

# 要約の自動評価手法としての Doc2Vec の利用

渋木 英潔<sup>†1</sup> 木村 泰知<sup>†2</sup> 乙武 北斗<sup>†3</sup> 石下 円香<sup>†1</sup> 秋葉 友良<sup>†4</sup>  
小川 泰弘<sup>†5</sup> 横手 健一<sup>†6</sup> 森 辰則<sup>†7</sup> 神門 典子<sup>††8</sup>

<sup>†1</sup>国立情報学研究所 <sup>†2</sup>小樽商科大学 <sup>†3</sup>福岡大学 <sup>†4</sup>豊橋技術科学大学  
<sup>†5</sup>名古屋大学 <sup>†6</sup>日立製作所 <sup>†7</sup>横浜国立大学 <sup>†8</sup>総合研究大学院大学

## 1 はじめに

自動要約の研究に置いて、要約の評価は改善の方向性を決める重要な指標である。しかしながら、人手による評価は多大な時間と労力が必要となり、研究サイクルを促進するためにも低コストで運用できる自動評価手法が望まれている。そのような背景から、2004年に Lin[1] が提案した ROUGE は、正解となる参照要約を用意するだけでよいという簡便性から、現在、要約の自動評価指標として広く用いられている。一方で、ROUGE によるスコアが人間の評価の代わりとして適切であるかどうかという議論は現在も続いており、ROUGE の適切性に関する研究 [2] がなされている。

2013 年に Mikolov ら [3, 4, 5] が提案した単語分散表現 (Word2Vec) は、その有用性から多くの分野で広く使われるようになり、単語から文書に拡張した Doc2Vec[6] も提案された。また、gensim[7]<sup>1</sup>などのライブラリを利用することで、Word2Vec や Doc2Vec を誰もが簡単に実装できるようになっている。Campr ら [8] は、ROUGE が類義語などの単語の意味を考慮しないことから、文書間類似度の観点から分散表現を含む5つの言語モデル (TfIdf, LSA[9], LDA[10], Word2Vec, Doc2Vec) との比較を行った。彼らは、レストランのレビューを対象として、人手で要約を作成し、また、要約間の類似度を5段階で人手で評価した。作成したデータに基づいて、ROUGE および5つの言語モデルを用いた結果と人手評価との相関を調査した。その結果、ROUGE-1 によるスコアが最も人出による評価との相関が高かった (0.7276) が、言語モデルの中では Doc2Vec による類似度が人手評価との相関が最も高い (0.6614) ことを示した。

我々は、2019年7月から2020年12月にかけて行われた評価型ワークショップ NTCIR-15<sup>2</sup>において、政治課題を解決する上で適切な情報を自然言語処理技術を用い

て提示することを目的とした QA Lab-PoliInfo-2 タスク [11]<sup>3</sup>を開催しており、Stance classification, Dialog summarization, Entity linking, Topic detection の4つのサブタスクを設計した。サブタスクの一つである Dialog summarization [12] は、議会における対話の構造を考慮しながら自動要約することを目的としたタスクであり、Formal run において7チームが参加して19の要約が提出された。そのうち12要約に対して参加者による人手評価がなされており、このデータセットを用いることで人手評価との相関を調査することができる。

以上から、本稿では、Dialog Summarization のデータセットを用いて、ROUGE や Doc2Vec による自動評価スコアが人手評価とどの程度相関があるかについて分析する。

## 2 関連研究

Campr ら [8] は、TfIdf, LSA, LDA, Word2Vec, Doc2Vec の5つの言語モデルの中で Doc2Vec による類似度が人手評価との相関が高いことを示した。Doc2Vec には、Distributed memory (DM) モデルと Distributed bag-of-words (DBOW) モデルの2つが存在する。両モデルは、Word2Vec における Continuous bag-of-words (CBOW) モデルと Skip-gram モデルの考え方を応用したものであり、DM は CBOW に、DBOW は Skip-gram にそれぞれ対応する。Campr らが用いた Doc2Vec がどちらのモデルであるかは明記されていない<sup>4</sup>ため、本稿ではその点を考慮した調査を行う。

<sup>1</sup><https://radimrehurek.com/gensim/>

<sup>2</sup><https://research.nii.ac.jp/ntcir/ntcir-15/index-ja.html>

<sup>3</sup><https://poliinfo2.net/>

<sup>4</sup>Word2Vec の説明において Skip-gram の優位を述べていることから、DBOW と思われる。

## 入力（都議会会議録）

百十五番小山くはひこ君  
(百十五番小山くはひこ君 答)

○百十五番（小山くはひこ君） 東京都議会第四回定例会に当たり、都民ファーストの会東京都団を代表して、小池知事及び教育長、関係局長に質問いたします。

いよいよ二〇二〇年の東京オリンピック・パラリンピック競技大会まで二年を切りました。一九六四年の東京大会は、戦後復興の象徴であり、首都高速道路や地下鉄の建設、東海道新幹線の開通など、各種インフラの整備が進みました。一九六四年大会後、日本は高度経済成長を駆け、その後高度成長の時代でありました。バブル崩壊から始まった長期的な経済停滞、経済のグローバル化、IT化の流れの中で、日本の国際的地位は低下しました。一九六四年大会後に増加を続けていた日本の人口は、二〇〇八年をピークに激減に転じており、東京都の人口も二〇二五年をピークに減少に転じると見込まれております。このような社会経済情勢の劇的な変化は、戦後日本の成長を生んだ社会モデルからの変革を迫っております。

平成の時代が明年幕を閉じ、新たな時代を迎える成熟都市東京は、今まさに大きな変革を必要としています。少子高齢化による生産年齢人口が減少する中で、次なる成長の源泉となる人、物、金、情報をめぐる世界の都市間競争、まさに熾烈をきわめています。このような状況下において多様なこそ成長の源泉であると、そういった認識に立ち、二〇二〇年の東京大会とその先を見据え、世界中で戦う東京の成長戦略を描き出さなければなりません。

そして、私たちは、一九六四年東京大会をきっかけに築き上げられてきた東京を二〇二〇年大会を契機として再構築し、東京と他の地域がともに栄える、東京の持続的成長を実現していかなければなりません。

私たち都民ファーストの会東京都団は、都議会最大会派となり一年余が経過しました。この間、議会改革を初め、受動喫煙防止条例の制定、待機児童の大幅減少、オリンピック・パラリンピック憲章人権条例の成立など、二〇二〇年の先を見据えた東京の成長と発展の礎となる施策が確実に推進されてきました。

本定例会でも、中小企業の振興条例、防災対策、省エネ対策を柱とする補正予算など、未来の東京の成長と発展のために必要不可欠な施策が取り上げられております。

このような東京の取り組みにもかかわらず、国はまた、不合理な都府県税の取返を繰り返そうとしています。今、都議会に求められているのは、都議会一丸となって、他の地域との共有共栄を可能とする首都東京の成長戦略を描き出し、着実に実行することであると改めて申し上げ、以下質問いたします。

平成三十一年度税制改正について伺います。

国は、いわゆる偏在是正の名のもと、都の税財源を地方へと配分すべく、さまざまな措置を講じてきました。この間、都としても対策を講じてきましたが、平成に入ってから三十年間で都が失った財源は六兆円に上り、平成三十一年度税制改正においても、さらなる措置が事実上予告されております。

こうした国の不合理な税制改正の動きに対して、先般、私たちの提案により上げられました東京と日本の成長を考える検討会の報告書が取りまとめられ、また、東京都税制調査会の審中も示されました。そして、それらを受けた東京都の見解も示されております。

都はこれまで、小池知事を先頭に、全国知事や東京都道出の国會議員、与党税制調査会の国會議員、都内区市町村との折衝を行ってまいりました。私たち都民ファーストの会東京都団も、東京都道出の国會議員や与党税制調査会の国會議員への要請活動、都民への啓発活動等に努めてまいりました。



## 出力（都議会だより）

図 1: Dialog Summarization のイメージ

### 3 QA Lab PoliInfo-2 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<sup>5</sup>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正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。そこで、Dialog Summarization で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。

東京都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<sup>6</sup>に掲載されている。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 1 の例では、左側が要約前の議会会議録であり、右側に黄色で示した範囲の文字列が正解データとなる。Dialog Summarization 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した。

図 2 に都議会だよりの例を示す。冒頭に、質問全体のメインピックと質問した議員名が記述され、その下に「東京大改革」や「子育て・女性・教育」といった複数のサブピックが並ぶ。各サブピックには、議員による質問の要約（質問要約）と、1人以上の回答者（知事、福祉保健局長、教育長、警視総監など）による回答の要約（回答要約）が記述される。都議会だよりでは一問一答方式でまとめられているが、実際の会議では一括質問一括答弁方式がとられている。そのため、同じサブピックの質問と回答の発言同士では重複する部分があり、都議会だよりでは、どちらか一方、例えば回答要約のみに記述して質問要約では記述しないといったまとめ方をする場合がある。このような場合、質問の発言と質問要約の単語が重複せず、ROUGE スコアが不当に低くなるという問題があった。したがって、サブピックごとの質問要約と回答要約を合わせて評価することで、より適切な評価になると考えた。本稿では、サブピックごとの質問要約と回答要約をまとめたものを対話要約と定義し、個々の質問要約ま

<sup>5</sup><https://www.gikai.metro.tokyo.jp/record/>

<sup>6</sup><https://www.gikai.metro.tokyo.jp/newsletter/>

全力で待機児童対策を進めよ  
 高齢運転者の事故対策速やかに

メイントピック

議員名  
 荒木ちはる (都ファースト)

サブトピック1  
 東京大改革

質問要約1  
 回答要約1 } 対話要約1

回答者 1  
 実現に向けた知事の見解は。  
 知事 運動喫煙の防止等数々の条例に結実し、数字に表れる成果も出ている。更なる高みを目指した挑戦を重ね、東京大改革に邁進する。

サブトピック2  
 子育て・女性・教育

質問要約2  
 回答要約2-1 } 対話要約2

回答者 2-1  
 知事 (1) 量と質の観点から全力で待機児童対策を進めるべき。(2) 大江戸線の一部車両に試験的に設置する子育て応援スペースの内容と進捗は。(3) 児童虐待のLINEによる相談が受けやすい環境づくりを。(4) LINEによるいじめ相談体制の強化を。(5) 登下校時の子供の安全を守るための対策は。(6) 犯罪被害者等支援条例の一刻も早い制定を。

回答者 2-2  
 福祉保健局長 (3) 元年8月から本格実施予定。

回答者 2-3  
 教育長 (4) 元年度は時間を延長し毎日実施するとともに、対象を国公私立の中高生に拡大。

回答者 2-4  
 監視総監 (5) 登下校の際に子供が集まる可能性のある場所を点検し、制服警察官によるパトロールの強化等、見せる警戒活動を実施。

図 2: 都議会だよりの例

たは回答要約を文要約<sup>7</sup>と定義する。

の考えの歪曲、制限文字数など)は考慮にいれない。以下の4段階で評価した。

## 4 人手評価

以下の5つの観点から3段階(一部4段階)で評価する。参照要約(都議会だより)と比較して判断する。

- Content: 参照要約との内容の合致
- Well-formed: 表現の適切さ
- Non-twisted: 話者の考えが歪曲されていない
- Sentence goodness: それぞれの発言で見た場合の要約としての全体的な良さ
- Dialog goodness: 質問と回答のセットで見た場合の要約としての全体的な良さ

最初の3つ(Content, Well-formed, Non-twisted)は、ある観点到特化した評価であり、最後の2つ(Sentence goodness と Dialog goodness)は総合的に見た場合の評価である。以下は注釈時のインストラクションである。

### Content

参照要約の内容が、要約にどれだけ含まれているかの評価であり、それ以外の観点(表現の適切さ、話者

<sup>7</sup>質問要約や回答要約は必ずしも1文で表現されているわけではないが、対話要約との対比として文要約と定義する。

○ = 正解の内容をほぼ網羅している

△ = 正解の内容をそこそこ含んでいる

× = 正解の内容が含まれておらず、元文書の要約として不適切である

▲ = 正解の内容とは異なるが、元文書の内容を要約している

基本的に、参照要約の内容が含まれているかどうかで判断した。ただし、参照要約の内容ではないが元文書の要約としては適切な内容であるということも考えられるため、4番目の評価として「▲」を用意した。また、制限文字数を考慮せずに判断し、制限文字数を超えていることの影響は、Sentence goodness や Dialog goodness で総合的に判断した。

### Well-formed

要約の表現や文法がどれだけ正しいかの評価であり、それ以外の観点(内容の網羅性や話者の考えの歪曲など)は考慮に入れない。以下の3段階で評価した。

○ = 文法的に正しい日本語である

△ = 一部おかしい表現があるが理解できる

× = 日本語になっておらず理解できない

## Non-twisted

元文書を読んだときに感じる話者の考えと異なる考えを感じさせないか、元文書では存在しなかった曖昧性が生じていないかの評価であり、それ以外の観点（内容の網羅性や表現の適切さなど）は考慮に入れない。以下の3段階で評価した。

○=話者の考えと同じ印象を受ける

△=話者の考えとは少し異なる印象を受ける

×=話者の考えとは違う印象を受ける

### △の例

元文書：将来の発展のために、一時的な景気後退もやむを得ない  
要約：景気後退もやむを得ない

### ×の例

元文書：断固とした処置を取ることも選択肢に入れるべきだ  
要約：断固とした処置を取る

しかしながら、作業者が実際に Non-twisted を評価した際に、「歪曲されているかどうか以前に、話者の考えが分からない」という問題が発生した。この理由として、(1)述べている内容が異なる（Content が×）、または、(2)日本語がおかしい（Well-formed が×）ことが原因であると考えた。そこで、以下のような、そもそも話者の考えが分からないといったものは、Non-twisted を「×」とするように指示をした。

- 発言者が異なる（おそらく、Content が×）
- 議題が異なる（おそらく、Content が×）
- 意味が通じない文章（おそらく、Well-formed が×）

## Sentence goodness

発言ごとに見た場合の、内容、表現、歪曲性などを含めて、総合的に要約が元文書の要約としてどの程度適切かの評価であり、以下の3段階で評価した。

○=元文書の要約として適切である

△=元文書の要約としてまあまあである

×=元文書の要約として不適切である

制限文字数と比較して、あまりに長すぎる（短すぎる）といった評価も、この項目に含めた。

## Dialog goodness

質問と回答のセットで見た場合の、内容、表現、歪曲性などを含めて、総合的に要約が元文書の要約としてどの程度適切かの評価であり、以下の3段階で評価した。

○=元文書の要約として適切である

△=元文書の要約としてまあまあである

×=元文書の要約として不適切である

制限文字数と比較して、あまりに長すぎる（短すぎる）といった評価も、この項目に含め、質問と回答のセットで1つの評価を与えた。

## 5 分析

我々は、対話要約における人手評価の値が Dialog Summarization における目標値であると考えているが、文要約における人手評価の値も重要であると考えている。そのため、文要約である Sentence goodness と対話要約である Dialog goodness の値をそれぞれ人手評価の値として相関関係を調査することとした。それぞれ、○を1点、△を0.5点、×を0点として平均した値を人手評価の値として、ピアソンの積率相関係数を求めた。

ROUGE スコアは、文献 [13] と同じく、ROUGE-N1, -N2, -N3, -N4, -L, -W1.2, -SU4 の7通りの手法を用い、各手法に対し、再現率を用いた場合と F 値を用いた場合のスコアを計算した。また、計算する形態素列として、分割された原文の表現をそのまま用いた場合（以下、「表層形」）、原形に戻した形態素列を用いた場合（以下、「原形」）、内容語とみなした形態素のみを用いた場合（以下、「内容語」）の3通りを計算した。内容語とみなした形態素とは、(1)品詞が「助詞」、「助動詞」、「感動詞」、「空白」、「補助記号」、「記号一般」以外であること、(2)動詞の場合は「為る」、「居る」、「成る」、「有る」以外であること（かな表記を含む）、(3)名詞の場合は、「所」、「為」、「くらい」、「の」、「事」、「物」、「積り」、「訳」以外であること（かな表記を含む）、の3つの条件を満たした形態素とした。

文要約と対話要約における ROUGE スコアと人手評価との相関係数を表1と表2にそれぞれ示す。文要約と対話要約のどちらにおいても、[13]と同じく、原形に戻した形態素列を用いた場合の再現率による ROUGE-N4 が人手評価と最も高い相関（文要約 0.991、対話要約 0.985）を示した。ROUGE の種類による差は、文献

表 1: 文要約における ROUGE スコアと人手評価の相関係数

	N1	N2	N3	N4	L	SU4	W1.2
再現率							
表層形	0.979	0.979	0.982	0.981	0.984	0.983	0.987
原形	0.981	0.986	0.989	0.991	0.984	0.989	0.985
内容語	0.987	0.988	0.981	0.945	0.985	0.989	0.984
F 値							
表層形	0.987	0.984	0.985	0.983	0.986	0.987	0.988
原形	0.987	0.987	0.989	0.989	0.984	0.990	0.983
内容語	0.976	0.975	0.957	0.922	0.971	0.967	0.971

表 2: 対話要約における ROUGE スコアと人手評価の相関係数

	N1	N2	N3	N4	L	SU4	W1.2
再現率							
表層形	0.969	0.979	0.981	0.982	0.970	0.979	0.972
原形	0.968	0.981	0.983	0.985	0.967	0.981	0.964
内容語	0.965	0.966	0.958	0.911	0.961	0.968	0.959
F 値							
表層形	0.977	0.980	0.981	0.979	0.973	0.981	0.973
原形	0.976	0.979	0.980	0.980	0.967	0.981	0.963
内容語	0.949	0.948	0.927	0.884	0.944	0.939	0.942

[13]と比較して小さく、殆どなかったが、対話要約の値が文要約よりも全体的に低くなる傾向があった。

Doc2Vec のモデル学習には、Campr らと同じく gensim を利用した。学習データには、地方議会会議録コーパス [14] から 47 都道府県における平成 23 年 4 月から平成 27 年 3 月までの 4 年間の定例会 (4.7GB) を用いた。DM モデルでは、文脈となる単語ベクトルの合計を用いた場合 (DM-sum)、平均を用いた場合 (DM-mean)、連結を用いた場合 (DM-concat) の 3 つのモデルを調査した。DBOW モデルでは、文書ベクトルの学習と同時に単語ベクトルを学習した場合 (DBOW+skip) としない場合 (DBOW) を調査した。全てのモデルにおいて、単語予測に用いる単語数 (DM の場合) または予測する単語数 (DBOW の場合) が、5 単語 (window-5)、10 単語 (window-10)、15 単語 (window-15) の場合を調査した。また、要素ベクトルの次元数は 300 とし、学習率の初期値は 0.001 とした。文書ベクトル間の類似度には cos 類似度を用いた。

文要約と対話要約における Doc2Vec 類似度と人手評価との相関係数を表 3 と表 4 にそれぞれ示す。文要約と対話要約のどちらにおいても、window-10 の DM-mean モデルが人手評価と最も高い相関 (文要約 0.990, 対話要約 0.990) を示した。ROUGE と異なり、対話要約の値が文要約よりも低くなることはなかった。DM-sum と DM-mean の間に殆ど差はなかったが、

window-10 と window-15 の DM-concat は人手との相関が見られなかった。DBOW と DBOW+skip を比較すると、DM+skip の方が高い値となった。DM-concat を除き、全体的に Doc2Vec は ROUGE に匹敵する高い相関を示した。これは、Campr らの結果と異なり、ROUGE において対話要約の値が文要約よりも低くなる傾向があったことを考慮すると Doc2Vec の方が適している可能性があると考えられる。

## 6 まとめ

本稿では、NTCIR-15 QA Lab-PoliInfo-2 Dialog summarization タスクで構築されたデータセットを用いて、ROUGE や Doc2Vec による自動評価スコアが人手による評価とどの程度相関があるかについて分析した。文要約と対話要約のどちらにおいても、ROUGE の中では原形に戻した形態素列を用いた場合の再現率による ROUGE-N4 (文要約 0.991, 対話要約 0.985) が、Doc2Vec の中では単語予測に用いる単語数 10 のベクトル平均を用いた distributed memory モデル (文要約 0.990, 対話要約 0.990) が最も高い値を示した。Campr らの先行研究と異なり、Doc2Vec は ROUGE に匹敵する人手評価との高い相関を示し、対話要約においては ROUGE よりも適している可能性が示された。今後、

表 3: 文要約における Doc2Vec 類似度と人手評価の相関係数

	DM-sum	DM-mean	DM-concat	DBOW	DBOW+skip
window-5	0.984	0.985	0.930	0.969	0.971
window-10	0.987	<u>0.990</u>	-0.490	0.896	0.969
window-15	0.989	0.987	-0.221	0.912	0.970

表 4: 対話要約における Doc2Vec 類似度と人手評価の相関係数

	DM-sum	DM-mean	DM-concat	DBOW	DBOW+skip
window-5	0.989	0.985	0.938	0.961	0.976
window-10	0.989	<u>0.990</u>	-0.539	0.875	0.972
window-15	0.988	0.987	-0.170	0.900	0.971

さらなる調査をする予定である。

## 参考文献

- [1] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [2] Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada, June 2012. Association for Computational Linguistics.
- [3] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR 2013)*.
- [4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [5] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.
- [6] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org, 2014.
- [7] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.
- [8] Michal Campr and Karel Ježek. Comparing semantic models for evaluating automatic document summarization. In Pavel Král and Václav Matoušek, editors, *Text, Speech, and Dialogue*, pages 252–260, Cham, 2015. Springer International Publishing.
- [9] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

- [10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 601–608. MIT Press, 2002.
- [11] 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-2 task. *Proceedings of The 15th NTCIR Conference*, 12 2020.
- [12] 小川泰弘, 木村泰知, 渋谷英潔, 秋葉友良, 横手健一, 乙武北斗, and 石下円香. Ntcir-15 qa lab-poliinfo-2 における dialog summarization. 言語処理学会第 26 回年次大会 (*NLP2020*), pages 945–948, 3 2020.
- [13] 渋谷英潔, 木村泰知, 乙武北斗, 内田ゆず, 高丸圭一, 阪本浩太郎, 石下円香, 三田村照子, 森辰則, and 神門典子. Ntcir-14 qa lab-poliinfo formal run における summarization task の結果と評価. 言語処理学会第 25 回年次大会 (*NLP2019*), pages 1431–1434, 3 2019.
- [14] 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)*, pages 78–85, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.