令和２年度

修士論文

表整理技術を用いた文書群からのテンプレート生成

指導教員

村田真樹

鳥取大学大学院 持続性社会創生科学研究科

博士前期課程 工学専攻 情報エレクトロニクスコース
自然言語処理研究室

M19J4053U 守優太朗

概要

近年，インターネット上の文書から情報を取捨選択することが多い．しかし，文書の量は膨大であり，情報の取捨選択を効率的にする手法が求められている．

過去に岡﨑ら [1] は，文書群から重要な情報を文単位で抽出し，表の形に整理する手法を提案した．文書群から文単位で抽出して文をベクトルで表現した後，得られたベクトルを X-means 法 [2] でクラスタリングし，文書ごとに表に整理し，表示していた．ここで文書群とは，同じ種類の文書を集めたものである．例えば，異なる人に関する情報の記事において，「人名」や「生年月日」などの情報が種類別に表に整理される．このように，同種の文書群の情報が種類ごとに整理されることで，情報の取捨選択が効率的になり，文書間の情報の比較にも役に立つ．さらに，岡﨑ら [3] はクラスタリング手法の改案として階層クラスタリングによる表整理の手法を提案した．

そこで本研究ではこの技術の「文書群の情報が種類ごとに整理される」という点に着目し，大量の文書群からのテンプレート生成する手法を提案する．例えば，人に関する情報をクラスタリングして表生成する場合，「人名」や「生年月日」といった各重要な情報がクラスタとして分けられる．この重要な情報の各クラスタを変数に置き換えれば，「人名」と「生年月日」の情報が入った文を生成する時に，変数「人名」，変数「生年月日」を任意の単語に変えることで，様々なパターンの文が生成できるため，文書作成支援への応用に期待ができる．提案手法ではまず，文書群を階層クラスタリングでクラスタリングし，表に整理する．このクラスタリング結果について各列をテンプレートの変数の各グループとする．一列目に含まれる単語を変数 X1、2 列目を X2... と置換し元文に当てはめることでテンプレートが作成され，本研究ではこのテンプレート生成を試みる．

150 件の記事の入力データを 2 種類用意し実験を行った結果，文章レベルでの評価結果は，データ 1 を正解データ，データ 2 を実験データとした時のカバー率とデータ 2 を正解データ，データ 1 を実験データとした時のカバー率がそれぞれ 0.13，0.11 と共に低い結果となった．また，文レベルでの評価の結果，カバー率が最も高いもので「血液型」の列の 0.78，最も低いもので「本名」の列の 0.04 となり，一部ではカバー率の高いテンプレートが確認できた．

目 次

第 1 章　はじめに **1**

第 2 章　先行研究 **3**

2.1　テンプレートに関する先行研究 3

2.1.1　スポーツ要約生成のためのテンプレート生成 3

2.1.2　料理レシピの手順書の自動生成のためのテンプレート生成 3

2.2　表整理に関する先行研究 4

2.2.1　関連研究の手法の手順 4

2.2.2　手順 2：文ベクトルの計算 6

2.2.3　手順 3：階層クラスタリング 8

2.2.4　手順 4：各クラスタ数でのクラスタリング 9

2.2.5　手順 5：クラスタリング結果を表に整理 10

2.2.6　手順 6：列の項目名の求め方 12

第 3 章　提案手法 **13**

3.1　テンプレート 13

3.2　提案手法 1：文章レベルでのテンプレート生成 15

3.3　提案手法 2：文レベルでのテンプレート生成 16

3.4　原文への変数の置き方 18

第 4 章　実験環境 **19**

4.1　実験データ 19

4.2　MeCab 21

4.3　単語ベクトルモデル 21

第 5 章　実験 **23**

5.1　評価方法 23

5.1.1 文章レベルのテンプレートの評価方法 23
5.1.2 文レベルのテンプレートの評価方法 23
5.2 実験結果 25
5.2.1 文章レベルでの結果 28
5.2.2 文レベルでの結果 30
5.3 評価結果 32
5.3.1 文章レベルでの評価結果 32
5.3.2 文レベルでの評価結果 35

第 6 章 考察 **38**
6.1 文章レベルでの実験結果について 38
6.2 文レベルでの実験結果について 40
6.2.1 カバー率の結果 40
6.2.2 実際に生成されたテンプレートの例 42
6.3 頻出頻度の多い単語について 44
6.4 形態素解析について 45

第 7 章 今後の課題 **46**

第 8 章 おわりに **47**

表 目 次

2.2.1 文の密集率が高いクラスタの例 10
2.2.2 文の密集率が低いクラスタの例 10

4.1.1 Wikipedia のデータの例（下線部の内容を実験に用いる） 20
4.2.1 辞書による違いの例 21
4.3.1 学習データの例 22

5.2.1 生成された表の一部 26
5.2.2 生成された表の一部 (続き) 27
5.2.3 文章レベルでのテンプレートの結果の一部 28
5.2.4 文章レベルでのテンプレートの結果の一部 (続き) 29
5.2.5 列「日本」のクラスタリング結果の一部 30
5.2.6 列「日本」から生成されたテンプレートの一部 31
5.3.1 データ 2 でも生成されたデータ 1 のテンプレート 32
5.3.2 データ 1 でも生成されたデータ 2 のテンプレート 33
5.3.3 文章レベルのテンプレートの評価結果 34
5.3.4 データ 2 でも生成されたデータ 1 のテンプレート (1/2) 35
5.3.5 データ 2 でも生成されたデータ 1 のテンプレート (2/2) 35
5.3.6 データ 1 でも生成されたデータ 2 のテンプレート (1/2) 36
5.3.7 データ 1 でも生成されたデータ 2 のテンプレート (2/2) 36
5.3.8 データ 1 のテンプレートの評価結果 37
5.3.9 データ 2 のテンプレートの評価結果 37

6.2.1 データ 1 での「血液型」の列のクラスタリング結果 40
6.2.2 データ 2 での「血液型」の列のクラスタリング結果 40
6.2.3 データ 1 での「卒業」の列のクラスタリング結果 41
6.2.4 データ 2 での「卒業」の列のクラスタリング結果 41

6.3.1 データ 1 での「出身」の列のクラスタリング結果 44

図 目 次

2.1 階層クラスタリングによる表生成の手順の例 5
2.2 文ベクトルの計算手順の例 7
2.3 密集率の計算の例 10
2.4 クラスタの項目名の求め方の例 12

3.1 生成するテンプレートの例 14
3.2 文章レベルでのテンプレート生成の例 15
3.3 文レベルでのテンプレート生成の手順 17
3.4 変数を原文に当てはめる手順 18

5.1 評価方法の例 24

6.1 文章レベルのテンプレートの良かった例 38
6.2 文章レベルのテンプレートの悪かった例 39
6.3 文レベルのテンプレートの例 (1/2) 42
6.4 文レベルのテンプレートの例 (2/2) 43

第1章　はじめに

近年，インターネット上の文書から情報を取捨選択することが多い．しかし，文書の量は膨大であり，情報の取捨選択を効率的にする手法が求められている．

過去に岡崎ら [1] は，文書群から重要な情報を文単位で抽出し，表の形に整理する手法を提案した．文書群から文単位で抽出して文をベクトルで表現した後，得られたベクトルを X-means 法 [2] でクラスタリングし，文書ごとに表に整理し，表示していた．ここで文書群とは，同じ種類の文書を集めたものである．例えば，異なる人に関する情報の記事において，「人名」や「生年月日」などの情報が種類別に表に整理される．このように，同種の文書群の情報が種類ごとに整理されることで，情報の取捨選択が効率的になり，文書間の情報の比較にも役に立つ．さらに，岡崎ら [3] はクラスタリング手法の改案として階層クラスタリングによる表整理の手法を提案した．

そこで本研究ではこの技術の「文書群の情報が種類ごとに整理される」という点に着目し，大量の文書群からのテンプレート生成する手法を提案する．例えば，人に関する情報をクラスタリングして表生成する場合，「人名」や「生年月日」といった各重要な情報がクラスタとして分けられる．この重要な情報の各クラスタを変数に置き換えれば，「人名」と「生年月日」の情報が入った文を生成する時に，変数「人名」，変数「生年月日」を任意の単語に変えることで，様々なパターンの文が生成できるため，文書作成支援への応用に期待ができる．提案手法ではまず，文書群を階層クラスタリングでクラスタリングし，表に整理する．このクラスタリング結果について各列をテンプレートの変数の各グループとする．一列目に含まれる単語を変数 X1、2 列目を X2... と置換し元文に当てはめることでテンプレートが生成され，本研究ではこのテンプレート生成を試みる．

本研究の主張点を以下に示す.

新規性

テンプレート生成の研究は多くあるが本研究では情報抽出を活かし，文書群から表を生成し，そこからテンプレートを生成する.

有用性

テンプレートの生成によって，文章の作成において参考になり文書作成支援に繋がる．また，様々な種類の入力データへの応用が期待できる.

性能

文章レベルでのテンプレートのカバー率の平均は 0.12，文レベルでのテンプレートのカバー率の平均は 0.39 であった.

第2章　先行研究

2.1　テンプレートに関する先行研究

2.1.1　スポーツ要約生成のためのテンプレート生成

田川ら [4] は野球のイニング速報に注目し，試合の状況を簡潔に伝えるテンプレート型の文生成を試みている．イニング速報ではそのイニングで起こったイベントが文章としてまとめられており，この文章をより簡潔にするために圧縮してテンプレート化し，要約文を生成している．文章を圧縮し，選手名が入る部分を「NAME」，得点の部分は「SCORE」，ホームランやタイムリーヒットといった野球専門用語は「ACTION」としてスロット化し，このスロットに任意のイニングのデータを補完することで文を生成している．

2.1.2　料理レシピの手順書の自動生成のためのテンプレート生成

山崎ら [5] は料理レシピに着目し，レシピの原文からテンプレートを生成し，そのテンプレートに重要語を代入して手順文を自動生成している．テンプレート生成の手順としては，まずレシピの原文を単語毎に分割し，品詞を付与する．単語列に対して，レシピ固有表現を表すタグを r-NE 認識器：PWNER[6] を用いて付与する．例えば，食材であったり道具，食材の動作といった部分に対してタグを付与している．それにより，タグ部分に任意の単語列を入力することで料理の手順書の自動生成を試みている．

2.2 表整理に関する先行研究

岡崎ら [1] は，関連する内容の文書群から情報を抽出し，表に整理する手法を提案していた．

2.2.1 関連研究の手法の手順

以下の手順で文書群から表を自動で生成する．手順の概要図を図 2.1 に示す．

手順 1 文書群に含まれる文を句点区切りで抽出する．

手順 2 手順 1 で分割された各文の文ベクトルを計算する．

手順 3 文ベクトルを Ward 法による階層クラスタリングでクラスタリングする．

手順 4 階層クラスタリングによって得られた各クラスタ数でのクラスタリング結果を基に表に整理する．

手順 5 手順 4 で採用されたクラスタ数でのクラスタリングの結果を，行を文書，列をクラスタとする表に整理する．

手順 6 表の各列について，項目名を付与する．

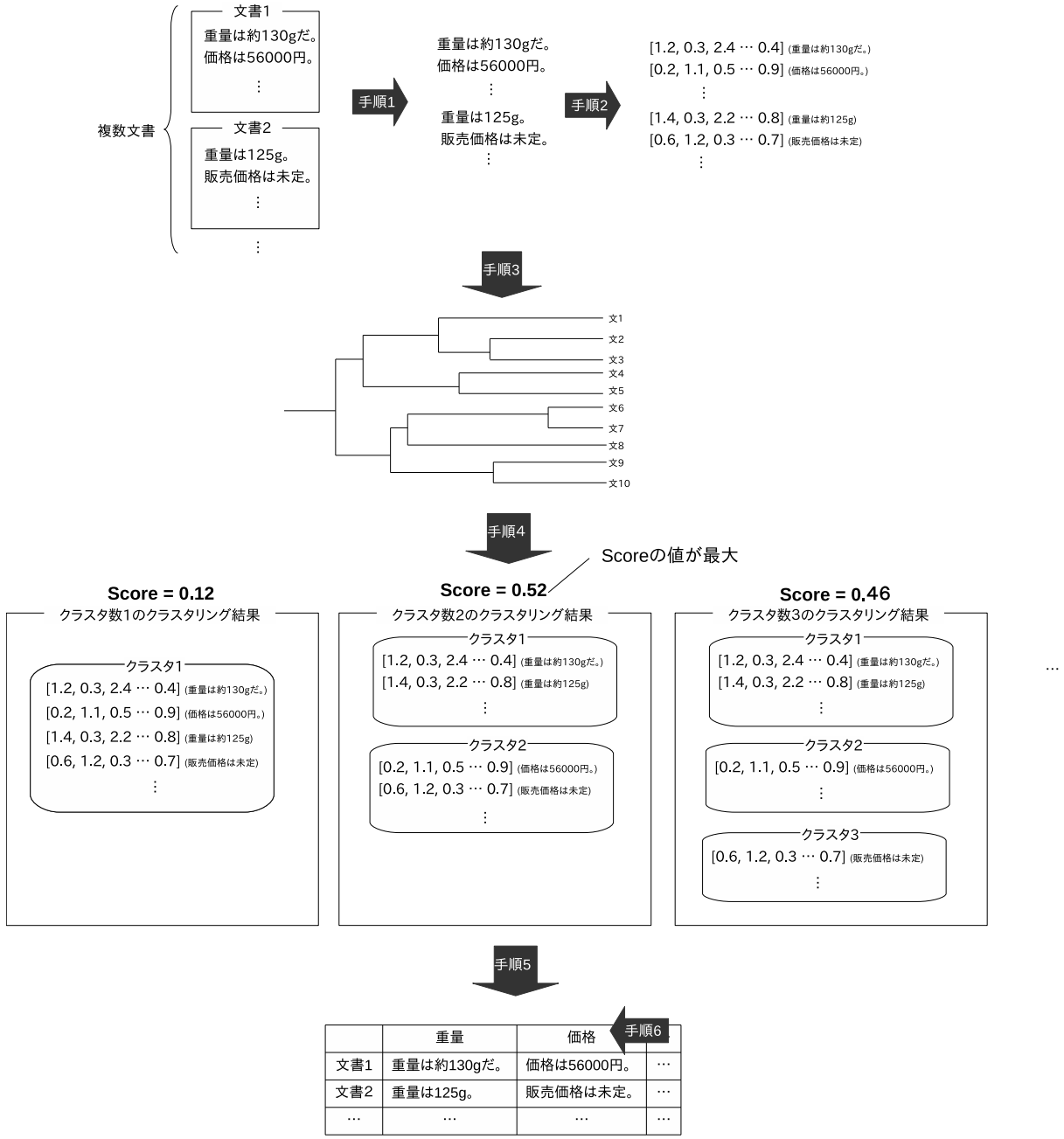

図 2.1: 階層クラスタリングによる表生成の手順の例

2.2.2 手順2：文ベクトルの計算

2.1節の手順2における文ベクトルの計算方法を説明する．文ベクトルは以下の手順で求める．図2.2に文ベクトルの計算手順の例を示す．

(1) 文を格要素ごとに分割する．

(2) 分割された格要素ごとに以下の手順で格要素ベクトルを求める．

  (a) 文をMeCab[1]を用いて形態素解析する．

  (b) 形態素解析結果のうち，品詞が名詞でかつ，品詞分類1が代名詞，数，非自立，副詞可能でない単語を抽出する．

  (c) 抽出した単語のベクトルの平均を格要素ベクトルとする．

(3) 格要素ベクトルの総和を文ベクトルとする．

[1] http://taku910.github.io/mecab/

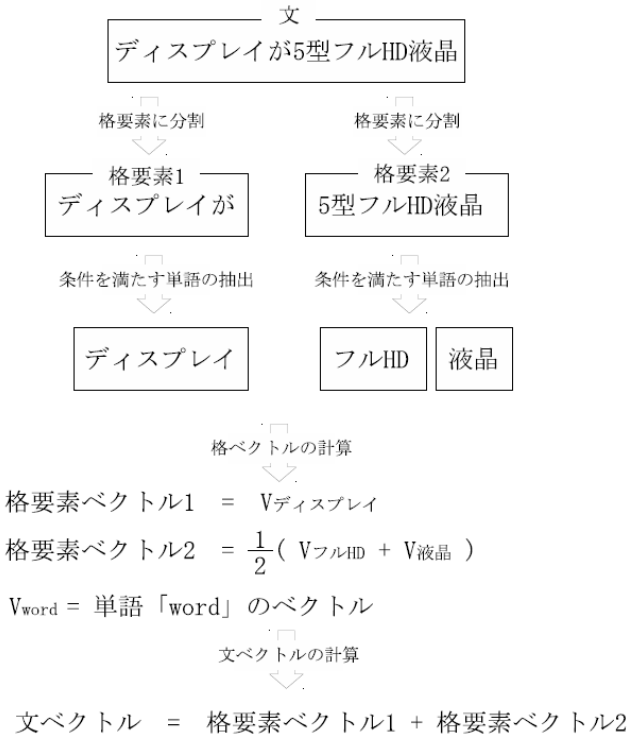

図 2.2: 文ベクトルの計算手順の例

2.2.3　手順 3：階層クラスタリング

階層クラスタリングは，距離の最も近いクラスタ同士の統合を繰り返すクラスタリング手法である．階層クラスタリングはクラスタ間の距離の定義の違いによっていくつかの手法が存在するが，岡崎らは Ward 法を用いている．Ward 法ではクラスタ間の距離 $D(C_1, C_2)$ を以下のように定義している．

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

$$\mathrm{E}(C_i) = \sum_{\mathbf{x} \in C_i} (d(\mathbf{x}, \mathbf{c}_i))^2$$

$$c_i = \sum_{\mathbf{x} \in C_i} \mathbf{x} / |C_i|$$

2.2.4 手順 4：各クラスタ数でのクラスタリング

階層クラスタリングによって得られた各クラスタ数でのクラスタリング結果を基に表に整理する．ここで，クラスタリングの際に，情報がどの文書に含まれていたかは考慮されないため，1 つのセルに複数の情報が含まれる場合がある．クラスタ数 k での表の埋まり具合を式 (2.1) から，情報の密集度を式 (2.2) からそれぞれ求める．$|c_{k,i}|$ はクラスタ数 k での表の i 番目の列に含まれる文の総数，$d_{k,i,j}$ はクラスタ数 k での表の i 番目の列の j 番目の文のベクトル，C_k はクラスタ数 k での表の列の総数，$cosine(x, y)$ は x，y のコサイン類似度を求める関数を表す．

$$cover_k = \frac{\text{クラスタ数 } k \text{ での表の埋まっているセルの数}}{\text{クラスタ数 } k \text{ での表のセルの総数}} \tag{2.1}$$

$$density_k = \min_{j \neq h}(cosine(d_{k,i,j}, d_{k,i,h})) \quad i = 1, \cdots, C_k \quad j, h = 1, \cdots, |c_k, i| \tag{2.2}$$

ここで, 全てのクラスタ数での $cover_k$ の集合を $COVER$, $max(COVER)$ を集合 $COVER$ の最大値，$min(COVER)$ は集合 $COVER$ の最小値とする．各クラスタでの $cover_k$ を式 (2.3) で 0〜1 の範囲に正規化する．

$$norm(cover_k) = \frac{cover_k - min(COVER)}{max(COVER) - min(COVER)} \tag{2.3}$$

同様に，全てのクラスタ数での $density_k$ の集合を $DENSITY$，$max(DENSITY)$ を集合 $DENSITY$ の最大値，$min(DENSITY)$ は集合 $DENSITY$ の最小値とする．各クラスタでの $density_k$ を式 (2.4) で 0〜1 の範囲に正規化する．

$$norm(density_k) = \frac{density_k - min(DENSITY)}{max(DENSITY) - min(DENSITY)} \tag{2.4}$$

クラスタ数 k での表の $Score_k$ を式 (2.5) より求める．$Score_k$ が最大となるときのクラスタ数 k を最適なクラスタ数として採用する．

$$Score_k = norm(cover_k) \times norm(density_k) \tag{2.5}$$

2.2.5　手順 5：クラスタリング結果を表に整理

2.1 節の手順 4 におけるクラスタごとの重要度の計算方法を説明する. クラスタリング結果には表 2.2.1 のように関連する文だけで構成される密集率の高いクラスタもあれば, 表 2.2.2 のように関連性のない文で構成される密集率の低いクラスタもある. 密集率の高いクラスタは重要であると考えられる. よって, k 番目のクラスタの密集率 d_k を式 2.6 のように定める. ここで, N_k は k 番目のクラスタに含まれる文の総数であり, $S_{k,l}$ は k 番目のクラスタに含まれる l 番目の文のベクトルであり, $S_{k,mean}$ は k 番目のクラスタに含まれる文のベクトルの平均である. 密集率の計算の例を図 2.3 に示す.

$$d_k = \frac{1}{N_k} \sum_{l=1}^{N} \frac{S_{k,l} \cdot S_{k,mean}}{|S_{k,l}||S_{k,mean}|} \tag{2.6}$$

表 2.2.1: 文の密集率が高いクラスタの例

	クラスタ 1
文書 1	重量は約 130g
文書 2	重量は 125g
文書 3	重量は 140g
文書 4	重量は 138g

表 2.2.2: 文の密集率が低いクラスタの例

	クラスタ 2
文書 1	重量は約 130g
文書 2	価格は 49800 円
文書 3	メモリーは 4GB
文書 4	12 月 9 日に発売予定

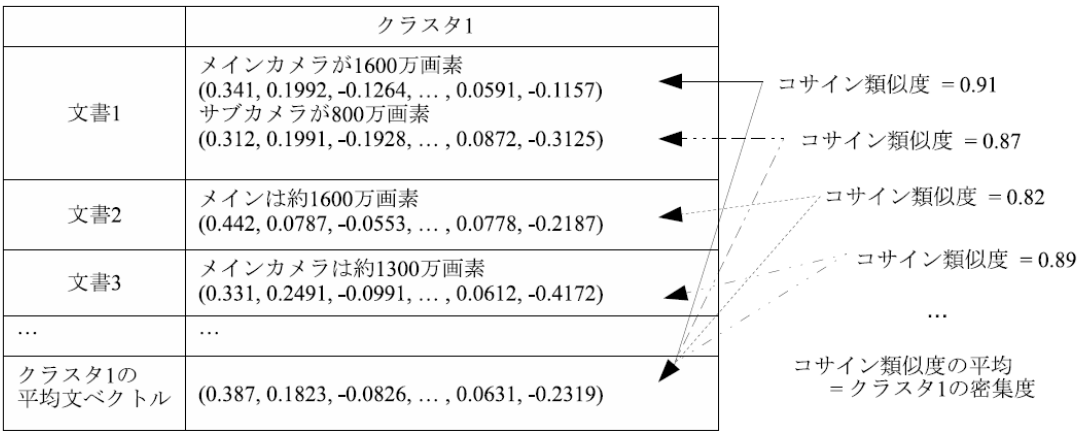

図 2.3: 密集率の計算の例

式 2.6 で求めたクラスタの密集率 d_k を, 式 2.7 を用いて, 最小値が 0, 最大値が 1 になるように正規化する. ここで, nd_k は k 番目のクラスタの正規化されたクラスタの密集率であり, K はクラスタの総数である.

$$nd_k = \frac{d_k - d_{min}}{d_{max} - d_{min}} \tag{2.7}$$

$$d_{min} = \min_{1 \leq k \leq K} d_k \tag{2.8}$$

$$d_{max} = \max_{1 \leq k \leq K} d_k \tag{2.9}$$

多くの文書の情報を含むクラスタほど重要であると考えられる. よって, k 番目の文書カバー率 c_k を式 2.10 のように定める. p_k は k 番目のクラスタにおいて文を抽出できた文書の数であり, P は文書の総数である.

$$c_k = \frac{p_k}{P} \tag{2.10}$$

式 2.10 で求めた文書カバー率 c_k を, 式 2.6 を用いて, 最小値が 0, 最大値が 1 になるように正規化する. ここで, nc_k は k 番目のクラスタの正規化された文書カバー率であり, K はクラスタの総数である.

$$nc_k = \frac{c_k - c_{min}}{c_{max} - c_{min}} \tag{2.11}$$

$$c_{min} = \min_{1 \leq k \leq K} c_k \tag{2.12}$$

$$c_{max} = \max_{1 \leq k \leq K} c_k \tag{2.13}$$

k 番目のクラスタの重要度 i_k を式 2.14 のように定義する.

$$i_k = nd_k \times nc_k \tag{2.14}$$

2.2.6　手順 6：列の項目名の求め方

2.1 節の手順 6 におけるクラスタごとの項目名の求め方の概要を図 2.4 に示す. 生成された表の各クラスタについて, 以下の手順でクラスタの項目名を付与する.

(1) クラスタに含まれるの各文について, 文に含まれる単語のうち品詞が名詞のものを抽出する.

(2) 1 で抽出した各単語について, 文書頻度を求める.

(3) 文書頻度が最大の単語をクラスタの項目名として付与する.

(4) 文書頻度が最大の単語が複数ある場合は, 読点で区切って全て付与する.

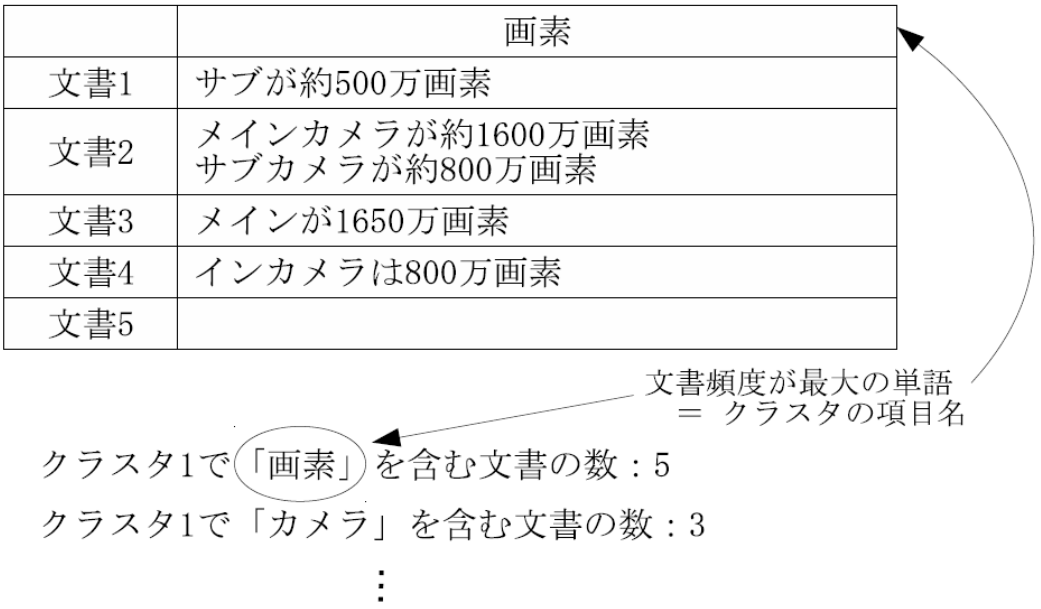

図 2.4: クラスタの項目名の求め方の例

第3章　提案手法

本研究では，文書群を階層クラスタリングでの表生成を利用し，テンプレートを生成する．表生成は 2.2 節の手法を用いる．提案手法では先行研究と違い，入力データの種類にとらわれず様々な入力データに対応できると考えている．

3.1　テンプレート

本研究で生成を取り組むテンプレートとは，文書作成支援になるような文の雛形のようなものであり，変数の部分を任意の文字列で補完することで様々な文章が生成可能となる．また，先行研究のように特定の種類のデータだけでなく，様々なデータに対してテンプレートの生成が期待できる．また，文章を書く際の構成や順番といった情報も得ることができる．以下に本研究で生成を取り組むテンプレートの例を図 3.1 示す．

例えば,「日本の首都は東京で、人口は約 900 万人です。」という文について考える．この文の名詞部分である「日本」,「首都」,「東京」,「人口」,「900 万人」という部分を X1, X2，X3，X4，X5 と変数化する．この変数部分を任意の文字列で補完することで,「ドイツの首都はベルリンで、面積は約 891km^{2}です。」という文を生成できる．このようなテンプレートの生成を本研究で取り組む．

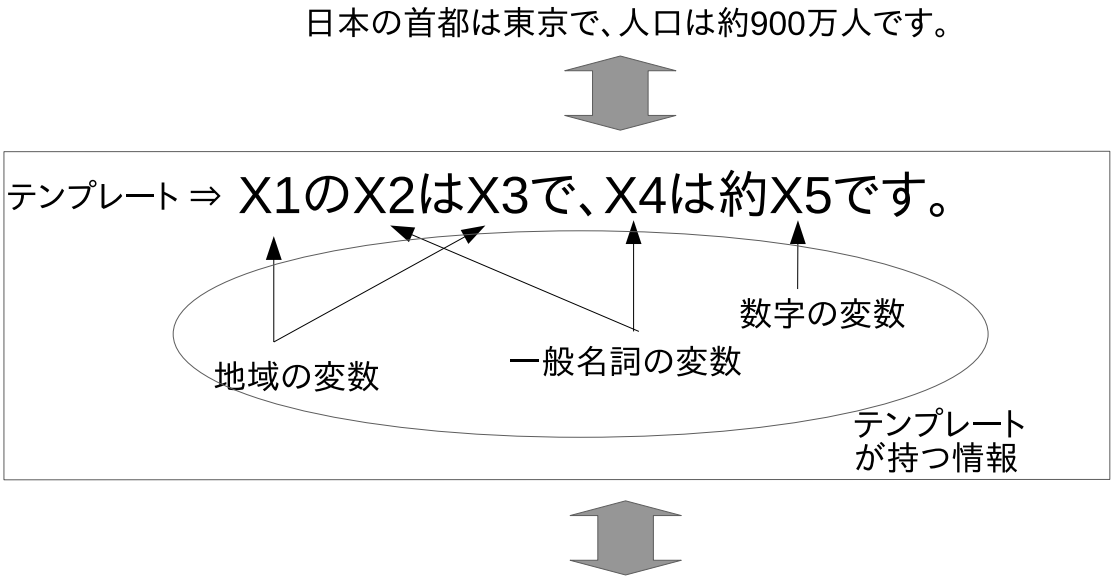

図 3.1: 生成するテンプレートの例

3.2　提案手法1：文章レベルでのテンプレート生成

この方法では文章単位のテンプレートを生成する．文書群を階層クラスタリングによってクラスタリングし表を生成する．その表に対して各列を変数のグループとし，左の列から順に変数 X0，X1... のグループとする．手順の概略を図 3.2 に示す．

表の列は情報のクラスタ毎に，行は文書として整理されている．列「重量」に含めれている文字列は変数 X1 となり，列「価格」に含まれる文字列は変数 X2 となる．よって，「重量は約 130g だ。」が X1 に，「価格は 56000 円。」が X2 に置換され，「X1。X2」というテンプレートが生成される．

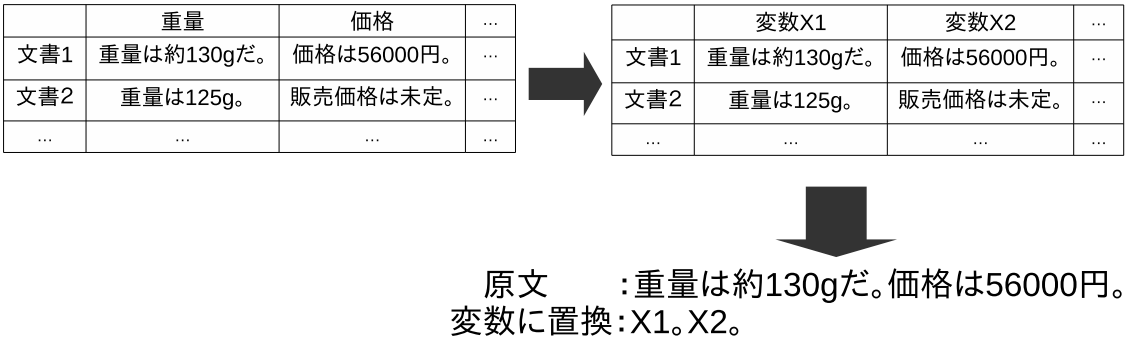

図 3.2: 文章レベルでのテンプレート生成の例

3.3　提案手法2：文レベルでのテンプレート生成

この手法では句点までで終わる短い文レベルでのテンプレートを生成する．作成された表に対して，文を全て名詞のみの状態にし再度列ごとでクラスタリングを行い，テンプレートを生成する．2 章の方法で生成された表からテンプレート生成までの手順を以下に示す．また，手順の概略を図 3.3 に示す．

手順 1 生成された表の全文を名詞のみの状態にする．テンプレートを生成するにあたり，本研究では名詞をテンプレートの変数と設定した．表中の文章に対して MeCab で形態素解析を行い，名詞以外の品詞は除去する．

手順 2 名詞のみの状態で列ごとに再度クラスタリングを行う．

手順 3 手順 2 で生成された表の各列をテンプレートの各変数とする．列の左側から順に X1，X2 と設定し，その列に含まれる単語が変数となる．

手順 4 元の文章に対して上記の変数を適応させ，テンプレートを生成する．

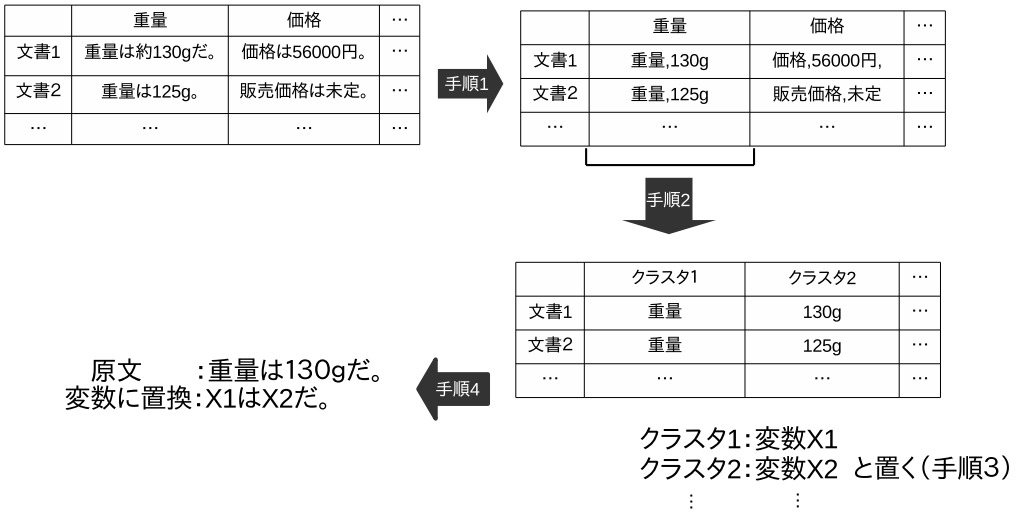

図 3.3: 文レベルでのテンプレート生成の手順

3.4　原文への変数の置き方

3.2 節，3.3 節にて共通する，テンプレート生成において原文への変数の置き方について説明する．表のクラスタを変数として，その変数を原文に当てはめる際，変数の文字列を全て原文と一致するか比較し，一致した場合変数に置換される．例を図 3.4 に示す．原文に「鳥取大学卒業です」という原文があった場合，表内の各変数を X1 から順に原文と比較する．この時，変数内に「鳥取」という単語と「鳥取大学」という単語があるが長い方である「鳥取大学」の部分を変数とする．これにより，より適切なテンプレートを生成できる．

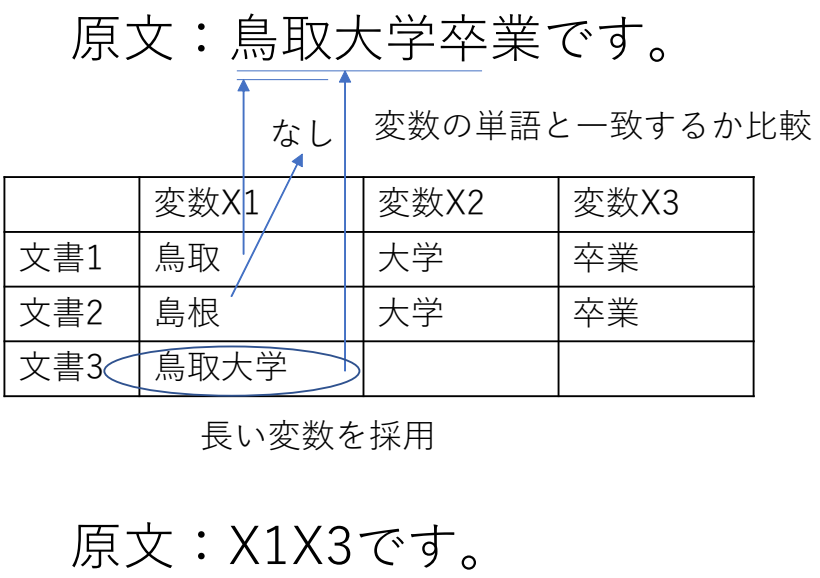

図 3.4: 変数を原文に当てはめる手順

第4章　実験環境

4.1　実験データ

関連性のあるデータを得るために，大量のデータに対して，K-means 法でクラスタ数を 2500 と指定してクラスタリングを行い，その中から密集度が高く件数も適度に多いクラスタを選択して実験データとする．今回は Wikipedia の記事 75,249 件に対してクラスタ数 2,500 でクラスタリングをかける．その結果のうち記事 561 件，密集度 0.941 のクラスタ (主に芸能人についての記事が集まったクラスタ) からランダムで 300 件を抽出した．ここで，密集度とはクラスタ内の情報の関連具合を表したものであり，似たような情報が詰まったクラスタは密集度が高くなる．

ここで，片方を正解データ，もう片方を実験データとして相互に性能を比較するために，この 300 件を 150 件ずつ分割し，それぞれ入力データとした．また，記事の内容は全文ではなく，その記事の概要部分に該当する初めの 3 行分を抽出している．以下に入力データの例を示す．

表 4.1.1: Wikipedia のデータの例（下線部の内容を実験に用いる）

<doc id=”148764” url=”https://ja.wikipedia.org/wiki?curid=148764” title=”原田昌樹” >
原田昌樹

原田 昌樹（はらだ まさき、1955 年 3 月 9 日 - 2008 年 2 月 28 日）は、日本の映画監督。長野県出身。長野県屋代高等学校卒業。血液型は A 型。

趣味は競馬。主に特撮テレビ番組やオリジナルビデオ作品の演出を手がけた。

同期は『半落ち』の佐々部清、『樹の海』の瀧本智行等。

小学生時代に一時住んでいた松本市で映画のロケ隊を見て映画製作に興味を抱いた。学生時代にはアルバイトとして映画の制作現場に入るようになり、教育映画の現場でついていた助監督の誘いで『宇宙鉄人キョーダイン』（1976 年）のサード助監督として本格的に制作に携わる。

助監督時代には東映、大映、三船プロダクション、フィルムリンク・インターナショナル等を渡り歩いた（この時期は長石多可男や蓑輪雅夫といった助監督の下で現場に従事）。また『この胸のときめきを』、『さらば愛しのやくざ』などの作品では和泉聖治監督のチーフ助監督として多くの映画につく。角川春樹が監督を手がけた『REX 恐竜物語』ではチーフ助監督を務めた。

(中略)
</doc>

4.2 MeCab

文の単語への分割には形態素解析器の MeCab を使用した．また, MeCab のシステム辞書には，2017 年 8 月 28 日時点での mecab-ipadic-NEologd[7, 8, 9] を使用した．mecab-ipadic-NEologd では, MeCab の標準のシステム辞書には含まれない固有名詞などの新語を形態素として認識できる．「全国学力テストが行われた」という文を MeCab の標準のシステム辞書と mecab-ipadic-NEologd のそれぞれを用いて分かち書きした結果を表 4.2.1 に示す．

表 4.2.1: 辞書による違いの例

標準のシステム辞書の場合	全国　学力　テスト　が　行わ　れ　た
mecab-ipadic-NEologd の場合	全国学力テスト　が　行わ　れ　た

4.3 単語ベクトルモデル

2.2.1 節の文ベクトルの計算で用いる単語のベクトルには, fastText[10, 11] によって学習させたものを使用した．fastText は隠れ層と出力層からなる 2 層のニューラルネットワークで，隠れ層が単語の分散表現に相当する．

今回は学習データとして, Wikipedia の全 1,061,375 記事を使用した．学習データは前処理としてアルファベットとカタカナは全角に, 英数字は半角に統一した．学習データの例を表 4.3.1 に示す．また，単語ベクトルの次元数は 300 次元とした．

表 4.3.1: 学習データの例

<doc id=”5” url=”https://ja.wikipedia.org/wiki?curid=5”
title=”アンパサンド” >
アンパサンド
アンパサンド (, &) とは「…と…」を意味する記号である。英語の に
相当するラテン語の の合字で、(et cetera = and so forth) をと記
述することがあるのはそのため。Trebuchet MS フォントでは、と表示
され”et”の合字であることが容易にわかる。
その使用は 1 世紀に遡ることができ (1)、5 世紀中葉 (2,3) から現代 (4-6) に
至るまでの変遷がわかる。
Z に続くラテン文字アルファベットの 27 字目とされた時期もある。
アンパサンドと同じ役割を果たす文字に「の et」と呼ばれる、数字の
「7」に似た記号があった (, U+204A)。この記号は現在もゲール文字で
使われている。
記号名の「アンパサンド」は、ラテン語まじりの英語「& はそれ
自身”and” を表す」(& per se and) のくずれた形である。
英語以外の言語での名称は多様である。
日常的な手書きの場合、欧米でアンパサンドは「ε」に縦線を引く単純化
されたものが使われることがある。
また同様に、「t」または「+（プラス）」に輪を重ねたような、
無声歯茎側面摩擦音を示す発音記号「」のようなものが使われることもある。
プログラミング言語では、C など多数の言語で AND 演算子として用いられる。
以下は C の例。
PHP では、変数宣言記号（$）の直前に記述することで、参照渡しを行うこと
ができる。
</doc>

第5章　実験

5.1　評価方法

150 件ずつの入力データをデータ 1，データ 2 とする．生成されたテンプレートを互いに比較し，以下の手順で評価する．

5.1.1　文章レベルのテンプレートの評価方法

データ 1，データ 2 でそれぞれ生成されたテンプレートについて，データ 1 を正解データ，データ 2 を実験データとした時と，データ 2 を正解データ，データ 1 を実験データとした時のカバー率をそれぞれ式 5.1 で求める．カバー率はテンプレートの種類の数ではなく，生成された頻度で計算する．生成されたテンプレートがもう片方の入力データでも出現するかを調べることで，テンプレート生成において一定のカバー率が得られ，テンプレートの汎用性を評価することが可能だと考える．

カバー率は次のようにして求める．

$$\text{カバー率} = \frac{\text{実験データに出現した正解データのテンプレートの数}}{\text{正解データのテンプレート数}} \tag{5.1}$$

5.1.2　文レベルのテンプレートの評価方法

文レベルでのテンプレートの評価方法を説明する．

(1) 2 つのデータで 1 回目のクラスタリング結果の表について，内容が共通する列に注目する．2 種類のデータでそれぞれ生成された表について，列の順番はそれぞれのクラスタリング結果において式 2.14 のクラスタの重要度の順番に並んでいるため，2 つの表で順番が異なる．そのため評価するクラスタの内容の対応付けを改めて設定する必要がある．図 5.1 にて例を示す．

データ 1 で生成された表に「首都」,「面積」,「人口」,「言語」の 4 列が存在し，データ 2 で生成された表に「首都」,「人口」,「首相」の 3 列が存在する場合,「首都」と「人口」の列が共通しているため，この 2 列からそれぞれ生成されたテンプレートを評価の対象とする.

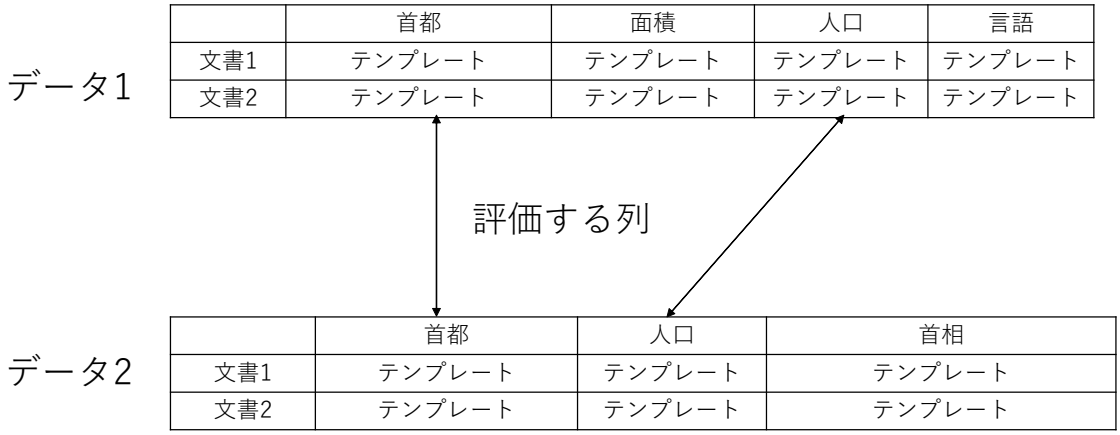

図 5.1: 評価方法の例

(2) 内容が共通している列でそれぞれ生成されたテンプレートについて，データ 1 を正解データ，データ 2 を実験データとした時のカバー率と，データ 2 を正解データ，データ 1 を実験データとした時のカバー率をそれぞれ式 5.1 で求める.

(3) 2 を各列で行う.

5.2　実験結果

2.2 節の表整理の方法で実際に生成された表の一部を表 5.2.1 から表 5.2.2 にて示す.

文章が句点毎に分割されクラスタ毎に表の列としてまとめられている．表の列はクラスタの重要度が高い順に，一番左のタイトルの列の次の列から，順に並べられている.

表 5.2.1: 生成された表の一部

10 c:0.80557 p:0.39267 nc×np:0.29433	出身 c:0.84369 p:0.73333 nc×np:0.73547	卒業 c:0.86312 p:0.44667 nc×np:0.48355	日本 c:0.80946 p:0.49333 nc×np:0.40273	所属 c:0.851 p:0.36667 nc×np:0.37077
原田昌樹	長野県出身			
西山喜久恵	長野県屋代高等学校卒業 広島県尾道市出身で、実家は地元で著名な老舗旅館『西山別館』である	上智大学文学部英文学科を卒業している		
竜崎勝	東京都世田谷区出身 出生地は高知県	法政大学経済学部卒業	竜崎勝は、日本の俳優・グルメレポーター	
山田パンダ	佐賀県神埼郡千代田町大字崎村字黒津生まれ	のち福岡に移り、久留米大学附設高等学校、福岡県立嘉穂高等学校で学ぶ 高校卒業後、八幡大学中退、明治大学工学部卒業。	山田パンダは、フォークシンガー、元かぐや姫メンバーである	
清水圭	京都府京都市出身	同志社香里中学校・高等学校を経て、内部進学にて同志社大学商学部卒業	清水圭は、日本のお笑いタレント	吉本興業所属
中牟田俊男	福岡県福岡市出身	武田は福岡県立筑紫中央高等学校の同級生		
石田ゆり子	東京都出身 母親は沖縄県石垣島出身	東京都立桜町高等学校に入学、東京都立青山高等学校を経て、女子栄養大学短期大学部卒業	石田ゆり子は日本の女優・エッセイスト・ナレーター・タレント 女優の石田ひかりは実妹	
山下真司	山口県下関市出身 下関市幡生町で育つ		山下真司は、日本の俳優、タレント	フロム・ファーストプロダクション所属
四家秀治	四家秀治は、千葉県松戸市出身のフリーアナウンサー	海城高等学校、同志社大学工学部卒業 ラグビーが好きだった父親の影響を受け、千葉県松戸市立小金小学校、東京都中央区立第一中学校、海城高等学校在学中は秩父宮ラグビー場などで、各種大学ラグビーを観戦		
日比野朱里	静岡県浜松市出身	日本工学院専門学校演劇科卒業	日比野朱里は、日本の元女性声優	
佐々木正洋 (1954年生)	福岡県北九州市出身	福岡県立小倉高等学校、慶應義塾大学法学部卒業後の1977年、テレビ朝日にアナウンサーとして入社した		所属事務所は株式会社ICH
冬馬由美	東京都出身		冬馬由美は、日本の女性声優、ナレーターである	ALLURE&Y所属・代表
よこざわけい子	新潟県新潟市中央区出身	中学校の英語教師であった横沢久八の娘として、新潟市に生まれる 新潟県立新潟高等学校卒業 日本大学藝術学部放送学科中退	よこざわけい子は、日本の女性声優、女優	
草地章江	東京都調布市出身		草地章江は、日本の女性ロック歌手、声優	

表 5.2.2: 生成された表の一部 (続き)

フジテレビ c:0.74898 p:0.60667 nc×np:0.31234	血液型 c:0.8623 p:0.26 nc×np:0.27096	本名 c:0.81056 p:0.29333 nc×np:0.23414	日本 c:0.70456 p:0.46667 nc×np:0.13334	現在 c:0.91413 p:0.02 nc×np:0.0	愛称 c:0.64786 p:0.24 nc×np:0.0
主に特撮テレビ番組やオリジナルビデオ作品の演出を手がけた	血液型はA型		原田昌樹は、日本の映画監督		趣味は競馬
西山喜久恵は、アナウンス室部長を務めるフジテレビの女性チーフアナウンサーである					
		本名および旧芸名は高島史旭	長男は元俳優の高島郷、長女はフリーアナウンサーの高島彩で、その夫はフォークデュオゆずの北川悠仁		
		本名は山田嗣人、別名に山田つぐと			
	血液型はA型Rh−	本名、清水圭太			
1972年にデビューする			中牟田俊男は、日本のギタリスト、シンガーソングライターであり、海援隊のメンバー 西南学院大学在学中の1971年に武田鉄矢・千葉和臣を誘い、海援隊を結成		通称「ムーさん」
		本名・石田百合子			愛称は「りり」「ゆりちゃん」「ゆり」「ゆりっぺ」
	身長183cm		母が働き父が家事育児を担当 姉がいる		
大学は、「大学ラグビーの名門で、自分の好きな大学ラグビーチーム」との理由で、以前から憧れていた同志社大学に入学 大学では、体育会機関紙『同志社スポーツアトム』を通して、ラグビー部の活躍を伝える事に熱中した	血液型AB型				
『キャプテン翼』ではオープニングとエンディング曲を歌っている	血液型はA型	旧芸名は小粥よう子	夫は彼女の主演作でもあるテレビアニメ『キャプテン翼』の原作者で漫画家の高橋陽一		
元テレビ朝日アナウンサー 慶應義塾大学では落語研究会の部長として活躍 妻は元フジテレビアナウンサーの古賀万紀子			佐々木正洋とは、日本のフリーアナウンサー、タレント 元NHKアナウンサーの宮本隆治とは中学校から大学まで出身校が全て同じであり、佐々木は宮本の4期後輩である また、「3年B組金八先生」乾友彦役を務めた俳優の森田順平は高校の同級生である		左利き
女性の声を演じる機会が多いが、キャリア初期には少年役も複数担当している		本名：吉田由美、旧姓：中川			
芸能プロダクションゆーりんプロ代表取締役 声優の勝田久の勧めで俳協付属養成所に入り、養成所在籍中の1975年に『タイムボカン』で声優デビュー		旧芸名は横沢啓子			
1989年、歌手として芸能界にデビュー 1990年代には主に声優で活躍し『クレヨンしんちゃん』鳩ヶ谷ミッチー役で知られるようになる	血液型B型	本名同じ			イメージカラーは赤

5.2.1　文章レベルでの結果

3 章の方法によって，表 5.2.1，表 5.2.2 から実際に生成された文章レベルのテンプレートの結果の一部を表 5.2.3，5.2.4 にて示す．表 5.2.1 の最初の列から表 5.2.2 の列の最後の列まで，順に X0 から X10 までの変数のグループとし，原文と変数のグループを照らし合わせる際に，そのグループに含まれる文は変数に置換されている．

表 5.2.3: 文章レベルでのテンプレートの結果の一部

原文	テンプレート
原田昌樹 原田 昌樹（はらだ まさき、1955年3月9日 - 2008年2月28日）は、日本の映画監督。長野県出身。長野県屋代高等学校卒業。血液型はA型。 趣味は競馬。主に特撮テレビ番組やオリジナルビデオ作品の演出を手がけた。	X0。X8。X1。X1。X6。X10。X5。
西山喜久恵 西山 喜久恵（にしやま きくえ、1969年6月22日 - ）は、アナウンス室部長を務めるフジテレビの女性チーフアナウンサーである。 広島県尾道市出身で、実家は地元で著名な老舗旅館『西山別館』である。上智大学文学部英文学科を卒業している。	X0。X5。X1。X2。
竜崎勝 竜崎 勝（りゅうざき かつ、1940年3月25日 - 1984年12月18日）は、日本の俳優・グルメレポーター。本名および旧芸名は高島 史旭（たかしま ふみあき）。長男は元俳優の高島郷、長女はフリーアナウンサーの高島彩で、その夫はフォークデュオゆずの北川悠仁。 東京都世田谷区出身。出生地は高知県。法政大学経済学部卒業。	X0。X3。X7。X8。X1。X1。X2。
山田パンダ 山田 パンダ（やまだ ぱんだ、1945年5月13日 - ）は、フォークシンガー、元かぐや姫メンバーである。本名は山田嗣人（やまだ つぐと）、別名に山田つぐと。 佐賀県神埼郡千代田町大字崎村字黒津（現：神埼市）生まれ。のち福岡に移り、久留米大学附設高等学校（学力不振で中退）、福岡県立嘉穂高等学校で学ぶ。高校卒業後、八幡大学中退、明治大学工学部（現：理工学部）卒業。	X0。X3。X7。X1。X2。X2。
清水圭 清水 圭（しみず けい、1961年6月24日 - ）は、日本のお笑いタレント。吉本興業所属。本名、清水圭太（しみず けいた）。京都府京都市出身。血液型はA型 Rh-。 同志社香里中学校・高等学校を経て、内部進学にて同志社大学商学部卒業。	X0。X3。X4。X7。X1。X6。X2。
中牟田俊男 中牟田 俊男（なかむた としお、1949年7月21日 - ）は、日本のギタリスト、シンガーソングライターであり、海援隊のメンバー。通称「ムーさん」。福岡県福岡市出身。 西南学院大学在学中の1971年に武田鉄矢・千葉和臣を誘い、海援隊を結成。1972年にデビューする。武田は福岡県立筑紫中央高等学校の同級生。	X0。X8。X10。X1。X8。X5。X2。
石田ゆり子 石田 ゆり子（いしだ ゆりこ、1969年10月3日 - ）は日本の女優・エッセイスト・ナレーター・タレント。本名・石田 百合子（いしだゆりこ）。愛称は「りり」「ゆりちゃん」「ゆり」「ゆりっぺ」。女優の石田ひかりは実妹。 東京都出身。母親は沖縄県石垣島出身。東京都立桜町高等学校に入学、東京都立青山高等学校を経て、女子栄養大学短期大学部卒業。	X0。X3。X7。X10。X3。X1。X1。X2。
山下真司 山下 真司（やました しんじ、1951年12月16日 - ）は、日本の俳優、タレント。身長183cm。フロム・ファーストプロダクション所属。山口県下関市出身。 下関市幡生町で育つ。母が働き父が家事育児を担当。姉がいる。	X0。X3。X6。X4。X1。X1。X8。X8。
四家秀治 四家 秀治（よつや ひではる、1958年8月18日 - ）は、千葉県松戸市出身のフリーアナウンサー。血液型AB型。海城高等学校、同志社大学工学部卒業。 ラグビーが好きだった父親の影響を受け、千葉県松戸市立小金小学校、東京都中央区立第一中学校（現銀座中学校）、海城高等学校在学中は秩父宮ラグビー場などで、各種大学ラグビーを観戦。大学は、「大学ラグビーの名門で、自分の好きな大学ラグビーチーム」との理由で、以前から憧れていた同志社大学に入学。大学では、体育会機関紙『同志社スポーツ アトム』を通して、ラグビー部の活躍を伝える事に熱中した。	X0。X1。X6。X2。X2。X5。X5。

表 5.2.4: 文章レベルでのテンプレートの結果の一部 (続き)

原文	テンプレート
日比野朱里 日比野 朱里（ひびの あかり、1959年7月5日 - ）は、日本の元女性声優。静岡県浜松市出身。血液型はA型。旧芸名は小粥よう子。 日本工学院専門学校演劇科卒業。夫は彼女の主演作でもあるテレビアニメ『キャプテン翼』の原作者で漫画家の高橋陽一。『キャプテン翼』ではオープニングとエンディング曲を歌っている。	X0。X3。X1。X6。X7。X2。X8。X5。
佐々木正洋 (1954年生) 佐々木 正洋（ささき まさひろ、1954年7月17日 - ）とは、日本のフリーアナウンサー、タレント。元テレビ朝日アナウンサー。左利き。所属事務所は株式会社ICH。 福岡県北九州市出身。福岡県立小倉高等学校、慶應義塾大学法学部卒業後の1977年、テレビ朝日（当時：全国朝日放送）にアナウンサーとして入社した。慶應義塾大学では落語研究会の部長として活躍。妻は元フジテレビアナウンサーの古賀万紀子。元NHKアナウンサーの宮本隆治とは中学校から大学まで出身校が全て同じであり、佐々木は宮本の4期後輩である。また、「3年B組金八先生」乾友彦役を務めた俳優の森田順平は高校の同級生である。	X0。X8。X5。X10。X4。X1。X2。X5。X5。X8。X8。
冬馬由美 冬馬 由美（とうま ゆみ、1966年12月20日 - ）は、日本の女性声優、ナレーターである。本名：吉田 由美（よしだ ゆみ）、旧姓：中川（なかがわ）。東京都出身。ALLURE&Y所属・代表。 女性の声を演じる機会が多いが、キャリア初期には少年役も複数担当している。	X0。X3。X7。X1。X4。X5。
よこざわけい子 よこざわ けい子（よこざわ けいこ、本名：難波 啓子〈なんば けいこ〉、1952年9月2日 - ）は、日本の女性声優、女優。芸能プロダクションゆーりんプロ代表取締役。新潟県新潟市中央区出身。旧芸名は横沢啓子（読み同じ）。 中学校の英語教師であった横沢久八の娘として、新潟市に生まれる。新潟県立新潟高等学校卒業。日本大学藝術学部放送学科中退。声優の勝田久の勧めで俳協付属養成所に入り、養成所在籍中の1975年に『タイムボカン』で声優デビュー。	X0。X3。X5。X1。X7。X2。X2。X2。X5。
草地章江 草地 章江（くさち ふみえ、1969年11月24日 - ）は、日本の女性ロック歌手、声優。1989年、歌手として芸能界にデビュー。1990年代には主に声優で活躍し『クレヨンしんちゃん』鳩ヶ谷ミッチー役で知られるようになる。本名同じ。 東京都調布市出身。血液型B型。イメージカラーは赤。	X0。X3。X5。X5。X7。X1。X6。X10。

5.2.2　文レベルでの結果

2.2 節の表整理の方法を用いて，表 5.2.1，表 5.2.2 の結果から 4 列目「日本」の列をクラスタリングし表に整理した結果の一部を表 5.2.5 にて示す．テンプレート生成する時，左の列から変数 X0，X1... となる．

表 5.2.5: 列「日本」のクラスタリング結果の一部

7 c:0.78444 p:0.55405 nc×np:0.32056	日本 c:0.99317 p:0.87838 nc×np:1.0	タレント c:0.78209 p:0.75676 nc×np:0.4547	俳優 c:0.92893 p:0.5 nc×np:0.44455	実妹 c:0.6241 p:0.82432 nc×np:0.17932	歌手 c:0.74484 p:0.37838 nc×np:0.16537	真佐雄 c:0.88192 p:0.09459 nc×np:0.0	元アイドル c:0.53606 p:0.44595 nc×np:0.0
竜崎勝	日本	グルメ レポーター	俳優	竜崎勝			
清水圭	日本	お笑いタレント		清水圭			
石田ゆり子	日本	ナレーター タレント	女優 女優	石田ゆり子 石田ひかり 実妹			エッセイスト
山下真司	日本	タレント	俳優	山下真司			
日比野朱里	日本	女性声優		日比野朱里			
冬馬由美	日本	女性声優 ナレーター		冬馬由美			
よこざわけい子	日本	女性声優	女優	よこざわけい子			
草地章江	日本	女性 声優		草地章江	ロック歌手		

3 章の方法によって，表 5.2.5 から生成されたテンプレートの一部を表 5.2.6 にて示す．

表 5.2.6: 列「日本」から生成されたテンプレートの一部

原文	テンプレート
竜崎勝は、日本の俳優・グルメレポーター	X0は、X1のX3・X2X2
清水圭は、日本のお笑いタレント	X0は、X1のX2
石田ゆり子は日本の女優・エッセイスト・ナレーター・タレント 女優の石田ひかりは実妹	X0はX1のX3・X7・X2・X2 X3のX4はX4
山下真司は、日本の俳優、タレント	X0は、X1のX3、X2
日比野朱里は、日本の元女性声優	X0は、X1の元X2
冬馬由美は、日本の女性声優、ナレーターである	X0は、X1のX2、X2である
よこざわけい子は、日本の女性声優、女優	X0は、X1のX2、X3
草地章江は、日本の女性ロック歌手、声優	X0は、X1のX2X5、X2

5.3 評価結果

5.3.1 文章レベルでの評価結果

文章レベルでの評価結果を示す．テンプレート数はデータ 1 とデータ 2 共に，150 個生成された．データ 2 でも生成されたデータ 1 のテンプレートを表 5.3.1 にて，データ 1 でも生成されたデータ 2 のテンプレートを表 5.3.2 にて示す．変数については，表 5.2.1 の左端の列に含まれる文字列を X0 とし，そこから順に X1，X2... と表 5.2.2 の右端の列 (X10) まで置いている．

データ 2 でも生成されたデータ 1 のテンプレートは 19 個，データ 1 でも生成されたデータ 2 のテンプレートは 17 個生成された．

表 5.3.1: データ 2 でも生成されたデータ 1 のテンプレート

頻度数	テンプレート
4	X0。X5。X5。
3	X0。X5。X5。X5。
2	X0。X8。X8。X8。X8。
2	X0。X5。
1	X0。X3。X1。X4。X2。
1	X0。X3。X1。X2。
1	X0。X3。X1。X4。X7。X2。
1	X0。X5。X1。X5。
1	X0。X5。X1。X2。
1	X0。X3。X4。X10。
1	X0。X5。X1。X5。X2。
1	X0。X3。X1。X4。X8。X8。

表 5.3.2: データ 1 でも生成されたデータ 2 のテンプレート

頻度数	テンプレート
3	X0。X8。X8。X8。X8。
2	X0。X3。X1。X2。
2	X0。X3。X1。X4。X2。
2	X0。X5。X1。X2。
1	X0。X5。X5。X5。
1	X0。X3。X4。X10。
1	X0。X3。X1。X4。X7。X2。
1	X0。X3。X1。X4。X8。X8。
1	X0。X5。X1。X5。X2。
1	X0。X5。X5。
1	X0。X5。
1	X0。X5。X1。X5。

式 5.1 で求められたカバー率の結果を表 5.3.3 にて示す．カバー率は 0.13，0.11 と共に低い結果となった．文章の場合は，文が長いことで完全に似通ったデータが少ないため，このような結果となった．

表 5.3.3: 文章レベルのテンプレートの評価結果

	実験データ：データ 2 正解データ：データ 1	実験データ：データ 1 正解データ：データ 2
テンプレート数	19	17
カバー率	0.13	0.11
カバー率平均	0.12	

5.3.2 文レベルでの評価結果

データ 1 で生成された表の列は 10 列，データ 2 で生成された表の列は 9 列であり，その内 2 つの間で類似している列は 6 列ありその列で評価を行った．データ 2 でも生成されたデータ 1 のテンプレートを表 5.3.4，表 5.3.5 にて示す．

表 5.3.4: データ 2 でも生成されたデータ 1 のテンプレート (1/2)

出身		日本		卒業	
頻度数	テンプレート	頻度数	テンプレート	頻度数	テンプレート
67	X2X1	5	X0 は、X1 の X2	8	X3X1
19	X2	2	X0 は、X1 の X2、X5	3	X3X3X3X3X1
5	X2X2X1	1	X0 は、X1 の X2、X2 である	2	X3X3X1
2	X2X2X2X1	1	X0 は X1 の X2	1	X3、X3X1
2	X2X1、X2X1	1	X0 は、X1 の X2、X2		
		1	X0 は、X1 の X2、X5、X2		

表 5.3.5: データ 2 でも生成されたデータ 1 のテンプレート (2/2)

血液型		本名		所属	
頻度数	テンプレート	頻度数	テンプレート	頻度数	テンプレート
9	X3X7X7	4	X1 同じ	7	X2X1
5	X1 はA X2	4	X1、X3	2	X3X1
5	X1X2	1	X1：X0		
4	X1 A X2				
4	X1 はO X2				
3	X1 は X2				
2	X1 O X2				
1	X3X3				

次にデータ 1 でも生成されたデータ 2 のテンプレートを表 5.3.6，表 5.3.7 にて示す．

表 5.3.6: データ 1 でも生成されたデータ 2 のテンプレート (1/2)

出身		日本		卒業	
頻度数	テンプレート	頻度数	テンプレート	頻度数	テンプレート
58	X2X1	10	X0 は、X1 の X2	10	X3X1
21	X2	8	X0 は、X1 の X2、X2	1	X3X3X3X3X1
4	X2X2X1	2	X0 は、X1 の X2、X2 である	1	X3X3X1
2	X2X2X2X1	2	X0 は、X1 の X2、X5	1	X3、X3X1
1	X2X1、X2X1	1	X0 は X1 の X2		
		1	X0 は、X1 の X2、X5、X2		

表 5.3.7: データ 1 でも生成されたデータ 2 のテンプレート (2/2)

血液型		本名		所属	
頻度数	テンプレート	頻度数	テンプレート	頻度数	テンプレート
11	X3X7X7	1	X1 同じ	28	X3X1
9	X1 はO X2	1	X1、X3	2	X2X1
5	X1 はA X2	1	X1：X0		
4	X1 は X2				
3	X1 O X2				
3	X1 A X2				
1	X1X2				
1	X3X3				

次に，データ 1 を正解データ，データ 2 を実験データとした時のカバー率と，データ 2 を正解データ，データ 1 を実験データとした時のカバー率をそれぞれ式 5.1 で求める．結果を表 5.3.8，表 5.3.9 に示す．

6 列でのカバー率の平均の結果は共に 0.33，0.44 と低い結果だが，「出身」の列や「血液型」では高いカバー率となっており，共通して生成されたテンプレートが高い頻度で生成されることが確認できた．

表 5.3.8: データ 1 のテンプレートの評価結果

	出身	日本	卒業	血液型	本名	所属	平均
テンプレート数	117	70	67	47	64	56	70.2
データ 2 にも出現したテンプレート数	86	24	13	37	3	30	32.2
カバー率	0.74	0.34	0.19	0.78	0.04	0.53	0.44

表 5.3.9: データ 2 のテンプレートの評価結果

	出身	日本	卒業	血液型	本名	所属	平均
テンプレート数	129	88	83	50	52	64	77.7
データ 1 にも出現したテンプレート数	95	11	14	33	9	9	28.5
カバー率	0.73	0.13	0.17	0.66	0.17	0.14	0.33

第6章　考察

6.1　文章レベルでの実験結果について

データ 1 を正解データ，データ 2 を実験データとした時のカバー率とデータ 2 を正解データ，データ 1 を実験データとした時のカバー率がそれぞれ 0.13，0.11 と共に低い結果となった．文章の場合は，文が長くなる程複雑化し，他の文章と構成も異なってくるため，このような結果となった．

人手で確認し，比較的良かったテンプレートの例を図 6.1 にて示す．変数については，表 5.2.1 の左端の列に含まれる文字列を X0 とし，そこから順に X1，X2... と表 5.2.2 の右端の列 (X10) まで置いている．図 6.1 に示すテンプレートでは，文毎に情報が分かれており，それぞれ別のクラスタとして適切にクラスタリングされているため，テンプレートも変数がそれぞれ別になっている．また，「X0 はタイトル」，「X3 は概要文」などこのテンプレートが持つ情報通りに「X0。X3。X4。X7。X1。X6。X2。」として変数に適切な文字列を補完すれば比較的流暢で様々な文章が生成可能なため，有用なテンプレートと考える．

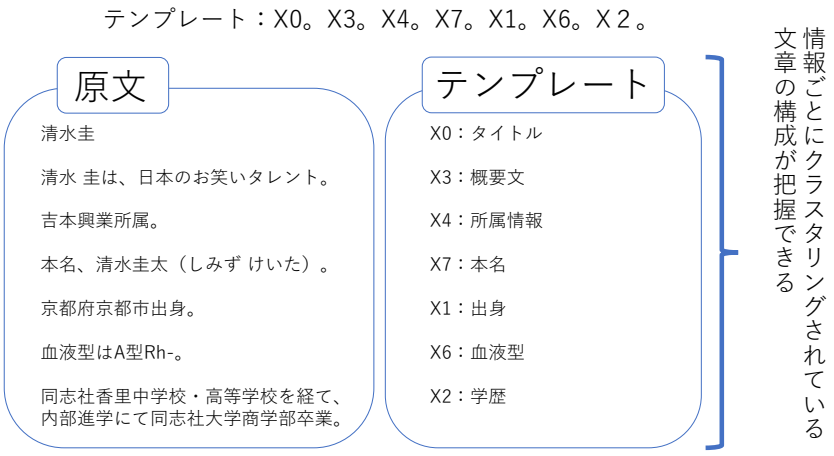

図 6.1: 文章レベルのテンプレートの良かった例

次に，図 6.2 にて比較的悪かった例を示す．文毎に情報が異なっているがクラスタリングが適切にされず，X0 以外が全て 1 つのクラスタとなっている．そのためテンプレートも「X0。X8。X8。X8。X8。」となり X8 が持つ情報が重複しているため，ここから文生成を行う場合，どの情報を変数に補完するべきか分からず，比較的悪いテンプレートである．

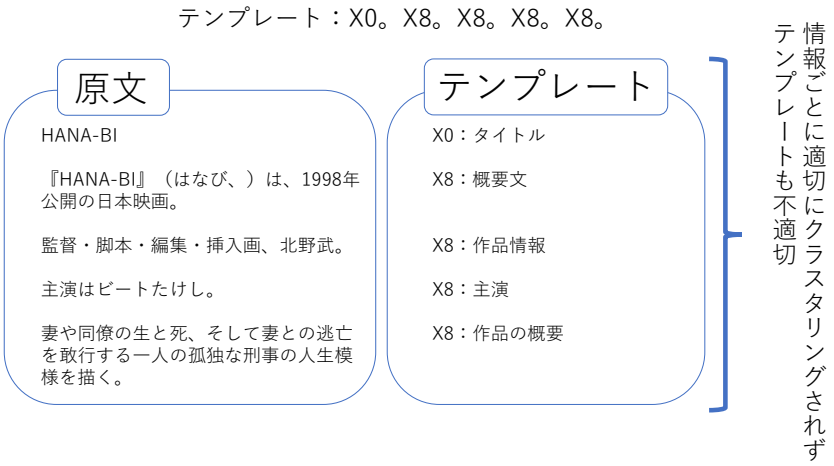

図 6.2: 文章レベルのテンプレートの悪かった例

6.2　文レベルでの実験結果について

6.2.1　カバー率の結果

文レベルでの評価の結果，カバー率が最も高いもので「血液型」の列の 0.78，最も低いもので「本名」の列の 0.04 であった．結果が列ごとで異なってくる理由は，2 回目のクラスタリング結果に依存するためであると考える．

データ「血液型」の列のクラスタリング結果を表 6.2.1，表 6.2.2 に示す．2 つの表共に，綺麗にクラスタリングされており，統一感があるため，テンプレートのカバー率も高い結果となった．

表 6.2.1: データ 1 での「血液型」の列のクラスタリング結果

12 c:0.98511 p:0.25427 nc×np:0.27077	血液型 c:1.0 p:0.69231 nc×np:1.0	型 c:1.0 p:0.66667 nc×np:0.96154	身長 c:0.98522 p:0.48718 nc×np:0.61442	体重 c:1.0 p:0.23077 nc×np:0.30769	kg c:1.0 p:0.23077 nc×np:0.30769	スリーサイズ c:0.96757 p:0.07692 nc×np:0.05794	AB c:0.86858 p:0.53846 nc×np:0.0
原田昌樹	血液型	型					
清水圭	血液型	型					
四家秀治	血液型	型					AB
日比野朱里	血液型	型					
草地章江	血液型	B型					
植田朝日	血液型	B型					
鈴木真仁	血液型	型					
松浦亜弥	血液型	B型	身長 156cm				

表 6.2.2: データ 2 での「血液型」の列のクラスタリング結果

7 c:0.96461 p:0.41224 nc×np:0.39128	血液型 c:1.0 p:0.74286 nc×np:1.0	型 c:1.0 p:0.74286 nc×np:1.0	身長 c:0.98522 p:0.54286 nc×np:0.62855	体重 c:0.93833 p:0.2 nc×np:0.11044	AB c:0.86875 p:0.54286 nc×np:0.0	靴 c:1.0 p:0.05714 nc×np:0.0	サイズ c:0.95996 p:0.05714 nc×np:0.0
小島武夫	血液型	型					
山崎弘士	血液型	型					
大島さと子	血液型	型	身長		161 cm		
小林聡美	血液型	型	身長 156cm		AB		
石原慎一	血液型	型					
園山俊二	血液型	B型					
斉藤暁	血液型	型	身長	体重 kg	165 cm 77		
山野さと子	血液型	型					

次に結果の悪かった「卒業」の列のクラスタリング結果を，表 6.2.3，表 6.2.4 に示す．

データ 1 ではクラスタ数が 3 つとなっているが，データ 2 では 5 つとなっている．また，データ 2 の方は「卒業」という単語が独立して 1 つのクラスタとなっているが，データ 1 の結果では学校の名前と「卒業」が一緒にクラスタリングされている場合が多く，このクラスタリング結果の違いが再現率の低下に直結している．

表 6.2.3: データ 1 での「卒業」の列のクラスタリング結果

3 c:0.73843 p:0.67662 nc×np:0.20603	卒業 c:0.76276 p:0.97015 nc×np:0.61809	のち c:0.88081 p:0.16418 nc×np:0.0	高等学校 c:0.57172 p:0.89552 nc×np:0.0
西山喜久恵	上智大学 文学部 科 卒業		英文学
竜崎勝	法政大学 経済学部 卒業		
山田パンダ	卒業 大学 中退 明治大学 工学部 卒業	のち 後	福岡 久留米大学附設高等学校 福岡 県立 嘉穂 高等学校 高校 八幡
清水圭	進学 同志社大学 商学部 卒業		同志社香里中学校・高等学校 内部
石田ゆり子	入学 女子栄養大学短期大学部 卒業		東京都立桜町高等学校 東京都立青山高等学校
四家秀治	同志社大学 工学部 卒業 在学中		海城高等学校 ラグビー 好き 父親 影響 千葉県 松戸市立小金小学校 東京都中央 区立 第一中学校 海城高等学校 秩父宮ラグビー場 各種 大学ラグビー 観戦

表 6.2.4: データ 2 での「卒業」の列のクラスタリング結果

5 c:0.76008 p:0.61455 nc×np:0.37543	卒業 c:0.90676 p:0.98182 nc×np:1.0	学科 c:0.79476 p:0.70909 nc×np:0.4883	高等学校 c:0.73232 p:0.70909 nc×np:0.38885	そこ c:0.87838 p:0.16364 nc×np:0.0	入社 c:0.48816 p:0.50909 nc×np:0.0
吉野公佳	卒業	富士短期大学	日出女子学園 高等学校		
山崎弘士	卒業	立命館大学法学部	津山 市立 北中学校 岡山 県立 津山 高等学校		
大島さと子	卒業 卒業	成城大学	フェリス女学院中学校・高等学校		
あいはら友子	卒業	関西学院大学法学部	兵庫県立御影高等学校		
雨森雅司	卒業	日本大学芸術学部 学科			映画 劇団 戯曲 座 入団
春日井静奈	卒業		青森市立浜田小学校 青森市立南中学校 青森 県立 青森 高等学校		
山内雅人	卒業	早稲田大学法学部			1950年
園山俊二	卒業	島根大学教育学部附属小学校	附属 中学校 島根 県立 松江 高等学校		

6.2.2　実際に生成されたテンプレートの例

データ 1 の「出身」の列で生成されたテンプレートについて図 6.3，図 6.4 にて示す．この列からはテンプレート「X2X1」が 67 個生成され，「○○県○○市」という地名を表す文字列が変数 X2 となり，「出身」や「生まれ」といった単語が変数 X1 となっている．しかし，図 6.4 のように，「○○出身」が一つの変数 X2 となった場合や，「○○県○○市出身」が県と市を分けて変数 X2X2X1 となった場合が存在する．このように，どこまでを一つの変数にした場合が正解なのか，テンプレートの精度の評価は今後の課題である．

「出身」の列で生成されたテンプレート

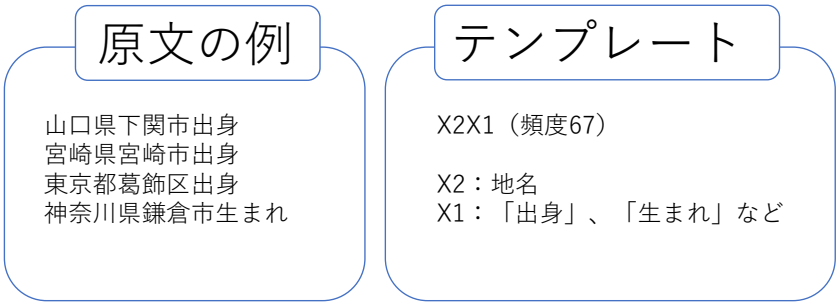

図 6.3: 文レベルのテンプレートの例 (1/2)

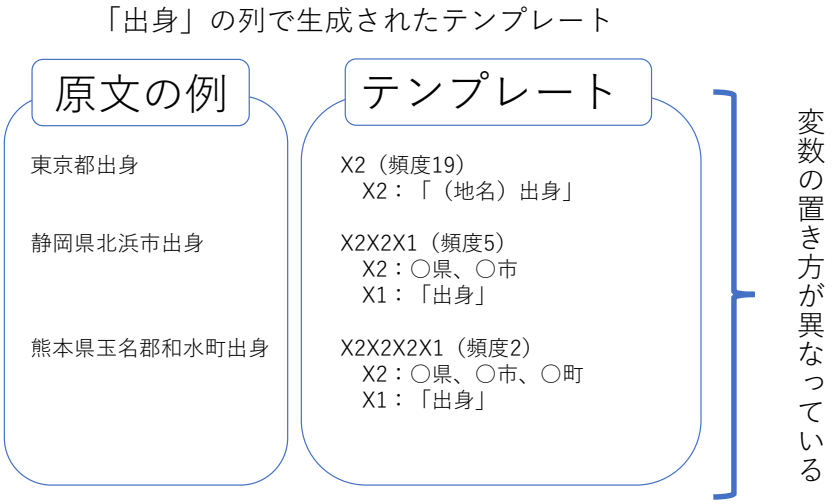

図 6.4: 文レベルのテンプレートの例 (2/2)

6.3　頻出頻度の多い単語について

本研究では，名詞部分を全てテンプレートの変数と設定しているため，一見分かりづらいテンプレートとなっている．しかし，一部の変数は変数化せずにそのままにする方がテンプレートとして見やすくなる場合がある．例を表 6.3.1 にて示す．表において 2 列目「出身」のクラスタ内の単語頻度をカウントしたところ「出身」という単語が 107 個中 78 個も存在している．このようにある程度頻出頻度が高い単語は変数化せずそのままにすることで「X2X1」というテンプレートが「X2 出身」となり，視覚的に分かりやすいテンプレートになると考える．

表 6.3.1: データ 1 での「出身」の列のクラスタリング結果

4 c:0.76087 p:0.51818 nc×np:0.38452	出身 c:0.89623 p:0.8 nc×np:0.79412	東京都出身 c:0.78766 p:0.99091 nc×np:0.74396	現在 c:0.88738 p:0.06364 nc×np:0.0	タレント c:0.47222 p:0.21818 nc×np:0.0
西山喜久恵	出身 実家	広島県尾道市		地元 著名 老舗 旅館 西山別館
竜崎勝	出身 出生	東京都世田谷区 高知県		地
山田パンダ	生まれ	佐賀県 郡 千代田町 大字		神埼 崎村 宇 黒津
清水圭	出身	京都府京都市		
中牟田俊男	出身	福岡県福岡市		
石田ゆり子	母親 出身	東京都出身 沖縄県石垣		島
山下真司	出身	山口県下関市 下関市幡生町		
四家秀治	出身	千葉県松戸市		四家秀治 フリーアナウンサー
日比野朱里	出身	静岡県浜松市		
佐々木正洋 (1954	出身	福岡県北九州市		
よこざわけい子	出身	新潟県新潟市中央区		
草地章江	出身	東京都調布市		
伊藤さおり	出身 在住	埼玉県 岩槻市 静岡県三島市	現在	
野崎昌一	出身	埼玉県 浦和市		
松尾スズキ	生まれ 父親 母親 出身	福岡県北九州市八幡西区 佐賀県 鹿児島県阿久根市		
金月真美	出身	兵庫県明石市		

6.4 形態素解析について

本研究ではクラスタリング時の形態素解析に MeCab を用いているが，単語によって分割のされ方が異なる．例えば，「慶應義塾大学法学部」という文字列はそのまま一つの固有名詞として認識されるが，「同志社大学商学部」という文字列は「同志社大学」と「商学部」で分割される．このように同種の文字列でも分割のされ方が異なることでテンプレートの種類が増えてしまい，同じテンプレートが出現しにくくなると考える．そのため，MeCab 以外の形態素解析ツールの性能を調査し，最適なツールを選択する必要がある．

第7章　今後の課題

文章レベルでの評価結果は，データ 1 を正解データ，データ 2 を実験データとした時のカバー率とデータ 2 を正解データ，データ 1 を実験データとした時のカバー率がそれぞれ 0.13，0.11 と共に低い結果となった．また，文レベルでの評価の結果，カバー率が最も高いもので「血液型」の列の 0.78，最も低いもので「本名」の列の 0.04 であった．カバー率の向上には，クラスタリングによる表生成の精度の向上が課題と考える．そのために，クラスタリング手法の改良や，形態素解析のツールを最適なものを使用するなどの工夫が必要である．

また，現在は名詞を全てテンプレートの変数としているが，頻度の高い単語については変数化せずそのままにするなど文書作成支援に繋がるようなテンプレートの生成ができるよう，手法の改良が必要である．

テンプレート自体の精度についても本研究では評価していないため精度の評価の調査と，要約文の生成などに応用ができるかの調査も行いたいと考える．

第8章　おわりに

本研究では，階層クラスタリングによる表生成の技術を用いてテンプレートの生成を行った．

提案手法では，文書群に対して階層クラスタリングを行い表を自動で作成し，文章レベルのテンプレート生成では，その表の各列をテンプレートの変数のグループとしてその列に含まれる単語は変数となるよう原文に置換してテンプレートを生成した．文レベルでのテンプレートでは，最初に出力された表中の文を全て名詞のみの状態にし，各列で再度クラスタリングを行い，出力された表の各列をテンプレートの変数のグループとしてその列に含まれる単語は変数となるよう原文に置換してテンプレートを生成した．

150 件の記事の入力データを 2 種類用意し実験を行った結果，文章レベルでのテンプレートのカバー率の平均は 0.12，文レベルでのテンプレートのカバー率の平均は 0.39 であった．結果はあまり高くないものの，一部の列のテンプレートのカバー率は 0.78 と高いものもあり，有効なテンプレートが生成できたと考える．

クラスタリング時の形態素解析で分割のされ方が異なる問題や，現在は名詞部分を全て変数としているが，重要部分だけを変数とするなど，テンプレートの生成方法の改良が今後の課題と考える．また，階層クラスタリングの分類の精度の向上も今後の課題である．

謝辞

最後に, 3 年間に渡りご指導いただきました鳥取大学工学部知能情報工学科自然言語処理研究室の村田真樹教授, 村上仁一准教授, 岡﨑健介氏をはじめ, 自然言語処理研究室の方々に厚く御礼申し上げます. また, 本研究における議論・検討に当たって, 有益な議論と情報交換をして頂いた木村周平教授をはじめとする方々, ならびに本論文にて参考にさせていただいた論文の著者の方々に対して深く感謝申し上げます.

参考文献

[1] 岡崎健介, 村田真樹, 馬青. 複数文書からの重要情報の抽出と表の作成. 言語処理学会第 24 回年次大会, pp. 240–243, 2018.

[2] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. *In ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 727–734, 2000.

[3] 岡崎健介, 村田真樹. 複数文書に含まれる情報を表に整理する手法の改良. 2019 年度修士論文, pp. 240–243, 2020.

[4] 田川裕輝, 嶋田和孝. スポーツ要約生成におけるテンプレート型手法とニューラル型手法の提案と比較. 自然言語処理, Vol. 25, No. 4, pp. 357–391, 2018.

[5] 山崎健史, 吉野幸一郎, 前田浩邦, 笹田鉄郎, 橋本敦史, 船卓哉, 山肩洋子, 森信介. フローグラフからの手順書の生成. 情報処理学会論文誌, pp. 849–862, 2016.

[6] S Mori, T Sasada, Y Yamakata, and K Yoshino. A Machine Learning Approach to Recipe Text Processing. *1st Cooking with Computer Workshop*, pp. 29–34, 2012.

[7] 奥村学, 佐藤敏紀. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017), pp. 875–878, 2017.

[8] 奥村学, 佐藤敏紀. 単語分かち書き用辞書生成システム neologd の運用 — 文書分類を例にして—. 自然言語処理研究会研究報告, Vol. 2016-NL-229, pp. 1–14, 2016.

[9] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab. *https://github.com/neologd/mecab-ipadic-neologd*, 2015.

[10] Piotr Bo-janowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *In Transactions of the Association for Computational Linguistics, Vol. 5*, pp. 135–146, 2017.

[11] Armand Joulin, Edouard Grave, Piotr Bo-janowski, and Tomas Mikolov. Bag of tricks for efficient text classi cation. *In arXiv preprint arXiv:1607.01759*, 2016.