対象として，予算項目と議論を結びつける **Budget Argument Mining** タスクの具体的なデータ形式について述べる．国と地方自治体では，公開されている会議録のフォーマットが異なる．国会では，地方議会では，特に，お金のある自治体では，また，国では表記揺れが存在しないが，都道府県，市町村，になるにつれて，表記揺れや誤りも含まれやすくなる[*12]．

どのように予算項目と議論を結びつけるのか

本研究では，国会，および，自治体で予算審議の予算項目を「予算概要」から取得し，会議録に含まれる予算項目に関する政治家の議論と結びつける．ここで，予算項目に関する議論とは，(全ての発言を対象にするのではなく) 会議録に含まれる金額表現を手がかりに，**金額表現を含む発言を議論の構成要素**として，議論ラベルを付与するとともに，予算項目と対応づける．

具体的には，国会，および，地方議会の会議録に含まれる金額表現と予算項目を対応づけるために，下記の項目を追加する．

- moneyExpression：(金額表現)
- relatedID：(関連する ID)
- argumentClass：(議論ラベル)

国会会議録，および，地方議会会議録のデータ形式については，次節以降で説明する．

5.2 国会会議録

国会会議録 API による収集

国会会議録は，会議録を取得するための外部提供インターフェイスである検索 API を用いて収集する[*13]．検索用 API は，(1) 会議単位簡易出力，(2) 会議単位出力，(3) 発言単位出力　の 3 種類がある．検索用 API を利用することで，検索リクエストに対し，XML 形式，または，JSON 形式でデータを取得することができる．本研究では，議長や議員の発言を全て取得する必要があるため，(2) 会議単位出力を用いて JSON ファイルを取得し，Budget Argument Mining に必要な情報を追加することとした．

図 2 は，国会会議録のデータ形式である．SpeechRecord は，SpeechID ごとに議員の発言 (speech) が記録されている．その発言 (speech) から GINZA[10] により金額表現を抽出し，moneyExpressions という金額表現に関する新たな項目を追加している．

5.3 地方議会会議録

本研究では，全自治体への拡張を見据えて，ウェブ公開形式が異なる 3 つの自治体（北海道小樽市，茨城県，福岡県福岡市）の議会会議録を対象とする．茨城県，福岡県福岡市の議会会議録は，検索システムを用いて HTML ファイルで公開されているが，北海道小樽市の議会会議録は，PDF ファイルで公開されており，HTML ファイルで公開されていない．本タスクでは，3 つの自治体の会議録を表 2 のようなデータ形式に変換して，統一したデータ形式の

[*12] 国会には表記揺れがないように，辞書が存在する．
[*13] https://kokkai.ndl.go.jp/api.html

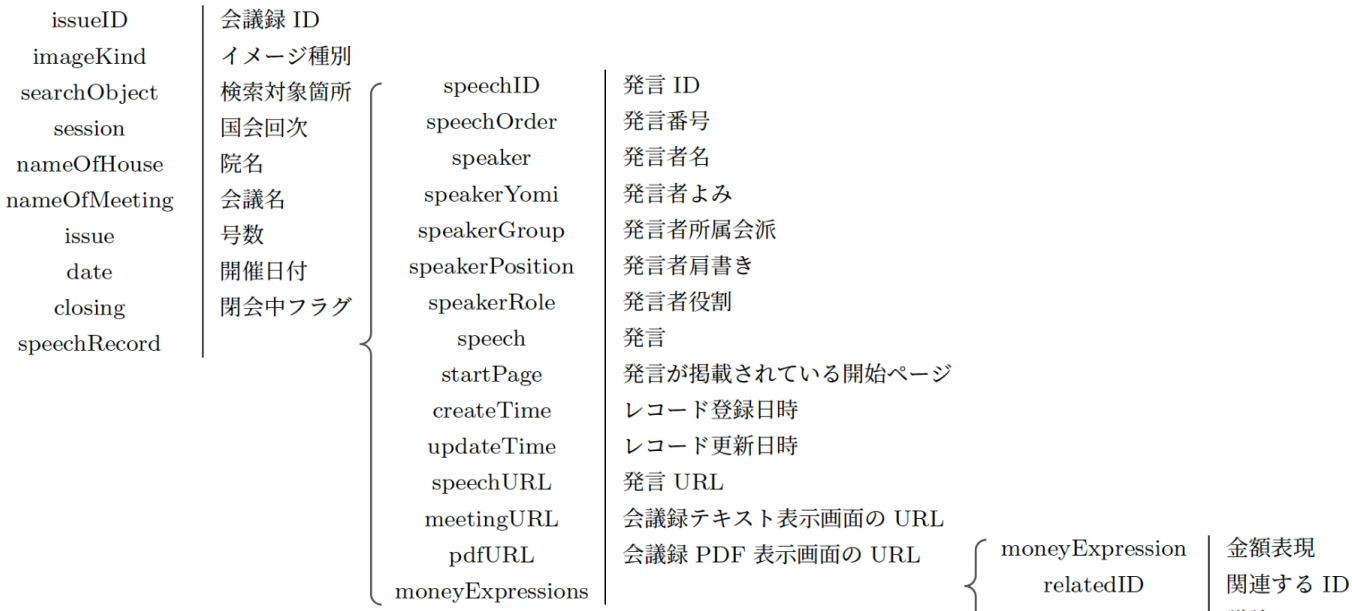

図 2 国会会議録のデータ形式

JSON ファイルとする．

表 2 地方議会会議録のデータ形式

フィールド名	説明
date	日付
localGovernmentCode	自治体コード（6 桁）
localGovernmentName	自治体名
proceedingTitle	議会名
url	URL
proceeding	発言者と発言内容
└ speakerPosition	役職
└ speaker	発言者
└ utterance	発言内容
└ moneyExpressions	発言に含まれる金額表現
└ moneyExpression	金額表現
└ relatedID	関連する予算 ID リスト
└ argumentClass	議論ラベル

6. アノテーション

6.1 アノテーションの目的

Budget Argument Mining のアノテーションは，会議録を対象として，政治家の発言から予算に関する発言（金額表現を含む発言）をみつけ，2 つの議論ラベル「Claim（主張）」「Premise（根拠）」を付与すること，そして，予算項目と対応づけることである．

予算に関する発言とは

本研究では，予算に関する発言の候補ととして，会議録に含まれる全ての発言を対象とせず，**金額表現を含む発言**を対象とする．金額表現を含む発言を対象とすることで，予算項目と議論の対応付けを効率的に行うことが可能となる．金額表現は，固有表現抽出器 GINZA[10] を用いて，予め，MONEY とラベル付けされた箇所を自動で抽出する．議会会議録に含まれる金額表現は，必ずしも，予算項目に記載された表現と完全に一致しない．例えば，図 1 のように，予算総括表に記載されている金額「287,108 千円」と議会会議録に記載されている金額「約 2 億 8700 万円」のように表現が異なる．他にも，異なる金額表現で同じ予算項目となる表現としては，合計金額，事業単位による表現，日割，月割，一人あたりによる負担金などがある．本タスクでは，金額表現の表記揺れ，曖昧性の問題を解決し，予算に関する発言を予算項目と対応づける．

議論ラベル (argumentClass) とは

議論ラベルとは，議会会議録に含まれる議員の発言に対して，議論の流れを理解するために必要となるラベルのことである．argumentClass (議論ラベル) は，「Claim（主張）」「Premise（根拠）」に加えて，「その他」と金額表現ではない場合の「金額表現ではない」ラベルを準備した．事前の調査において，主張よりも，根拠となる Premise が数多く含まれることを確認したため，Premise(根拠) は，時間軸で「過去」と「未来」に分け，どちらにも当てはまらない場合を「その他」とした．Claim(主張) は，意見・提案・質問による主張をまとめており，それ以外を「その他」とした．

アノテーションの基準とは

本節は，どのように argumentClass を付与するのかについて，説明する．議論の構成要素は，文，あるいは，節の単位で付与される．議会会議録に含まれる発言は，ひとつの発言が長いことから，複文が数多く存在することが確認されている [17]．そのため，ひとつの発言に「Claim（主張）」「Premise（根拠）」が含まれることもある．一方で，

「節[*14]」は，一つの述語ごとで区切ることになり，短くなりすぎることがある．また，注釈者自身に節を判断してもらうことは，判断に揺れが生じて，作業負担が大きくなる．そこで，本研究における「argumentClass」を付与する単位は，「**読点による区切り**」を目安として，「節」以上「文」以下の範囲とする．

下記に，argumentClass の説明と例文を示す．

(1) **Premise：過去**は，前年度予算，あるいは，執行済みの金額表現を含む根拠が記述されている箇所である．

小樽市平成 31 年度第 1 回定例会

次に, ふるさと納税に伴う本市の個人市民税の減収額につきましては，平成 29 年 1 月から 12 月に行われた寄附により，30 年度の課税に反映された額で申しますと**約 4,500 万円**となります．

(2) **Premise：未来**は，今年度予算，あるいは，見積の金額表現を含む根拠が記述されている箇所である．

小樽市平成 31 年度第 1 回定例会

次に，ふるさと納税関係経費が **4,670 万円**計上されていますが，...

(3) **Premise:その他**は，過去や未来に含めることができない根拠が記述されている箇所であり，例示・訂正事項などである．

小樽市平成 31 年度第 1 回定例会

大体，賦課限度額の国のモデル世帯というのは，収入で言うと，年金でさえ **1,000 万円**ですよ．

(4) **Claim ： 意見・提案・質問**は，金額表現を含む主張（意見・提案・質問）が記述されている箇所である．

小樽市平成 31 年度第 1 回定例会

子供は年齢が低いほど病院にかかることが多く，最低でも小学校卒業までの外来分を**無料**にすることで大いに子育て支援になります．

(5) **Claim ： その他**は，上記以外の金額表現を含む主張が記述されている箇所である．

(6) **金額表現ではない**

小樽市平成 31 年度第 1 回定例会

ただ小樽市全体では今，**18 万立方メートル**しかありませんので，...

(7) **その他**

[*14] 節は「複文を構成するところの，述語を中心とした各まとまり」である [18][19]．複文とは (1つの述語で1つの文を構成する単文でなく) 2つ以上の述語が含まれる文である．

6.2 アノテーションの進め方

アノテーションは，会議録の JSON ファイルに含まれる「relatedID」「argumentClass」の値を埋めることである．下記の手順で，アノテーションを行う．

(1) 国会，地方議会の会議録を対象として，金額表現を含む発言に「**argumentClass**（Premise, Claim, 金額表現ではない, その他）」を注釈付けする．

(2) 会議録に含まれる金額表現に対応する予算の識別子「**budgetId**」を予算情報からみつけ，会議録の「**relatedID**」の欄に予算の識別子を記述する．

会議録は，JSON 形式になっており，日付，自治体コード（6 桁），自治体名，議会名，URL，発言者と発言内容に分けられている．発言に金額表現含まれる場合，金額表現 moneyExpression を GINZA[10] により自動で抽出し，releted ID と argumentClass を空欄としている．

6.3 アノテーションツール

アノテーションツールは，第 3 著者の乙武氏が作成しており，本タスク専用である．本ツールはブラウザ上で動作する形態であるが，データを管理するサーバは存在せず，注釈者自身がファイルを管理する特徴がある．また，ツールの配置先で差し替えるだけで，注釈者全員にツールの更新版をリアルタイムに提供できる利点がある．アノテーション作業を進めていく過程で，注釈者の要望に応える形で，アノテーションのラベルの変更に加えて下記に示すツールの機能改善を行った．

- アノテーションの総数，および，未処理数の表示
- フィルタリングによる検索
- 重要箇所のハイライト，および，詳細情報の非表示
- アノテーション結果の確認[*15]

図 3 はアノテーションツールを用いて「ふるさと納税関係経費」にタグ付けする例である．アノテーションツールは，会議録の JSON ファイルと予算情報の JSON ファイルの 2 つを読み込み，**左側に会議録**，**右側に予算情報**を表示する．この例では，「ふるさと納税関係経費」に関する金額表現「4670 万円」に議論ラベル「Premise：未来（現在以降）・見積」を付与している．金額表現と結びつける予算項目の数は多く，budgetTitle だけでなく description まで見ないと判断できない例も少なくないことから，本アノテーションツールでは予算情報内の budgetTitle と description を対象としたフィルタリング・ハイライト機能を設けている．図 3 において，上部メニューバー右にあるテキストボックスに文字列を入力することで，その文字列を含む budgetTitle または description の予算項目のみが表示され，当該文字列がハイライトされる形で表示される．

[*15] relatedID をクリックすることにより，予算項目を確認可能となる

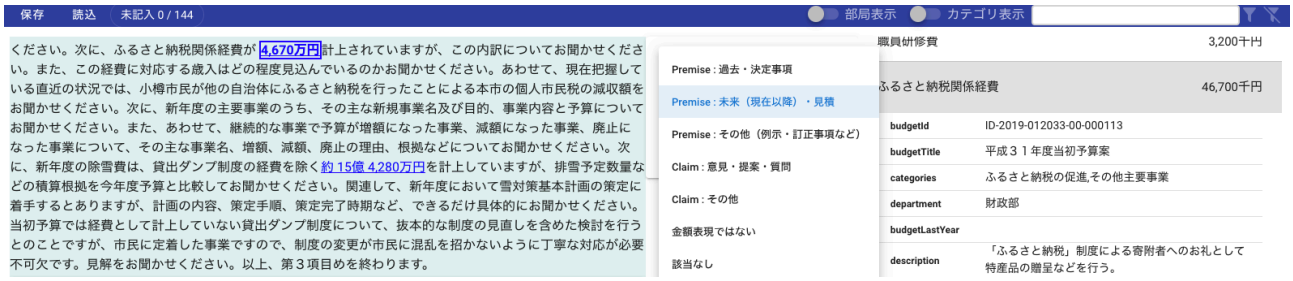

図 3 アノテーションツールを用いて左側に表示されている会議録の金額表現「4,670 万円」に「Premise : 未来（現在以降）・見積」のラベルを付与して，右側の予算情報に含まれる予算項目「ふるさと納税関係経費」を対応づける例

6.4 対象データ

対象期間は，2019 年と 2020 年とする．対象年を 2019 年と 2020 年にすることにより，「コロナ前」と「コロナ後」の比較が可能となる．地方議会会議録は，小樽市，茨城県，福岡市であり，2019 年，および，2020 年の第 1 回定例会とする．地方議会の予算情報は，各自治体の予算概要である．国会会議録は，第 201 回衆議院予算委員会において，第 2 次補正予算についての審議をしている 2020 年 6 月の会議録を対象とする．国会は，予算情報が省庁ごとに存在する．本研究では，最初に，厚生労働省の予算情報を対象にする．

6.5 アノテーションの結果

本研究では，2 名，あるいは，3 名の注釈者により，同一の会議録を対象にアノテーションを行った．複数人によるアノテーション結果は，Kappa 係数を用いて一致率を計算する．

Kappa 係数は，アノテーションをした場合のデータ間の一致度を評価する指標として用いられる [20][21]．Kappa 係数は，実測値と期待値の比較によって，偶然による一致の可能性を排除した上で，下記の式で算出される．

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o は，データの判定が実際に一致した割合であり，P_e は，データ間の独立を仮定した上で偶然に一致が期待される割合である．

Kappa 係数の見方としては，低い一致率の Poor(<0.00)，ごく軽度の一致の Slight(0.00-0.20)，軽度の一致の Fair(0.21-0.40)，中程度の一致の Moderate(0.41-0.60)，高度の一致の Substantial(0.61-0.80)，ほぼ完璧な一致の Almost Perfect(0.81-1.00) という基準がある [22]．本研究では，R の irr ライブラリを用いて Kappa 値を計算している[*16]．

注釈者は，大学生，大学院生，および，大学教員である．Kappa 値を計算するために，注釈者を A〜J で識別する．

[*16] https://cran.r-project.org/web/packages/irr/irr.pdf

表 3 に注釈者間の Kappa 値を示す．アノテーションの結果として，自治体，年度，対象議会，金額表現が含まれる対象数，注釈者，Kappa 値を記述している．下記に，国会，および，地方自治体を対象とする注釈者示す．

- 小樽市：A, B, C
- 茨城県：D, E, F, G, H
- 福岡市：I, B, C
- 国会　：A, J

表 3 アノテーションの結果

対象	年度	対象議会	対象数	注釈者	Kappa
小樽市	2019	定例会第 1 回	334	A,B	0.539 ⋆
小樽市	2019	定例会第 1 回	334	B,C	0.591 ⋆
小樽市	2019	定例会第 1 回	334	C,A	0.568 ⋆
小樽市	2020	定例会第 1 回	208	A,B	0.506
小樽市	2020	定例会第 1 回	208	B,C	0.561
小樽市	2020	定例会第 1 回	208	C,A	0.469
茨城県	2019	定例会第 1 回	181	D,E	0.656 ⋆⋆
茨城県	2019	定例会第 1 回	181	E,F	0.577 ⋆⋆
茨城県	2019	定例会第 1 回	181	F,D	0.497 ⋆⋆
茨城県	2020	定例会第 1 回	163	D,G	0.755 ⋆⋆
茨城県	2020	定例会第 1 回	163	G,H	0.885 ⋆⋆
茨城県	2020	定例会第 1 回	163	H,D	0.747 ⋆⋆
福岡市	2019	定例会第 1 回	334	I,B	0.314
国会・衆議院 (厚労省予算)	2020	予算委員会 6 月 8 日	20	A,J	Nan ⋆⋆⋆

⋆ 第 1 回定例会の 1 日目のみ，注釈者 A,B,C は協議しながら進めた．
⋆⋆ 注釈者 D, E, F, G, H は協議しながら進めた．
⋆⋆⋆ 注釈者 A,J が全て Premise : 未来と付与した．

注釈者ごとのラベルの付与数

ここでは，注釈者ごとの argumentClass ラベルの付与の違いを明らかにする．表 4 には，小樽市を対象にした注釈者による argumentClass と relatedID を付与した Training data の結果を示す．本タスクでは，第 1 回定例会の 3 日目，4 日目を Test data として，残りの日を Training data として公開する．

表 4　小樽市議会会議録を対象にした各注釈者の argumentClass および relatedID の注釈数 (Training data)

	注釈者	argumentClass								relatedID
		Premise			Claim		その他	金額表現ではない	合計	注釈数
		未来	過去	その他	意見・提案・質問	その他				
2019	A	88	29	2	18	0	0	7	144	27
	B	107	31	2	0	0	0	4		28
	C	103	19	17	1	0	0	4		29
2020	A	56	10	4	13	0	0	2	85	6
	B	68	9	4	3	0	0	1		16
	C	71	5	7	1	0	0	1		13

小樽市	Training		Test		Training
	1 日目	2 日目	3 日目	4 日目	5 日目
2019	2/20	2/25	2/26	2/27	3/14
2020	2/19	2/25	2/26	2/27	3/13

6.6 考察

Premise と Claim について

表 4 からわかるように，金額表現を含む箇所は，Claim に比べて，Premise が多い．これは，過去の執行額，予算の見積額などの**金額表現を含む箇所が Premise となり，その後の文が Claim になる**傾向にあるためである．

例えば，下記の例では，金額表現を含む発言「**百七十九億円**」が Premise となり「本当ですか.」が Claim となる．

国会会議録の例

先ほど，**百七十九億円**が電通ライブに残っているとおっしゃいましたね．本当ですか.

今後，Premise と Claim の組み合わせを考慮して，金額表現を含まない Claim の抽出を検討する．

知事の発言について

1. はじめにでも述べたとおり，地方自治体の予算は，首長により予算案が作成され，議会で審議された後に成立する．そのため，首長が予算案を成立させたいと考えているのは，明らかであり，共通認識といえる．このような状況では，首長の表現に主張している表現がほとんど存在せず，**Claim 表現の省略が行われている可能性がある**．つまり，知事の発言は，金額を説明しつつ，予算案の提案をしていることから，主張を含んでいるとも考えられる．今後，知事の発言に対するラベルについて検討する．

アノテーションの判断に必要な範囲について

argumentClass を付与する場合に，判断するために，どの程度読むべきか決める必要があり，スコープの問題といえる．スコープは，argumenClass のようなラベルの影響が及ぼす範囲である [23]．本稿で述べたアノテーションでは，スコープの範囲が曖昧で，関連する範囲を読んで判断することとしていた．今後，議員の発言が切り替わるまでというように範囲を明確にすることもできるが，会議録における首長や代表質問者の発言がとても長いことから，議会の特徴を踏まえて，検討する必要がある．

予算と会議録に含まれる金額表現について

予算情報に含まれる予算額と会議録に含まれる金額が異なる場合がある．これは，議論において，予算項目の合計金額を述べたり，逆に，詳細な金額を述べたりしているためである．

例えば，下記の**非常時停電対策関係経費**は，「非常時停電対策関係経費（指定避難所）」のような予算項目が 10 項目あり，それらの合計金額が議会で述べられている．

小樽市議会会議録の例

防災力の強化を図る主な事業としましては，大規模停電に備えて，指定避難所となる小・中学校などへの非常用発電機の配備等を行う**非常時停電対策関係経費**が**1,949 万 6,000 円**，...

本タスクは，金額表現の異表記を考慮しつつ，予算項目と議論を結びつけることが特徴といえる．

金額表現の抽出について

表 4 からわかるように，金額表現の抽出が失敗していることがわかる．下記に，「金額表現ではない」と付与された例を示す．

- 50 万立法メート
- 1000 万円単位
- 要介護 2

一方で，金額表現を抽出してない例も存在する．本タスクでは，金額表現を自動抽出した表現だけを対象とすることにしている．

7. おわりに

本稿では，NTCIR16 QA Lab-PoliInfo-3 のサブタスクで

ある Budget Argument Mining のデータセット構築に向けた，データ形式の設計，アノテーションの方法，および，結果について述べた．Budget Argument Mining は，国，あるいは，自治体のウェブサイトで公開されている予算審議の項目に関連する議論をみつけて，結びつけることを目的としたタスクである．データセット構築のためのアノテーションでは，予算審議の情報（予算項目，金額，管轄省庁・部局名，説明）が与えられたときに，議会会議録に含まれる政治家の予算関連の発言（金額表現を含む発言）をみつけだし，３つの議論ラベル「Claim（主張）」「Premise（根拠）」「その他」を推定するために必要な注釈付けを行った．

今後，構築したデータセットは NTCIR16 QA Lab-PoliInfo-3 **Budget Argument Mining** タスクにおいて公開する．

謝辞

本研究は JSPS 科研費 21H03769，および，セコム科学技術振興財団の助成を受けたものである．

参考文献

[1] 横田茂. [研究ノート] 日本の予算制度と予算過程：その特質の形成. 關西大學商學論集, Vol. 64, No. 1, pp. 77–103, jun 2019.
[2] 名取良太, 田中智和, 岡本哲和, 石橋章市朗, 梶原晶, 坂本治也, 秦正樹. 地方議会の審議過程：テキスト分析による定量化の試み. mar 2020.
[3] 増田正. 計量テキスト分析によるわが国地方議会の審議内容を可視化する方法について. 地域政策研究, Vol. 19, No. 3, pp. 161–175, feb 2017.
[4] 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, Kenji Araki, Satoshi Sekine, and Noriko Kando. Overview of the ntcir-15 qa lab-poliinfo task. *Proceedings of The 15th NTCIR Conference*, 12 2020.
[5] 栗林樹生, 大内啓樹, 井之上直也, 鈴木潤, Paul Reisert, 三好利昇, 乾健太郎. 論述構造解析におけるスパン分散表現. 自然言語処理, Vol. 27, No. 4, pp. 753–779, 2020.
[6] 近藤崇宏, 鷲尾光樹, 林克彦, 宮尾祐介. 議論の構造化と妥当性評価のための bayesian argumentation-scheme networks の提案とアノテーションデータ作成. Technical report, 2021.
[7] 関根聡, 井佐原均. Irex：情報検索、情報抽出コンテスト. Technical report, 1998.
[8] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, jan 2002.
[9] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, May 2002.
[10] 松田寛. Ginza - universal dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.
[11] Akitaka Matsuo and Kentaro Fukumoto. Legislators' sentiment analysis supervised by legislators. Technical report, 2020.
[12] Keiichi Takamaru, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, and Noriko Kando. Extraction of the argument structure of Tokyo metropolitan assembly minutes: Segmentation of question-and-answer sets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2064–2068, Marseille, France, May 2020. European Language Resources Association.
[13] S. E. Toulmin. *The Use of Argument*, 1958.
[14] James B. Freeman. Dialectics and the macrostructure of arguments : a theory of argument structure. Technical report, 2011.
[15] John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, Vol. 45, No. 4, pp. 765–818, December 2019.
[16] 高丸圭一. 地方議会会議録コーパスと地方議会会議録を用いた学術研究の現状 (特集政治・経済とコトバ). 知能と情報 = Journal of Japan Society for Fuzzy Theory and Intelligent Informatics : 日本知能情報ファジィ学会誌, Vol. 31, No. 2, pp. 25–33, 4 2019.
[17] 葦原史敏, 木村泰知, 荒木健治. 地方議会会議録における節単位による議員の要望抽出. 電子情報通信学会論文誌, Vol. J98-D, No. 11, pp. 1390–1401, 11 2015.
[18] 益岡隆志, 田窪行則. 基礎日本語文法. くろしお出版, 1989.
[19] 丸山岳彦, 佐藤理史, 夏目和子. 現代日本語における節の分類体系について. 言語処理学会第 22 回年次大会予稿集, pp. 1113–1116, 3 2016.
[20] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, p. 37, 1960.
[21] 宮内拓也, 浅原正幸, 中川奈津子, 加藤祥. 『現代日本語書き言葉均衡コーパス』への情報構造アノテーションとその分析. 国立国語研究所論集, No. 16, pp. 19–33, oct 2018.
[22] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, 1977.
[23] 成田和弥, 水野淳太, 上岡裕大, 菅野美和, 乾健太郎. 誤り分析に基づく日本語事実性解析の課題抽出. 自然言語処理, Vol. 22, No. 5, pp. 397–432, 2015.