3N-8

そば口上データの自動分類とその教育システムの構築 *

○阿部 智恵, 杉山 雅英 (会津大学 大学院)

1 まえがき

本報告では会津地方に伝わる芸能「そば口上」の内容による分類と文化継承者育成の教育システムの構築について述べる. そば口上とは「そばのほめことば」であり, 昔は結婚式後の披露宴の席上でそばを振る舞う際に, それに節をつけておもしろおかしく演じる風習があった. そば口上の伝播や成立を調べるときに内容の分類が参考になる. 地域の文化はその担い手がいなくなるとすたれ消滅する. そば口上も例外ではなく, 披露する機会が減り口上を出来る演者は少ない.

ネットワーク上の膨大な数の文書中から必要な情報を含む文書を取り出す技術 [1], 複数の文書の類似性の計算の研究 [2], 自動分類の研究では, テキスト中の文字列の出現頻度に基づく方法 [3], グラフで表す用語関係から用語とクラスの関連性を計算する方法 [4] がある.

本報告では, そば口上テキストデータ (「会津そば口上」[5], ビデオ収集したそば口上 [6]) に着目した自動分類法の改良について述べその性能評価結果を報告する. また現在構築中のビデオデータ [6] を用いたそば口上教育システムとそば口上の自動分類結果の利用方法について述べる.

2 そば口上テキストの自動分類

「会津そば口上」に収録されている 74 のテキストデータを「茶筌」(ChaSen)[7] を用いて形態素解析して索引語を求め, ベクトル空間法を用いてテキストを式 (1) の文書ベクトルで表現する.

$$\boldsymbol{d}_j = (d_{1j}, d_{2j}, \cdots, d_{mj})^T, \tag{1}$$

ここで T は行列の転置を表し, d_{ij} は索引語 $w_i(i = 1, \cdots, m)$ のテキスト文書 $D_j(j = 1, \cdots, n)$ に対する重みである. 本報告での文書ベクトルの重みは式 (2) の TF×IDF を用いる [6].

$$d_{ij} = f_{ij} \times g_i, \left(g_i = \log_2 \frac{n}{n_i} \right). \tag{2}$$

f_{ij} は w_i の D_j における出現頻度である. n_i は w_i を含む文書数を表し, g_i は DF の逆数である. 対数化は IDF の値の変化を小さくするためである. $1 \leq n_i \leq n$ であるので, $0 \leq g_i \leq \log_2 n$ である. $n_i = n$ 即ち w_i が全ての文書に含まれる時, $g_i = 0$ となる.

2.1 未知語処理

ChaSen を用いて, テキストデータの品詞解析を行うと, 次に示す例のように未知語が出現し, 形態素解析性能が悪くなることがある.

(例)

先づ (まず 副詞)		
先	サキ	名詞-一般
づ	づ	未知語

きくらげ (木耳 名詞)		
きく	キク	名詞-一般
ら	ラ	名詞-接尾-一般
げ	ゲ	名詞-接尾-一般

そこで未知語処理として次の 3 つの処理を行う.

a ChaSen が認識可能な字体に統一 (17 箇所)
「ー」→ 「一」
「〇」→ 「○」
「ニ」→ 「二」

b テキスト表記の統一 (327 箇所)
(例)

変換前	変換後
あげざる	揚笊
掲ざる	揚笊
板かげらう	板影ろう
板かげろう	板影ろう
きくらげ	木耳

c 辞書への単語登録 (36 単語)
(例)
...(副詞 一般)) (見出し語 (ホウホケキョ 2875))...
2875 はコスト値を表し, ChaSen の同一品詞辞書の他のコスト値を参考にし, 一番大きい値に設定

これら 3 つの処理の形態素解析結果の性能, 分類への効果を検討する.

2.2 文書のクラスタリング手法

文書クラスタリングに K-means 法を使用する. ただし文書ベクトル間のユークリッド距離ではなくコサイン尺度で計算する類似度を用いた. 予備実験の結果は距離を用いた結果の方がクラス内における正解に安定性はあるが, 全体の正解数が減少する問題があった. クラスの重心ベクトル (centroid) をそのクラスに属する文書ベクトルの相加平均で定義する. 本報告ではクラスの重心ベクトルの変化量 D が閾値よりも小さくなるまで K-means 法を繰り返し処理を行う. 変化量の計算には式 (3) を用いる.

$$D = D(\boldsymbol{C}^t, \boldsymbol{C}^{t+1}) = \sum_{k=1}^{K} ||\boldsymbol{c}_k^t - \boldsymbol{c}_k^{t+1}||^2, \tag{3}$$

K はクラス数を表し, $\boldsymbol{C}^t = \{\boldsymbol{c}_k^t\}$ は K-means 法の t 回目の繰り返し処理で得られる k 番目のクラスの重心ベクトルである. $D \leq \theta$ となるとき処理を停止することとし $\theta = 0.00001$ とした.

2.3 新規テキストのクラスタリング方法

収録したビデオデータの書き起こしテキストを「会津そば口上」の 74 データから作成したクラスターに分類する. 収録したそば口上ビデオのテキスト 8 データをクラスタリングのために類似度を用いる. 「会津そば口上」74 データのクラスタリングが終了したときのクラスの重心ベクトルとのコサイン尺度が最大のクラスに属するものとする. テキストの索引語や重み付け方法は「会津そば口上」74 データと同一とする.

*Classification of "Soba Kojo" and Construction of Its Education Support System by T.Abe, M.Sugiyama (Graduate School, The Univ. of Aizu)

3　テキスト分類の評価実験

分類評価実験では索引語を形態素解析で得られる「名詞」とした．名詞は話し言葉でも変化することが少なく, 文書の特徴を最も表す品詞であると考えられる．

K-means 法において初期シードをランダムに与えるのではなく, 事前に分類したクラスの中から各々1 つの文書に対応するベクトルをシードとした．初期シードを固定したときに TF×IDF を用いて 2.1 で述べた 3 つの未知語処理の組み合せによる分類正解率を表 1 に示す．

表 1: 未知語処理の組み合せによる分類正解率の関係

未知語処理	未知語	索引語	最高 (%)	最低 (%)	平均 (%)
処理前	159	2247	78.378	39.189	61.327
a	157	2247	70.270	33.784	54.185
b	34	2200	82.432	51.351	70.908
c	132	2252	78.378	39.189	61.353
a, b	32	2200	82.432	51.351	70.789
a, c	130	2252	78.378	39.189	61.219
b, c	3	2206	82.432	51.351	71.065
a, b, c	0	2206	82.432	51.351	70.856

表 1 から, 処理 b を行うことで性能が良くなることが分かる．b は該当個所が他に比べると多いことが理由として考えられる．全ての未知語処理を行った後の結果は処理前と比べて分類正解率が 9.529%向上した．処理 a を行うと処理前よりも結果が悪くなるのは, 索引語に変化はないが全名詞数が変化し, 該当する単語の重みが変化したためだと考えられる．

ビデオデータの新規テキストの分類結果を表 2 に示す．ここで未知語処理は a, b, c 全てを行い, 8 データで未知語は 19 単語 (44 箇所) あり, 索引語は 74 データと同一の 2206 単語である．「会津そば口上」74 データを分類したときに与えた初期シード 2700 通り全ての結果から分類した．新規の 8 データの正解クラスは事前に主観的に与えた．この結果から新規のテキストに対してもほぼ同等の分類性能を与えることが判った．

表 2: 新規テキストデータ (8 文書) の分類正解率

	最高	最低	平均
TFxIDF	100.000%	25.000%	70.357%

4　そば口上教育システムの構築

そば口上の文化継承のために, 教育システムの構築を行う．要求項目は, 2 画面表示, 切り替え, 読み仮名付きの字幕表示, 方言意味解説の表示である．JAVA Media Framework (JMF) を用いて, 字幕付きビデオ再生教育システムの構築を行っている．動画再生に用いられるデータは, 2 つのアングルからの動画データ, 書き起こし字幕データ, そして, 方言 (表現) 解説データである．

そば口上には様々な動きがあり, 顔の表情や手足の動き, 会場内の移動もある．演者の動きや移動が記録された 2 つの動画データ (上半身・全体) があり, この 2 つの動画を表示することで全体の流れや細かい動きを学ぶことが可能となる．口上には方言があり, 聞き取れないことが考えられるため字幕表示をし, そのときに方言やそば口上特有の表現が現れたらその解説を表示する．

そば口上をネットワーク上で公開するためホームページを作成する．そば口上データはデータベースシステムで管理する．管理するデータは, 題目, 演者名, テキストデータや字幕データ等のファイル名と K-means 法や類似度を用いて得られた分類クラスの番号である．今後データが増えたときに対応可能となるように, データベースシステムによるページ表示を行うように HTML 文書を作成する．サーバークライアント技術を用いて, サーバーにデータファイルを置き, クライアント側で動画再生アプリケーションを起動するように整備する．

現状では, まだネットワークで利用可能となるように整備をしていないが, ローカルでシステムを起動した画面を図 1 に示す．

図 1: そば口上教育システム (動画再生)

アングルの異なる 2 つのビデオデータの再生, 字幕表示, 解説表示が可能である．2 つのビデオデータの同期が充分でない．動画データが大きいため起動までに多少時間を要する．また, 字幕表示に折り返し機能がついていないので, 長い文の場合は, 最後まで読めない等の問題がある．今後はこれらの改善を行い, さらにネットワーク上で利用可能となる様に整備をする．

5　むすび

本報告ではそば口上のテキストデータを自動分類する際のテキスト表記の統一, 形態素解析における未知語登録などの処理を行うことにより, 分類性能の改善効果, さらに, 新規テキストの分類性能について述べた．また, 収録したビデオデータを用いた字幕つきの教育システムについて述べた．

今後の課題としては, 構成要素 (段落) を考慮した文書ベクトル間の類似度の定義の検討, 教育システムの機能改善とネットワーク利用上で可能とすることである．

参考文献

[1] 金田, 他, “コーパスからのキーワード自動抽出,” 情報処理学会研究報告, Vol.2004, No.47, pp.1-7 (2004-05).
[2] 深谷, 他, “頻度統計と概念辞書を用いた文書の類似性の定量化,” 情報処理学会自然言語処理研究会, No.153, pp.73-79 (2003-01).
[3] 湯浅, 他, “大量文書データ中の単語間の共起を利用した文書分類,” 情報処理学会論文誌, Vol.36, No.8, pp.1819-1827 (1995-08).
[4] 相澤, 景浦, “グラフ的類似度による学術文献の自動分類に関する検討,” 言語処理学会第 4 回年次大会原稿 (1999-03).
[5] 元木, “会津そば口上,” 歴史春秋社 (1994-11).
[6] 阿部, 杉山, “そば口上ビデオデータの収集と自動分類,” 情報処理学会第 67 回全国大会, Vol.2, pp.401-402 (2005-03).
[7] 松本, “形態素解析システム「茶筌」,” 情報処理学会誌, Vol.41, No.11, pp.1208-1214 (2000-11).