小説の内容と類似した楽曲検索方法の提案

2020SC037 児玉侑己
指導教員：河野浩之

1　はじめに

若者のアナログ, 活字離れは進んでいると考えられる. 実際, 高校 2 年生の 44.9 %が本を読まないと回答している [1]. また原因の多くにメールを返す等の携帯電話に当てられるものがある [1]. それと対照的にオンラインサービスが普及し音楽ストリーミングサービスは年々増加傾向にある [2]. 以上のことから本に触れる機会を増やすためには, 電子媒体から本をとって読んでみようと思えるよな動機づけや感動体験が必要であると考えられる. そのアプローチとして本研究では歌詞の文章と小説の内容をベクトル化し, 類似性の高い文書をコサイン類似度を用いて検索することができるシステムを提案する.

2　コサイン類似度を用いた文書の類似性に関する先行研究

本章では文書の類似性に関する先行研究について記述する.2.1 節では Word2Vec について, 2.2 節では Doc2Vec,BERT について記述する.

2.1　Word2Vec を用いた先行研究

韓ら [3] は歌詞サイトと観光地レビューから取得したデータを用いて実験を行った.Word2Vec を用いてベクトル生成を行い, コサイン類似度での検索を行った. 成果として, 観光地レビューは個人の主観である感想が多く観光地の情報量が少ないためコサイン類似度の検索結果は偏った結果となった.

河原ら [4] の研究ではコメントと歌詞を用いて類似楽曲検索を行った.Word2Vec によるベクトル化から c-means によるクラスタリングを行い, コサイン類似度での検索を行った. 結果としてある程度の検索精度は得ることができたが, 一部のカテゴリでは精度が低下した. 原因として文脈の意味を捉えることのできない Word2Vec に問題があると考えられる.

2.2　BERT を用いた先行研究

長ら [5] は BERT を用いて日本の法と外国の法の対応付けを行った. 提案手法として BERT を用いたベクトル生成を行い, 日本の法と外国の法をコサイン類似度を計算し, 類似している法律を示した. 本研究では生成されたベクトルの精度の比較のために Jaccard 係数による対応付けの結果と比較した.

3　類似性検索のための提案手法

本章では小説と類似した歌詞の検索の提案手法について記述する.3.1 節では先行研究の課題を含めた改善点について,3.2 節では楽曲検索システムの提案について示す.

3.1　先行研究を踏まえた改善案

本節は先行研究の課題を考慮し, 本研究の改善案を記述する.

- すべてのクチコミを扱うことによるデータの偏り
- 扱うモデルによるベクトル生成の偏り

以上の課題を考慮し, 本研究では 2 つの改善案を示す.

- 扱うデータを小説の内容を含むクチコミに限定することによりデータの偏りを軽減する
- 文章ごとに高精度のベクトル生成を行う BERT を用いる

3.2　類似性検索システムの構成

本節では BERT を用いて類似性検索システムを作成する. 本研究で用いるデータを歌ネット [6], 読書メーター [7] からスクレイピングを行いデータを抽出する. 抽出したデータを必要に応じて形態素解析を行う. また歌詞データを BERT を用いてベクトル生成し,2 次元に t-SNE を用いて次元圧縮し, 生成されたベクトルを用いてコサイン類似度を用いて検索する. また本研究では BERT を用いたベクトル生成やコサイン類似度の有用性を示すために Doc2Vec を用いた対称実験を行う. 図 1 は提案手法のフローチャートである.

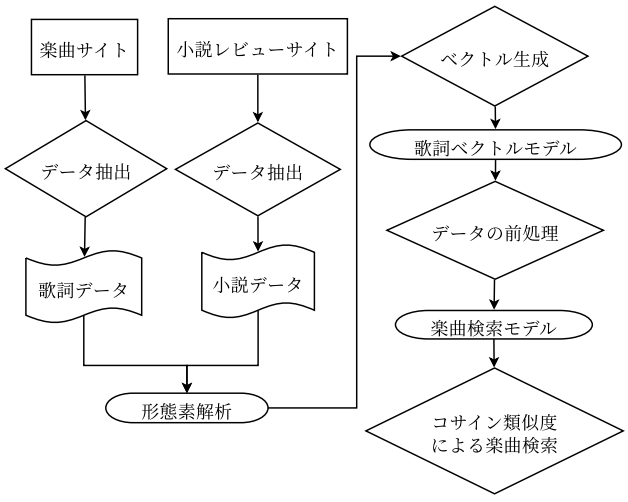

図 1　提案手法

4　実験結果

本章では実験結果について示す.3.1 節では t-SNE を用いたベクトル生成, 3.2 節ではコサイン類似度の結果について示す.

4.1　ベクトル生成についての結果

本節では BERT,Doc2Vec でベクトル生成をし, 生成された 768 次元,300 次元のベクトルを t-SNE を用いて 2 次元に圧縮し可視化ができるようにし, ベクトル生成の精度について評価を行う. 次元圧縮の結果の一部をモデルごとに図に示す.

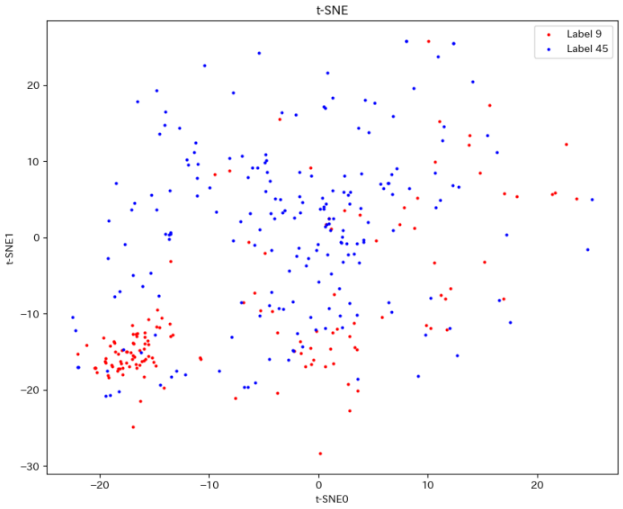

図 2　T-SNE を用いた Doc2Vec の次元圧縮の結果の一部

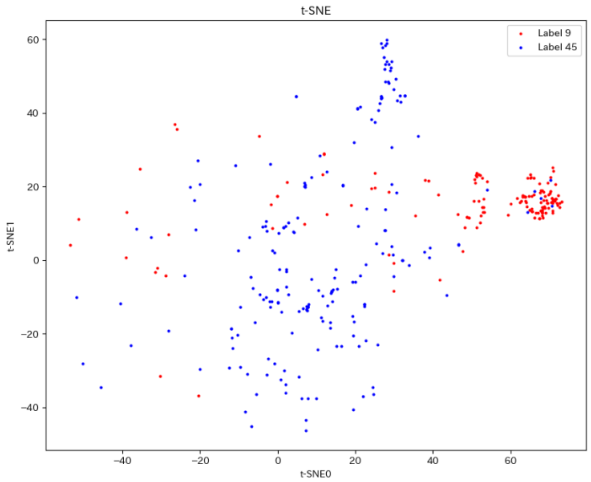

図 3　t-SNE を用いた BERT の次元圧縮の結果の一部

4.2　コサイン類似度に基づく歌詞検索結果

本節ではコサイン類似度で小説のクチコミに類似した歌詞を検索する. 生成したベクトルからコサイン類似度で検索をしていく. 本研究では恋愛小説である「桜のような僕の恋人」を用いて検索を行う. 結果を表 1, 表 2 に示す.

5　まとめ

本研究の目標である活字離れを抑えるという目標は達成できなかった. しかし, 各モデルを用いて t-SNE では BERT の方がより密度の高いクラスターが発生していること. コサイン類似度の数値では BERT の方が高い数値を出していること. という 2 つの点から BERT の有用性を示すことはできた. この理由として BERT 特有の Attention 機構が影響していると考えられる. 今後は推薦した楽曲が聞きたくなったかを MRR という手法を行い, どれだけ活字離れを抑えることを可能にするかを数値化したい.

表 1　「桜のような僕の恋人」(宇山佳佑) における BERT での類似性検索結果

順番	歌手	曲名	コサイン類似度
1	コブクロ	NOTE	0.9531
2	コブクロ	手紙	0.9513
3	コブクロ	同じ窓から見てた空	0.9489
4	コブクロ	シルエット	0.9461

表 2　「桜のような僕の恋人」(宇山佳佑) における Doc2Vec での類似性検索結果

順番	歌手	曲名	コサイン類似度
1	松任谷由美	Oh Juliet	0.7571
2	HY	いつまでたっても女の子	0.7392
3	Mr.Children	Another Mind	0.7336
4	ポルノグラフィティ	小規模な敗北	0.7103

参考文献

[1] 株式会社浜銀総合研究所 ,“子供の読書活動の推進等に関する調査研究 ”, 文部科学省委託調査 ,2016-06-10

[2] 日本レコード協会,“音楽配信売上実績 過去 10 年間 全体”, 日本レコード協会,https://www.riaj.or.jp/f/data/annual/dg_all.html

[3] 韓 毅弘, 山西 良典, 西原 陽子, 奥 健太,“単語分散表現を用いた観光地レビューからのクロスドメイン歌詞検索”,ARG WI2 No.15,2019,WI2-2019-14

[4] 河原 大智, 松本 和幸, 吉田 稔, 北 研二,“コメントデータと歌詞に基づく楽曲動画検索システム”, 人工知能学会全国大会論文集,36,pp01-pp04,2022

[5] 小関龍也, 長裕樹, 中村誠,“Doc2Vec と BERT を用いた比較法研究における類似条項の対応付け”, 言語処理学会年次大会発表論文集 (Web),29,pP01-10,2023-03

[6] 株式会社ページワン,“ 歌詞検索サービス 歌ネット”, 歌ネット,https://www.uta-net.com/

[7] ブックウォーカー,“ 読書メーター”,https://bookmeter.com/,2021-10-01