7T-08

文分散表現を用いた議会発言の類似度に関する分析

樋口 雄也† 乙武 北斗‡ 吉村 賢治‡
福岡大学大学院 工学研究科† 福岡大学 工学部‡

1. はじめに

近年，政治や議会に対する関心が高い人は少ない． 特に地方政治はその傾向が強く，議会の活動は，公開されている会議録などを通して知ることができるにも関わらず，それらを閲覧している人は多くない．議会で行われている議論を視覚的に表示できれば，地方政治に関心を持つ人も増加することが考えられる．

そこで筆者らは議会における発言内容の可視化や議員のマッピングを行うシステムの開発を検討している．このシステムを実現するためには，最初に発言内容を分析する必要がある．発言内容を分析する方法としては文分散表現の利用が考えられるが，分散表現の取得にはいくつかの方法がある．

本研究ではシステム実現の第一段階として，表記ゆれや同義語が含まれていても，内容が類似している文なら類似度が高くなる手法の選定を議会の発言内容を用いて分析する．具体的には，４手法(Bag of Words，Word2Vec， BERT/CLS トークン， BERT/各トークンの平均)で得られる文分散表現に対してコサイン類似度を用いて，議会発言の類似度の特性の分析を行った．

2. 議会会議録

本研究では，公開されている福岡県の議会会議録データ[1]を対象に分析を行った． データの内容は，2011 年 5 月から 2015 年 2 月の間に開催された議会の会議録で，会議名・年月日・発言者名・発言内容等の情報が CSV 形式で格納されている．ファイルの行数は 84,918 行(１行に１発言)，データサイズは 30.4MB となっている．

3. 各手法による文分散表現の取得方法

3.1. Bag of Words(BoW)

議会会議録データ全てを Juman++[2]で形態素解析し，ユニークな形態素に id をつけた辞書を作成する．文分散表現として表したい文を形態素解析し，各形態素の頻度を要素とするベクトルで文分散表現として表す．

An Analysis of Similarity Scoring Methods for Assembly Minutes' Records Using Sentence Distributed Representation
Yuya Higuchi † Hokuto Ototake‡ Kenji Yoshimura‡
Grad. Sch. of Electronics Eng. & Computer Sci, Fukuoka Univ.†
Dept. of Electronics Eng. & Computer Sci, Fukuoka Univ.‡

3.2. Word2Vec

gensim[3]による Word2Vec の実装を使用した．モデルの学習には wikipedia 全文(2019 年 11 月時点)を Juman++で形態素解析したものを使用した．モデル作成のパラメータは，次元数は 200，単語の必要最低限の出現回数は 15，ウィンドウサイズは 20 とした．

文分散表現として表したい文を形態素解析し，各形態素の分散表現を平均したものを文分散表現として扱う．

3.3. BERT

事前学習モデルは BERT 日本語 Pretrained モデル[4]を使用し，出力は 11 層目を使用した．BERT の出力は，形態素解析と BPE(Byte Pair Encoding)が施されたものに[CLS]と，[SEP]という特殊なトークンが付与されたものからなる．文分散表現の候補としては，

(1) 文全体を表す[CLS]トークン
(2) [CLS]と[SEP]を除いた各トークンの分散表現の平均

の２通りがある．今回は，両者の文分散表現の取得方法を試した．

4. 発言内容の分析

4.1. 分析の概要

福岡県の議会会議録データ中の各発言文と，他の発言文の文分散表現のコサイン類似度を計算する． 議会会議録データに含まれる議長の発言は，定型的な文が多いので，議員の発言のみで分析を行う．議員の発言のみに絞ると 45,594 発言となった．以下の文を入力したときの結果を例にして考察を述べていく．

入力文: また，国民体育大会においては，デモンストレーション等のスポーツ行事として開催がされています。

入力文との類似度の算出結果は以下のように示す．

順位/コサイン類似度/入力文

4.2. BoW の結果

出現する形態素が一致しているものが上位に来るため，その単語から他のものが連想されるような文(スポーツからサッカーやオリンピックなど)は取得できなかった．そのため，表記ゆれ

や同義語も同様に取得できないと言える．

#1/0.624/今月二十八日から東京で国民体育大会が開催され、国体の後には、慣例として開催地で全国障害者スポーツ大会が開催されます。

辞書を全品詞で作成しているため，どのような文でも使われる「また」「〜ます」が含まれる内容の関係ない文も上位に出現した．

#2/0.438/また，最近では全国においてジビエの取り組みが盛んに報道されています。

4.3. Word2Vec の結果

入力文に「スポーツ」や「大会」という単語が含まれることから，結果には「オリンピック」や他のスポーツの名前が含まれる文が上位に出現した．また，「開催」と意味が近い「実施される」を含む文が出現した．

#1/0.905/ところで、二〇一二年のロンドン大会で女子ボクシングが競技として採用されたことにより、オリンピックでは全ての競技が男女ともに実施されるようになっています。

しかし，順位を下げて見ていくにつれて，「講演会」や「総会」といった「大会」を拡大解釈したものが含まれる文が出現した．

#10/0.883/在校生に対して開催される母校での講演会や、東京を初め各地で開催される総会、そして本校での総会と、当番期としてのさまざまな活動があります。

4.4. BERT CLS の結果

Word2Vec で出現していた「開催」と意味が近い「実施される」の他に「開かれる」「行われる」といった単語を含む文が上位に出現した．

しかし，他の手法と比較して文の形の影響が強い結果となった．上位 10 文中，7 文で「また」「さらに」「次に」などの接続詞ではじまっており，また文の長さ，句読点の位置も似ている文が多い．

#1/0.837/また，障害者スポーツ大会は，それぞれの地域の大勢のボランティアの皆さんの力によって運営されています。

また，Word2Vec と違い，上位にもスポーツではない話題の文が出現した．そのため，文の内容よりも文の形の影響が強いと考えられるため，内容の類似性を取る場合は BERT CLS は不適切だと言える．

#3/0.819/また、例年七月には全国知事会議が開かれております。

4.5. BERT トークン平均の結果

BERT CLS ほど文の形が似ているものの類似度が高くなるという特徴は顕著ではないが，Word2Vec と比べるとその傾向は見られる．

Word2Vec/BERT CLS と同じように，「開催」と意味が近い「実施される」「開かれる」「行われる」を含む文が上位に出現した．また，順位を下げて見ていくと，Word2Vec と同じように「大会」を拡大解釈したものを含む文が出現した．

1/0.769/次に，夏休み期間中，市大会や県大会などさまざまな競技大会が開かれます。

#10/0.706/学校の PTA や地域の子ども会等の行事では、グラウンドゴルフはよく採用されていますが、ペタンクが採用されているとは余り聞きません。

5. おわりに

本論文では，4 手法の文分散表現とコサイン類似度を用いて，議会発言の類似度の特性を分析した．BoW では取得できなかった表記ゆれや同義語を含む文を Word2Vec/BERT で取得できることが確認できた．

しかし，Word2Vec/BERT のどちらでも，「大会」を拡大解釈したような「講演会」や「総会」などを含む文が出現した．これは表記ゆれや同義語を扱えるようになったことの欠点だと言える．

「開催」を含む文の入力から「実施される」「行われる」「開かれる」を含む文が出現したことから，Word2Vec より BERT CLS と BERT 各トークンの平均のほうが表記ゆれ/同義語を含む文が幅広く取得できた．

しかし，BERT CLS では，内容よりも文の形の影響が強くなってしまうことがある．このことから，内容の類似性だけ見たい場合，Word2Vec または BERT トークンの平均を用いる方法が有用だと言える．

参考文献

[1] 木村 泰知，渋木 英潔，高丸 圭一，乙武 北斗，森 辰則：地方議会会議録コーパスの構築とその利用，第 26 回全国大会(2012)
[2]Juman++, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++
[3]gensim, https://radimrehurek.com/gensim/models/word2vec.html
[4] BERT 日本語 Pretrained モデル, http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB