概要

文章作成の際に重要情報を書き漏らす場合がある．書き漏れがあると不明瞭な文になる場合や読者が知りたい情報が書かれておらず情報取得において不便になる場合がある．重要項目を文章から抽出し表にまとめ，文章に重要項目が書いていない場合に書き漏れを指摘することで，文章作成を支援することが考えられる．この技術に関する研究はいくつかある．藤原ら [1] の研究では上位下位知識を用いて Wikipedia の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，先行研究では抽出された重要項目の種類が少なかった.

そこで本研究では先行研究の改善を目的として Wikipedia からの情報抽出における重要項目の取り出し技術の改良に焦点をあてて研究を行う．具体的には重要項目の種類を増やす目的で研究を行う.

実験の結果，情報抽出の実験においては，word2vec [2] を用いたクラスタリングを利用した実験では正解率は 0.82 で，先行手法の上位下位知識を利用した実験では正解率は 0.72 と提案手法の方が精度が良かった．また文章作成支援においても，提案手法の F 値は 0.92 で先行手法での F 値は 0.85 と提案手法の方が精度が高かった．また，重要項目も先行研究では 4 個しかなかったが，提案手法では 20 個に重要項目を増やすことができた.

目 次

表 目 次

図 目 次

第1章　はじめに

文章作成の際に重要情報を書き漏らす場合がある．書き漏れがあると不明瞭な文になる場合や読者が知りたい情報が書かれておらず情報取得において不便になる場合がある．重要項目を文章から抽出し表にまとめ，文章に重要項目が書いていない場合に書き漏れを指摘することで，文章作成を支援することが考えられる．本研究では，重要と思われる項目のことを重要項目と呼ぶ.

本研究はWikipedia内から情報抽出を行い，抽出した情報から多くの記事で共通している項目を重要情報として表にまとめることを目的とする．また抽出した文章に欠落部分があれば空白として指摘を行い，ユーザーに追加記載を促すことで，文章作成支援の役に立つ．また，情報抽出と文章作成支援の例を表1.1に示す.

表1.1の例はWikipediaの城ページから「地名」,「人名」,「組織名」に属する単語を抽出したものである．この単語の抽出処理が，本研究で言う情報抽出に相当する．また，表の「地名」,「人名」,「組織名」が重要項目に相当する．表1.1の空白部分はWikipediaの城ページに「地名」,「人名」,「組織名」に属する単語が記載されていないことを示している．記載がないことを空白で記載することでユーザーに記載すべきものが欠落していることを指摘を行うことができる．このように空白を利用して欠落の指摘をすることで文章の作成を支援することが，本研究で言う文章作成支援に相当する.

表 1.1: 情報抽出と文章作成支援の例

城名	地名	人名	組織名
大坂城	大阪	豊臣	
二条城		徳川家康	織田軍
仙台城	仙台	伊達政宗	
熊本城		出田	飯田
岐阜城	岐阜		

本研究の主張点を以下に示す.

- 情報抽出
  - 情報抽出では先行研究の手法の上位下位知識と提案手法であるクラスタリングで行う.
  - 抽出した重要情報を表の形に可視化する.
  - 上位下位知識に基づく手法では正解率は 0.72 であるのに対して，クラスタリングを用いた手法では 0.82 となった.
  - 重要項目は先行研究が 4 個であるのに対して提案手法では 20 個に増やすことができた.
- 文章作成支援
  - 記載されていない項目を空白で示すことで追加記載を促すことができる.
  - 文章作成支援の性能は上位下位知識に基づく手法ではＦ値は 0.85 に対して，クラスタリングに基づく手法では 0.92 のＦ値であった.

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 章では情報抽出の手法と文章作成支援の手法を提案する．第 4 章では実験環境の説明を行う．第 5 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 6 章では今後の課題を述べる．第 7 章では本稿をまとめる.

第2章　関連研究

藤原ら [1] は情報抽出と文章作成支援の観点で研究を行っていた．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．また重要情報の抽出で作成する表の空欄箇所を情報が欠けている項目と判定し，そのことをユーザーに知らせ記載の追加を促すことで文章作成支援をした．

岡田ら [3] 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文章作成支援を行っていた．

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．宮崎らは Wikipedia を遠距離教師あり学習で情報抽出を行っていたが，本研究ではクラスタリングで情報抽出を行い，文章作成支援も行うという新規性がある．

村田ら [5] の研究 では，論文内から YamCha と教師あり機械学習を用いて「精度表現」「主要な 分野」「言語名」「組織人名」の取り出しを行った．

第3章　提案手法

本章では，本研究の手法を説明する．

3.1　情報抽出

本研究では word2vec を用いて重要項目の取り出し技術の改良を行う．重要項目の選定方法としては word2vec 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う．単語のクラスタリングは類似度の高い単語をまとめて単語のクラスタを作るものである．各クラスタにはクラスタ番号を割り当てる．重要項目の選定を行い，表にまとめる方法を以下に示す．

1. 抽出したい事柄を決定し，Wikipedia から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．(例えば，1 のクラスタ番号のクラスタには地名の単語群が属し，2 のクラスタ番号のクラスタには人名の単語群が属する．例を表 3.1 と表 3.2 に示す．)

表 3.1: 1 を地名とした単語群

地名
京都
大阪
宮城

表 3.2: 2 を人名とした単語群

人名
伊達政宗
徳川家康
豊臣秀吉

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その箇所に埋める．

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．選ばれなかった列を削除して表を作る．このようにして作成する表の例を表 3.3 に示す．

表 3.3: 表にまとめたもの

城名	地名	人名
大阪城	大阪	豊臣秀吉
二条城	京都	徳川家康
仙台城	宮城	伊達政宗

3.2 文章作成支援

作成する表の空欄箇所を情報が欠けている項目と判定し，そのことをユーザーに知らせ記載の追加を促すことで文章作成支援をする．文章作成支援が必要な表の例を表 3.4 に示す．この表について「二条城」の「地名」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ユーザーに知らせることで Wikipedia での文章作成の支援に役に立つ．

表 3.4: 文章作成支援が必要な表例

城名	地名	人名	組織名
大坂城	大阪	豊臣	
二条城		徳川家康	織田軍
仙台城	仙台	伊達政宗	
熊本城		出田	飯田
岐阜城	岐阜		

第4章　実験環境

4.1　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4.1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 4.1: Wikipediaの記事の例

4.2 mecab

本研究では word2vec のクラスタリングを使用する．word2vec の入力データでは，記事の文章の単語の境目に空白を入れる必要がある．そこで単語ごとに空白をいれるために「mecab」の分かち書きを使用する．以下の図 4.2 が分かち書き前のものであり，図 4.3 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 4.2: 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 4.3: Wikipedia の記事に mecab を使用した結果の例

4.3 クラスタリング

本研究は word2vec 内のツールであるクラスタリングを使用する．

まず，word2vec は単語をベクトル変換するものである．作者の Mikolov ら [2] 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 [6]．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．このように入力された文章から似たような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．

Wikipedia の「大学」に関するデータ (2014 年 11 月) を入力として，1,000 個のクラスタにクラスタリングした結果の一部 (3 つのクラスタ) を例として表 4.1，表 4.2，表 4.3 に示す．ここで言う，Wikipedia の「大学」に関するデータは，タイトルが「大学」を含む Wikipedia のページのことである．

表 4.1 は芸術大学という点でで同じような単語が集まっている．表 4.2 は短期大学という点で同じ単語が集まっている．表 4.3 は点数関係が集まっている．

表 4.1: クラスタリングの抽出例 1

愛知県立芸術大学
沖縄県立芸術大学
京都市立芸術大学
女子美術大学
多摩美術大学
東京芸術大学
東京造形大学
武蔵野音楽大学
武蔵野美術大学
.....

表 4.2: クラスタリングの抽出例 2

宮城県農業短期大学
京都経済短期大学
京都市立看護短期大学
京都文化短期大学
京都文教短期大学
共栄学園短期大学
九州造形短期大学
九州大谷短期大学
駒沢女子短期大学
.....

4.4 類似度

本節では word2vec 内にあるツールの類似度の説明を行う.

類似度は人手で入力した単語に似た単語を任意で選んだ上位の単語数を出力されるものである. 例えば, 単語を「大学」と入力すると表 4.4 のように「大学」に似た単語とその類似度が表示される.

表 4.3: クラスタリングの抽出例 3

スコア
テスト
最低
習熟
上回り
値
適性
点数
到達
倍率
平均
偏差
満点
.....

4.5 上位下位知識

先行手法では上位下位関係の抽出に ALAGIN の上位下位関係抽出ツールを用いる．上位下位関係抽出ツールは，Wikipedia から上位下位関係となる用語ペアを数百万対のオーダーで抽出できるツールである．上位下位関係とは，”X は Y の一種 (一つ) である”と言える X と Y の関係を言う． X のことを下位語，Y のことを上位語と呼ぶ．上位下位関係の抽出例を表 4.5 に示す．Wikipedia 全体を入力とした場合上位語の種類は 43,987 単語，下位語の種類は 422,223 単語であった．

表 4.4: 類似度の例

法学部 0.752459
科 0.742712
卒業 0.718339
学科 0.690399
工学部 0.674038
文科 0.650838
専任 0.636888
講師 0.629301
農学 0.618454
学部 0.609659
卒 0.607958
入学 0.606177
助教授 0.595031
教授 0.593078
千葉大学 0.589457
大学院 0.588689
文学部 0.582074
.....

表 4.5: 上位下位関係の抽出例

上位語	下位語
仏像	七面大明神像
楽器	カンテレ
文房具	スティックのり
神楽団体	川平神楽社中
プログラミング言語	prolog
戦争映画	ハワイ・ミッドウェイ大海空戦
AOC ワイン	ラ・グランド・リュー　ブルゴーニュ
ゲーム	ファイナルファンタジー XI
研究所	情報通信研究機構

第5章　実験

5.1　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．

5.2　評価方法

5.2.1　情報抽出による評価実験

先行手法である上位下位知識と提案手法であるクラスタリングで正解率を求めるために，4.1 節の 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．

上位下位知識の実験では，4 つの上位語の「県名」,「時代」,「地名」,「元号」を重要項目とする．重要項目として決定した 4 つの上位語の下位語が城ページに検出されれば城ページの行の表にそれを出力する.「県名」の項目はその城が存在する県名が抽出された場合正解,「時代」の項目では築城されてから廃城するまでの時代のいずれかが抽出された場合正解,「地名」の項目では城の所在地が抽出された場合正解,「元号」の項目では築城されてから廃城するまでの元号のいずれかが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．出現した全ての重要情報をまとめた表では，1 つでも正解が抽出された場合正解とする．

クラスタリングの実験では，クラスタリングを行った結果において頻度計算から人手で重要項目を決定して，その中から人手で選んだクラスタ 3 つの「クラスタ 401」,「クラスタ 407」,「クラスタ 765」を使って評価実験を行う．重要項目として決定した 3 つのクラスタ結果が城ページに検出されれば城ページの行の表に出力する.「クラスタ 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,「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,「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．

5.2.2 文章作成支援による評価実験

文章作成支援の実験において，上位下位知識とクラスタリングの結果の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．

$$F = (\frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}) \tag{5.1}$$

$$\text{適合率} = \frac{\text{空欄かつ Wikipedia 内に正解がないもの}}{\text{空欄のもの}} \tag{5.2}$$

$$\text{再現率} = \frac{\text{空欄かつ Wikipedia 内に正解がないもの}}{\text{Wikipedia 内に正解がないもの}} \tag{5.3}$$

本研究において，適合率はシステムにより空欄になったものの中に，正しく空欄と検出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5.2，5.3 において「空欄のもの」というのは重要情報の抽出実験で作成した表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」というのは，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．

5.3 実験結果

5.3.1 クラスタリングを用いた情報抽出の結果

Wikipedia に関する城ページにおいて mecab で分かち書きを行ったものを入力として，1,000 個のクラスタを作るクラスタリングを行った．クラスタリング結果の一部を表 5.1 に示す．表 5.1 の左の数字はクラスタ番号を示しており，右の単語はクラスタ番号に属する単語である．半数以上の城ページでクラスタ内の単語が検出されたクラスタを重要項目の候補とし，そこから人手で重要項目を選んだ．

表 5.1: クラスタリング結果

435 筑前
435 長門
435 那珂
435 能登
435 両国
436 一存
436 家臣
436 虎丸
436 高屋
436 十河
437 一向
437 一乗寺城
437 越中
437 加賀
437 吉江
437 拠る
.....

人手で選んだ重要項目を表 5.2 に示す．重要項目のクラスタ内の単語の一部を表 5.3 に示す．

先行研究では重要項目が 4 個であったのに対して，提案手法は重要項目を 20 個に増やすことができた．

表 5.2: 重要項目名

戦い状況，城の造り，交通路，堀の種類
塀の種類，策略，天守，敗戦
豪族，血筋，天皇，改築
防壁，寺，神社，物流
統治，藩，砦，権力

表 5.3: 重要項目のクラスタ内の単語例

クラスタ名	クラスタ内の単語 1	クラスタ内の単語 2	クラスタ内の単語 3	クラスタ内の単語 4	クラスタ内の単語 5
戦い状況	攻め寄せ	攻め落とす	惨敗	銃撃	...
城の造り	東丸	東大手	内門	東門	...
交通関係	中山道	東海道	要所	要衝	...
堀の種類	水堀	内堀	堀割	用水路	...
塀の種類	土塀	矢倉	櫓門	塀	...
策略	内通	破棄	奮闘	捕縛	...
天守	天守	天守	天守閣	復興	...
敗戦	逃れ	逃れる	逃亡	味方	...
豪族	関白	御家人	改姓	郡司	...
血筋	分家	本家	末裔	本領	...
天皇	皇后	皇子	皇太子	皇族	...
改築	修築	増設	増築	築造	...
防壁	横堀	横矢	空堀	堅	...
寺	釈迦	歌碑	供養	句碑	...
神社	祈願	貴船	祇園	宮司	...
物流	基地	拠点	水運	物流	...
統治	天領	統括	領地	領有	...
藩	藩士	藩主	兵学	攘夷	...
砦	難攻不落	要害			
権力	逝去	遷都	即位	退位	...

次に情報抽出の性能を調べる．20 個のクラスタのうちから選んだ表 5.5，表 5.6，表 5.7 の 3 つのクラスタを評価に利用した．このクラスタ内の単語が各欄に正しいものが１つでも得られて出力されれば正しく抽出したとする．

情報抽出に基づき表を作成した結果を表 5.8，表 5.9 に示す．表 5.8，表 5.9 において太字で表記されているものは正解と判断したものである．また，○と表記されているものは Wikipedia 内に正解の記載が無く，空欄が正しく抽出されたと判断したものである．

3 つのクラスタで１つでも正しく抽出された正解率は 0.82 となった．

また，表に抽出された単語の正解率も求めた．例えばクラスタ 407 である「門」が抽出されたとする．この場合 wikipedia ページ内に「門」と記述されていれば正解とす

表 5.4: 重要項目の個数

手法	重要項目の個数
上位下位知識 (先行手法)	4
クラスタリング (提案手法)	20

表 5.5: クラスタ 401

おびき出し，ひい，引き返し，援軍 ，炎上 ，加わっ ，壊滅 ，開城
勧告，陥落，頑強，奇襲，喫し，救援，窮地，屈服，迎え撃つ ，向かわ
抗戦，攻める，攻め寄せ，攻め落とす，惨敗，持ちこたえ，銃撃，出撃
出陣，少数，焼か，焼き討ち，焼き払い，焼き払わ，申し入れ，進軍
占拠，全滅，阻止，総崩れ，遭い，態勢，退け，退却，大敗，着陣，駐留
直ぐ，抵抗，撤退，転戦，逃走，派兵，破っ，敗戦，敗走，敗退，敗北
迫り，不完全，伏兵，奮戦，兵糧，放火，防戦，本隊，明け渡し，戻り
夜襲，落ち延びる，落城，篭城

るが,「五右衛門」の中の「門」だけが抽出された場合は不正解としている．また，クラスタリングを行った段階でクラスタ内に関係のない単語が抽出されその関係のない単語が表に抽出された場合は不正解としている．例えば，表 5.6 の中の「医」の単語が城ページで出力されたとする．この場合クラスタ 407 は城の造りに関する単語が集まっている．だが,「医」という単語は城の造りに関係がないことは明白である.「医」のようにクラスタと関係のない単語が出力された場合は不正解とする．単語の正解率を求めた結果，単語の正解率は 0.71 となった．

表 5.6: クラスタ 407

）]]，くろがね，移築，医，一ノ門，円城寺，外丸，外門，官衙，歓会
丸，丸内，祈念，亀甲，喰違，御殿，御門，高麗，三ノ丸，山里，仕切
鐘，政庁，正門，西丸，西大手，西門，前門，総門，大手門，大門，中仕切
中門，長屋門，追廻，追手，天球，土蔵，東丸，東大手，東門，撞堂，内門
二の丸，二ノ，二之，日出，納屋，番所，表門，北御門，北門，本丸，門
門跡，門扉，役所，薬，裏門，蓮池

表 5.7: クラスタ 765

ほど近い，ロマンティック，伊勢湾，碓氷，越え，越える，奥大道，往還，押さえ 押さえる，海道，街道，幹線，関所，経路，繋がる，結ぶ，古道，交差，交通 国境，作手，参宮，参詣，山陰，山陽，水上，水陸，瀬戸内，生野，中山道，中道 通ずる，東海道，峠，分岐，並行，便，便利，北国，北陸，要所，要衝，要地 抑える，霊場，連絡

5.3.2 類似度を用いた情報抽出の結果

本節はクラスタリングを行った結果から「城」と類似度の高い単語以外を削除して情報抽出を行った．各クラスタには関係のない単語が混ざっていることが多く,「城」と類似度が高い単語以外を削除すると，関係のない単語を削除できるのではないかと考えて行ったものである．例として表 5.11 のクラスタにおいて「城」と類似度の高い単語以外を削除した結果を表 5.12 に示す.「城」と関係のない単語が削除された．また,クラスタ 407 では「ニノ」や「薬」などクラスタと関係のない単語を削除することができた．しかし,「救援」や「惨敗」などの重要情報である単語が多く削除されているクラスタが多かった．このため，本研究では類似度に基づく単語の削除は行わないことにした．

表 5.8: クラスタリングを使った情報抽出

城名	クラスタ 401(戦い状況)	クラスタ 407(城の造り)	クラスタ 765(交通関係)
宇和島城		**門, 大手門**, 山里, **三ノ丸, 追手**, **移築**, 二ノ, 丸, **本丸**	**交通**
筑後十五城	**抵抗, 大敗, 篭城, 頑強**, 少数	門	○
岡崎城	○	門, **二の丸, 大手門**, **三ノ丸, 北門, 移築, 本丸**, 丸	海道, **東海道, 交通**
松尾城	○	○	○
リンダーホーフ城	○		**街道**
小峯城	○	○	○
高橋城	○	○	○
川田城	○	○	○
長森城	○	丸	**中山道**
石神井城	加わっ, **進軍**, 喫し, **落城, 出撃, 惨敗, 敗走**, 戻り, **引き返し, 放火, 救援**	**大門**, 門, 丸	○
鴨山城	**出陣**, 破っ	門	越え, **要衝, 山陽, 瀬戸内**
安濃津城	○	○	○
省城	○	○	○
打吹城		**土蔵, 本丸**	
バルモラル城		○	○

表 5.9: クラスタリングを使った情報抽出

城名	クラスタ 401(戦い関係)	クラスタ 407(城の造り)	クラスタ 765(交通関係)
道本城	○	○	○
荊の城	○	○	○
白雲の城	○	○	○
三田城	**落城**	門, **御門, 丸内, 大手門, 移築**, 二ノ, **番所, 土蔵**, 丸, **本丸**	結ぶ, **要衝**
門司城	**敗戦, 壊滅**	門, 丸, **本丸**	○
下大留城	○	○	○
作山城	**落城**	○	○
溝口城	○	○	○
新屋城	**落城**		○
浦賀城		○	○
幻想水滸伝 V 黎明の城	○	○	○
寒河江城	**本隊**, 敗北, **撤退, 攻め寄せ, 退け, 救援**	薬, 門, **二の丸, 移築, 丸内**, 丸, **本丸**	○
鏡島城	○	門	○
河渡城		○	**要衝, 中山道**
田幡城		○	○

表 5.10: 単語の正解率の結果

手法	単語の正解率
クラスタリング (提案手法)	0.71

表 5.11: 類似度削除前のクラスタ 401

おびき出し，ひい，引き返し，援軍 ，炎上 ，加わっ ，壊滅 ，開城
勧告，陥落，頑強，奇襲，喫し，救援，窮地，屈服，迎え撃つ ，向かわ
抗戦，攻める，攻め寄せ，攻め落とす，惨敗，持ちこたえ，銃撃，出撃
出陣，少数，焼か，焼き討ち，焼き払い，焼き払わ，申し入れ，進軍
占拠，全滅，阻止，総崩れ，遭い，態勢，退け，退却，大敗，着陣，駐留
直ぐ，抵抗，撤退，転戦，逃走，派兵，破っ，敗戦，敗走，敗退，敗北
迫り，不完全，伏兵，奮戦，兵糧，放火，防戦，本隊，明け渡し，戻り
夜襲，落ち延びる，落城，篭城

5.3.3 情報抽出の比較

5.2.1 節で述べた方法で情報抽出を行った結果と先行研究の上位下位知識で情報抽出を行った実験において，word2vec を用いたクラスタリングを利用した実験では正解率は 0.82 で，先行手法の上位下位知識を利用した実験では正解率は 0.72 と提案手法の方が精度が良かった.

表 5.12: 類似度削除後のクラスタ 401

引き返し，援軍 ，開城，陥落，奇襲，喫し，迎え撃つ
攻める，攻め寄せ，攻め落とす，持ちこたえ，出撃
出陣，焼き討ち，焼き払い，焼き払わ，進軍
退却，大敗，着陣，敗走，伏兵，奮戦，兵糧，防戦
本隊，夜襲，落ち延びる，落城，篭城

表 5.13: 比較実験の結果

手法	情報抽出における正解率
上位下位知識 (先行手法)	0.72
クラスタリング (提案手法)	0.82

5.3.4 文章作成支援の性能評価

5.2.2 節で述べた方法で文章作成支援の実験を行った．その結果と，先行研究の上位下位知識で文章作成支援を行った結果の比較を行った．

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，文章作成支援が適切に行えたと考える．また，城ページにクラスタ内の類似単語の記述があり，それが表に出力されてなかった場合は不適切とする．具体的には，河渡城のページに「敗れる」という記載があった．クラスタ 401 の戦い関係には「敗北」や「敗走」などの単語が出力されていた．だが河渡城のクラスタ 401 には空白として検出されていた．この場合「敗れる」が「敗北」の類似単語とすることができる．このため，この場合は不適切に空白と検出したとする．

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 5.14 に示す．上位下位知識に基づく手法では F 値は 0.85 で，クラスタリングに基づく手法では F 値は 0.92 でクラスタリングの結果の方が性能が良かった．

表 5.14: 文章作成支援の性能評価

手法	再現率	適合率	F 値
上位下位知識	0.89 (37/33)	0.83 (40/33)	0.85
クラスタリング	1.00 (55/55)	0.85 (65/55)	0.92

第6章　今後の課題

6.1　抽出周辺の単語の抜き出し

情報抽出を行った表をみると,「改築」や「築城」とあっても何年に「改築」したものなのか，また「築城」は何年に「築城」したものか，また誰が「築城」したものかがわからない．この点を改善するために,「改築」や「築城」の周辺の単語を抜き出し利用することが考えられる．

例えば大坂城のページの一部を以下に示す．

> 初代「築城」総奉行、黒田孝高が縄張を担当。輪郭式平城であり、本丸を
> 中心に大規模な郭を同心円状に連ね、間に内堀と外堀を配する。

上述の文中には「築城」の周辺に築城に関わった「黒田孝高」が記述されている．このように抽出された単語の周辺には抽出されたものに関連のある単語が出現することが多い．それらの単語を抽出し表示することで「築城」に関するより詳しい情報をユーザーに示すことができる．

6.2　城データ以外の抽出内容

先行研究の精度比較を行うために城データでの実験を行ったが，城データ以外でも適切に情報抽出ができ，文章作成の支援ができるかを確かめる必要がある．そのために現段階では,「国」,「日本の内閣総理大臣」,「観光地百選」のページを用いる実験を検討している．

- 国
- 日本の内閣総理大臣
- 観光地百選

6.3 評価方法について

現段階のクラスタリング結果の評価方法では Wikipedia の城ページにクラスタ内の単語が 1 つでも正しく取り出せれば正解としている．この評価方法は甘い評価方法だと考えている．よって，評価方法をかえて実験を行いたい．

6.4 重要項目の数

本研究ではクラスタリングを用いて重要項目の選定を行い重要項目を 20 個に増やすことができた．事前の予想ではもう少し増えると考えていた．今後はもう少し重要項目を増やしたいと考えている．

6.5 名詞連続

word2vec は英単語を基準に作られている．英単語では単語毎に空白が入っていてクラスタリングをしやすい．日本語では，word2vec を利用するには，単語毎に空白を入れる処理を行わなければならない．本研究では mecab を使って単語毎に空白を入れた．ただ，mecab の処理結果において未知語となっている名詞連続は分割されていることが多い．例えば,「熊本城」という単語は「熊」,「本城」と分割されることがある．この場合だと「熊本城」と正しく検出されることが望ましい．このように未知語の分割を正しく分割することで，クラスタリング結果の精度の向上が考えられる．

6.6 類似度

類似度の高い単語以外を削除することによって重要情報の一部が削除されるという問題が生じた．この問題が生じた原因は，類似度計算における入力単語に「城」を固定して使用したためであると考えられる．このことから，以上の問題を解決する方法として，以下のような新たな方法が考えられる．

- 重要項目を「戦い関係」「城の造り」「交通関係」の 3 つとする．

- 「戦い関係」「城の造り」「交通関係」の類似単語を取得する．(この場合,「戦い関係」の類似単語の取得には，類似度計算の入力に「戦い」を用いる．同様に,「城の造り」には類似度計算の入力に「造り」を,「交通関係」には類似度計算の入力に「交通」を用いる．これらの入力単語の類似単語を取得する.)

- 「戦い関係」「城の造り」「交通関係」のクラスタリング結果に対して，上記で得た類似単語を利用して関係のない単語を削除する.「戦い関係」のクラスタリング結果の単語群からは,「戦い」の類似単語以外の単語を削除する.「城の造り」のクラスタリング結果の単語群からは,「造り」の類似単語以外の単語を削除する.「交通関係」のクラスタリング結果の単語群からは,「交通」の類似単語以外の単語を削除する．以上の削除をした単語群を，各クラスタの単語群として用いて以降の処理を行う.

以上のように，各クラスタの概念を示す単語を選定し，その単語の類似単語以外の単語を削除することで，各クラスタ内の関係のない単語を削除できるのではないかと考える．この方法を利用することで，クラスタ内で関係のない単語の数が減少するので，そういうクラスタの単語群を利用すると，より適切な単語からなる表を作成できるのではないかと考える.

第7章　おわりに

文章作成の際に重要情報を書き漏らす場合がある．書き漏れがあると不明瞭な文になる場合や読者が知りたい情報が書かれておらず情報取得において不便になる場合がある．重要項目を文章から抽出し表にまとめ，文章に重要項目が書いていない場合に書き漏れを指摘することで，文章作成を支援することが考えられる．

これらを改善するために本研究では重要項目の選定と改良や文章作成支援を行うため，クラスタリングの手法を提案した．

情報抽出の実験においては，上位下位知識と提案手法であるクラスタリングで正解率を求めるためにランダムに決定した城ページ 30 件を用いて評価を行った．word2vec を用いたクラスタリングを利用した実験では正解率は 0.82 で，先行手法の上位下位知識を利用した実験では正解率は 0.72 と提案手法の方が精度が良かった．類似度を使った情報抽出は重要情報が多く削除される結果になったのでクラスタリング結果のみを使うことを決定した．

また文章作成支援は 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，文章作成支援が適切に行えたと考える．また，城ページにクラスタ内の類似単語の記述があり，それが表に出力されてなかった場合は不適切とする．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．提案手法の F 値は 0.92 で先行手法での F 値は 0.85 と提案手法のほうが精度が高かった．また，重要項目も先行研究では 4 個，提案手法は 20 個と重要項目を増やすことができた．

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，鳥取大学工学部知能情報工学科計算機工学 C 講座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計算機工学 C 講座研究室の皆様方に感謝の意を表します．

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] 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.

[3] 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.

[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] 村田真樹, Stijn De Saeger, 橋本力, 風間淳一, 山田一郎, 黒田航, 馬青, 相澤彰子, 鳥澤健太郎. 論文データからの重要情報の抽出と可視化. 第 23 回人工知能学会全国大会, pp. 1–4, 2009.

[6] 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.