2ZC-03

落語の演目検索

中尾　大†　　秋岡明香‡
明治大学総合数理学部ネットワークデザイン学科

1. はじめに

近年、テレビや雑誌、YouTube などの媒体において落語や講談などの古典演芸が取り上げられる機会が増加している。それに伴って、寄席やホール落語へと足を運ぶ人、特に落語にあまり造詣の深くない初心者が増加している。ホール落語では、こういった落語の初心者向けに公演終了後の演目張り出しを行う場合が多い。しかし、寄席ではそういった張り出しを行わない所も多い。そのため、聞いた演目の名前が分からないままという客は多い。落語の演目は本編に出てこない言葉がタイトルになっていたり、現代社会では既に死語となっている言葉が使われていたりと初心者には自力で調べるのは難しい。指定した条件やキーワード等からユーザーの好みに合う小説を推薦する研究[1][2]等は存在するが、内容のみを知っている状態からタイトルを特定するような研究はなされていない。本研究では、登場人物や噺の舞台など、客の印象に残りやすい落語の要素や、特徴的なキーワードなどから、落語の演目を検索・推定し、演目の名前を特定することを目指す。

2. 落語データベースの作成

落語について、演目ごとにあらすじや解説をまとめているウェブサイト[3]からあらすじをスクレイピングする。このテキストから半角・全角スペースなどを取り除いて体裁を整え、演目ごとにテキストファイルで保存する。この中から寄席や落語会などで目にする機会の多い代表的な演目 225 個を選定する。次に、文中の単語の索引を作成する。単語の索引を作成するためには文章の分かち書きを行わなければならない。本研究ではウェブ上のオープンソース日本語形態素解析エンジンである kuromoji[4][5]を用いた形態素解析によって文章の分かち書きを行う。あらすじのテキストを kuromoji によって分かち書きをしたのち、文章の特徴を判別する際に重要である名詞・動詞・形容詞のみを抽出する。活用による変化などを kuromoji の辞書によって基本形へと統一し、保存する。この形態素解析後のテキストファイル名、及びその中で用いられている各単語の登場した箇所を把握する。この位置情報を単語ごとにまとめなおすことで転置インデックスを作成し、保存する。このインデックスを用いることで完全一致検索や部分一致検索を行うことが出来る。

しかし、単に検索語がデータに含まれているかどうかのみで検索を行うと、その検索語の重要性を加味することが出来ない。そのため、検索語により関係する演目が検索結果の上位に表示されるようにするために単語の重みづけを行う。本研究では単語の重要度を評価する際に tf-idf[6]を用いる。作成した転置インデックスの単語ごとに各文書における tf-idf 値を算出して、検索に用いる。

3. 演目検索

演目の検索を行う際には一般的なウェブ検索と同様に、1 つあるいはスペースで区切った複数の語句、もしくは文章を検索ボックスに入力する。入力された文章に対しては kuromoji による分かち書き、名詞・動詞・形容詞の抽出、活用形の基本形への修正を行う。ここで、入力された検索語に対する各演目の関連度を評価する。演目ごとに、入力された各検索語の tf-idf 値をデータベースから取得し、足し合わせた値を算出する。こうして検索語に対する各演目の関連度が算出される。これらから関連度が 0 の演目を除いたものを数値が高い順に並び変え、結果として出力する。

4. 条件による絞り込み

本研究では、キーワード検索だけでなく条件を指定し、その条件に合致する演目を絞り込む検索も併せて行う。そのために、あらかじめ演目の持つ性質などからメタデータを作成する。

本研究で演目に付与するメタデータは「分類」「古典・新作」「噺の長さ」「登場人物」「噺の舞台」「不適切な表現」の 6 つとする。「分類」は落語を内容によって滑稽噺、人情噺、怪談噺に分類したデータである。「古典・新作」はその落語の歴史によって古典、新作に分類したデ

Rakugo program search application.
†Dai Nakao, Meiji University
‡Sayaka Akioka, Meiji University

ータである。「噺の長さ」は演目を長さごとに小噺、普通、大ネタに分類したデータである。「登場人物」は演目に登場人物等の持つ性別、年代、職業、名前などの属性のうちどれが含まれているかどうかを登録したデータである。「噺の舞台」は演目の舞台となった場所の属性を登録したデータである。「不適切な表現」は江戸時代が続く古典芸能である落語には現代の価値観では不適切な表現が含まれている場合があり、そういった表現を含む演目を判別するデータである。

5. 評価

本研究ではキーワード検索、条件による絞り込み検索、2 つを合わせた複合検索の 3 機能を Java 上で実装した。図 1 に検索画面を示す。

図 1　複合検索の動作画面

このアプリケーションを 16 人のユーザーに利用してもらい、利用感についてアンケート調査を行った。なお、落語の知識レベルの指標として、落語会に足を運ぶ頻度についても併せて調査を行った。結果は「全く行かない」が 2 人、「たまに行く（数年に 1〜2 回）」が 4 人、「ときどき行く（2、3 か月に 1 回）」が 6 人、「よく行く（1 か月に 1〜2 回以上）」4 人がとなった。

キーワード検索に対して寄せられた意見として、「検索キーワードがあらすじの本文中の単語でなければならないので、古典落語を調べる際に昔の表現をしなくてはならないのが難しい」、「弔いで検索して出たものが葬式では出ないなどしたので、同義語も併せて検索できるとなお良い」などがあった。これらの意見から考えられる改善点として、同義語など似た意味を持つ言葉による検索が出来るようにする必要があると考える。そのためには潜在的意味解析などの手法を用いる必要がある。

条件検索に対して寄せられた意見として、「分類や登場人物、舞台などで検索をかけられる点が、噺をあまり知らない初心者に対して有用」、「噺の長さに関して基準をどこにおいていいか分からない」、「同じ演目でも演者や口演によって時間は振れ幅が大きいので基準が必要」、「一言だけ台詞があったり存在だけが確認できる人物を登場人物に含めていいか分からなかった」などがあった。これらの意見から、落語という演者や口演によって内容が変化することもある不安定な物語を定量的に評価、判別することは難しいという点が問題であると考える。このことから、演目の長さや登場人物の定義をガイドラインとしてしっかりと定義づけする必要があると考える。

6. おわりに

本研究では自然言語処理を用いた落語の演目検索システムを扱った。前項で記したように、キーワード検索、条件検索ともに改善の余地があることから、本稿の考察をもとに改善していきたいと考える。また、検索のシステムだけでなく GUI などアプリケーションそのものに対する意見も寄せられたので、そちらの方にも改善の余地があると考える。

● 参考

[1] 小坂 直輝, 小林 哲則, 林 良彦:隠れた良作の発掘を助ける Web 小説推薦システムの構成と評価, 人工知能学会全国大会論文集, 2020. https://doi.org/10.11517/pjsai.JSAI2020.0_3Rin477 (2021 年 01 月 05 日)

[2] 神谷 美希, 當間 愛晃:小説検索システムのためのプロット作成に関する基礎研究, 第 77 回全国大会講演論文集, 2015. http://id.nii.ac.jp/1001/00164245/ (2021 年 01 月 05 日)

[3] web 千字寄席「落語あらすじ事典」 https://senjiyose.com/archives/category/index (2020 年 12 月 18 日)

[4] atilika「Kuromoji」 https://www.atilika.com/ja/kuromoji/ (2020 年 12 月 18 日)

[5] GitHub https://github.com/atilika/kuromoji (2020 年 12 月 28 日)

[6] Juan Ramos:Using TF-IDF to Determine Word Relevance in Document Queries, In ICML, 2003. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.1424 (2021 年 01 月 05 日)