1N2-OS-39a-5

地方議会議事録の探索的閲覧のための自動タグ付け手法の開発

Developing Automatic Tagging Method for Exploratory Browsing of Local Council Minutes

成瀬 雅人[*1] Masato Naruse
白松 俊[*1] Shun Shiramatsu
松島 格也[*2] Kakuya Matsushima

[*1] 名古屋工業大学 大学院工学研究科 情報工学専攻
Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology

[*2] 京都大学 大学院工学研究科 都市社会工学専攻
Department of Urban Management , Graduate School of Engineering, Kyoto University

Although there are minutes of local councils that are published to citizens, they are unstructured and too long to be read. We are developing an automatic tagging method that enables citizens to browse remarks in the minutes in exploratory way. Specifically, our system automatically assigns field name tags and place name tags to remarks in the minutes. For tagging the field names, we use similarity of field names and text in remarks based on Paragraph Vector. For tagging place names, we use the LOD dataset called Linked Open Address Japan. We also implement a prototype of user interface for exploratory browsing of council minutes.

1. はじめに

地方議会の議事録が市民に向けて公開がなされているが，公開されている議事録は非常に長い内容になっており，またその構造も分かりにくくなっているため，市民に読んでもらうために公開されているのに，実際は市民が気軽に閲覧できる形にはなっておらず，市民が見たいと思うような情報を簡単に探索することが難しい．また探索の補助のため発言にタグを付ける事が考えられるが，発言の内容に沿って人手でタグを付ける事は難しく時間もかかってしまうため，自動的にタグを付与する方法が必要になってくる．

そこで，本研究では地方議会議事録に含まれる発言を，探索の手掛かりとなるような外部データセットと紐づけし，発言に対し探索がしやすいよう適切なタグを自動で付けられる手法の開発を行う．本研究では閲覧の手掛かりになるような外部データセットの例として，自治体の Web ページで公開されているような分野の階層構造や，地図上の地理データを用いる．また，発言と紐づけた外部データセットを用いて手軽に探索が行えるようなインターフェイスの提案を行う．なお，本研究では，地方議会の議事録の例として名古屋市会の議事録[1]を利用する．

2. データセットの成形

地方議会の議事録に紐づけるための外部データセットとして，発言と関連した内容の多い同地方の自治体の Web ページを利用する．Web ページの階層構造からタグ候補として分野名を取り出し，発言と紐づける事でその発言と内容の近い分野名を得る事ができる．本研究では名古屋市 Web ページの分野階層[2]を用いる．

また，地方議会の議事録の発言にはその地方に関連した地名が含まれている場合が多いため，発言から地名を抜き出して

表 1：タグとして適さない分野名と理由の例

タグとして適さない理由	実際の例
一目で内容が想像できない	「主な事業内容」 「制度の概要」
同じ名前が何度も現れる	「組織と業務のご案内」 「交通案内」
内容が適さない	「投稿者のコメント」 「意見募集期間」

その発言に地名をタグとして追加することで一目でその発言がどの地域について話しているのかが判別でき，閲覧者が特定の地域，例えば自分が住んでいる地域に関連する内容の議事録を見たい場合に地名のタグから議事録を検索するという事が可能になる．本研究では地図情報として，Linked Open Addresses Japan[3]を利用する．

研究に用いる地方議会の議事録だが，これは議会での発言をそのまま文章に起こしたもののため，他のデータと紐づけるには適さない発言も存在する．具体的には挨拶だけの発言や，議会を進行させるための発言である．そのため本研究では一つの発言がある程度の内容を持っているためのおおよその基準として，170 文字以上の発言であるという基準を経験的に定め，これより文字数の少ない発言を切り捨てて議事録のデータとして使用している．

また，Web ページの階層構造においても，取得したデータの中にあるタグ候補にタグとして適していないものが存在する．表 1 にタグとして適さない理由と，その例をまとめたものを示す．これらをタグ名として利用しないようにすることで，発言に対してより近い内容をタグとして付け，あまり適さないタグを付ける数を減らすことができるだろう．

連絡先： 成瀬雅人 〒466-8555 名古屋市昭和区御器所町 名古屋工業大学つくり領域 白松研究室
naruse@srmtlab.org

[1] http://www.kaigiroku.net/kensaku/nagoya/nagoya.html
[2] http://www.city.nagoya.jp/shisei/category/53-0-0-0-0-0-0-0-0-0.html
[3] http://uedayou.net/loa/

3. 発言への自動タグ付け

地図と議事録を紐づける為，議事録の発言に現れた地名をタグとして発言に追加する．タグとして付与する地名は Linked Open Addresses Japan を用いて統一された形態で付与する．

また発言にタグを付ける為には，それぞれの発言の内容とタグとして使われる分野名の内容がどれだけ類似しているかを検証する必要がある．本研究では自治体 Web ページの階層構造からタグとして使われる分野名以下に含まれる文章と各発言を比較することで比較を行う．

それぞれの文章の比較には sentence2vec[1]を使用し，各分野と発言の内容の類似度を計算する．まず，比較したい分野名以下の階層に存在する文章を一つの巨大な文として扱う．議事録の全発言と全分野をコーパスとして各文章の Paragraph vector[1]を求める．分野と発言の Paragraph vector が求められたら，分野と発言を一つずつ選んで比較を行う．比較には Cos 類似度を用いる．文書aの Paragraph vector の i 番目の要素を$s(a,i)$とおくと，文書aと文書bの Cos 類似度は以下のように求められる．

$$Cos(a,b) = \frac{\sum_{i=1}^{m} s(a,i)S(b,i)}{\sqrt{\sum_{i=1}^{m} s(a,i)^2} \cdot \sqrt{\sum_{i=1}^{m} s(b,i)^2}}$$

この値は-1 から 1 の範囲をとり，1 に近いほど二つの文書は内容が類似しているとみなされる．本研究では文書が類似しているとする閾値を経験的に 0.4 と設定し，この値を越えた組み合わせの場合発言に対して類似している分野名をタグとして追加する事で分野と発言を紐づける．

類似度計算によって紐づけられた分野と発言のデータの保存は，本研究では RDF を記述するため Turtle 型を用いる．具体的には次節で必要な情報として，分野の階層構造と，分野ごとに類似度が高かった発言，またその類似度を Triples として持たせる．

4. 探索的閲覧のためのインターフェイス

探索的閲覧(Exploratory browsing)とは，ユーザー自身が必要な情報を十分に把握していない場合にリンクを辿りながら見つかった情報をヒントにして必要な情報を探す行動である[2]．

探索的閲覧の支援のため，自動付与されたタグを利用して分野名と発言を結び付けて作成した上記 Turtle 型データを元に，閲覧者が分野名の階層構造から発言を検索するためのシステムのインターフェイスを図 1 に示す．

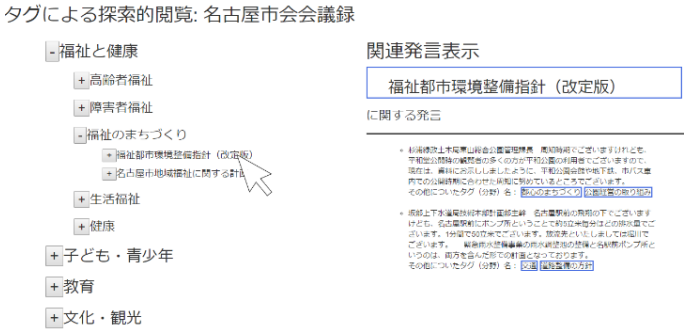

図 1: 分野名の階層を起点にして
発言を探索的に探すための閲覧インターフェイス

図 1 では，画面の左側に分野名の階層構造が示され，閲覧者は分野名の左についたボタンを押すことで，その分野名の直下の分野名が展開され，分野の中でも深く掘り下げて見たい分野を階層的に辿ることができる．また同様に展開されている時にボタンを押せば展開されている部分を隠す事ができる．また，各分野名にカーソルを合わせてクリックすることでその分野名がタグとして登録された発言の一覧が画面の右側に表示される．図 1 では福祉と健康，福祉のまちづくりを展開し，福祉都市環境整備指針(改訂版)という分野名をクリックしたことで，その分野に関連するような発言が現れている．

このインターフェイスによって，閲覧者は手軽に閲覧したい情報あるいは気になる分野の探索ができるだろう．

5. 考察

本研究では地方議会の議事録のデータと同地方の自治体の Web ページの類似度を求める事でタグ付けを行い，その結果から閲覧者の探索的閲覧を支援するようなインターフェイスを提案した．しかしこの自動タグ付け手法にはいくつか問題点が挙げられる．

まず，発言ごとに付けられるタグの数に偏りが見られ，付けられるタグの平均が多くてもその中でタグがあまり付けられない発言が存在する．そのため分野と発言が類似しているとするための閾値を全体で統一するのではなく，組み合わせごとに適切な値を求める事や，付けられるタグが多すぎる場合は類似度が高いものから数えて一定数残す，といった方法で付けられるタグの数を調節する事が考えられる．

また，本研究では分野と発言のデータを Turtle 型で記述したが，現在の比較的単純な関連付けだけでなくより詳細な関連付けを必要とする場合には，Web Annotation Data Model[2]など他のデータ記述形式を用いる事も検討される．

探索的閲覧のためのインターフェイスにおいては，現在は試作的なインターフェイスのため，実験等を行って実際に閲覧者が使いやすいインターフェイスを模索する必要があるだろう．

6. 関連研究

木村ら[3]は，地方議会の会議録のフォーマットが自治体ごとに異なっている事や，各研究者が重複するようなデータの電子化を行っている事を背景に，地方議会会議録コーパスの構築を行っている．本研究では名古屋市の会議録と名古屋市の Web ページのみを対象としているが，名古屋市以外の会議録，あるいは同自治体の Web ページなどの階層構造などを扱う際にはそれらも扱う事ができるようなフォーマット作りをどうするかという課題があげられる．

渡辺ら[4]は，社会問題に関する文章を与えた時に，その文章を表すのにふさわしいタグを自動で付与するシステムの構築のため DBpedia を用いた社会問題のタグ候補の抽出や，社会問題に関する記事に対し最適なタグを候補から選択するという研究を行っている．社会問題と地方議会の議事録という範囲の差異はあるが，どちらもとある文章に対してタグを自動で付与するという目標がある，今後の研究への利用を検討していく予定である．

[1] https://github.com/klb3713/sentence2vec

[2] https://www.w3.org/TR/annotation-model/

7. おわりに

本稿では地方議会議事録の探索的閲覧のための自動タグ付け手法の開発について述べた. また, 自動付与されたタグを使って探索的な閲覧を支援するためのインターフェイスを提案した. 今後はインターフェイスで発言と表示されたタグ, そのタグ名からも発言が検索可能であったりと, より探索的な閲覧が支援できるようなインターフェイスへの拡張が考えられる.

謝辞 本研究の一部は JICE 研究開発助成, JSPS 科研費 (25870321), および JST CREST の支援を受けた.

参考文献

[1] Le, Q. V., et al. “Distributed Representations of Sentences and Documents.” In ICML, Vol. 14, pp. 1188-1196, 2014.

[2]White, R. W., et al. ″Exploratory search: beyond the query-response paradigm.″ Morgan and Claypool, 2009.

[3]木村 他：地方議会会議録コーパスの構築とその利用. 第 26 回人工知能学会全国大会, 3B3-NFC-4-3, 2012.

[4]渡辺 他：DBpedia を用いた社会課題タグ自動付与 API の試作 2017 年度 人工知能学会全国大会（第 31 回）, 2017.