D-004

階層性に着目したレビューコーパスの特徴可視化

Visualizing Features of the Review Corpus Focusing on the Hierarchy

林 沙也加 †　　松下 光範 †
Sayaka Hayashi †　Mitsunori Matsushita †

1 はじめに

大量の自然言語文を集約した言語コーパスは，機械学習や言語処理などのリソースとして活用されており，近年では複数のコーパスを組み合わせて単一のコーパスでは難しい課題の解決に利用する手法が注目されている．例えば山西らは，飲食店のレビュー文中に含まれているホスピタリティと料理の観点を機械的に分類するためにホスピタリティに関する特徴を持つホテルレビューコーパスと，料理に関する特徴を持つレシピレビューコーパスを組み合わせた擬似コーパスを用いて分類器を作成している [2]．この手法は，異なる複数のレビューの偏りを利用したものであるが，単一のコーパス内のカテゴリ毎の情報の偏りは扱っていない．

多様なカテゴリが含まれるコーパスを扱う場合には，カテゴリ毎の情報の偏りに着目することで単一のコーパスであっても同様の活用が可能だと考えられる．例えば，EC サイトのように多様な商材を扱うサービスのレビューから構築されたコーパスの場合，商品のカテゴリ分類に代表されるデータの部分集合を取り出してその差異を比較することにより，各集合の組み合わせを擬似コーパス（以下，部分コーパスと記す）として活用できる．

そこで本研究では，コーパスが持つ階層性に着目し，階層ごとに含まれる情報を比較できるように可視化することで，ユーザがカテゴリ毎の情報の偏りを容易に把握し，単一のコーパス内で部分コーパスを作成する支援を試みる．提案する可視化ツールを用いて 2 つのカテゴリ情報の差異やデータ量を探索的に比較できるようにすることで，ユーザがカテゴリに含まれる情報を確認する作業が簡便になり，自身の目的に沿ったレビュー集合を見つけるコストの低下が期待できる．

2 関連研究

谷口ら [1] は，2 つのテキストデータ間の比較を行うためのシステムを構築した．このシステムでは比較したい 2 つのテキストデータの独自単語（それぞれのデータセットで独自に使用されている単語）と共通単語（両者で共通して使用されている単語）を可視化し，独自単語をもとにテキスト集合から対応するテキストを表示している．

谷口らの提案手法ではカテゴリが階層的になっているデータセットのノード同士が親子関係になっている部分集合の比較ができない．ノード同士が親子関係になっている部分集合では子ノードに含まれる情報は親ノードにも含まれており，親ノードで提示する情報は子ノードの情報を含めることが望ましい．

3 デザイン指針

ユーザが階層性を持つコーパスから円滑に部分コーパスを作成するためには，階層的なコーパスの部分集合に含まれる情報を比較することが望ましい．1 章で述べたように，ユーザがコーパス内の階層ごとの部分的な特徴を容易に峻別することで 1 つのコーパス内であっても部分コーパスを作成することが可能である．コーパスの特徴を比較することによって，ユーザは任意のコーパスを分類するための部分コーパス作成に使用するコーパスの部分集合を探すことができる．

ユーザがコーパス内の部分集合を比較するパターンとして，同一階層にあるノードの比較と親子ノード間の比較の 2 つが想定される．

3.1 同一階層間の比較

ユーザが同一階層にあるノードを比較する場合，比較対象が兄弟ノードである場合と，異なる親ノードに接続されたノードである場合が想定される．比較対象が兄弟ノードである場合，可視化ツールではそれらの親ノードに接続しているすべての子ノードに共通して含まれる語彙（共通語）を除き，各々のノードに固有の語彙を提示する．共通語を除くことでユーザがノードの情報を把握しやすくなることが期待できる．また，比較対象が兄弟ノードでない場合，可視化ツールでは共通語を含めた各々のノードに含まれる全ての語彙を提示する．これは，ノードの階層的な位置関係が遠くなり共通語の数が減少することが想定されるためである．

共通語を特定する基準として，本システムでは兄弟ノードを対象とした inverse document frequency(IDF) を用いた．IDF の高い語彙はあるノードにおいて特徴となる語彙であるため，IDF の高い語彙をノードの情報とすることでユーザがノードの特徴を把握しやすくなると期待できる．なお，親子関係に

関西大学大学院総合情報学研究科, Graduate School of Informatics, Kansai University (†)
〒 569-1095 大阪府高槻市霊仙寺町 2 丁目 1-1

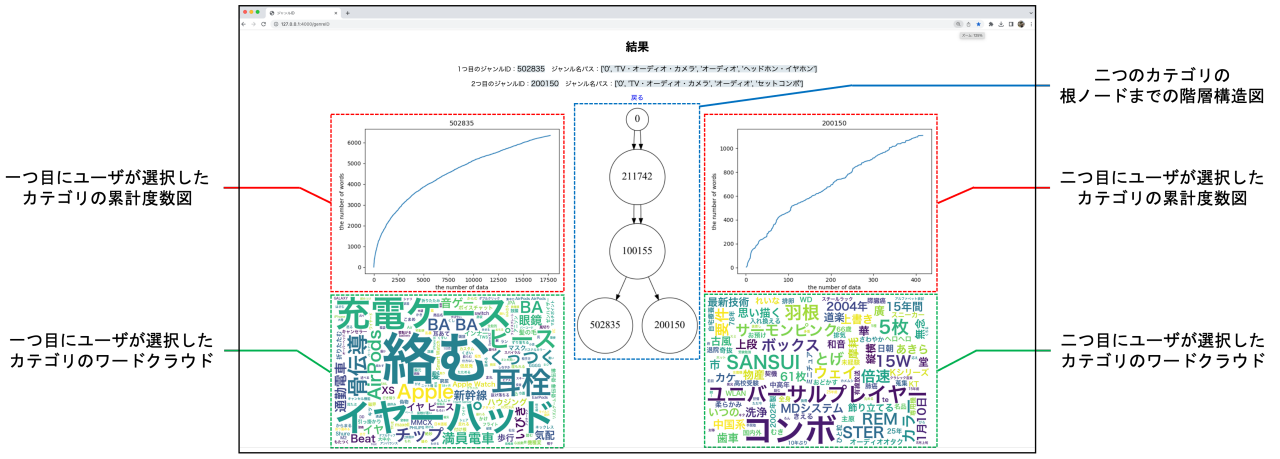

図 1: システム画面

ない中間ノード同士を比較する場合は，その中間ノードに接続された葉ノードに含まれるレビュー文を集約して扱う．

3.2 親子ノード間の比較

ユーザが親子ノード間を比較する場合，親ノードは接続されている全ての子ノードの語彙を親ノードに集約して提示し，子ノードは子ノードに含まれる全ての語彙を提示する．この比較では，ユーザの目的に応じて適切な深さでノードを選定する際に用いることが想定される．親ノードには複数の子ノードが接続しているためデータ量は多くなると考えられる反面，複数のトピックが混在することで単一の子ノードを対象とするよりもその対象の子ノード以外の語彙がノイズとなる可能性が高まる．一方で，子ノードに含まれる情報はカテゴリが細分化されているため，そのノードのトピックに関わる特徴語が集中して現れると考えられる反面，そのノードに含まれるデータ量が少なくなる．そのため，こうしたトレードオフを勘案して目的に応じた適切な階層のノードを選択するために，ユーザは親ノードの語彙と子ノードの語彙を探索的に比較する必要がある．

4 可視化ツールの詳細

試作した可視化ツールを図 1 に示す．可視化ツールでは，2 つのカテゴリが指定されるとカテゴリの根ノードまでの階層構造図，各カテゴリ中のレビューに出現する語彙の累計度数図，および語彙のワードクラウドを表示する．階層構造図（図 1 の青枠部分）では，ユーザは 2 つのカテゴリの階層的な位置関係を把握することができる．累計度数図（図 1 の赤枠部分）は，テキストが一文増えるごとに異なり語彙数がどのように増加しているかをグラフ表示したものである．コーパスの利用においてそのカテゴリに関わる語彙が十分に存在しているかがしばしば問題になる．このグラフを見ることで，ユーザは対象カテゴリに含まれている語彙数が飽和状態にあるかを観察することができる．ワードクラウド（図 1 の緑枠部分）では，カテゴリの特徴となる代表的な語彙を一覧できる．

なお，本稿ではレビューコーパスとして，楽天データセット [3] を用いた．

5 おわりに

本稿では，コーパスが持つ階層性に着目し，ユーザがコーパスの部分集合に含まれる情報を比較できるように可視化し，部分コーパスを作成する支援を試みた．試作したツールは，階層構造図，累計度数図，ワードクラウドをユーザに提示することで部分集合の特徴把握を容易にする．このとき，比較対象のノード間の関係性に応じて提示する語彙の選定基準を変更している．今後は，ツールの利用を通じて機能の改良と有用性の評価を行う．

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用した．記して謝意を表す．

参考文献

[1] 谷口 拓紀, 砂山 渡, 服部 峻: テキスト集合間の差分を表す単語の可視化による独自情報の把握支援システム, JSAI2023, 2K1-GS-9-04 (2023).

[2] 山西 良典, 藤岡 寛子, 西原 陽子: 擬似コーパスを用いた飲食店レビューの観点の自動分類, 人工知能学会論文誌, 36(1), WI2-A_1-8 (2021).

[3] 楽天グループ株式会社： 楽天データセット，国立情報学研究所情報学研究データリポジトリ（データセット),https://doi.org/10.32130/idr.2.0 (2014).