概要

長文から知りたい情報を取得するためには長い文を読む必要がある．よって，情報取得に時間がかかる．そこで，長文を表に示すことで長い文を読む必要がなくなり，情報を容易に取得することができる．情報を抽出して表に整理する方法として，藤原の研究 [1] や Akano の研究 [2] がある．藤原の研究では上位下位知識を利用して抽出データから下位語の頻度分析を行い，頻度が高かった下位語の上位語を重要項目と選定して，Wikipedia の抽出データから重要項目の下位語を取り出し，表にまとめていた．しかし，藤原の研究では抽出された重要項目の種類が少ないという問題点がある．Akano の研究では word2vec [3] 内の単語クラスタリングを利用して，表生成に使用する抽出データの類似している単語をまとめて単語のクラスタを作り，頻度の高いクラスタを重要項目として人手で表にまとめていた．しかし，Akano らの研究では 1 つのクラスタを重要項目としていたため，単語の網羅性が低いという問題点と，クラスタリングを行うデータは，Wikipedia から抽出した事柄を含むページのデータのみで単語クラスタリングを行っていたため，クラスタリングの精度が低くなる問題点の 2 つあった．

そこで，本研究では単語クラスタリングの改良や分類語彙表を用いて表生成を行い，重要項目の選定を行った．重要項目の選定方法としては,「単語クラスタリング」による手法と「類似度」による手法と「分類語彙表」による手法の 3 つを提案する.「単語クラスタリング」は Akano の研究と同様に抽出データの単語から類似している単語をまとめて単語のクラスタを作り重要項目の選定を行う手法である．ただし，Akano らの研究では Wikipedia から抽出した事柄を含むページのデータのみで単語クラスタリングしていたが，本研究では Wikipedia 全ページを利用して単語のクラスタリングを行う.「類似度」は入力した単語のベクトルと近いベクトルの単語 (類似した単語) を取得できる．入力した単語と入力データと類似した単語を使用して重要項目の選定を行う手法である.「分類語彙表」は分類語彙表よって分類・整理したシソーラス (類義語集) を利用して重要項目の選定を行う手法である．Akano らの研究と提案手法 3 つで評価実験を行う．評価実験としては情報抽出と記載不足の指摘の 2 点で評価を行う．

情報抽出の評価実験は，表抽出における正解率と単語抽出における正解率の 2 つで

評価を行った．表抽出における正解率は表に 1 つでも正しく情報を抽出したものを正解とした．また，空欄を正しく空欄として検出できれば正解とした．表抽出における正解率を評価した結果，先行手法の表抽出における正解箇所の割合は 0.68 であり，提案手法「Wikipedia 全ページでクラスタリング」の表抽出における正解箇所の割合は 0.71 であり，提案手法「類似度」の表抽出における正解箇所の割合は 0.88 であり，提案手法「分類語彙表」の表抽出における正解箇所の割合は 0.81 であった．このように，先行手法より提案手法の方が精度が高い結果になった．また,「Wikipedia 全ページでクラスタリング」と「分類語彙表」よりも「類似度」の結果の方が精度が高い結果になった．

また，単語抽出における正解率を評価した結果，先行手法の単語抽出における正解箇所の割合は 0.73 であり，提案手法「Wikipedia 全ページでクラスタリング」の単語抽出における正解箇所の割合は 0.89 であり，提案手法「類似度」の単語抽出における正解箇所の割合は 0.82 であり，提案手法「分類語彙表」の単語抽出における正解箇所の割合は 0.82 であった．このように，先行手法より提案手法の方が精度が高い結果になった．また,「Wikipedia 全ページでクラスタリング」の方が「類似度」と「分類語彙表」より精度が高かった．

記載不足の指摘の評価実験は，F 値を用いて正しく空欄として検出できたかを評価した．F 値を評価した結果，先行手法の F 値は 0.77 であり，提案手法「Wikipedia 全ページでクラスタリング」の F 値は 0.75 であり，提案手法「類似度」の F 値は 0.84 であり，提案手法「分類語彙表」の F 値は 0.81 であった．このように，提案手法「類似度」,「分類語彙表」の方が先行手法と提案手法「Wikipedia 全ページでクラスタリング」より精度が高い結果になった．

情報抽出と記載不足の指摘の 2 点で評価を行った結果，以下のことがわかった．単語クラスタリングに利用するデータを増やすことによって，類似した単語が違うクラスタに分割されにくくなり，１つのクラスタに属する単語数は増加した．よって，単語クラスタリングに利用するデータは増やしたほうが精度が上がると考える．単語抽出における正解率から，抽出単語の総数が多いほど，表抽出における正解率の精度は高くなる傾向にある．重要項目に属する単語数を増やすことによって精度の向上が見込める．しかし，単語抽出における正解率から，重要項目に属する単語数が多くなると，重要項目と関係にない単語が表に検出され，単語抽出における正解率は下がるという問題点があることがわかった．また，F 値における記載不足の指摘の評価実験を行った結果，F 値の結果は抽出単語数が多いと高くなる傾向にあることがわかった．

目 次

表 目 次

図 目 次

第1章　はじめに

長文から知りたい情報を取得するためには長い文を読む必要がある．よって，情報取得に時間がかかる．そこで，長文を表に示すことで長い文を読む必要がなく情報を容易に取得することができる．情報を抽出して表に整理する方法として藤原の研究 [1] や Akano らの研究 [2] がある．藤原の研究では上位下位知識を利用して抽出データから下位語の頻度分析を行い，頻度が高かった下位語の上位語を重要項目と選定して，Wikipedia の抽出データから重要項目の下位語を取り出し，表にまとめていた．Akano らの研究では word2vec [3] 内の単語クラスタリングを利用して，表生成に使用する抽出データの類似している単語をまとめて単語のクラスタを作り，頻度の高いクラスタを重要項目として人手で表にまとめていた．しかし，藤原の研究では抽出された重要項目の種類が少なく，Akano らの研究では 1 つのクラスタを重要項目としていたので単語の網羅性が低いという問題点と，クラスタリングを行うデータは，Wikipedia から抽出した事柄を含むページのデータのみで単語クラスタリングを行っていたため，クラスタリングの精度が低くなる問題点があった．

そこで，本研究では先行手法の問題点を元に単語クラスタリングの改良と分類語彙表を用いて表生成を行うことで重要項目の選定と精度の向上を行う．

本研究の主張点を以下に示す.

- 情報抽出
  - 本研究では「単語クラスタリング」による手法と「類似度」による手法と「分類語彙表」による手法の 3 つを提案する．(以降,「単語クラスタリング」による手法と「類似度」による手法と「分類語彙表」による手法を単に,「単語クラスタリング」と「類似度」と「分類語彙表」と表記する場合がある．)「単語クラスタリング」と「類似度」による手法は Akano らの研究 [2] の改良を行う．Akano ら [2] の研究は Wikipedia から抽出した事柄を含むページのデータのみで単語クラスタリングを行っていた．本研究は Wikipedia 全ページを利用して単語クラスタリングを行う．クラスタリングに利用するデータが異なるという新規性が本研究にある．また,「類似度」による手法は，類似度を使用して重要項目に対応する単語群を抽出する点に新規性がある.「分類語彙表」による手法は，分類語彙表よって分類・整理したシソーラス (類義語集) を利用して重要項目の選定を行う手法である．本研究は情報抽出における重要項目の選定するという観点で，分類語彙表を利用することに新規性がある.
  - 情報抽出において，先行手法 [2] と提案手法「単語クラスタリング」と「類似度」と「分類語彙表」の 4 つの手法の評価実験を行う．情報抽出における評価実験としては，表抽出における正解率と単語抽出における正解率の精度を求めた．表抽出における正解率を評価した結果，先行手法の表抽出における正解箇所の割合は 0.68 であり，提案手法「Wikipedia 全ページでクラスタリング」の表抽出における正解箇所の割合は 0.71 であり，提案手法「類似度」の表抽出における正解箇所の割合は 0.88 であり，提案手法「分類語彙表」の表抽出における正解箇所の割合は 0.81 であった．このように，先行手法より提案手法の方が精度が高い結果になった．また，単語抽出における正解率を評価した結果，先行手法の単語抽出における正解箇所の割合は 0.73 であり，提案手法「Wikipedia 全ページでクラスタリング」の単語抽出における正解箇所の割合は 0.89 であり，提案手法「類似度」の単語抽出における正解箇所の割合は 0.82 であり，提案手法「分類語彙表」の単語抽出における正解箇所の割合は 0.82 であった．このように，先行手法より提案手法の方が精度が高い結果になった．よって，提案手法の有用性が

確認できた.

- 表抽出における抽出単語数は，先行手法は 94 単語を表に抽出し，提案手法「単語クラスタリング」は 200 単語を表に抽出し，提案手法「類似度」は 869 単語を表に抽出し，提案手法「分類語彙表」476 単語を表に抽出した．先行手法より提案手法 3 つの方が抽出単語数が多く単語の網羅性が向上した.

- 記載不足の指摘
  - 情報抽出において，入力した文章に記載されてない項目は空欄で表に抽出する．記載されていない項目を空欄として示すことで書き手に追加記載を促すことを目的とする．表の空欄箇所の検出性能を F 値を用いて評価した．先行手法と提案手法「単語クラスタリング」と「類似度」と「分類語彙表」の 4 つの手法で評価実験を行った．先行手法の F 値は 0.77 であり，提案手法「Wikipedia 全ページでクラスタリング」の F 値は 0.75 であり，提案手法「類似度」の F 値は 0.84 であり，提案手法「分類語彙表」の F 値は 0.81 であった．このように提案手法「類似度」と提案手法「分類語彙表」の方が先行手法より精度が高い結果になり,「類似度」と「分類語彙表」の手法においては空欄指摘の有用性が確認できた.

本論文の構成は以下の通りである．第 2 章で関連研究の紹介をする．第 3 章で word2vec と分類語彙表の説明を行う．第 4 章では情報抽出の手法と記載不足の指摘の手法を提案する．第 5 章では実験環境の説明を行う．第 6 章では実験条件や評価方法や実験結果と性能評価を行う．第 7 章では本稿をまとめる．

第2章　関連研究

2.1　情報抽出における関連研究

Akano ら [2] は word2vec [3] 内にある「単語のクラスタリング」を利用して，抽出データに関連した重要項目の選定を行っていた．実験環境と単語のクラスタリングを利用した表生成の手順を以下に示す．

- 実験環境
  - ベクトルの次元は 200 次元
  - 文脈は最大 8 単語
  - ネガティブサンプリングは 25
  - 学習を 20 スレッド並列
  - クラスタ数 1000 個生成
- 表生成の手順
  1. 抽出したい事柄を決定する．Wikipedia から，先に決定した抽出したい事柄を含むページを抽出する．
  2. word2vec 内の単語のクラスタリングの機能を用いて，抽出したデータ内の単語をクラスタリングする．各クラスタにクラスタ番号をふる．各クラスタには類似した単語群が属することになる．
  3. クラスタリング結果に基づく単語のクラスタを表の列とし，抽出したデータのページを表の行とし，ページに出現するクラスタの単語を該当する行と列の箇所に埋める．クラスタの複数の単語がそのページに出力される場合は，それらすべての単語を表の該当する箇所に埋める．
  4. 表の各列にある単語の延べ数 (頻度 A と呼ぶ) を求める．頻度 A が大きい列が左にくるように表で列をソートする．頻度 A の少ないクラスタ番号の列を削除する．

5. 表のソート結果により頻度 A の大きいクラスタ番号の列の中から人手で城に関する情報として重要と思われる列 (重要項目) を選ぶ.

藤原ら [1] は Wikipedia の城に関するページ (対象データ) を抽出し，その中から城に関する重要情報を CaboCha(固有表現抽出ツール) を用いた固有表現抽出に基づく手法と ALAGIN の上位下位知識に基づく手法の 2 手法で抽出した．対象データから CaboCha を用いて,「人名」「地名」「組織名」に分類された語句を抽出し表にまとめた．同様に上位下位知識を用いて対象データで下位語の頻度分析を行い，頻度が高かった下位語の上位語を重要項目とした．対象データで重要項目の下位語を取り出し，表にまとめていた.

宮崎ら [4] は遠距離教師あり学習（distant supervision）を用いて，Wikipedia から得た用語をもとにコーパスに自動でアノテーションすることで専門用語を抽出する手法を行っていた.

近藤ら [5] の研究では大規模コーパスへの網羅的・系統的な語義情報付与を目的とした、分類語彙表・UniDic 見出し対応表の構築を行っていた.

Akano ら [2] の研究では Wikipedia から抽出した事柄を含むページのデータのみで単語クラスタリングを行っているが，本研究では Wikipedia 全ページを利用して単語クラスタリングを行う．単語クラスタリングを利用するデータの違いから本研究は新規性があると考える．藤原ら [1] と宮崎ら [4] の研究は Wikipedia のページを利用して表生成を行っていた．しかし，表抽出の方法が違うため新規性があると考える．近藤ら [5] の研究は分類語彙表の対応表を用いた大規模コーパスへの網羅的な語義情報付与を目的としている．本研究は情報抽出を目的として分類語彙表を利用する.

2.2 空欄指摘における関連研究

岡田ら [6] は，論文の研究成果や研究の有効性や必要性といった論文に記載必要な情報を「記載必要項目」として，論文内で記載必要項目が欠落しているか否かを自動で検出することで，記載不足の指摘を行っていた．岡田ら [6] の研究は論文の記載不足の指摘を行っていたが，本研究では Wikipedia の記載不足の指摘を行う.

第3章　word2vecと分類語彙表

3.1　word2vec

word2vecはTomas Mikolov [7] らによって提案されたニューラルネットワーク（Skip-gram）の手法である．Skip-gramは，文脈を利用して与えられた単語と与えられた単語の周辺に出現する単語を予測できるように，単語ベクトルの学習を行うモデルである．

Mikolovら [7] は、意味的に関連が強い単語はベクトルが近くなると主張している [3]．例えば,「Java」「Perl」「Ruby」などはプログラミング言語として似た単語としてベクトルが近くなる．単語をベクトルに変換することで，人手で入力した単語のベクトルと近いベクトルの単語 (類似した単語) を取得することができる．

また，類似した単語ベクトルを集めてクラス毎に分類することをクラスタリングという．本研究のクラスタリングのアルゴリズムとしてはk- means法を用いる．k-means法のアルゴリズムを以下に述べる．

1. 与えられたデータの中からランダムにk個の単語を取り出し，k個の単語をそれぞれ別の1つのクラスタに割り当て，k個のクラスタを作る．k個の単語がクラスタ中心となる．

2. 残りの単語を，ベクトルの距離が最小となるクラスタ中心のクラスタに割り当てる．

3. 割り当てられたクラスタ内の単語のベクトル平均値を求める．ベクトル平均値をクラスタ中心とする．

4. 上記の2，3を繰り返す．繰り返し処理で得られた前の計算から，クラスタ距離に変化がない，または，繰り返し回数の上限に達した場合，現時点で割り振られたクラスタを出力する．

3.2 分類語彙表

分類語彙表は,「語を意味によって分類・整理したシソーラス (類義語集) 」のことを言う．本研究では国立国語研究所のデータベースを利用して重要項目の選定を行う．

分類語彙表は，数字を利用した構造的な分類体系をとっている．例えば，分類番号は「1.4131」のように 5 桁の数字で表記され，各数字あるいはその組み合わせが「類」「部門」「中項目」「分類項目」の階層的意味づけの構造となっている [5].

最初の 1 桁目は，品詞分類に相当し「類」に対応する．また,「類」は以下の 4 つに分類され，意味的な分類と文法的な分類を両立している．

1. 名詞の仲間－体の類
2. 動詞の仲間－用の類
3. 形容詞・形容動詞・副詞等の仲間－相の類
4. その他の仲間（接続詞・感動詞など）

1〜3 の類は，日本語の文法的特性を考える上の基本的な枠組みで，意味的にはそれぞれ，もの・動き・ありさまという概念に対応する．

次に，分類番号の 2 桁目は,「部門」に相当し,「部門」は「類」の中を意味的に大きな概念で分けたものである．各類と部門に対応する例を表 3.1 に示す．列が「類」に相当し，行が「部門」に相当する．

表 3.1: 「類」と「部門」の関係

	体	用	相
抽象的関係	1.1	2.1	3.1
人間的主体	1.2	-	-
精神および行為	1.3	2.3	3.3
生産物および用具	1.4	-	-
自然物および自然現象	1.5	2.5	3.5

3 桁目は「中項目」に相当し,「部門」より小さな概念を示す．また,「中項目」より小さい概念で,「中項目」を細分化した項目を 4 桁目と 5 桁目で示し，1 桁から 5 桁の組み合わせた数字は「分類項目」相当する.【1.1 抽象的関係】に属する「中項目」の項目と「分類項目」を表 3.2 に示す.

数字あるいはその組み合わせが「類」「部門」「中項目」「分類項目」という 4 階層の意味的範疇を示す構造となり,「類」は 1 階層の項目に属し,「部門」は 2 階層の項目に属し,「中項目」は 3 階層の項目に属し,「分類項目」は 4 階層の項目に属する．本研究では「分類項目」(5 桁) に属する単語群を利用して表生成を行う.

表 3.2: 「中項目」の項目と「分類項目」

【1.1　抽象的関係】
1.10　事柄――――
1.1000　事柄
1.1010　こそあど，他
1.1030　真偽
1.1040　本体，代理
1.11　類――――
1.1100　類，例
1.1101　等級，系統
1.1110　関係
.....

第4章　提案手法

4.1　表生成における情報抽出

Akano ら [2] の研究では，Wikipedia から抽出した事柄を含むページのデータのみで単語クラスタリングを行っていた．しかし，Wikipedia から抽出した事柄を含むページだけではデータが少なく，クラスタリングの精度が低くなる問題点があった．また，1 つのクラスタを重要項目としていたため，単語の網羅性が低いという問題点もあった．

そこで，本研究では 2 つの問題点を解消するために word2vec と分類語彙表を用いて重要項目の取り出し技術の改良を行う．以下で，本研究の提案手法である,「単語クラスタリング」と「類似度」と「分類語彙表」の 3 つの手法を説明する．

4.1.1　単語クラスタリングに基づく情報抽出

先行手法 [2] では，Wikipedia から抽出した事柄を含むページのデータのみで単語クラスタリングを行っていたが，本研究では Wikipedia の全ページを利用して単語のクラスタリングする．表生成手順を以下に示す．

1. 抽出したい事柄を決定する．Wikipedia から，先に決定した抽出したい事柄を含むページを抽出する．

2. word2vec 内の単語のクラスタリングの機能を用いて，Wikipedia の全ページを利用して単語をクラスタリングする．各クラスタにはクラスタ番号をふる．各クラスタには類似した単語群が属することになる．

3. クラスタリング結果に基づく単語のクラスタを表の列とし，抽出したデータのページを表の行とし，ページに出現するクラスタの単語を該当する行と列の箇所に埋める．クラスタの複数の単語がそのページに出力される場合は，それらすべての単語を表の該当する箇所に埋める．

4. 表の各列にある単語の延べ数 (頻度 A と呼ぶ) を求める．頻度 A が大きい列が左にくるように表で列をソートする．頻度 A の少ないクラスタ番号の列を削除する．

5. 表のソート結果により頻度 A の大きいクラスタ番号の列の中から人手で抽出したい事柄に関する情報として重要と思われる列 (重要項目) を選ぶ．

4.1.2 類似度に基づく情報抽出

先行手法では，1 つのクラスタを重要項目としていたため，単語の網羅性が低いという問題点があった．そこで，本研究では単語の類似度に着目する．

word2vec [7] は単語を入力することによって，入力した単語のベクトルと近いベクトルの単語 (類似した単語) を取得することができる．word2vec [7] を用いることで類似度を算出する．類似度を算出した例を表 4.1 に示す．

表 4.1: 類似度算出例

文法	0.611805
語	0.584067
プログラミング	0.560733
語彙	0.559109
インタプリンタ	0.558252
単語	0.549504
コンパイラ	0.540684
アセンブラ	0.534999
日本語	0.534249
文法	0.519926
LISP	0.516639
プログラム	0.514491
諸語	0.514285
方言	0.513930
....	

表 4.1 では，人手で「言語」と入力すると,「言語」と近いベクトルの単語 (類似度の高い単語) の「文法」や「語」を取得することができる．これを利用して人手であらかじめ重要情報と設定した単語との間の類似度の高い単語を算出する．算出して得られ

た類似度の高い単語を重要項目の単語群とする．(「言語」を重要項目と設定し,「文法」,「プログラミング」を重要項目「言語」の単語群とする)

上記の方法を用いた表生成方法を以下に示す.

1. 抽出したい事柄を決定する．Wikipedia から，先に決定した抽出したい事柄を含むページを抽出する.

2. 人手であらかじめ重要情報と設定した単語を重要項目とし，重要項目と設定した単語との間の類似度が高い単語を重要項目の単語群とする．(図 4.1 の例：「言語」を重要項目,「文法」,「プログラミング」を重要項目「言語」の単語群とする)

3. 2 で得られた結果から，重要項目を表の列とし，抽出したデータのページを表の行とし，ページに出現する重要項目名の単語を該当する行と列の箇所に埋める.

4.1.3 分類語彙表に基づく情報抽出

本研究では分類項目を利用して表生成を行う．表生成を行う際は，分類番号の 5 桁を利用する．Wikipedia の大学のページを分類語彙表を利用して，表の生成を行った結果を表 4.2 に示す．また，大学における重要項目を「学科」,「国公立私立」と定義し，重要項目「学科」に対応する分類項目を { 学問・…} とし，重要項目「国公立私立」に対応する分類項目を { 確立 } とする．{ 学問・…} に属する単語を重要項目「学科」の単語群とし，{ 確立 } に属する単語を重要項目「国公立私立」の単語群とする．分類項目 { 学問・…}，{ 確立 } に属する単語の例を図 4.1 に示す.

1. 抽出したい事柄を決定する．Wikipedia から，先に決定した抽出したい事柄を含むページを抽出する．(例：大学)

2. 人手であらかじめ重要情報と設定した単語を重要項目とする．(重要項目を「学科」「国公立私立」と人手で決める)

3. 重要項目に対応する 5 桁の分類項目内の単語を取り出す.（分類項目の例を図 4.1 に示す．分類項目は { 学問・…}，{ 確立 } のことを指す．また，分類項目 { 学問・…} に属する単語を重要項目「学科」の単語群とする）

{学問・…} がく 学 学術 学芸 学問 学業 文武 科学 サイエンス 雑学 実学 官学 家学 専門 科 百科 学際 分科 課目 科目 全科 課業 学課 正課 課外 教科 学科 講座 実科 選科 専科 本科 予科 斯道 その道 テクノロジー 国学 漢学 儒学 経学 朱子学 道学 陽明学 心学 蘭学 洋学 英学 形而上学 美学 神学 語学 文学 史学 考古 考古学 ちり 地理 法学 地政学 商学 家政学 博物学 理数 理学 理化学 数学 代数 きか 幾何 トポロジー 物理 力学 量子力学 光学 化学 生化学 地学 天文学 工学 エレクトロニクス バイオテクノロジー 分子生物学 農学 林学 エコロジー 医学 生理学 疫学 法医学 漢方 歯学 薬学 本草学 文科 法科 商科 りか 理科 いか 医科 工科 文系 理系 算数 算術 図工 家庭科 ホームルーム 音楽 美術 ぎじゅつ 技術 家庭 体育 手工
{確立　　} 成る 成り立つ なす 成す 成就 成立 成業 完成 生成 つくりあげる 作り上げる 醸成 形成 作成 作製 プログラミング 仕上げ 成り立ち かなう みかん 未完 未成 しかけ 組み合わせる つくる 作る 造る 落成 竣工 かんこう 完工 結実 たつ 立つ 実り 大成 既成 速成 むすぶ 結ぶ 構える 設ける 仕掛け 仕組み 組織 構成 構図 組み立て 組み合わせ 組成 自立 独立 特立 存立 築く 孤立 中立 遊離 分立 共存 併存 並存 並立 両立 対立 乱立 確立 樹立 創立 創設 創建 創業 創部 結成 結団 結党 発会 発足 存置 配置 設置 もうけ 設け 新設 既設 未設 特設 こうせつ 公設 私設 仮設 併置 併設 ふち 付置 附置 付設 附設 廃藩置県 設立 りつ 国立 官立 公立 国公立 州立 都立 府立 県立 区立 市立 町立 村立 私立 王立 共立
　.....

図 4.1: 分類項目例

4. 2 の結果に基づく重要項目を表の列とし，抽出したページを表の行とし，抽出したページに出現する重要項目の単語を該当の行と列の箇所に埋める．重要項目の複数の単語がそのページに出力される場合は，それらすべての単語を表のその箇所に埋める．（表抽出の結果例を表 4.2 に示す）

表 4.2: 分類語義表を使った大学での情報抽出

大学名	学科	国公立私立
北 海 道 大 学	医学, 光学, いか, 工学, トポロジー, 化学, 学芸, 語学, 学問, 学術, エレクトロニクス, 学際, 工科, 科学, 物理, 実学, 予科, 文学, 技術, 法学, 法科, 数学, がく, 音楽, 農学, 科, 歯学, 理系, 薬学, 専門, サイエンス, 学課, 経学, 理学, 文系, 医科, 文科, 学, 学科, 科目	付設, 設け, 成り立ち, 発足, 成す, 配置, 構成, 設ける, 国立, 形成, 私立, 創立, 官立, 組織, 附置, 独立, 州立, 設置, 完成, 設立, 作成
京 都 大学	洋学, 医学, 体育, 工学, 化学, 学芸, 語学, 学問, 史学, 学術, 美術, 学際, 全科, 工科, 科学, 家学, 物理, 美学, 講座, 文学, 考古, 理化学, 技術, 課外, 法学, 法科, 数学, がく, 音楽, 農学, 科, 理系, 分科, 薬学, 専門, 経学, 理学, 文系, 医科, 文科, 学, 理科, 地理, 学科, 考古学, 科目	公立, 設け, 竣工, 発足, 仕組み, 生成, 結成, 成す, 王立, 創設, 構成, 設ける, 国立, 構える, 形成, 私立, 創立, 組織, 附置, 国公立, 併設, 州立, 独立, 設置, 完成, 設立, 新設, 作成
九 州 大学	医学, いか, 体育, 工学, 化学, 学芸, 学問, 力学, 史学, 学術, エレクトロニクス, 学際, 工科, 科学, 物理, 講座, 文学, 技術, 課外, 法学, 法科, 数学, がく, 音楽, 農学, 科, 歯学, 理系, 分科, 薬学, 専門, サイエンス, 理学, 文系, 医科, 学, 文科, 理科, 学科, 科目	設け, 竣工, 未完, 成す, 構成, 創設, 設ける, 国立, 形成, 創立, 県立, 組織, 附置, 独立, 完成, 設置, 設立, 新設, 作成

4.2 記載不足の指摘

作成する表の空欄箇所を情報が欠けている項目と判定する．空欄検出により，記載不足を書き手に促す．記載不足の指摘が必要な表の例を表 4.3 に示す．この表について，「郡上城」の「人名」のように空欄になっている箇所は情報抽出において，Wikipedia 内に正解の文章がなく空欄となっている．他の城には存在する重要な項目が Wikipedia において記載されていないことを書き手に知らせることを目的とする．

表 4.3: 記載不足の指摘が必要な表例

城名	地名	人名
大坂城	大阪	豊臣秀吉
二条城	京都	徳川家康
仙台城	仙台	伊達政宗
郡上城	岐阜	

第5章　実験環境

5.1　実験データ

本研究ではWikipedia(2014年11月)のうち，タイトルが城で終っているページ(2,665ページ)を利用する．Wikipediaの記事の例を図5.1に示す.

<title>根添城</title>
<ns>0</ns>
<id>546490</id>
<revision>
<id>52980461</id>
<parentid>50929209</parentid>
<timestamp>2014-09-23T10:41:18Z</timestamp>
<contributor>
<username>Terumasa</username>
<id>406998</id>
</contributor>
<minor />
<text xml:space="preserve">'''根添城(館)'''（ねぞえじょう）は、[[宮城県]][[仙台市]][[太白区]]坪沼地区にある、[[古墳]]跡を利用した[[日本の城]]（館）の跡である。[[陸奥国]]の豪族[[安倍氏(奥州)—安倍氏]]の[[支城]]として用いられた。

[[11世紀]]の[[前九年の役]]で[[源頼義]]に攻められ陥落した。現在は、[[空堀]]、[[土塁]]の跡は認められるが、大部分は[[畑]]となっている。城跡の南側には、源頼義が祀ったといわれる坪沼八幡神社が建っている。

図 5.1: Wikipediaの記事の例

5.2 クラスタリング

本研究におけるクラスタリングの実験環境を以下に示す.

- ベクトルの次元は 200 次元とする
- 文脈は最大 8 単語とする
- ネガティブサンプリングは 25 とする
- 学習を 20 スレッド並列で行う
- クラスタ数 2000 個生成（先行手法は 1000 個生成) する

また，word2vec に使用するデータは，単語毎に空白を入れる必要がある．本研究では日本語の文章を使用しているために文章を単語毎に分割する必要がある．そこで，単語毎に分割を行うため「mecab-0.993」を使用する.「mecab」で単語毎に分割した例を図 5.2，図 5.3 に示す．(以下の図 5.2 が分かち書き前のものであり，図 5.3 が分かち書き後のものである．)

大坂城は、[[上町台地]] の北端に位置する。かつて、この地のすぐ北の台地下には [[淀川]] の本流が流れる天然の要害であり、またこの淀川を上ると [[京都]] に繋がる交通の要衝でもあった。元々古墳時代の古墳があったと言われ、[[戦国時代]] 末期から [[安土桃山時代]] 初期には [[石山本願寺]] があったが、1580 年（天正 8 年）に [[石山合戦]] で焼失した。[[石山合戦]] 終了後、[[織田信長]] の命令で [[丹羽長秀]] に預けられ、後に [[四国攻め]] を準備していた [[津田信澄]] が布陣したこともあったが、信澄は [[本能寺の変]] の際に、丹羽長秀に討たれた。その後、[[清州会議]] で [[池田恒興]] に与えられるも、ただちに [[美濃国—美濃]] へ国替えとなり、秀吉によって領有された。そして秀吉によって大坂城が築かれ、豊臣氏の居城および [[豊臣政権]] の本拠地となったが、[[大坂の役—大坂夏の陣]] で [[豊臣氏]] の滅亡とともに焼失した。徳川政権は豊臣氏築造のものに高さ数メートルの盛り土をして縄張を改め再建した。その後、江戸幕府が [[大坂城代]] を置くなど [[近畿]] 地方、および [[西日本]] 支配の拠点となった。”’[[姫路城]]、[[熊本城]]”’ と共に”’ 日本 [[三名城]] の一つ”’ に数えられている。

図 5.2: Wikipedia の記事に mecab を使用する前の例

大 坂城 は 、 [[上町 台地]] の 北端 に 位置 する 。 かつて 、 この 地 の すぐ 北 の 台地 下 に は [[淀川]] の 本流 が 流れる 天然 の 要害 で あり 、 また この 淀川 を 上る と [[京都]] に 繋がる 交通 の 要衝 でも あっ た 。 元々 古墳 時代 の 古墳 が あっ た と 言わ れ 、 [[戦国 時代]] 末期 から [[安土 桃山 時代]] 初期 に は [[石山 本願寺]] が あっ た が 、 1580 年 （ 天正 8 年 ） に [[石山 合戦]] で 焼失 し た 。 [[石山 合戦]] 終了 後 、 [[織田 信長]] の 命令 で [[丹羽 長秀]] に 預け られ 、 後 に [[四国 攻め]] を 準備 し て い た [[津田 信 澄]] が 布陣 し た こと も あっ た が 、 信 澄 は [[本能寺 の 変]] の 際 に 、 丹羽 長秀 に 討た れ た 。 その後 、 [[清州 会議]] で [[池田 恒 興]] に 与え られる も 、 ただちに [[美濃 国 — 美濃]] へ 国替え と なり 、 秀吉 によって 領有 さ れ た 。 そして 秀吉 によって 大 坂城 が 築か れ 、 豊臣 氏 の 居城 および [[豊臣 政権]] の 本拠地 と なっ た が 、 [[大 坂 の 役 — 大 坂 夏 の 陣]] で [[豊臣 氏]] の 滅亡 とともに 焼失 し た 。 徳川 政権 は 豊臣 氏 築造 の もの に 高 さ 数 メートル の 盛り土 を し て 縄張 を 改め 再建 し た 。 その後 、 江戸 幕府 が [[大 坂 城代]] を 置く など [[近畿]] 地方 、 および [[西日本]] 支配 の 拠点 と なっ た 。 ”’[[姫路城]]、[[熊 本城]]”’ と共に ”’ 日本 [[三 名城]] の 一つ ”’ に 数え られ て いる 。

図 5.3: Wikipedia の記事に mecab を使用した結果の例

5.3 類似度

本研究における類似度の実験環境を以下に示す.

- ベクトルの次元は 200 次元とする
- 文脈は最大 8 単語とする
- ネガティブサンプリングは 25 とする
- 学習を 20 スレッド並列で行う
- 閾値 0.2 以上の類似した単語を 4000 個まで取得する.

また，クラスタリングと同様「mecab-0.993」を使用する.

5.4 分類語彙表

5 桁の分類項目内の単語群を利用して重要情報の抽出と表生成を行う.

第6章　実験

6.1　実験条件

実験データには，Wikipedia の 3,264,893 ページ (2014 年 11 月) を用いる．本研究では「城」というキーワードに基づき記事の抽出を行う．

6.2　評価方法

先行手法と提案手法の評価実験を行うため，2,665 件の城ページからランダムに抽出した城ページ 30 件を用いて評価を行う．

また，情報抽出と記載不足の指摘の観点で行う．情報抽出は表抽出における正解率と抽出単語の正解率を求める．記載不足の指摘における評価方法は空欄検出における F 値で評価する．

6.2.1　表抽出における正解率

先行手法の実験はクラスタリングを行い，頻度計算から重要項目の決定を行う．頻度計算から，人手で選んだクラスタ番号「401」「407」「765」を評価対象として評価実験を行う．クラスタ番号「401」では戦い関係の情報が１つでも正しく抽出された場合正解とし，クラスタ番号「クラスタ 407」では城の造りの情報が１つでも正しく抽出された場合正解とし，クラスタ番号「クラスタ 765」は交通関係の情報が１つでも正しく抽出された場合正解とする．また，空欄が抽出された場合は，Wikipedia 内に本当に正解の記載が無かった場合正解とする．クラスタ番号「401」を「戦い」とし，クラスタ番号「407」を「城の造り」とし，クラスタ番号「765」を「交通」として人手で重要項目名をふる．

提案手法の単語のクラスタリングの実験も同様に頻度計算から重要項目の決定を行う．頻度計算から，人手で選んだクラスタ番号「1556」「465」「1472」を評価対象として評価実験を行う．また，クラスタ番号「1556」を「戦い」とし，クラスタ番号「465」

を「城の造り」とし，クラスタ番号「1472」を「交通」として人手で重要項目名をふる．評価方法としては先行手法と同様に行う．

提案手法の類似度の実験は先行手法と提案手法の比較を行うため,「戦争」「文化財」「交通」を重要項目と決定し，それぞれ 3 つの単語との類似度が高い単語を重要項目の単語群とする.「戦争」は「戦い」に対応し,「文化財」は「城の造り」に対応づけて評価する．評価方法としては先行手法と同様に行う．

提案手法の分類語彙表は分類項目名「からだ」「時代」「火」「平和」「競争」「攻防」「勝敗」「軍事」「支配」「刑」「捕縛」に属する単語を重要項目「戦い関係」と定義し，分類項目名「社寺」「住居」「家屋」「門・塀」「へや」「屋根」「その他」に属する単語を重要項目「城の造り」と定義し，分類項目名「道路」「過程」「通行」「運輸」に属する単語を重要項目「交通関係」と定義し評価する．評価方法としては先行手法と同様に行う．

6.2.2 単語抽出における正解率

抽出単語の正解率は抽出された単語が重要項目の内容のものであれば正解とする．

6.2.3 記載不足の指摘による評価実験

記載不足の指摘の実験において，表の空欄の箇所について F 値を求める．F 値の算出方法を以下に示す．

$$F = (\frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}) \tag{6.1}$$

$$\text{適合率} = \frac{\text{空欄かつ Wikipedia 内に正解がないもの}}{\text{空欄のもの}} \tag{6.2}$$

$$\text{再現率} = \frac{\text{空欄かつ Wikipedia 内に正解がないもの}}{\text{Wikipedia 内に正解がないもの}} \tag{6.3}$$

本研究において，適合率は表抽出において空欄となった場所が正しく空欄とした割合を表したものである．再現率はWikipedia内に正解の記載がなかったもののうち，正しく空欄を抽出できた割合である．F値は適合率と再現率の調和平均である．式6.2，式6.3において「空欄のもの」は表抽出において空欄の部分のことである．また「Wikipedia内に正解がないもの」はWikipedia内にもともとその項目に関する事柄の記載がなされていないもののことである．F値が大きいほど，Wikipediaでの記載の欠如をシステムがより正しく抽出できたことを意味する．

6.3 実験結果

6.3.1 表生成

2,665 の全城ページの中からランダムに選んだ 30 個の城ページを使用する．空欄部分は空欄として正しく検出できたことを示す.（×）は重要項目の単語としてふさわしくないものに（×）をつけている．また，×は重要項目の重要情報とする単語が Wikipedia の城ページに記載されているが，提案された手法では単語を抽出できなかったことを示す.

先行手法での表生成結果を表 6.1 と表 6.2 に示す．また，提案手法（単語クラスタリング）での表生成結果を表 6.3 と表 6.4 に示し，提案手法（類似度）での表生成結果を表 6.5 から表 6.7 に示し，提案手法（分類語彙表）での表生成結果を表 6.8 と表 6.9 に示す.

表 6.1: 先行手法の表生成結果 1

	戦い	城の造り	交通
1	×		×
2	敗走, 退却, 放火	門 (×), 御殿, 二の丸, 御門, 門跡, 鐘, 大手門, 二ノ (×), 土蔵, 丸 (×), 本丸	北陸, 要衝, 便 (×), 交通, 街道
3			
4	阻止, 攻める, 奮戦	役所, 門 (×), 二の丸, 移築, 大手門, 丸, 本丸	越え (×)
5	×	役所, 日出 (×), 門 (×), 二の丸, 御門, 門跡, 大手門, 二ノ (×), 丸 (×), 東大手, 本丸, 御殿, 表門, 西丸, 東門, 西門, 番所, 正門	北陸, 海道 (×)
6			
7	撤退, 援軍	門 (×), 丸 (×), 本丸	×
8	攻める, 出陣, 落城	門 (×)	×
9			×
10	落城	×	北陸
11			
12	落城	×	×
13			
14			
15			

表 6.2: 先行手法の表生成結果 2

	戦い	城の造り	交通
16	攻める, 撤退, 落城, 援軍	門 (×), 丸 (×)	要所
17	落城, 援軍, 奮戦	二ノ (×), 丸 (×), 本丸	×
18	撤退, 壊滅	門 (×)	交通
19			
20	×	門 (×), 移築, 大手門, 高麗	×
21	焼か, 落城	×	×
22	×	×	×
23	ひい (×)	門 (×)	抑える (×), 交通
24		門 (×)	便 (×), 交通
25	篭城, 落城		
26	×	×	×
27			
28	防戦	×	×
29			
30			

表 6.3: 提案手法（単語クラスタリング）の表生成結果 1

	戦い	城の造り	交通
1	×		至る (×)
2	率い, 配下, 退却, 残党, 入城, 包囲, 敗走, 戦い, 自刃, 戦死	天主, 屋敷, 石垣, 二の丸, 御門, 大手門, 城内, 城跡, 郭内, 東向 (×), 白壁, 二ノ (×), 天守, 本丸, 二条城 (×), 御殿, 櫓門, 多聞, 曲輪, 却後 (×), 城郭, 城下町, 搦手 (×), 城下, 縄張 (×)	曲がりくねっ(×), 要衝, 通り (×), 至る (×), 街道, ルート
3			
4	追い返し, 攻める, 率い, 攻め, 降伏, 内応, 入城, 包囲, 戦い, 自害, 奮戦, 大軍	石垣, 二の丸, 見城 (×), 大手門, 城内, 多聞, 城跡, 城郭, 城下町, 城下, 移築, 三の丸, 天守, 本丸	沿い (×)
5	×	天主, 楽市 (×), 石垣, 二の丸, 物見, 御門, 大手門, 城内, 城跡, 郭内, 天守閣, 二ノ (×), 天守, 江戸城 (×), 本丸, 二条城 (×), 御殿, 櫓門, 表門, 多聞, 曲輪, 姫路城 (×), 西丸, 城郭, 正殿, 北の丸	×
6			
7	差し向け, 攻め, 降伏, 援軍, 包囲, 戦い, 撃退, 大軍, 加勢	枡形, 城郭, 修築, 石垣, 城内, 曲輪, 本丸	×
8	攻める, 攻め, 討伐, 落城, 戦い, 自害, 平定, 出陣, 戦死	城郭, 石垣, 物見, 城跡	×
9			×
10	平定, 落城	城郭	×
11			
12	攻め, 落城, 大軍	天守閣, 城跡, 天守	×
13			
14			
15			

表 6.4: 提案手法（単語クラスタリング）の表生成結果 2

	戦い	城の造り	交通
16	寝返り,攻める,攻め,落城,援軍,入城	陣屋, 城郭, 曲輪, 城跡	要所
17	自害, 軍勢, 急襲, 落城, 撃退, 援軍, 寝返っ, 奮戦	城郭, 屋敷 (×), 城下, 二ノ (×), 本丸	車道
18	率い, 軍勢, 降伏, 包囲	城郭, 城跡, 居館, 別邸	×
19			
20	×	屋形, 石垣, 櫓門, 大手門, 城内, 城跡, 城郭, 城下町, 城下, 移築, 天守	×
21	落城, 対峙	城郭, 城内, 曲輪, 城跡	×
22	軍兵, 戦い, 本営 (×)	×	×
23	×	城跡	道路, 標識
24		見城 (×)	×
25	篭城, 落城		
26	×	城郭, 空堀, 城跡	×
27			
28	防戦, 抗す		
29	入城 (×)	縄張 (×)	
30			

表 6.5: 提案手法（類似度）での表生成結果 1

	戦い	城の造り	交通
1	大戦, 共和, 同国,1945(×), 朝鮮, 人民 (×), 独立, 中国, 中華人民共和国, 戦後, 統治, 併合, 世界 (×), 終戦		新聞 (×), 首都, 京都
2	集団, 率いる, 有様, 事実 (×), 歴史, 八月 (×), 包囲, 戦い, 敗走, 混乱, 略奪, 会議, 謀反, 宣教, 証言, 横死, 留守, ヨーロッパ, 警戒, 賢人, 戦死, 謙信, 思想 (×), 将軍, 人々(×), 続く (×), 建造, 勢力, ローマ, 退去, 成立, 軍事, 率い, 退却, イタリア, 政治, 残党, 入城, 重要, 彼ら (×), 権力, 一揆, 本国, 侵入, 近代, 離日, 吉次, 尽く	寄贈, 築造, バチカン (×), 唐破風, 巨石, 発掘, 宝塔, 焼失, 舎利, 本丸, 遺構, 史料, 仏堂, 修復, 出土, 建て (×), 史蹟, 国定, 建築, 由緒 (×), 焼け, 造営, 名城, 近江八幡, 復元, 本堂, 皇居, 仏塔, 城山, 入母屋, 公園, 博物館, 建物, 建造, 御所, 総見, 母屋, 白壁, 二ノ (×), 御物, 展示, 座敷, 名勝, 開館, 資料, 住居, 文化, 所在 (×), 観音, 復原, 史跡, 境内, 現存, 絵図, 観音寺, 堂塔, 選定 (×), 仁王門, てら (×), 焼け落ち, 壁画, 八角 (×), 金箔, 城郭, 城址, 遺物, 礎石, 土蔵, 石垣, 二の丸, 文書, 夢殿, 保管, 城跡, 障壁, 金閣, 保存, 天守, 指定, 石碑, 木造, 櫓門, 御殿, 多聞, 屏風, 城下町, ギャラリー (×), 寺院	整備, 中央 (×), 実施 (×), 徒歩, 計画 (×), 名古屋, 施設, 建設, 西日本, 管理, 水運, 概要 (×), 京都, 熊本, 設置, 置場, 旅客, 総務 (×), 利用 (×), 新聞 (×), 鉄道, 観光, 事業, 平成 (×), 公園, 利便, 国土 (×), 北陸, 航空, 情報 (×), ルート, 工事, 指定, アクセス, 近畿, 周辺, 住宅
3			
4	隆盛, 漸く (×), 歴史 (×), 包囲, 支配, 戦い, 問題, 恨み, 大軍, 敗れ, 征伐, 攻め, 出来事, 戦功, 攻撃, 交渉, ヨーロッパ, 謀殺, 攻める, 関ヶ原, 降伏, 起こっ(×), 臣従, 建造 (×), 戦闘, 率い, 城壁, 脅かさ, 入城 (×), 敗れる, 独立, 幕末, 近代 (×), 西郷	現存, 選定, 大火, 本丸, 遺構, 神社, 市役所, 出土, 鎮魂 (×), 城郭, 城址, 移築, 焼け, 名城, 復元, 二の丸, 石垣, 城跡, 民家 (×), 公園, 建物, 建造, 保存, 再建, 建立, 天守, 指定, 石造り, 石造, 多聞, 開館, 城下町, 文化, 三の丸, 史跡	整備 (×), 中央 (×), 利用 (×), 函館, 地方, 平成(×), 公園, 国土 (×), 九州, 施設, 航空, 指定 (×), 市内, 市役所, 概要 (×), 熊本
5	長篠, 歴史, 半島, 重要 (×), 中国, ヨーロッパ, 本国 (×), 世界 (×)	民俗 (×), 札所, 天守閣, 本丸, 史料, 神社, 表門, 市役所, 図書館, 宝物, 遺跡, 彦根城, 建築, 町並 (×), 番所 (×), 宝物殿, 名城, チャシ (×), 近江八幡, 復元, 本堂, 道府県 (×), 城山 (×), 景観, 公園, 博物館, 建物, 遺産, 二ノ (×), 御物, 郷土, 多賀城, 展示, 名勝, 重文, 資料, 旧宅, 文化, 観音, 所在 (×), 鑁阿, 復原, 史跡, 宮城 (×), 現存, 観音寺, 選定, 城郭, 城址, 二の丸, 石垣, 城跡, 保存, 天守, 旧跡, 図書, 指定, 御殿, 櫓門, 多聞, 姫路城, 国宝, 正殿, 栃木 (×)	東北, 都道府県, 北海道, 県民 (×), 四国, 函館, 東海, 東京, 九州, 広島, 名古屋, 施設, 岩国, 管理, 金沢, 市役所, 仙台, 京都, 大阪, 熊本, 事務, 設置, 置場, 唐津, 総合 (×), 観光, 那覇, 富山, 事業, 地方 (×), 公園, 料金, 北陸, 八王子, サービス, 情報(×), 関東, 指定, 近畿, 大連 (×), 都市, 太宰府, ビジターセンター, 駅前

表 6.6: 提案手法（類似度）での表生成結果 2

	戦い	城の造り	交通
6	議論 (×)	保存 (×)	利用 (×)
7	背景 (×), 北方, 歴史, 包囲, 戦い, 和議, 鎮圧, 大軍, 加勢, 再興, 征伐, 攻め, 堅固, 援軍, 将門, 禁令, 争い, 続い (×), 不和, 孤立, 関ヶ原 (×), 降伏, 謙信, 起こっ(×), 将軍, 両国 (×), 勢力, 世紀 (×), 成立, 撃退, 苦悩, 決別, 賞賛, 一方 (×), 一掃	名城 (×), 石垣, 公園, 下野 (×), 建立, 本丸, てら (×), 遺構, 神社, 祀る, 城郭, 建て, 鎮座, 所在 (×), 栃木 (×)	市街地, 新宿, 市街, 概要 (×), 地方, 平成 (×), 公園, 関東
8	武功, 脅威, 落城, 歴史 (×), 戦い, 合戦, 塹壕, 攻め, 帰国, 戦功, 重臣, 参戦, 中国, 断絶, 決意, 戦死, 攻める, 関ヶ原, 討伐,1527(×), 戦果, 意識 (×), 掌握, 建造 (×), 平定, 勢力, 出陣, 領土, 武力, 城壁, 離反, 説き (×)	石垣, 城跡, 建造, 再建, 町 (×), 遺構, 指定, 屋久, 城郭, 称名寺, 城址, 文化 (×)	大阪, 中央 (×), 指定 (×), 神戸, 水路, 新聞 (×), 概要 (×)
9			富山
10	失敗, 反乱, 調停, 占領, 落城, 歴史 (×), 一揆, 平定, 勢力, 侵攻	城郭, 所在 (×), 修復	東海, 北陸, 概要 (×), アクセス (×), 鉄道, 旅客, 運用
11			
12	乗り出す, 征伐, 攻め, 滅ぼさ, 落城, 歴史 (×), 支配, 勢力, 和議, 批判, 侵攻, 大軍	復元, 城跡, 公園, 天守閣, 再建, 天守, てら (×)	整備, 四国, 公園
13			
14			
15			

表 6.7: 提案手法（類似度）での表生成結果 3

	戦い	城の造り	交通
16	攻める, 攻め, 落城, 援軍, 領内, 歴史 (×), 入城 (×), 戊辰戦争, 末弟, 独立, 対立, 勢力, 途絶, 中東 (×)	地区 (×), 神社, 城跡, 公園, 城郭, 遺跡, 焼失	行政, 地区, 概要 (×), 地方, 公園
17	不和, 北方 (×), 落城, 援軍, 重臣, 歴史, 攻撃, 破綻, 軍勢, 合戦, 撃退, 同盟, 裏切っ, 寝返っ	文書, 城郭, 城址, 二ノ (×), 本丸	東北, 車道, 市街地, 市街, バス, 概要 (×), アクセス, 鉄道, 徒歩, 自動車, 営業
18	率いる, 殺害, 農民, 降伏, 起こっ, 事件, 歴史 (×), 大義, 包囲, 支配, 大義名分, 軍勢, 襲っ, 勃発, 不満, 乗じ, 壊滅, 率い, 征伐, 思惑, 凶作, 攻撃, 騒動, 独立, 一揆, 断絶, 長州, 動機, 経済, 統治, 苦しん	遺構, 指定 (×), 大明神, 資料, 城跡, 居館, 別邸, 小字 (×), 発掘, 城郭, 遺跡, 文化, 栃木 (×), 史跡	指定 (×), 近隣, 成田, 用地, 平成 (×), 国土 (×), 周辺, 交通省, 地元, ホームページ, 情報, 関東, 工業団地
19	0088(×)		
20	文禄・慶長の役,1813(×), 戦後, 朝鮮, 通信使, 歴史 (×)	民俗 (×), 発掘, 焼失, 大火, 遺構, 史料 (×), 神社, 城郭, 対馬, 移築, 名城, 文化庁 (×), 復元, 文書, 石垣, 城跡, 馬市 (×), 遺産, 保存, 再建, 天守, 庭園, やかた, 指定, 木造, 名勝, 櫓門, 城下町, 文化 (×), 史跡, 厳原	整備, 指定 (×), 九州, 市街, 新聞 (×), 概要 (×), 事業, 平成 (×)
21	敗れ, 落城, 回復, 争乱, 続く, 歴史 (×), 権力, 帰し, 統制, 対峙, 死す, 権威	史料 (×), 興福寺, 城跡, 城郭, 妙法, 再建	熊本, 九州, 金沢, 概要 (×), 北九州
22	出征, 艦長, 軍需, 大本営, 作戦, 中将, 軍人, 除隊, 歴史 (×), 帝国, 戦時, 軍兵, 太平洋戦争, 戦い, 少将, 終戦, 各国 (×), 艦隊, 連合, 以後 (×), フィリピン, 陸戦, 参謀,1945(×), 太平, 参戦, 軍務, 海軍, 戦艦	墓所, 保存 (×)	関東, 情報 (×)
23	共和, 北方 (×), 防衛, 派遣, 長城, 懐柔, 民族, 歴史 (×), 重要 (×), 人民, 戦略, 中華人民共和国, 建造, 成立, 世界 (×), 万里 (×)	指定 (×), 修復, 城跡, 遺産, 文物, 建造, 建築	乗車, 観光, 標識, 建設, 自動車, 指定 (×), 市内, 路線, 高速, 直行, バス, 市街, 道路, 設置, 首都, 地下鉄
24	平和 (×), ナント (×)	所在 (×)	バス停, 那覇, 徒歩, 施設, 主要 (×), 管理, 路線, 開業, アクセス, バス, 供用 (×), 商圏, 企業, 産業, 郵便, 駐車, 隣接 (×), ショッピング, 営業
25	征伐, 落城, 歴史 (×), 従属, 攻撃, 攻略, 脱出, 篭城, 離反, 与し (×)	栃木 (×)	
26	攻撃, 半島, 征伐, 世紀 (×), 歴史 (×)	城郭, 空堀, 境内, 神社, 城跡	概要 (×), 東浦, 水道
27			
28	共和, 防戦, 攻防, 要塞, 城壁, 歴史 (×), 重要 (×), 支配, ヨーロッパ, ハンザ, 世紀 (×), 同盟, 陣地	資料 (×)	建設, 都市
29	議論 (×), 問題 (×), 入城 (×)	保存 (×)	利用 (×), 管理 (×)
30			

表 6.8: 提案手法（分類語彙表）での表生成結果 1

	戦い	城の造り	交通
1	大戦, 共和, 明治, 独立, 時代, 戦後, 統治, 併合, 年代 (×), 終戦, 昭和	城府 (×)	×
2	政事, 戦国, 戦い, 敗走, 時代 (×), 焼失, 古代 (×), 長命 (×), 安政, 大正, 延焼, 謀反, 出火, 護法 (×), 天正, 横死, 一統, 防御, 走り, 守護, 戦死, 自治 (×), 阻む, 財政, 中世, 平成 (×), 天下, 自刃, 放火, 軍事, 昭和, 退却, 近世, 政治, 現代, 近代, 内務	伽藍, 破風, 物置, 大手門, 堂塔, 内裏, 文庫 (×), 宝塔, 施設, 焼失, 仁王門, 霊廟, 延焼, 本丸, 暗渠, 仏堂, 出火, 部屋 (×), 小座敷, 書院, 城郭, 吹き抜け, 出窓, 石段, 土蔵, 欄干, 名城 (×), 屋敷, 邸宅, 石垣, 二の丸, 本堂, 皇居, 仏塔, 台所, 地階, 建物, 障壁, 四畳半, 御所, 研究所 (×), 母屋, 白壁, 放火, 井戸, 座敷, 本城 (×), 御殿, 住居, 居城, 離れ, 古城, ギャラリー (×), 望楼, 寺院, 山城, 屋根, 住宅	通り (×), 流し (×), 途中 (×), 鉄道, 通り抜け, 徒歩, 航空, 街道, ルート, 回路, 水運, 滑り (×), 渡し, 横滑り (×), 交通, 走り (×)
3			
4	明治, 戦国, 殺し, 支配, 戦い, 時代, 全焼, 大火, 奮戦, 火事, 攻め, 征伐, 出火, 天正, 付き (×), 自害, 生まれ (×), 謀殺, 廃藩置県, 攻める, 降伏, 慶長, 平成, 戦争, 戦闘, 昭和, 近世, 治める, 現代 (×), 敗れる, 独立, 幕末, 近代 (×)	名城, 石垣, 二の丸, 大手門, 民家 (×), 建物, 基礎, 全焼, 施設, 大火, 本丸, 火事, 国大 (×), 本城, 出火, 神社, 付き (×), 城壁, 城郭, 山城, 三の丸	途中 (×), 流れ (×), 航空
5	時代 (×), 戦国, 発達 (×), 昭和	宮城 (×), 宝物殿, 名城, 石垣, 舞台, 二の丸, 札所, 本堂, 大手門, 台所, 大学 (×), 建物, 天守閣, 茶屋, 施設, タワー, ハウス (×), 本丸, 女子大 (×), 根城, 神社, 本城, 表門, 御殿, 旧宅, 書院, 城郭, 正殿, 山城, 正門	交流 (×), 海道 (×), 都道 (×)
6	侵害 (×)		
7	明治, 戦国, 降伏, 中世, 慶長, 平成 (×), 戦い, 鎮圧, 時代, 撃退, 世紀 (×), 昭和, 攻め, 征伐, 付き (×), 天正, 撤退	名城 (×), 神社, 本城, 石垣, 付き (×), 居城, 城郭, 古城, ビル, 山城 (×), 鎮守, 井戸, 本丸	早馬
8	攻める, 討伐, 近衛, 明治, 落城, 追討, 平定, 戦い, 時代 (×), 出陣, 合戦, 歴代 (×), 攻め, 武力, 天正, 死去, 自害, 守護, 戦死	城郭, 塹壕, 水路, 石垣, 平城, 城壁, 山城 (×)	水路
9		山城 (×)	×
10	反乱, 天正, 戦国, 落城, 平定, 時代 (×), 侵攻, 大正, 年代 (×)	城郭, 小学 (×), 山城	本線, 鉄道
11		山城 (×)	
12	征伐, 攻め, 天正, 戦国, 落城, 制覇, 支配, 時代 (×), 侵攻, 昭和	天守閣, 東屋, 新居	流れ (×), 難航
13		山城 (×)	
14		山家 (×)	
15			

表 6.9: 提案手法（分類語彙表）での表生成結果 2

	戦い	城の造り	交通
16	歴代 (×), 攻める, 攻め, 近世, 天正, 落城, 中世, 独立, 行政, 戦争, 撤退, 焼失, 時代 (×)	城郭, 古城, 神社, 焼失, 山城	×
17	天正, 戦国, 落城, 慶長, 徳政 (×), 自害, 急襲, 時代 (×), 撃退, 合戦, 奮戦	離れ, 城郭, 築地, 屋敷, 山城 (×), 本丸	車道, 流れ (×), 鉄道, インターチェンジ, 徒歩, 経由
18	殺害, 降伏, 没する, 慶長, 平成 (×), 支配, 刑罰, 時代 (×), 分け (×), 歴代 (×), 慶応, 征伐, 天正, 独立, 島流し, 撤退, 統治, 年代 (×)	城郭, 本城 (×), 平城, 別邸, 居城	流し (×), 流れ (×), 交通, 経緯
19			
20	近世, 慶長, 平成 (×), 焼失, 戦後, 大正, 大火, 昭和	名城, 石垣, 大手門, 文庫, 小学 (×), 焼失, 大火, 中学 (×), 神社, 平城, 学校 (×), 居城, 城郭, 噴水, 山城, 望楼, 中学校 (×)	×
21	応仁, 時代 (×), 立法, 落城, 中世, 年代 (×), 死する	城郭, 小学 (×), 馬屋, 山城, 学校 (×), 小学校 (×)	動向 (×)
22	明治, 死亡, 生誕 (×), 生まれる (×), 戦時, 太平洋戦争, 戦い, 戦争, 大正, 終戦, 昭和, 陸戦, 太平, 軍令, 警備, 生まれ (×), 死没	大学, 基礎, 鎮守, 学校 (×)	×
23	共和, 防御, 時代 (×), 防衛, 守備		路線, 道路, 高速, 交通, 直行
24		施設 (×)	路線, バイパス, 交通, 徒歩
25	攻略, 征伐, 時代 (×), 戦国, 落城, 中世		
26	征伐, 時代 (×), 戦国, 世紀 (×)	城郭, 空堀, 神社, 山城, 水道	×
27			
28	支配, 攻防, 共和, 防戦, 中世, 世紀 (×)	城塞, 要塞, 舞台 (×), 城壁	×
29	侵害 (×)		
30			

6.3.2 表抽出における正解率

表に 1 つでも正しく情報を抽出したものを正解とし，また空欄を正しく空欄と検出できれば正解とする．先行手法と提案手法の単語クラスタリングと分類語彙表の表抽出における正解率を表 6.10 に示す.

表 6.10: 表抽出における正解率

手法	正解率
先行手法	0.68 (61/90)
提案手法 (Wikipedia 全ページでクラスタリング)	0.71 (64/90)
提案手法（類似度）	0.88 (79/90)
提案手法（分類語彙表）	0.81 (73/90)

表抽出における正解率を評価した結果，先行手法の表抽出における正解箇所の割合は 0.68 となり，提案手法「Wikipedia 全ページでクラスタリング」の表抽出における正解箇所の割合は 0.71 となり，提案手法「類似度」の表抽出における正解箇所の割合は 0.88 となり，提案手法「分類語彙表」の表抽出における正解箇所の割合は 0.81 となった．このように，先行手法より提案手法の方が精度が高い結果になった．また，「Wikipedia 全ページでクラスタリング」と「分類語彙表」よりも「類似度」の結果の方が精度が高い結果になった.

「Wikipedia 全ページでクラスタリング」と「分類語彙表」が低かった原因としては,「Wikipedia 全ページでクラスタリング」では 1 つのクラスタを重要項目としていたことが挙げられる.「Wikipedia 全ページでクラスタリング」の重要項目 1 つあたりの単語数は約 190 単語を網羅し,「類似度」の重要項目 1 つあたりの単語数は 4000 単語を網羅し,「分類語彙表」の重要項目 1 つあたりの単語数は約 516 単語を網羅していた.「Wikipedia 全ページでクラスタリング」は他の提案手法 2 つよりも大幅に少ない.よって，他の提案手法 2 つよりも単語の網羅生が低く重要情報の抽出ができなかったことが原因と考える.

また,「分類語彙表」は「源義家」,「吉田」といった固有名詞や「岐阜」,「五日市」といった地名は分類語彙表に登録されてない単語で未定義に分類される．未定義に属する単語は単語同士の関連性が低く重要項目の候補から外していた．未定義を重要項目から外していたため「分類語彙表」の「交通」の列に地名を抜き出すことができなかったことが原因だと考える.

6.3.3 単語抽出における正解率

先行手法と提案手法における「戦い」「城の造り」「交通」の抽出単語数を表 6.11 に示す.

表 6.11: 抽出単語の総単語数

手法	抽出単語数
先行手法	94
提案手法 (Wikipedia 全ページでクラスタリング)	200
提案手法（類似度）	869
提案手法（分類語彙表）	476

先行手法より提案手法の方が抽出できる単語数が増加した．また，抽出単語の正解率を表 6.12 に示す.

表 6.12: 抽出単語の正解率

手法	抽出単語の正解率
先行手法	0.73 (69/ 94)
提案手法 (Wikipedia 全ページでクラスタリング)	0.89 (177/ 200)
提案手法（類似度）	0.82 (710/ 869)
提案手法 (分類語彙表)	0.82(391/ 476)

単語の不正解の例としては，提案手法（類似度）の城ページ番号 28 の「城の造り」が「資料（×)」となっている．これは重要項目「城の造り」として単語「資料」は内容としてふさわしくないので×としている．また，提案手法（分類語彙表）の城ページ番号 11 の「城の造り」が「山城（×)」となっている．これは城の造りで山を利用して建てられた城という意味で抽出されたのではなく，瀬戸山城の「山城」の部分が抽出されていたので (×) としている.

先行手法の単語抽出における正解箇所の割合は 0.73 となり，提案手法「Wikipedia 全ページでクラスタリング」の単語抽出における正解箇所の割合は 0.89 となり，提案手法「類似度」の単語抽出における正解箇所の割合は 0.82 となり，提案手法「分類語彙表」の単語抽出における正解箇所の割合は 0.82 となった．このように，先行手法より提案手法の方が精度が高い結果になった．また,「Wikipedia 全ページでクラスタリング」の方が「類似度」と「分類語彙表」より精度が高かった.

「類似度」と「分類語彙表」が低かった原因としては，重要項目と関係のない単語が表に検出されることに問題があると考える．表 6.11 より，単語の網羅性は上がっている．しかし,「類似度」の表生成では重要項目「文化財」の列に「バチカン」,「道府

県」などが検出されている．また,「分類語彙表」の表生成でも重要項目「交通」の列に「通り」,「流し」など「交通」とは関係のない単語が検出されている．

結果として，単語取得の増加で網羅性は上がっているが，重要項目と関係のない単語が表に検出されることが精度が低くなったと考えられる．

6.3.4 F 値における記載不足の指摘の評価

6.2.3 節で述べた方法で記載不足の指摘の実験を行った．また，先行手法と提案手法で記載不足の指摘を行った結果を比較した．

Wikipedia の城ページにおいて実際に情報が欠落していた項目を，情報抽出の実験で適切に空欄として検出できると，記載不足の指摘が適切に行えたと考える．空欄箇所に基づく情報の欠落項目の検出性能を再現率，適合率，F 値で評価した．その結果を表 6.13 に示す．

表 6.13: 文章作成支援の結果の評価

手法	再現率	適合率	F 値
先行手法	0.97 (36/ 37)	0.66 (36/ 61)	0.77
提案手法（Wikipedia 全ページでクラスタリング）	0.95 (35/ 37)	0.62 (34/ 55)	0.75
提案手法 (類似度)	0.73 (27/37)	1.00 (27/27)	0.84
提案手法 (分類語彙表)	0.81 (30/37)	0.81 (30/37)	0.81

再現率は Wikipedia 内に正解の記載がなかったもののうち，正しく空欄を抽出できた場合に正解とする．また，重要項目の単語が Wikipedia のページに表記されていない場合に，表の箇所に間違えて単語を抽出した場合は不正解とする．例としては，提案手法「分類語彙表」の城ページ番号 6 の重要項目「戦い」で「議論 (×)」となっている．城ページ番号 6 のページで重要項目「戦い」に関連する単語は表記されていないが「議論」が間違えて抽出された．このような場合不正解とする．

適合率は表抽出において空欄となった場所が正しく空欄とした場合正解とする．また，城ページにおいて重要項目の単語の表記されている表の箇所に間違えて空欄を検出した不正解とする．例としては，先行手法の城ページ番号 1 の重要項目「交通」で×となっている．しかし，城ページ番号 1 のページでは交通に関連する地名の「京都」が表記されていた．このように重要項目に関する単語が城ページに表記されているが，空欄と検出した場合は不正解とする．

F 値を評価した結果，先行手法の F 値は 0.77 となり，提案手法「Wikipedia 全ページでクラスタリング」の F 値は 0.75 となり，提案手法「類似度」の F 値は 0.84 となり，提案手法「分類語彙表」の F 値は 0.81 となった．このように，提案手法「類似度」，「分類語彙表」の方が先行手法と提案手法「Wikipedia 全ページでクラスタリング」より精度が高い結果になった．提案手法「Wikipedia 全ページでクラスタリング」が低かった原因としては，重要項目「交通」の単語をあまり抽出できなかったことで適合

率が低くなってしまったことが挙げられる．1つのクラスタを重要項目としているため，移動手段の方法や地名といった単語は別のクラスタに属する単語を抽出することが出来ず，単語の網羅性が低かったことが原因だと考えられる．

6.3.5 評価実験のまとめ

Akano ら [2] の研究の改良を行い，単語クラスタリングに利用するデータを増やすことによって，１つのクラスタに属する単語数は増加した．また，情報抽出における精度も向上した．よって，単語クラスタリングに利用するデータを増やしたほうが精度が上がると考える．

単語抽出における正解率から，抽出単語の総数が多いほど，表抽出における正解率の精度は高くなる傾向にある．重要項目に属する単語数を増やすことによって精度の向上が見込める．しかし，重要項目に属する単語数が多くなると，重要項目と関係がない単語が表に検出され，抽出単語における正解率は下がるのではないかと考える．

F 値における記載不足の指摘の評価実験を行った結果，抽出単語数が少ないほど再現率が高く，抽出単語が多いほど適合率が高くなる結果になった．F 値の結果は抽出単語数が多いと高くなる傾向にあることがわかった．

第7章　おわりに

長文から知りたい情報を取得するためには長い文を読む必要がある．よって，情報取得に時間がかかる．長文を表に示すことで長い文を読む必要がなく情報を容易に取得することができる．そこで，本研究では単語クラスタリングの改良や分類語彙表を用いて表生成を行い，重要項目の選定を行った．

情報抽出の評価実験は表抽出における正解率と単語抽出における正解率の 2 つで評価を行った．

表抽出における正解率は表に 1 つでも正しく情報を抽出したものを正解とした．また，空欄を正しく空欄と検出できれば正解とした．表抽出における正解率を評価した結果，先行手法の表抽出における正解箇所の割合は 0.68 となり，提案手法「Wikipedia 全ページでクラスタリング」の表抽出における正解箇所の割合は 0.71 となり，提案手法「類似度」の表抽出における正解箇所の割合は 0.88 となり，提案手法「分類語彙表」の表抽出における正解箇所の割合は 0.81 となった．このように，先行手法より提案手法の方が精度が高い結果になった．また,「Wikipedia 全ページでクラスタリング」と「分類語彙表」よりも「類似度」の結果の方が精度が高い結果になった．

また，単語抽出における正解率を評価した結果，先行手法の単語抽出における正解箇所の割合は 0.73 となり，提案手法「Wikipedia 全ページでクラスタリング」の単語抽出における正解箇所の割合は 0.89 となり，提案手法「類似度」の単語抽出における正解箇所の割合は 0.82 となり，提案手法「分類語彙表」の単語抽出における正解箇所の割合は 0.82 となった．このように，先行手法より提案手法の方が精度が高い結果になった．また,「Wikipedia 全ページでクラスタリング」の方が「類似度」と「分類語彙表」より精度が高かった．

また記載不足の指摘の評価実験は F 値を用いて空欄指摘の調査を行った．F 値を評価した結果，先行手法の F 値は 0.77 となり，提案手法「Wikipedia 全ページでクラスタリング」の F 値は 0.75 となり，提案手法「類似度」の F 値は 0.84 となり，提案手法「分類語彙表」の F 値は 0.81 と提案手法「類似度」となった．このように,「分類語彙表」の方が先行手法と提案手法「Wikipedia 全ページでクラスタリング」より精度が

高い結果になった.

情報抽出と記載不足の指摘の 2 点で評価を行った結果，単語クラスタリングに利用するデータを増やすことによって，１つのクラスタに属する単語数は増加した．また，情報抽出における精度も向上した．よって，単語クラスタリングに利用するデータは増やしたほうが精度が上がると考える．単語抽出における正解率から，抽出単語の総数が多いほど，表抽出における正解率の精度は高くなる傾向にある．重要項目に属する単語数を増やすことによって精度の向上が見込める．しかし，単語抽出における正解率結果から，重要項目に属する単語数が多くなると，重要項目と関係にない単語が表に検出され，抽出単語における正解率は下がるという問題点がある．また，F 値における記載不足の指摘の評価実験を行った結果，F 値の結果は抽出単語数が多いと高くなる傾向にあることがわかった.

謝辞

本研究を進めるにあたり，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました，鳥取大学情報エレクトロニクス自然言語処理研究室の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授に心から御礼申し上げます．その他様々な場面で御助言を頂きました自然言語処理研究室の皆様方に感謝の意を表します．

付 録A　Wikipedia以外の実験

本研究では実験データをWikipediaの「城」ページを利用して情報抽出を行った．しかし，Wikipediaのページ以外の実験はしていない．そこで本付録では，Wikipediaのページ以外の「小説データ」を使って実験を行う．

A.1　先行研究

馬場ら [8] は「人名抽出」と「特徴表現の抽出」の観点で人物抽出を行っていた．

「人名抽出」は人名辞書を利用し，辞書に載っている人名を小説テキストから抽出する．また，人名辞書に載っていないものは形態素解析を利用して抽出する．人名とする条件は人名として抽出された語のテキスト全体における出現回数をf，小説テキストに含まれる形式段落数をLとした場合，f/Lが閾値よりも大きい場合にその語を人名と選定していた．

また,「特徴表現の抽出」は人手で作成した特徴「性別」「年齢」「年代」「職業」「身体的特徴」「性格」を重要項目として評価を行っていた.「性別」は性別がわかる語（男性，母，叔父など）や性別固有の一人称（俺，わしなど）が含まれていれば正解としていた.「年齢」は「17歳」や「三十五才」といった表記がなされていれば正解としていた.「年代」は人間の一生を「乳幼児期」「少年期」「青年期」「中年期」「老年期」の特徴ができていれば正解としていた.「職業」は「世界樹の下」を参考に「剣士・騎士・戦士」といった職業リストとその特徴語が一致した場合正解としていた.「身体的特徴」は髪や瞳の色，声，体格など，容姿に関する特徴を抽出できれば正解としていた．

本研究では「特徴表現の抽出」の観点で先行研究と比較実験を行う．

A.2　実験環境

本研究では小説「怪人二十面相」を利用する．また，提案手法は「類似度」と「分類語彙表」で実験を行う．

提案手法「類似度」は 5.3 節の実験環境を使用する．ただし，5.3 節では Wikpedia 全ページを利用して類似度算出を行っていたが，本実験では 8,227 タイトルの小説データを利用して類似度の算出する．

提案手法「分類語彙表」は 5.4 節の実験環境を使用する．

A.3　評価方法

先行研究 [8] は重要項目「性別」「年齢」「年代」「職業」「身体的特徴」「性格」を「再現率」と「精度」で評価していた.「再現率」はあらかじめ人手で作った特徴表の単語が抽出できた割合を示し,「精度」は抽出された単語の正解率で示していた．本研究では先行研究と同様の評価方法で行う．

提案手法の類似度は「性別」「年齢」「職業」「身体的特徴」「性格」を重要項目と決定し，それぞれ 5 つの単語との類似度が高い単語を重要項目の単語群とし，5 つの重要項目の評価を行う．ただし，先行研究の「年齢」は「17 歳」や「三十五歳」といった歳の年齢を正解としていた．また，提案手法の「年齢」は「少年」や「老人」といった「乳幼児期」「少年期」「青年期」「中年期」「老年期」の特徴を示す単語が抽出された．これらは先行研究の重要項目「年代」に相当する単語である．よって，先行研究の「年代」と提案手法の「年齢」を対応づけて評価する．

提案手法の分類語彙表は分類項目名「男女」に属する単語を重要項目「性別」と定義し，分類項目名「老少」「夫婦」「親・先祖」「子・子…」「兄弟」「親戚」に属する単語を重要項目「年齢」と定義し，分類項目名「社会階…」「人物」「成員」「専門的」「支配的」「販売な」「運輸」「生産工…」「保安サ…」「サービ…」「反社会…」「軍人」「長」「相対的」「臨時的」に属する単語を重要項目「職業」と定義し，分類項目名「動物」「衣服」「雨着」「下着」「袖・衿…」「帽子」「ネクタイ」「はき物」「鏡・レンズ」に属する単語を重要項目「職業」と定義し，分類項目名「安心・…」「対人感…」「表情（…」「声」「自我・…」「自信・…」「欲望・…」に属する単語を重要項目「職業」と定義し評価する．提案手法の類似度と同様に先行研究の「年代」と提案手法の「年齢」を対応づけて評価する．

また，先行研究の実験結果 [8] を利用する．ただし，本実験は「怪人二十面相」の小説データで評価実験を行い，先行研究の評価する小説データとは違うデータで比較する．

A.4　表生成

提案手法は「怪人二十面相」の章タイトル 31 個からランダムに 5 個の章タイトルと章タイトルの文章を利用して表生成を行う．また，あらかじめ「怪人二十面相」を読み，重要項目「性別」「年齢」「年代」「職業」「身体的特徴」「性格」に対応する単語が,「怪人二十面相」に記載されていた場合，その単語は重要項目の単語リストとする．提案手法「類似度」と「分類語彙表」の表生成に抽出できた単語が単語リストの単語の場合,（○）をつける．あらかじめ人手で作成した重要項目「性別」「年齢」「職業」「身体的特徴」「性格」の単語リストを表 A.1 から表 A.5 に示す．また，表生成結果を表 A.6 と表 A.7 に示す．

表 A.1: 性別の単語リスト

無頼漢, 女, 少年, 男, 彼, おじいさん主人, おかみさん, 野郎, 兄貴, 夫人, 紳士, •••

表 A.2: 年齢の単語リスト

少年, 若者, 老人, 学生, 子ども, おじいちゃん, 坊ちゃん, 小僧, •••

表 A.3: 職業の単語リスト

富豪, 盗賊, 探偵, 怪人, 怪盗, 学者, 警察, 秘書, 悪人, 学生, 泥棒, コック, 事務 ホテル, 魔法使い, ルンペン, 先生, 博士, 係長, 警官, 館長, 影武者 首領, 警部, 名人, 総監, •••

表 A.4: 身体的特徴の単語リスト

かわいらしい, 着物, 白い顔, ピストル, しわ, しらが, きたない, さるぐつわ, 髪の毛 鳥打帽, シャツ, 背広, 帽子, 美しい, きたならわしい, 白髪, 笑顔, かつら つけひげ, 若々しい, …

表 A.5: 性格の単語リスト

おそろしい, ぐずぐず, あらくれもの, 考えぶかい, 毒口, ひもじい, ずうずうしい, …

A.5　再現率と精度の評価

重要項目「性別」「年齢」「年代」「職業」「身体的特徴」「性格」の 5 つの「再現率」と「精度」を求めた．評価した結果を表 A.8 と表 A.9 に示す．

「再現率」では，先行研究の手法より提案手法「類似度」と「分類語彙表」の結果の方が 5 つの重要項目のうち 4 つの項目で良いという結果になった．

また,「精度」では，先行研究の手法より提案手法「分類語彙表」の結果の方が重要項目５つのすべての項目で良い結果になった．しかし，提案手法「類似度」の結果は先行研究の結果よりすべて低い結果になった．提案手法「類似度」が低くかった原因は，単語リストの単語と関係のない単語を抽出したことが考えられる．

表 A.6: 類似度による表生成

	性別	年齢	職業	身体的特徴	性格
はしがき	無頼漢（〇）, 闘争, 自身, リス, 日本, 美し, 活動, 以上, 女（〇）, 美しく, 一つ	少年（〇）, 自身, 老人（〇）, 天才, 若者（〇）, 分の, 以上	自分, 新聞, 富豪（〇）, 品物, 日本, 警察（〇）, 無頼, 心がけ, 美術, 闘争, 自身, 天才, 以上, 活動, 場所, 興味, 不公平, 公平, 学者（〇）	少し, 自分	自分, 闘争, 自身, 天才, 美し, 活動, 一つ, 興味, 公平
仏像の奇跡	男（〇）, 三つ, ジン, 学生, 人間, 円満, 二つ, 無視, 多い, ピッタリ, 方向, 時代, 問題	三つ, メートル, 学生（〇）, 人間, 円満, 時代, 問題, 老人（〇）, 十分, 大き, 子ども（〇）, 読者, 小さ, 多い, 柔和, 想像, 親子	学生（〇）, 品物, 人間, 機械, 仕事, 無視, 警察（〇）, 虐待, 時代, 問題, 考え, 美術, 十分, 読者, 仲間, 多い, 方向	ドア, 片手, 機械, 両手, ガラス, 小さく, 荷物, ピストル（〇）, 右手, エンジン, グーッ, 今にも, グルッ, 着物（〇）, 手箱, 部屋, ピッタリ, 少し, ニュッ, 背中, 動く, トントン, 思わず	人間, 読者, 無視, 方向, 時代, 問題, 考え, 想像
奇妙な取りひき	心づか, 身分, 三つ, 自身	身分, 十分, 三つ, 自身, 大き, 時分, 子ども（〇）, 小さ, 少年（〇）	身分, 十分, 事務（〇）, 警察（〇）, 自身, 考え, 応じ	少し, 片腕, 鉄棒, 部屋, 天井, ピストル（〇）	自身, 考え, 信じ
二十面相の新弟子	男（〇）, 自身, 若く, 位置, 習慣, 美し, シャ, 一つ, 人物, 人がら, 美しい	低い, 鳥打ち帽, 自身, 苦労, 若く, 友人, 習慣, 質素, 考える, 少年（〇）, 時間, 人物, 人がら, 想像, 分の	商売（〇）, 自身, 仕事, 友人, 位置, 習慣, ルンペン（〇）, 考える, 浮浪（〇）, 考え, 人物, 非難, 住宅	髪の毛（〇）, 地面, ドア, ヤッ, 動き, 身ぶるい, 電柱, 身動き, ブルブル（〇）, 着物（〇）, ヨロヨロ, ツブ, 起き, シャツ（〇）, グッ, 帽子（〇）, 起きあがっ, 片足, 少し, 組みつい, 動く, ソッ	位置, 事件, 動き, 一つ, 考え, 信じ, 自身, 策略, 習慣, 美し, 考える, 身辺, 想像, 人がら, 非難, 人物
怪盗の捕縛	自由, 学生, 人々, 家庭, 分ち, 順序, 寸分, 観念, 男（〇）, 若々しい, 自身, 完全, 紳士（〇）, 場面, 美し, 事情, 服装, 文学, 場合, 人物, 美しい	, メートル, 学生（〇）, 家庭, 志願, 寸分, 白髪, 小学, 親密, 若々し, 老人（〇）, 分の, 十分, 若々しい, 正銘, 自身, 大き, 苦労, 子ども（〇）, 読者, 紳士, 事情, 小学生（〇）, 服装, 少年（〇）, 場合, 人物, 想像	自由, 自分, 手段, 所有, 世間, 機会, 美術, 成功, 十分, 自身, 完全, 紳士, 事務（〇）, 文学, 場合, 新聞, 学生（〇）, 品物, 仲間入り, 志願, 家庭, 人々, 仕事, 事業, 順序, 観念, 小学, 警察（〇）, 説明, 考え, 信用, 読者, 仲間, 浮浪（〇）, 事情, 目的, 服装, 人物	ハッ, 髪の毛（〇）,, 自分, 麻酔, 手首, 左手, 両手, 動き, 視線, 右手, 身動き, 今にも, 気持, 元気, 部屋, 少し, 落ち, 満身, 苦し, 思わず	, 自由, 自分, 人々, 家庭, 手段, 動き, 所有, 観念, 世間, 親密, 若々し, 説明, 考え, 気持, 自身, 完全, 皮肉, 読者, 場面, 美し, 事情, 文学, 目的, 場合, 人物, 想像

表 A.7: 分類語彙表による表生成

	年齢	職業	身体的特徴	性格	性別
はしがき	少年（○）, 老人（○）, 若者（○）	無頼漢, 富豪（○）, 天才, 盗賊（○）, 助手, 探偵（○）, 怪人（○）, 怪盗（○）, 学者（○）	殺し, 面相	悪い, 見え	女（○）
仏像の奇跡	坊ちゃん（○）, 老人（○）	秘書（○）, 悪人（○）, 学生（○）, 太郎, 盗人（○）, 泥棒（○）, 読者, 手下, 名匠, 部下, 工夫	十徳, 着物（○）, 生き, 両手, 円満, 面相, 右手, 甲冑, 背中, 走り	待ち, 笑い, 心配, 見え	男（○）
奇妙な取りひき	少年（○）, 主人	探偵（○）, コック（○）	面相	安心, 心配, 平気, 見え	
二十面相の新弟子	夫人, 兄貴, 少年（○）	凶賊, 親分, 魔法使い（○）, ルンペン（○）, 手下, 先生（○）, 弟子, 探偵（○）, 助手, 部下, 使い	髪の毛（○）, 着物（○）, 鳥打ち帽（○）, シャツ（○）, 背広（○）, 帽子（○）, 面相, 美しい（○）	無精, 苦労, 見え, 悪い, 所望, 心配, 笑い, 憤慨	男（○）, 野郎（○）
怪盗の捕縛	少年（○）, 老人（○）	博士（○）, 係長（○）, 学生（○）, 警官（○）, 館長（○）, 部下, 怪盗（○）, 団長, 影武者（○）, 使い, 悪人（○）, 館員, 首領（○）, 読者, 武者, 先生（○）, 弟子, 小僧, 探偵（○）, 小学生（○）, 警部（○）, 名人（○）, 総監（○）	髪の毛（○）, 手首, 両手, 左手, 白髪（○）, 右手, 元気, 完全, 白髯（○）, 不自由, 背広（○）, 面相, 服装, 走り, 満身, 美しい	, 待つ, 志願, 笑顔, 親密, 感謝, 安心, 信用, 待ち, 苦労, 見え, 親切（○）, 尽力, 神妙, 心配, 笑い	男（○）, 紳士（○）

表 A.8: 単語抽出における再現率

手法	先行手法	分類語彙表	類似度
性別	0.62(13/21)	0.35 (6/17)	0.35 (6/17)
年齢	0.25(5/20)	0.50 (9/18)	0.83 (15/18)
職業	0.25(8/32)	0.68 (32/47)	0.32 (15/47)
身体的特徴	0.07(6/91)	0.38 (12/32)	0.28 (9/32)
性格	0.03(4/117)	0.13 (1/8)	0.13 (1/8)

表 A.9: 単語抽出における適合率

手法	先行手法	分類語彙表	類似度
性別	0.62(13/21)	1.0 (6/6)	0.09 (6/68)
年齢	0.63(5/8)	0.75 (9/12)	0.20 (15/75)
職業	0.30(9/30)	0.57 (32/56)	0.16 (15/94)
身体的特徴	0.12(22/183)	0.32(12/37)	0.12 (9/73)
性格	0.03(4/117)	0.03 (1/33)	0.01 (1/62)

参考文献

[1] 藤原隆太. Wikipedia からの城情報の取り出しと文章作成支援. 卒業論文, 鳥取大学知能情報工学科, 2015.

[2] Hokuto Akano, Masaki Murata, and Qing Ma. Detection of inadequate descriptions in Wikipedia using information extraction based on word clustering. In *IFSA-SCIS*, pp. 1–6. 2017.

[3] 西尾泰和. word2vec による自然言語処理. 株式会社オライリー・ジャパン, 2014.

[4] 宮崎亮輔, 小町守, 疋田敏朗, 柏倉俊樹. Wikipedia を用いた遠距離教師あり学習による専門用語抽出. 言語処理学会 第 21 回年次大会 発表論文集, pp. 87–90, 2015.

[5] 近藤明日子, 田中牧郎. 分類語彙表・unidic 見出し対応表の構築ーコーパスへの網羅的・系統的な語義情報付与を目指してー. 言語処理学会, pp. 90–93, 2017.

[6] 岡田拓真, 村田真樹, 徳久雅人, 馬青. 論文からの記載必要項目の抽出と文章作成支援. 言語処理学会第 21 回年次大会, pp. 988–991, 2015.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[8] 馬場こづえ, 藤井敦, 石川徹也. 小説テキストを対象としたジャンル推定と人物抽出. 言語処理学会 第 11 回 年次大会予稿集, 2005.

研究成果リスト

Hokuto Akano, Masaki Murata, and Qing Ma. Detection of inadequate descriptions in Wikipedia using information extraction based on word clustering. In IFSA-SCIS, pp. 1—6. 2017.