

議会会議録と予算表を紐づける Minutes-to-Budget Linking タスクの提案

木村泰知¹ 梶縁¹ 乙武北斗² 門脇一真³ 佐々木稔⁴ 小林暁雄⁵
¹小樽商科大学²福岡大学³株式会社日本総合研究所⁴茨城大学⁵農研機構
kimura@res.otaru-uc.ac.jp

概要

予算編成に関する議論の背景・過程・結果を理解するには、複数の文書や表を参照する必要がある。NTCIR-17 QA-Lab PoliInfo-4 のサブタスクの一つである **Minutes-to-Budget Linking** (以下 MBLink) では、議会における予算審議に焦点を当て、ある予算に関する発言に対して、関連する予算表のセルを紐づけるタスクを実施する。MBLink は、議会会議録のような非構造化データから、表形式データの予算表に含まれる特定のセルを紐づけることで、発言の根拠となる数値をみつけ、予算増減の理由をわかりやすく提供することを目指している。本稿では、MBLink のデータ形式、評価方法、アノテーション方法について述べる。

1 はじめに

地方自治体の役割のひとつに予算編成がある。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。しかしながら、審議結果である予算は、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。予算作成の議論を理解するには、参照しなければならない文書や表が数多く存在する [1][2]。例えば、対象年度の予算は、前年度の予算と比較しながら、説明される [3]。特に、正確な金額を確認する場合には**予算表**を参照することになる。

小樽市議会における予算説明の例 (図 1 も参照)

まず、**歳入**についてであります。市税につきましては、**個人市民税**、**法人市民税**などで減収が見込まれるものの、**固定資産税**、**都市計画税**などで増収が見込まれることから、**2.7%**、**3億 5,280万円増**の**135億 7,350万円**を見込みました。

ここで、小樽市議会における市長の説明を例に、予算に関する金額表現処理の難しさについて説明する。上記の例にあるように、市長は「歳入」となる「市税」「個人市民税」「法人市民税」「固定資産税」「都市計画税」の税収入の説明をしつつ、見込み額を述べている。「市税」「個人市民税」などの額は、文中に記載されていないものの、別の資料に記載されており、予算表を参照することで、正確な金額を確認できる。「**135億 7,350万円**」という金額表現は、前年度の市税の「**2.7%**」である「**3億 5,280万円**」の金額が増えることを見込んで計算された対象年度の見積額である。

議会における発言は、予算編成の経緯が述べられることもあるが、予算関連の書類が配布されたことを前提にしているため、詳細な数値や正確な金額が明言されないこともある。このような場合には、予算表を参照することで、金額や数値を確認することが可能である。また、予算の増減額や割合についても、予算表の金額と紐づけて、計算することができる。さらに、予算編成に至る経緯については、提案された予算金額とともに、理由を加えて説明していることが多い [3]。

そこで、本研究では、議会における予算審議の発言と予算表の関連項目を紐づける MBLink (Minutes-to-Budget Linking) タスクを提案する。MBLink の目的は、予算の根拠となる金額を明らかにするとともに、どのような理由で予算額が決定したのかを確認できるようにすることである。MBLink は、NTCIR-17 QA-Lab PoliInfo-4¹⁾ のサブタスクのひとつである [4]。NTCIR-17 は、NII が主催する評価型ワークショップであり、2022年7月から2023年12月まで開催される²⁾。本稿では、MBLink のデータ形式、評価方法、アノテーション方法について述べる。

1) <https://sites.google.com/view/poliinfo4>

2) <https://research.nii.ac.jp/ntcir/ntcir-17>

2 議会会議録と予算表

2.1 予算編成の流れ

自治体の予算は、次年度(4月以降)に実現すべき方向性を考えて、前年度の議会で決定される。予算編成のおおまかな流れは、前年度の後半から始まる。最初に、次年度に実現すべき方向性を考えて、組織内部の意思統一を図りつつ、予算編成方針が決定される。予算編成方針に基づいて、各部からの予算要求がある。それらの要求に対して、社会状況や国の動向を踏まえ、要求内容の追加・変更が行われる。1月頃に、各部から要求された内容について査定が行われ、予算案が公表される。2月上旬に開催される第一回定例会に予算案が上程され、議決を受ける。議会では、首長から予算案の説明された後で、さまざまな観点から審議が行われ、予算案を可決すべきか否かが決定される。

2.2 予算の説明箇所

首長による予算の説明箇所は、第一回定例会の1日目の冒頭に行われることが多い。例えば、平成31年(2019年)の小樽市議会³⁾の第1回定例会の議案は「予算」と「条例」に分けられ、第1号議案から第19号議案までが予算であり、第20号議案から第38号議案までが条例である。第1回の本会議⁴⁾の冒頭は、予算の提案から始まり、市長が予算の説明を行っている。予算説明されている文は、2ページ程度であり、35文であった。本研究では、首長が予算を説明している発言に焦点を当て、関連する予算表のセルと紐づける。

2.3 予算表の種類

前節で述べた通り、予算の審議は、第一回定例会において行われ、首長による提案から始まる。首長は、予算関連書類に記載されている金額を根拠としながら、説明することが多い。例えば、小樽市の予算情報は下記の8つの書類が公開されている^[3]。

- 1) 予算編成方針について
- 2) 予算総括表
- 3) 予算案のポイント
- 4) 主要事業
- 5) 部別要求額・予算額

3) <https://www.city.otaru.lg.jp/docs/2020112500241/>

4) <https://www.city.otaru.lg.jp/docs/2020113000634/>

- 6) ふるさと応援寄附金の使い道
- 7) 予算書
- 8) 予算説明書

梶らの調査^[3]によれば、予算情報は「予算総括表」「予算案のポイント」「部別要求額・予算額」「予算書」と対応づけられることが多い。予算総括表は、一般会計、特別会計、企業会計に対する前年度と今年度の予算額が記述されている表である。予算案のポイントは、財政部が予算編成のポイントを図や表を含めて説明した文書である。部別要求額・予算額は、小樽市の総合計画に沿って、各事業に対する予算額をまとめた表である。予算書は、一般会計、特別会計、企業会計に分けて、歳入と歳出が表として記載されており、約40ページである。

3 Minutes-to-Budget Linking

3.1 目的

MBLinkの目的は、議会における発言と予算表の数値を紐づけることで、根拠となる金額を明らかにするとともに、どのような理由で予算額が決定したのかを確認できるようにすることである。

3.2 データセット

MBLinkの入力、出力、評価方法を下記に示す。

入力	予算表 (HTML) 議会会議録 (HTML)
出力	会議録の発言と予算項目の紐付け (JSON)
評価	議会会議録から予算表への連結、および、理由の有無についての評価 (F値)

MBLinkでは、議会会議録中のある文が予算表の特定のセルの金額に関連するものであれば、その文にそのセルのIDを付与し、さらにその文に予算の主張に関する理由が含まれているか否かの判定結果も付加する。入力は、予算表と議会会議録であり、HTMLファイルとして与えられる。与えられるHTMLは、図1の例にあるように、IDが付与されたHTMLである。出力は、予算表に関連するセルがあれば、議会会議録に含まれる文IDと予算表に含まれるセルIDを紐づけて、JSONファイルとして返す。また、理由が述べられている場合には“containsReason”を“True”とする。評価は、議会会議録から予算表への連結、および、理由の有無についての性能評価としてF値を用いる。

まず、歳入についてであります、

市税につきましては、個人市民税、法人市民税などで減収が見込まれるものの、

固定資産税、都市計画税などで増収が見込まれることから、2.7%、3億5,280万円増の135億7,350万円を見込みました。

「理由」

× 100.0% = 132億2千万円

× 2.7% = 3億5千万円

2018年 予算書			2019年 予算書								
款	項	金額	款	項	金額						
1	市	税	13,220,700	1	市	税	13,573,500				
	1	市	民	1	市	民	税	減	5,600,700		
	2	固	定	2	固	定	産	税	増	5,773,100	
	3	軽	自	3	軽	自	動	車	税	180,700	
	4	た	ば	4	た	ば	こ	税	914,500		
	5	特	別	5	特	別	地	保	有	1,000	
	6	入	湯	6	入	湯		税	22,800		
	7	都	市	7	都	市	計	画	税	増	1,081,500

入力（配布ファイル：HTMLファイルにIDが付与されたもの）

```
<p>
<span data-mblink-sentence-id="011002-2019-doc54321-sent30">まず、歳入について
ありますが、市税につきましては、個人市民税、法人市民税などで減収が
見込まれるものの、固定資産税、都市計画税などで増収が見込まれることから、
2.7%、3億5,280万円増の135億7,350万円を見込みました。</span>
</p>
```

```
<table data-mblink-table-id="011002-2019-doc12345-tab2">
</tr>
<tr>
<td data-mblink-cell-id="011002-2019-doc12345-tab2-r2c1">市税</td>
<td data-mblink-cell-id="011002-2019-doc12345-tab2-r2c2"></td>
<td data-mblink-cell-id="011002-2019-doc12345-tab2-r2c3">-13,573,500</td>
</tr>
</table>
```

出力（提出ファイル：IDを紐付けた JSONファイル）

```
{ "sentenceID": "011002-2019-doc54321-sent30", "linkedCellIDs": [ "011002-2019-doc12345-tab2-r2c1", "011002-2019-doc12345-tab2-r2c3" ], "containsReason": true }
```

図1 市税について「対象年度」「前年度」の税収入を参照しつつ説明している箇所に対して、アノテーションする例

4 アノテーション

本節では、アノテーションの流れを説明する。

4.1 アノテーションの準備

自治体の公開している予算表は、PDFで公開されていることが多い。そのため、アノテーションを行う前に、PDFからHTMLへ変換する。HTMLへの変換は、変換ツールとしてABBYY FineReader PDF⁵⁾を用いて自動で変換した後で、下記の処理をすることで、人手でデータ整形をしている。

1. 結合されているセルを分割する
2. 必要があれば、セル内で改行する
3. 余計なスペースを削除する
4. 明らかな文字認識の誤りを訂正する

4.2 対象データ

対象データは、小樽市、福岡市、茨城県の議会とする予定である。現時点では、小樽市の2019年および2020年の第一回定例会の市長の予算説明に関する発言を対象としてアノテーションを進めている。

- 2019年 第1回定例会 第1日目⁶⁾

5) <https://pdf.abbyy.com/>

6) https://www.city.otaru.lg.jp/docs/2020113000634/file_contents/r01-02.pdf

- 2020年 第1回定例会 第1日目⁷⁾

予算表は、予算説明で参照されることが多い「予算総括表」「予算案のポイント」「部別要求額・予算額」「予算書」に含まれる表とする。

4.3 アノテーションの方法

アノテーションは、図2の例にあるような独自で開発したツールを用いて行う。アノテーションツールは、左側に会議録、右側に予算表が見える状態になっており、予算に関する発言と関連のある表のセルを紐づけることができる。梶らの調査[3]と同様に、予算編成の根拠となる「金額」「割合」「増減」「参照」に着目したアノテーションが可能であり、理由の有無も付与できることを確認した。

5 関連研究

5.1 テキストと表の連結

テキストと表を紐づける研究には、FEVEROUS (Fact Extraction and VERification Over Unstructured and Structured information) [5]がある。FEVEROUSは、Wikipediaに記述されたClaim(主張)の真偽を検証するために、Evidence(根拠)をWikipediaの表から

7) https://www.city.otaru.lg.jp/docs/2020113000634/file_contents/02-01.pdf

の当初予算と比較して説明申し上げます。

まず、**歳入**についてであります。市税につきましては、**個人市民税、法人市民税**などで**減収**が見込まれるものの、**固定資産税、都市計画税**などで**増収**が見込まれることから、**2.7%、3億5,280万円増の135億7,350万円**を見込みました。

地方交付税につきましては、国の地方財政計画の伸び率などを基本に、本市の特殊事情を勘案しながら積算し、**臨時財政対策債**を加えた実質的な地方交付税では、**1.7%、2億9,500万円減の170億5,100万円**を見込みました。

地方消費税交付金につきましては、**3.2%、7,800万円増の25億700万円**を見込みました。

また、**歳出**の主なものについて経費別に申し上げますと、いわゆる**義務的経費**では、**人件費が0.6%の減、公債費が4.3%の減**となりましたが、**扶助費において、市内幼稚園の新制度幼稚園・認定こども園への移行に伴う教育・保育給付費負担金の増**などにより、**2.1%の増**となったことから、合計では**0.3%の増**となり、**歳出合計に占める義務的経費の割合は、前年度を1.9ポイント下回る56.1%**となりました。

市税		歳入		
参照	▼	款	項	金額
				千円
sent85		1市	税	13,573,500
メモ			1市	民
			税	5,600,700
			2	固
			定	資
			産	税
			5,773,100	
			3	軽
			自	動
			車	税
			180,700	
			4	た
			ば	こ
			税	914,500
			5	特
			別	土
			地	保
			有	税
			1,000	
			6	入
			湯	税
			22,000	
			7	都
			市	計
			画	税
			1,081,500	
		2	地	方
			譲	与
			税	327,501
			1	地
			方	揮
			発	油
			譲	与
			税	81,000
			2	自
			動	車
			重	量
			譲	与
			税	215,000
			3	地
			方	道
			路	譲
			与	税
			1	
			4	森
			林	環
			境	護
			与	税
			7	500

図2 MBLinkのアノテーション画面の例

みつけるタスクである。例えば、Claimに「Roberto Ficoは、総投票数の57.6%で57,119票を獲得しました。」と入力文が与えられた場合、表から投票数が「61,819」であることをみつけ、Verdict(評決)としてはRefuted(反論)に分類する。

議会会議録と異なる言語資源を連結する研究は、NTCIR-15 QA-Lab PoliInfo-2のEntity Linkingで行われている[6]。Entity Linkingは、議会で議論されている法律名、あるいは、条例名に対して、関連するWikipediaの記事と紐づけるタスクである。例えば、「働き方改革を推進するための関係法律の整備に関する法律」という正式名称が会議録中に出てきた場合、Wikipediaの「働き方改革関連法」という記事と紐づけている。

他にも、NTCIR-16 QA Lab-PoliInfo-3[7]には、予算を対象にした議論マイニングとして、議会会議録のテキストに含まれる金額表現を予算表と紐づけるBudget Argument Miningタスクがある[1][2][8]。しかしながら、Budget Argument Miningは、議論ラベルの分類に対する比重が大きく、会議録と予算表を紐づける点に研究の余地がある。例えば、紐づける対象が金額表現のみとされていたことや複数存在する予算表のうち一つのみを対象としていたことにより、紐づけることが困難なこともあった。

5.2 表の理解

表構造解析に関するshared taskには、ICDAR 2013 Table competition[8]⁹⁾がある。Table competitionでは、

8) <https://github.com/polinfo3>

9) <https://iapr.org/archives/icdar2013/>

2つの政府機関(.europa.eu/.gov)の発行した文書PDFから表を収集したデータセットを作成している。

表の理解に関する研究は現在も注目されており、国際会議KDD21では、表の理解に関するチュートリアルとして“From Tables to Knowledge: Recent Advances in Table Understanding”が開催された¹⁰⁾。表の理解には、表のタイプ分類[9]、表のエリア抽出[10]、表の意味構造の解析[11]、セルの表記統一などの課題がある。

NTCIR-17 Understanding of non-Financial Objects in Financial Reports (以下、UFO)¹¹⁾では、有価証券報告書を対象として、表形式のデータや文書から構造化情報を抽出する技術を開発することを目的としたタスクが提案されている[12]。UFOタスクには2つのサブタスクとして、表データ抽出(TDE; table data extraction)サブタスク、ならびに、テキストと表の関係抽出(TTRE; text-to-table relationship extraction)サブタスクがある。

6 おわりに

本稿では、議会における予算審議の発言と予算表の関連項目を紐づけるMBLinkタスクを提案した。また、MBLinkのデータセットの形式、アノテーションの方法について述べた。

今後は、NTCIR-17 QA-Lab PoliInfo-4のサブタスクとして、MBLinkを進めるとともに、有価証券報告書を対象にしたNTCIR-17 UFOタスクとの連携も考えている。

10) <https://usc-isi-i2.github.io/KDD21Tutorial/>

11) <https://sites.google.com/view/ntcir17-ufo>

謝辞

本研究は JSPS 科研費 21H03769, 22K12740 の助成を受けたものである。

参考文献

- [1] Yasutomo Kimura, Hokuto Ototake, and Minoru Sasaki. Budget argument mining dataset using Japanese minutes from the National Diet and local assemblies. **LREC 2022**, 2022.
- [2] 木村泰知, 永渕景祐, 乙武北斗, 佐々木稔. 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. 情報処理学会 第 249 回自然言語処理研究会, 2021.
- [3] 梶縁, 木村泰知. 地方自治体の予算を対象にした金額表現の分析. 第 38 回ファジィシステムシンポジウム (FSS2022), 9 2022.
- [4] 小川泰弘, 木村泰知, 渋木英潔, 乙武北斗, 内田ゆず, 高丸圭一, 門脇一真, 秋葉友良, 佐々木稔, 小林暁雄. NTCIR-17 QA Lab-PoliInfo-4 のタスク設計. 言語処理学会第 29 回年次大会, 2023.
- [5] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) shared task. In **Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)**, pp. 1–13, Dominican Republic, 2021. Association for Computational Linguistics.
- [6] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyosi 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**, 12 2020.
- [7] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. Overview of the NTCIR-16 QA Lab-PoliInfo-3 task. **Proceedings of The 16th NTCIR Conference**, 6 2022.
- [8] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. ICDAR 2013 Table Competition. In **2013 12th International Conference on Document Analysis and Recognition**, pp. 1449–1453, 2013.
- [9] Kyosuke Nishida, Kugatsu Sadamitsu, Ryuichiro Hishinaka, and Yoshihiro Matsuo. Understanding the semantic structures of tables with a hybrid deep neural network architecture. In Satinder Singh and Shaul Markovitch, editors, **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA**, pp. 168–174. AAAI Press, 2017.
- [10] Kexuan Sun, Harsha Rayudu, and Jay Pujara. A hybrid probabilistic approach for table understanding. In **Conference on Artificial Intelligence (AAAI)**, 2021.
- [11] Binh Vu, Craig Knoblock, Pedro Szekely, Jay Pujara, and Minh Pham. A graph-based approach for inferring semantic descriptions of Wikipedia tables. In **International Semantic Web Conference**, 2021.
- [12] 木村泰知, 近藤隆史, 門脇一真, 加藤誠. UFO: 有価証券報告書の表を対象とした情報抽出タスクの提案. 人工知能学会第二種研究会資料, Vol. 2022, No. FIN-029, pp. 32–38, 2022.