5R-7

そば口上ビデオデータの収集と自動分類 *

○阿部 智恵, 杉山 雅英 (会津大学 大学院)

1 まえがき

本報告は会津地方に伝わる地方芸能の一つの「そば口上」の記録保存とその内容の分類を目的としている. 地域に伝わる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, 昔は結婚式で必ず行なわれたが今では式場で挙式をすることから披露する機会もなくなり, 口上を出来る演者は多くない. そば口上の伝播や成立を調べるために内容の分類が参考になる.

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が行われてきた. 自動分類の先行研究では, テキスト中の文字列の出現頻度に基づく方法 [2] や予め定義されたシソーラスを用いる方法 [3], また用語関係をグラフで表し, 用語とクラスの関連性を計算する方法がある [4].

本報告では「そば口上」データベースの収集とその書き起こしテキストに着目した自動分類について述べる.

2 そば口上データの収集

「そば口上」とは「そばの褒め詞 (ほめことば)」であり, 会津の伝統芸能の一つである. 会津地方では結婚式の後の披露宴の席でそばを振る舞い, 客人にそばをおいしく食べてもらうために「そば口上」に節をつけておもしろおかしく語る風習があった. 現在では, 各地域で行われる秋の新そば祭りでの一つのアトラクションとして行われている. そば口上の記録保存のためデータ収集を行った. データベースはテキストデータとビデオデータの 2 種類ある.

2.1 テキストデータ

テキストデータベースは元木氏による「会津そば口上」[5] のテキストの書き写しと, 2.2 で述べる収録したビデオデータの書き起こしテキストとがある.

元木氏は会津各地に伝わる 74 の口上を集めた. その一部は既に文書化されたものの再録であり, 一部はそば口上音声を書き起こしたものである. 各文書は内容に従って段落付けされている. ビデオデータの書き起こしは, ビデオ中の音声を聴いて書き起こし, 演者に書き起こしテキストの確認を行った.

2.2 ビデオデータ

実際のそば口上を演じているデータを収集するために, 会津の各地域で開催された新そば祭りを訪れ, そば口上演者の口上をビデオ収録した. 収録に使用した機材はデジタルビデオカメラ 2 台で, 1 台 (Panasonic NV-MX5000) で口上演者の上半身を, もう 1 台 (SONY DCR-VX1000) で口上演者や全体が入るように撮影した. 口上演者が会場内を移動する可能性があるため, その移動記録のために全体用カメラを用いた. そば祭りでは接話マイクやピンマイクを使用してもらうことが難しい. そのため, カメラの内蔵マイクのみを使用したので祭り参加者の話し声やざわめき声等が含まれている.

収集したビデオデータ数は 10 で, そば口上の演者数は 8 名 (男 6 名, 女 2 名) であり, 地域数は 4 である. 今後データを増やす予定である.

2.3 そば口上の構成要素

そば口上テキストデータを分析した結果, 構成要素は以下の 6 つとなった.

1. (W: Wedding) 祝い事に関する文章
2. (L: Life History) そばの生い立ちに関する文章
   - L_0: 「種を蒔いて」から「三角」まで
   - L_1: 「誕生」から「膳」まで
   - L_2: 「口」から「腹の町」まで
3. (S: Spices) そばの薬味に関する文章 (1 から 10 まで)
   - S_0: 部分的にある (1 から 3 まで等)
4. (A: Artisan) そば職人の宣伝に関する文章
5. (V: Vender) そば売りに関する文章
6. (O: Others) 上記以外の文章
   - そばを食べるときの状況
   - そばとラーメンの売り文句
   - そばの侍について
   - そばの食べかた
   - 種を蒔いてからそばができるまで (人から見た)

L_0, L_1, L_2 はこの順序で現れ, 一部が省略されることがある. S_0 は S の一部だけのものである. 構成要素の出現順番は, O - L_0 - W - L_1 - S (S_0) - L_2 - A - V である. W が L_0 の前や A の後, V が L_1 の後や A の前, A が S の前の様に移動することがある. これらの要素は省略が可能である. 文書ごとに含まれている構成要素を基に分類した結果を表 1 に示す.

表 1: そば口上テキストの構成要素による分類

クラス	テキスト数	主な構成要素
C1	30	L_1, S, L_2
C2	10	L_1, L_2, A, V
C3	4	L_0, A, V
C4	9	A, V
C5	4	W
C6	9	L_0
C7	8	O
計	74	

3 テキストの自動分類

74 のテキストを日本語形態素解析器「茶筌」[6] 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 $w_i (i = 1, \cdots, m)$ の文書 $D_j (j = 1, \cdots, n)$ における重みである.

*Construction of SobaKojo Video Data and Its Automatic Classification, by T.Abe, M.Sugiyama (Graduate School, The Univ. of Aizu)

3.1 索引語の重み

文書の内容を特徴付ける上で必要な単語 (索引語) を設定し, その単語の文書中に出現する頻度に応じて重みを付ける. 文書を特徴付ける上で, 索引語の設定とそれの出現頻度に応じた重み付けは非常に重要である.

重み付けには索引語頻度 (TF) と文書頻度の逆数 (IDF) を使用する. TF は索引語 w_i のテキスト D_j における出現頻度 f_{ij} で表される. 文書頻度 (DF) は文書による索引語出現の偏りを表す. DF は索引語 w_i を含む文書数で定義され n_i で表される. IDF は DF の逆数で式 (2) で定義される.

$$g_i = \log_2 \frac{n}{n_i}, \tag{2}$$

ここで n は文書の総数を表す. 対数化は IDF の値の変化を小さくするためである. $n_i \leq n$ であるので, $g_i \geq 0$ 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.

3.2 文書ベクトルの重み

索引語 w_i のテキスト D_j における重みを以下の 2 通りとする.

1. TF $\quad d_{ij} = f_{ij}$ (3)
2. TF×IDF $\quad d_{ij} = f_{ij} \cdot g_i$ (4)

3.3 文書のクラスタリング

文書クラスタリングに K-means 法を使用する. ただし、文書ベクトル間の距離ではなくコサイン尺度で計算する類似度を用いた. 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のだが、全体の正解数が減少する等, 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 (5) を用いる.

$$D = D(\boldsymbol{C}^t, \boldsymbol{C}^{t+1}) = \sum_{k=1}^{K} ||\mathbf{c}_k^t - \mathbf{c}_k^{t+1}||^2 \tag{5}$$

$\boldsymbol{C}^t$ 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 $\theta = 0.1$ とした.

4 テキスト分類の実験結果と評価

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. その名詞 (索引語) 数は 74 テキストに対して 2,096 である.

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, 表 1 に示した 7 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 と TF×IDF を用いた分類結果の正解文書数は表 2 のようになった.

表 2: 重み付けの違いによる正解文書数の比較

重み	クラス	C1	C2	C3	C4	C5	C6	C7
	正解数	30	10	4	9	4	9	8
TF	41	27	5	3	0	3	2	1
TF×IDF	47	24	5	3	8	4	2	1

TF×IDF の方が最大のクラス C1 に分類された正解数は減っている. しかし全体の正解数は若干増えている. また C4 に分類されているものがほぼ正解数と変わらない. 現状では, TF×IDF を用いたときの重み付けが有効である.

また初期シードと分類性能の関係を調べるため C1 の初期シードのみ変更し, 次のような結果を得た.

1. 平均的に TF よりも TF×IDF の方が正解数が多い.
2. TF は C1 と C2 と C6 に, TF×IDF は C1 と C6 における正解数の変動が大きい.
3. 正解数が TF は 33〜44, TF×IDF は 33〜49 の様に正解にばらつきがある.

初期シードの与え方によって, テキスト分類性能に影響があることが分かった.

5 むすび

本報告ではそば口上のビデオデータ及びテキストデータの収集について述べ, テキストデータを自動分類する実験を行った. テキストに含まれる名詞を索引語として出現頻度に応じて重みを計算し, ベクトル空間法を用いて類似度を求め, K-means 法による分類を行った. 結果として初期シードによる分類への影響はあるが, 現状は TF×IDF による重み付けでテキスト分類をすることが有効であることが判った.

今後の課題として, 自動分類に関して表記や読みの統一, 形態素解析における辞書登録, 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, テキスト分類に用いるアルゴリズムを検討する. またビデオに口上のテキストを字幕として表示することにより, そば口上教育システムとして整備する.

謝辞

研究を進める上で御支援頂いたヒューマンインタフェース学講座の皆様, そば口上ビデオデータ収録の撮影に協力していただきました皆様に感謝いたします.

参考文献

[1] 金田, 藤野, 斉藤, 上田, “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] 湯浅, 上田, 外川, “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).

[3] 福本, 鈴木, 福本, “辞書の語義文を用いた文書の自動分類,” 情報処理学会論文誌, Vol.37, No.10, pp.1789-1799 (1996-10).

[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).

[6] 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).