概要

文章に重要な情報が記載されていない場合，読者の知りたい情報が欠落しているため読みにくい文章となる．そこで，書き漏れのある文章であることを指摘したり，書き漏れのある文章に書き足すべき情報を提示する技術があれば文章の修正がしやすくなると考える．このような文章修正支援の研究はいくつかある．赤野 [2] の研究では word2vec [5] によるクラスタリングを用いて Wikipedia のデータから重要な情報を抽出し表にまとめた．表に空欄箇所があった場合，情報が欠けている記載欠落箇所と判定し，記載欠落箇所をユーザに知らせて記載の追加を促すことで文章修正支援を行った．しかし，先行研究では表の記載欠落箇所の指摘は行うものの，記載欠落箇所に埋めるべき情報をユーザに提示する手法の検討は行われていない．

そこで本研究では，検索エンジンを用いて表の記載欠落箇所に適切な情報を埋める研究を行う．記載欠落箇所に埋められた情報を参考にしながら文章の書き足しを行ってもらうことで書き漏れのある文章の修正支援に役立つと考える．具体的には検索エンジンにより Web 文書を取得し，Web 文書から重要な情報を抽出して表を作成する．表には単語の出現した記事数の上位 1 位から 5 位までを出力して，5 位正解率で情報抽出と記載欠落箇所の補完の性能を評価する．表に出力した 5 個の単語をユーザに見せただけではどれが正解かわからないが，5 個の単語を取り出した記事も見せることによってどれが正解かわかるようになり，ユーザも単語を 5 個見せられても困らないので文章の修正支援に役立つ．5 個の中に正解があれば役に立つため，5 位正解率によって評価を行った．また 5 位正解率ならば 1 位正解率で精度が低くても，1 位正解率よりも高い精度を出すことができる．

実験の結果，Web 文書からの情報抽出の実験において 5 位正解率で正解率を求めたところ，固有表現抽出を用いた情報抽出の実験では 0.71 で，上位下位知識を用いた情報抽出の実験では 0.70 で，クラスタリングを用いた情報抽出の実験では 0.66 であった．また

表の記載欠落箇所のみでの情報抽出の実験において 5 位正解率で正解率を求めたところ，固有表現抽出では 0.45，上位下位知識では 0.45，クラスタリングでは 0.44 であった．

目 次

表 目 次

図 目 次

第1章　はじめに

文章に重要な情報が記載されていない場合，読者の知りたい情報が欠落しているため読みにくい文章となる．そこで，書き漏れのある文章であることを指摘したり，書き漏れのある文章に書き足すべき情報を提示する技術があれば文章の修正がしやすくなると考える．

本研究では，まず Wikipedia から多くの記事で共通して現れる項目を重要項目として，それに関わる情報を取り出し，表の形に整理して表示する．さらに，その表において空白になっている箇所があった場合は重要情報が欠落している記載欠落箇所とする．重要情報とは表の重要項目に入る具体的な情報のことである．例えば，表 1.1 において記載欠落箇所となっている重要項目「人名」と「組織名」には，対応する Wikipedia の城ページに「人名」や「組織名」に関する情報が記載されていないため，この記載欠落箇所を指摘することで不足していた情報を書き足すことができるようになる．表に記載欠落箇所があることを指摘し書き足すように促すだけでも文章の修正支援になるが，表 1.2 のように表の記載欠落箇所を埋めることができれば，それを参考にして文章の書き足しを行うことができる. 表 1.2 の括弧付きの箇所が補完された情報である．

そこで，本研究では検索エンジンを用いて記載欠落箇所に適切な情報を埋める研究を行う．このように，記載欠落箇所に適切な情報を埋めることで文章の修正を支援することが，本研究で言う文章修正支援に相当する．記載欠落箇所に埋められた情報を参考にしながら文章の書き足しを行ってもらうことで書き漏れのある文章の修正支援に役立つと考える．

表 1.1: 表の空白は記載欠落箇所

	地名	人名	組織名
宇和島城	宇和島		

表 1.2: 表の記載欠落箇所を埋める

	地名	人名	組織名
宇和島城	宇和島	(藤堂高虎)	(宇和島藩)

本研究の特徴を以下に示す.

- Wikipedia からの情報抽出
  - Wikipedia から先行研究の手法の固有表現抽出と上位下位知識とクラスタリングによってそれぞれ重要情報を抽出する.
  - 抽出した重要情報を表の形に可視化する.
  - 表に空白箇所があった場合は，記載欠落箇所と判定する.
  - 記載欠落箇所は検索エンジンによって取得した文書から補完する.
- Web 文書からの情報抽出
  - 検索エンジンを用いて記事を 50 件取得し，Web 文書から先行研究の手法の固有表現抽出と上位下位知識とクラスタリングによってそれぞれ重要情報を抽出する.
  - 抽出した重要情報を表の形に可視化する.
  - Web 文書での情報抽出の性能を 5 位正解率で求めると固有表現抽出に基づく手法では 0.71，上位下位知識に基づく手法では 0.70，クラスタリングに基づく手法では 0.66 であった.
- 記載欠落箇所の補完性能
  - Wikipedia から情報抽出をして作成した表の記載欠落箇所を，Web 文書から情報抽出をして作成した表を用いて対応する記載欠落箇所の補完ができたかを確認する.
  - 記載欠落箇所の補完の性能を 5 位正解率で求めると固有表現抽出に基づく手法では 0.45，上位下位知識に基づく手法では 0.45，クラスタリングに基づく手法では 0.44 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

文書からの情報抽出による文章作成支援として藤原 [1]，赤野 [2] の研究が挙げられる．藤原 [1] の研究では，Wikipedia の城に関するページを抽出し，その中から城に関する重要情報を CaboCha [3](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 [4] の上位下位知識に基づく手法の 2 手法で抽出する．2 つの手法によって重要情報を抽出したものを表にまとめる．表に空欄箇所があった場合，情報が欠けている記載欠落箇所と判定し，記載欠落箇所をユーザに知らせて記載の追加を促すことで文章作成支援を行った．赤野 [2] の研究では，藤原 [1] の研究で行った重要項目の取り出し技術の改良を目的として，word2vec [5] を用いたクラスタリングに基づく手法で Wikipedia から重要情報を抽出し表にまとめた．藤原 [1] と同様に，表に空欄箇所があった場合，情報が欠けている記載欠落箇所と判定し，記載欠落箇所をユーザに知らせて記載の追加を促すことで文章作成支援を行った．これらの先行研究では表の記載欠落箇所の指摘は行うものの，記載欠落箇所に埋めるべき情報をユーザに提示する手法の検討は行われていない．そこで本稿では検索エンジンを用いて，記載欠落箇所に埋めるべき情報を取得する手法の研究を行うことで文章修正支援に役立てる．

質問応答システムの精度向上として村田ら [6] [7] の研究が挙げられる．質問応答システムの精度向上を目的に，スコアを減らしながら複数の記事でのスコアを利用する方法を提案した．単純にスコアを加算するだけではシステムの性能が下がる場合があるため，この研究ではスコアの加算の際にスコアを減らしながら加算する手法を用いている．

Web を情報源とする質問応答システムとして北斗ら [8] の研究が挙げられる．質問文に含まれる「誰」や「どこ」などの疑問詞を手がかりに，人名，地名，組織名，人工名，数字の 5 種類の回答タイプの分類を行う質問解析と，Web ページ上の質問文に対する回答が書かれていそうな部分の抽出を行うための文書検索と，回答候補を固有表現抽出とスコアの加算処理によって抽出する回答選択の 3 つの手法を組み合わせたシステムを構

築した．

論文に記載すべき情報の自動検出による文章作成支援として岡田ら [9] の研究が挙げられる．論文の研究成果や研究の有効性・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た．

第3章　提案手法

本章では，本研究の手法を説明する．本研究の手法は，文書内における重要情報の抽出と，検索エンジンを用いた情報抽出の 2 つの段階からなる．

3.1　文書内における重要情報の抽出

Wikipedia の城名に関するページに対して，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と，word2vec のクラスタリングに基づく手法の 3 手法により城に関する重要情報を抽出する．抽出は城のページ単位で行う．

3.1.1　固有表現抽出に基づく手法

Wikipedia の城ページから CaboCha を用いて,「地名」「人名」「組織名」に分類される表現を抽出し，表の行を城名，表の列を重要項目として表にまとめる．表には城データの中で単語が出現した頻度の多い上位 5 つの単語を出力する．この手法では城に関わる人物や城の所在地などの重要情報が抽出される．Wikipedia の記事から「地名」「人名」「組織名」を抽出し表にまとめた例を図 3.1 に示す．図 3.1 は Wikipedia の根添城のページから「地名」「人名」「組織名」の表現を抽出し，表にまとめたものを示している．根添城の記事中に「地名」の表現である「宮城県」,「人名」の表現である「源頼義」,「組織名」の表現である「坪沼八幡神社」があるため，これらの表現が抽出されて表に出力される．

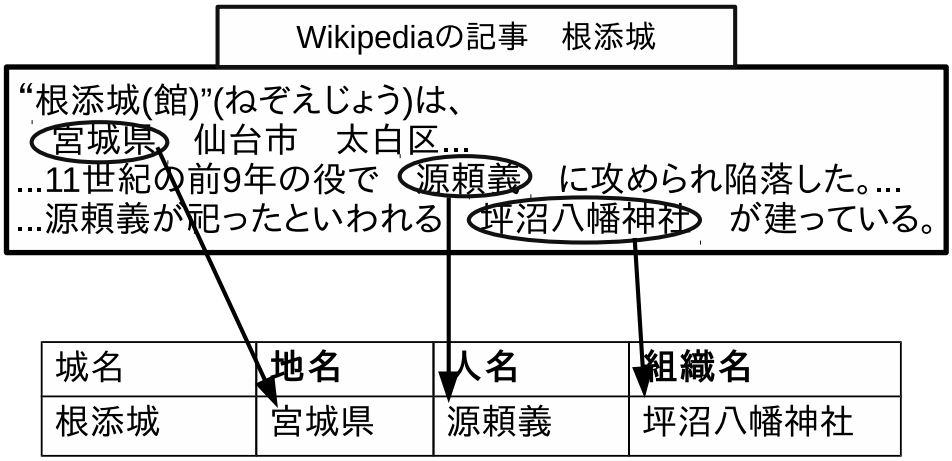
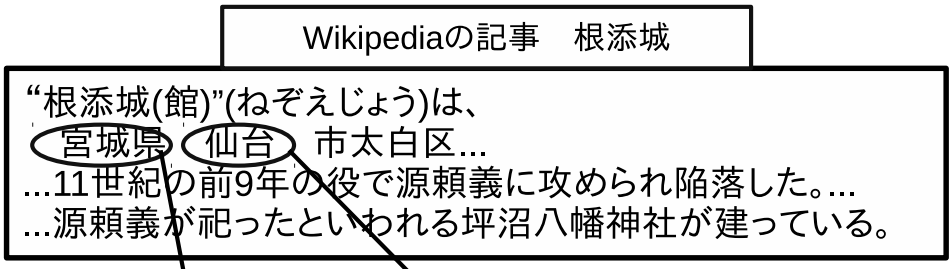


図 3.1: Wikipedia の記事に固有表現抽出を使用した結果の例

3.1.2　上位下位知識に基づく手法

上位下位知識を用いて Wikipedia の城ページ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する．Wikipedia の城ページから重要項目の下位語を取り出し，表にまとめる．固有表現抽出を用いた方法では抽出できなかった情報を抽出できる可能性がある．本研究では「県名」「時代」「地名」「元号」の 4 つの上位語を重要項目として選定して，表の行を城名，表の列を重要項目として表にまとめる．表には城データの中で単語が出現した頻度の多い上位 5 つの単語を出力する．

Wikipedia の記事から「県名」「時代」「地名」「元号」を抽出し表にまとめた例を図 3.2 に示す．図 3.2 は Wikipedia の根添城のページから「県名」「時代」「地名」「元号」の下位語となるものを抽出し，表にまとめたものを示している．根添城の記事中に「県名」の下位語である「宮城県」,「地名」の下位語である「仙台」があるため，これらの下位語が抽出されて表に出力される．図 3.2 では根添城の「県名」が「宮城県」,「地名」が「仙台」として情報抽出されているが,「時代」「元号」は空白になっている．このような空白がある場合は，Wikipedia のページに「時代」や「元号」に関する情報が記載されていないということであり，空白になっている箇所を埋めるように文章の書き足しを行えばより読みやすい文章になる．そこで本研究では 3.2 節のような手法による文章の修正支援を提案する．



城名	県名	時代	地名	元号
根添城	宮城県		仙台	

図 3.2: Wikipedia の記事に上位下位関係抽出を使用した結果の例

3.1.3　クラスタリングに基づく手法

word2vec 内のツールであるクラスタリングを用いて Wikipedia の城ページから重要項目の抽出を行う．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てその中から人手で重要項目の選定を行い，表にまとめる．表にまとめる方法を以下に示す．本研究ではクラスタリングを行った結果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」を重要項目として，表の行を城名，表の列を重要項目として表にまとめる.「クラスタ 401」は戦い関係の情報がまとまったクラスタで,「クラスタ 407」は城の造りの情報がまとまったクラスタで,「クラスタ 765」は交通関係の情報がまとまったクラスタである．

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．

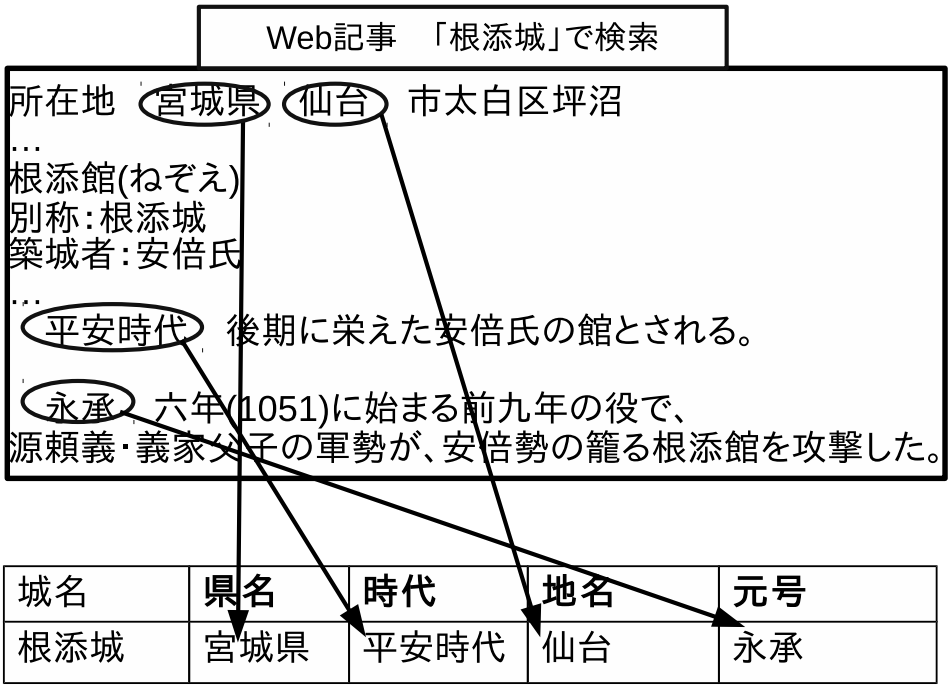
表 3.3: 表にまとめたもの

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 検索エンジンを用いた情報抽出

固有表現抽出に基づく手法と上位下位知識に基づく手法とクラスタリングに基づく手法の 3 手法により作成した表に空白があった場合，対応する城ページにはその重要情報が記載されていない．そこで，検索エンジンを用いて表の空白を埋めるべき情報を取得し，表を補完する．

検索エンジンを用いた表の補完方法として，まず城名を検索クエリとして検索エンジンに入力し 50 件の記事を取得する．取得した記事 50 件をまとめた文書に対し 3.1.1 節や 3.1.2 節や 3.1.3 節の手法を用いて城に関する重要情報の抽出を行う．抽出した重要情報のうち，記事 50 件の中で単語が出現した記事の数が多い上位 5 つの単語を表にまとめる．作成した表をユーザに提示することで文章の修正に役立つ．Web 文書から重要情報を抽出し表にまとめる例を図 3.3 に示す．図 3.3 では Web 文書からの情報抽出のために上位下位知識に基づく手法を用いている．根添城を検索クエリとして検索エンジンに入力し取得した Web 文書から「県名」「時代」「地名」「元号」の下位語となるものを抽出し，表にまとめたものを示している．図 3.3 中の文章は取得した 50 件の記事の中からランダムに選んだ記事 1 つを抜粋したものである．記事中に「県名」の下位語である「宮城県」,「時代」の下位語である「平安時代」,「地名」の下位語である「仙台」,「元号」の下位語である「永承」があるため，これらの下位語が抽出されて表に出力される．図 3.2 と見比べると，図 3.2 では表に出力できなかった「時代」「元号」を図 3.3 では出力できている．このように Wikipedia に記載されていなかった重要情報を Web 文書から取得することができるため，文章の修正支援に役立てることができる．



城名	**県名**	**時代**	**地名**	**元号**
根添城	宮城県	平安時代	仙台	永承

図 3.3: Web の記事に上位下位関係抽出を使用した結果の例

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 固有表現抽出

本研究では文書から固有表現を抽出するために CaboCha の固有表現抽出を用いる．以下の図 4.2 は Wikipedia の根添城の記事から固有表現を抽出した例である．活用型，活用形の後に固有表現タグが付与される．LOCATION は「地名」を，PERSON は「人名」を，ORGANIZATION は「組織名」をそれぞれ表す．本研究ではこの 3 つのタグのどれかが付与された表現を抽出し表にまとめる．

根添 名詞, 固有名詞, 地域, 一般,*,*, 根添, ネゾエ, ネゾエ B-LOCATION
宮城 名詞, 固有名詞, 地域, 一般,*,*, 宮城, ミヤギ, ミヤギ B-LOCATION
県 名詞, 接尾, 地域,*,*,*, 県, ケン, ケン I-LOCATION
仙台 名詞, 固有名詞, 地域, 一般,*,*, 仙台, センダイ, センダイ B-LOCATION
市 名詞, 接尾, 地域,*,*,*, 市, シ, シ I-LOCATION
太白 名詞, 固有名詞, 地域, 一般,*,*, 太白, タイハク, タイハク B-LOCATION
区 名詞, 接尾, 地域,*,*,*, 区, ク, ク I-LOCATION
坪沼 名詞, 固有名詞, 人名, 姓,*,*, 坪沼, ツボヌマ, ツボヌマ B-LOCATION
日本 名詞, 固有名詞, 地域, 国,*,*, 日本, ニッポン, ニッポン B-LOCATION
城 名詞, 一般,*,*,*,*, 城, シロ, シロ B-LOCATION
館 名詞, 接尾, 一般,*,*,*, 館, カン, カン I-LOCATION
安倍 名詞, 固有名詞, 人名, 姓,*,*, 安倍, アベ, アベ B-PERSON
奥州 名詞, 固有名詞, 地域, 一般,*,*, 奥州, オウシュウ, オーシュー B-LOCATION
安倍 名詞, 固有名詞, 人名, 姓,*,*, 安倍, アベ, アベ B-PERSON
源頼義 名詞, 固有名詞, 人名, 一般,*,*, 源頼義, ミナモトノヨリヨシ, ミナモトノヨリヨシ B-PERSON
坪沼 名詞, 固有名詞, 地域, 一般,*,*, 坪沼, ツボヌマ, ツボヌマ B-ORGANIZATION
八幡 名詞, 固有名詞, 地域, 一般,*,*, 八幡, ヤハタ, ヤハタ I-ORGANIZATION
神社 名詞, 一般,*,*,*,*, 神社, ジンジャ, ジンジャ I-ORGANIZATION
日本 名詞, 固有名詞, 地域, 国,*,*, 日本, ニッポン, ニッポン B-LOCATION
宮城 名詞, 固有名詞, 地域, 一般,*,*, 宮城, ミヤギ, ミヤギ B-LOCATION
県 名詞, 接尾, 地域,*,*,*, 県, ケン, ケン I-LOCATION
太白 名詞, 固有名詞, 地域, 一般,*,*, 太白, タイハク, タイハク B-LOCATION
区 名詞, 接尾, 地域,*,*,*, 区, ク, ク I-LOCATION

図 4.2: Wikipedia の記事に CaboCha を使用した結果の例

4.3 上位下位知識

本研究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 4.1 に示す．

表 4.1: 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

4.4 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．クラスタリングの説明を以下に示す．なお，以下の文章は赤野 [2] の論文から引用したものである．

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 [5] 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 [10]．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 4.2，表 4.3，表 4.4 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 4.2 は芸術大学という点で同じような単語が集まっている．表 4.3 は短期大学という点で同じ単語が集まっている．表 4.4 は点数関係が集まっている．

表 4.2: 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 4.3: 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
.....

表 4.4: 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
.....

第5章　実験

5.1　実験条件

本研究では，Wikipedia の城に関する記事を入力として 3.1 節の手法で情報抽出を行い，記事中の記載不備を検出するための表を作成する．なお，本研究ではその表として，先行研究である藤原 [1]，赤野 [2] の研究結果を用いる．

3.2 節の手法では検索エンジンにより得られた Web の情報を用いて，Wikipedia の記事中の記載不備を修正しやすくするための表を作成する．3.2 節の手法の性能評価のために，以下の 2 種類の実験を行う．

- 表の全ての箇所に対して検索エンジンを用いた情報抽出
- 表の記載欠落箇所のみに対して検索エンジンを用いた情報抽出

ただし，表の全ての箇所に対して検索エンジンを用いて情報抽出する実験は，Web 文書からの情報抽出の性能を見るために行う実験である．

Wikipedia(2014 年 11 月) の城ページ (2,665 ページ) に対して 3.1.1 節や 3.1.2 節や 3.1.3 節の手法を用いて情報抽出と表の作成を行った．また，2,665 件の城ページのうちランダムに選んだ 30 件の城名を検索エンジンにそれぞれ入力して記事を取得し，この 30 件の城データで 3.2 節の手法の実験と評価を行った．なお本研究で用いる検索エンジンは Microsoft 社の BingSearchAPI [11] である．

5.2　評価方法

3.2 節の手法の性能評価のために，Web 文書を対象に 3.1.1 節や 3.1.2 節や 3.1.3 節の手法を用いて作成した表に対して正解率を求める．本研究では，1 位正解率，5 位正解率，MRR の 3 つの方法で評価実験を行った．

5.2.1 n 位正解率を用いた評価方法

3.2 節の手法で作成した表について，表の頻度上位 5 つの出力のうち正解がどの順位に当てはまるかを利用して，n 位正解率を用いて評価する．n 位正解率とは，優先度の上位 n 個の候補において，1 位から n 位のいずれかに正解が含まれる場合，1 の得点を与え，その合計を問題数で割った値のことである．本研究では，1 位正解率と 5 位正解率を用いる．また，固有表現抽出に基づく手法で決定した重要項目「地名」「人名」「組織名」の 3 つそれぞれに対して評価を行う．同様に上位下位知識に基づく手法で決定した重要項目「県名」「時代」「地名」「元号」の 4 つそれぞれに対して評価を行う．クラスタリングに基づく手法で決定した重要項目「戦い状況」「城の造り」「交通関係」の 3 つそれぞれに対して評価を行う．

5 位正解率で評価する理由として，表に出力した 5 個の単語をユーザに見せただけではどれが正解かわからないが，5 個の単語を取り出した記事も見せることによってどれが正解かわかるようになり，ユーザも単語を 5 個見せられても困らないので文章の修正支援に役立つ．5 個の単語の中に正解があれば役に立つため，5 位正解率によって評価を行った．また 5 位正解率ならば 1 位正解率で精度が低くても，1 位正解率よりも高い精度を出すことができる．

5.2.2 MRR を用いた評価方法

n 位正解率と同様に表の出力のうち正解がどの順位に当てはまるかを利用して，MRR(Mean Reciprocal Rank) で評価する．MRR とは，以下の式 5.1 で表される評価値である．

$$MRR = \frac{\sum_{i=1}^{N} 1/r_i}{N} \tag{5.1}$$

N は評価する対象の総数，r_i は評価対象 i がもつ最も高い正解の順位である．今回は，Web から取得した記事 50 件の中で単語が出現した記事数が多い上位 5 つの単語を表に出力しているため，$1 \leq r_i \leq 5$ となる．また，n 位正解率と同様に，固有表現抽出に基づく手法と上位下位知識に基づく手法とクラスタリングに基づく手法の 3 手法それぞれの重要項目に対して評価を行う．

5.2.3　固有表現抽出に基づく手法の正解基準

3.2 節の手法で，Web 文書に対して固有表現抽出に基づく手法を用いて作成した表の評価方法について，n 位正解率，MRR で評価を行うための正解基準を説明する．

「地名」の項目は，県名または所在地が抽出された場合正解とする．本研究では，日本の実在する城の場合は「日本」が抽出された単語の中にあっても正解としないが，海外の城の場合は国名は正解とし，日本でつくられた架空の城の場合は「日本」でも正解とする.「人名」の項目は，築城主，城主のどちらかが抽出された場合正解とする.「組織名」の項目は，城に関すると思われる組織が抽出された場合正解とする．また，表に出力する 5 つの単語のうち 1 つでも正解があれば正解とする．

5.2.4　上位下位知識に基づく手法の正解基準

3.2 節の手法で，Web 文書に対して上位下位知識に基づく手法を用いて作成した表の評価方法について，n 位正解率，MRR で評価を行うための正解基準を説明する．

「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とする.「時代」の項目は，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とする.「地名」の項目は，城の所在地が抽出された場合正解とする.「元号」の項目は，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，表に出力する 5 つの単語のうち 1 つでも正解があれば正解とする．

5.2.5　クラスタリングに基づく手法の正解基準

3.2 節の手法で，Web 文書に対してクラスタリングに基づく手法を用いて作成した表の評価方法について，n 位正解率，MRR で評価を行うための正解基準を説明する．例えばクラスタ 407「城の造り」として「門」が抽出されたとする．この場合 Web 文書内に「門」と記述されていれば正解とするが,「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．また，表に出力する 5 つの単語のうち 1 つでも正解があれば正解とする．

5.2.6 表の記載欠落箇所のみに対して検索エンジンを用いた情報抽出の評価条件

3.2 節の評価実験のうち，表の記載欠落箇所のみに対して検索エンジンを用いた情報抽出の評価実験の条件として，システムによって正解候補が記事頻度上位 1 位から 5 位までの単語にないと判定された場合は，評価実験のときは，記事頻度上位 6 位から 20 位までにも正解候補がないかを取得した Web 文書から人手で確認する．さらに 6 位から 20 位までの単語にも正解候補となるものが Web 文書内になかった場合は，Web を利用して本当に正解候補が存在しないものなのかを確かめる．Web 上にも正解候補となるものが本当に存在しないのであれば，その箇所は正解がないままでよいものとして，省いて評価実験を行う．

5.3 実験結果

5.3.1 表の全ての箇所に対して検索エンジンを用いた情報抽出

提案手法の 3.2 節より，Wikipedia(2014 年 11 月) の城ページ (2,665 ページ) を対象として，そのうちランダムに選んだ 30 件の城名を検索エンジンにそれぞれ入力する．検索エンジンより取得した Web 文書を用いて固有表現抽出に基づく手法と上位下位知識に基づく手法により表を作成する．

固有表現抽出に基づく手法を用いて得られた結果の例を表 5.1 に示す．表において太字で記載してあるものは正解と判断したものである．検索エンジンにより取得した記事 50 件の中で単語が出現した記事の数が多い上位 5 つの単語をそれぞれの重要項目に対し出力している．また抽出した表の正解率を 1 位正解率，5 位正解率，MRR で評価した結果を表 5.4 に示す．提案手法の 3.1.1 節により選定した重要項目全てで正解率を求めた結果を表 5.5 に示す．

上位下位知識に基づく手法を用いて得られた結果の例を表 5.2 に示す．表において太字で記載してあるものは正解と判断したものである．検索エンジンにより取得した記事 50 件の中で単語が出現した記事の数が多い上位 5 つの単語をそれぞれの重要項目に対し出力している．また抽出した表の正解率を 1 位正解率，5 位正解率，MRR で評価した結果を表 5.6 に示す．提案手法の 3.1.2 節により選定した重要項目全てで正解率を求めた結果を表 5.7 に示す．

クラスタリングに基づく手法を用いて得られた結果の例を表 5.3 に示す．表において太字で記載してあるものは正解と判断したものである．検索エンジンにより取得した記事 50 件の中で単語が出現した記事の数が多い上位 5 つの単語をそれぞれの重要項目に対し出力している．また抽出した表の正解率を 1 位正解率，5 位正解率，MRR で評価した結果を表 5.8 に示す．提案手法の 3.1.3 節により選定した重要項目全てで正解率を求めた結果を表 5.9 に示す．

また，固有表現抽出に基づく手法と上位下位知識に基づく手法とクラスタリングに基づく手法の 3 つ全ての重要項目に対して正解率を求めた結果を表 5.10 に示す．

なお，表 5.4 から表 5.10 で求めた正解率は，単に Web 文書からの情報抽出の性能を見るために行った評価実験である．

表 5.1: 固有表現抽出に基づく手法による情報抽出の結果

	地名	人名	組織名
宇和島城	**宇和島**	**藤堂高虎**	宇和島城
	日本	伊達	MS
	宇和島城	桑折	朝日新聞
	宇和島市	藤兵衛	**宇和島市観光協会**
	愛媛県	高虎	二ノ丸
筑後十五城	龍造寺	**蒲池**	**龍造寺軍**
	九州	大友	南朝
	龍造寺隆信	筑後	鎮実
	筑後	隆信	龍造寺家
	肥前	龍造寺隆信	中日
岡崎城	**岡崎**	岡崎	名鉄
	岡崎城	**家康**	備前曲輪
	日本	徳川家康	**能楽堂**
	岡崎市	岡崎城	岡崎城
	愛知	三河武士	ブリタニカ国際大百科事典小項目事典
桜尾城	桜尾城	**大内**	桜尾城
	尾城	桂元澄	サイトに埋め込む桜尾城
	日本	毛利	**毛利軍**
	桜尾	毛利元就	中日
	築城	桜尾城	篠尾城
リンダーホーフ城	**ドイツ**	城内	宿泊ホテル申込最大人数名予約期限ツアー
	ミュンヘン	**ワーグナー**	中日
	日	ルイ	ファミリーマート
	日本	城内リンダーホーフ	シンデレラ城
	バイエルン	ルイ世	サークル K サンクス
小田原城	**小田原**	**北条**	小田原城
	小田原城	小田原	日立
	日本	豊臣秀吉	新越
	小田原市	大久保	日経
	神奈川県	北条氏政	韮山
川田城	日	武田	北条軍
	日本	川田	薬師堂
	川田	山城	川島入道川田雅楽助
	天正	川田まみ	武田軍
	舘林	北条	朝日
長森城	**岐阜**	**土岐**	**切通陣屋**
	岐阜市	長森城	切通観音
	切通	長森	美濃国守護職
	日本	美濃	岐阜城
	長森城	加納	長森西小学校
石神井城	**石神井**	**豊島**	**三宝寺池**
	石神井城	道灌	平塚
	石神井公園	長尾景春	氷川神社
	日本	太田道灌	泰経
	石神井公園駅	景春	石神井城

表 5.2: 上位下位知識に基づく手法による情報抽出の結果

	県名	時代	地名	元号
宇和島城	福井県	**江戸時代**	石垣	文化
	長野県	戦国時代	**四国**	**慶長**
	静岡県	安土桃山時代	城山	明治
	香川県	現代	名城	昭和
	愛知県	戦前	大洲	寛文
築後十五城	**福岡県**	**戦国時代**	**田尻**	天皇
	佐賀県	南北朝時代	山下	文化
	大分県	江戸時代	毛利	**天正**
	三潴県	室町時代	天皇	平成
	熊本県	現代	河崎	天文
岡崎城	**愛知県**	**江戸時代**	**愛知**	平成
	静岡県	戦国時代	平成	文化
	長野県	安土桃山時代	名城	昭和
	新潟県	現代	城内	明治
	岐阜県	室町時代	城下	**天正**
桜尾城	山口県	**戦国時代**	毛利	**天文**
	島根県	室町時代	藤原	承久
	新潟県	江戸時代	鎌倉	文化
	鳥取県	現代	室町	慶長
	大分県	南北朝時代	**本町**	天正
リンダーホーフ城	新潟県	現代	中央	普通
	福岡県		城内	文化
	静岡県		東京	平成
	神奈川県		駅前	天皇
	ムジンバ		カナ	大統
小田原城	**神奈川県**	**江戸時代**	北条	平成
	葉県	戦国時代	石垣	文化
	千葉県	現代	**箱根**	明治
	静岡県	室町時代	平成	**天正**
	茨城県	安土桃山時代	城内	昭和
川田城	長野県	**戦国時代**	田町	**天正**
	愛知県	江戸時代	北条	文化
	岐阜県	南北朝時代	愛知	天文
	茨城県	室町時代	城内	長久
	山梨県	現代	東京	平成
長森城	**岐阜県**	江戸時代	**中山**	明治
	笠松県	**南北朝時代**	千石	**文和**
	愛知県	戦国時代	渋谷	文治
	静岡県	現代	鎌倉	享和
	新潟県	戦前	愛知	文化
石神井城	神奈川県	**室町時代**	**東京**	文明
	葉県	戦国時代	鎌倉	文化
	千葉県	江戸時代	江古田	昭和
	島根県	南北朝時代	室町	**応永**
	新潟県	現代	沼袋	明治

表 5.3: クラスタリングに基づく手法による情報抽出の結果

	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	**開城**	**本丸**	**交通**
	攻める	長屋門	海道
	落城	移築	街道
	直ぐ	御殿	瀬戸内
	出撃	門跡	北陸
筑後十五城	**大敗**	役所	**海道**
	抵抗	本丸	交通
	出陣	二ノ	連絡
	落城	**政庁**	結ぶ
	奮戦	御殿	街道
岡崎城	**出陣**	**本丸**	**交通**
	破っ	二の丸	海道
	戻り	大手門	東海道
	大敗	御殿	北陸
	防戦	役所	便利
桜尾城	**落城**	**本丸**	**瀬戸内**
	開城	二の丸	参詣
	出陣	日出	山陽
	占拠	役所	交通
	戻り	御門	便利
リンダーホーフ城	戻り	移築	**交通**
	放火	日出	街道
	直ぐ		便利
	少数		海道
			ロマンティック
小田原城	**開城**	**本丸**	**海道**
	落城	二の丸	東海道
	退け	正門	交通
	戻り	御殿	幹線
	撤退	蓮池	北陸
川田城	**落城**	**本丸**	**街道**
	迫り	二の丸	交通
	落ち延びる	御殿	要衝
	明け渡し	役所	便利
	伏兵	番所	交差
長森城	落城	本丸	**交通**
	出陣	二の丸	海道
	攻める	**御殿**	中山道
	開城	大門	北陸
	撤退	役所	便利
石神井城	**落城**	**本丸**	**連絡**
	敗走	長屋門	便利
	放火	役所	交通
	戻り	二の丸	海道
	大敗	大門	街道

表 5.4: 固有表現抽出で抽出した情報の正解率

評価方法	地名	人名	組織名
1 位正解率	0.63(19/30)	0.66(20/30)	0.16(5/30)
5 位正解率	0.80(24/30)	0.83(25/30)	0.50(15/30)
MRR	0.70	0.75	0.26

表 5.5: 固有表現抽出で抽出した情報のすべての重要項目での正解率

評価方法	地名・人名・組織名
1 位正解率	0.48(44/90)
5 位正解率	0.71(64/90)
MRR	0.57

表 5.6: 上位下位知識で抽出した情報の正解率

評価方法	県名	時代	地名	元号
1 位正解率	0.53(16/30)	0.66(20/30)	0.36(11/30)	0.33(10/30)
5 位正解率	0.56(17/30)	0.86(26/30)	0.50(15/30)	0.86(26/30)
MRR	0.54	0.75	0.40	0.50

表 5.7: 上位下位知識で抽出した情報のすべての重要項目での正解率

評価方法	県名・時代・地名・元号
1 位正解率	0.47(57/120)
5 位正解率	0.70(84/120)
MRR	0.55

表 5.8: クラスタリングで抽出した情報の正解率

評価方法	クラスタ 401	クラスタ 407	クラスタ 765
1 位正解率	0.56(17/30)	0.50(15/30)	0.53(16/30)
5 位正解率	0.70(21/30)	0.60(18/30)	0.70(21/30)
MRR	0.62	0.52	0.57

表 5.9: クラスタリングで抽出した情報のすべての重要項目での正解率

評価方法	クラスタ 401・407・765
1 位正解率	0.53(48/90)
5 位正解率	0.66(60/90)
MRR	0.55

表 5.10: 3 手法すべての重要項目での正解率

評価方法	固有表現抽出・上位下位知識・クラスタリング
1 位正解率	0.49(149/300)
5 位正解率	0.69(208/300)
MRR	0.55

5.3.2　表の記載欠落箇所のみに対して検索エンジンを用いた情報抽出

Wikipedia からの情報抽出によって得られた表の記載欠落箇所に対して，検索エンジンによって得られた文書から作成した表を用いて，記載欠落箇所に対応する箇所のみでの検索エンジンの情報抽出の正解率を集計した．その結果を表 5.11 と表 5.12 と表 5.13 に示す．

表 5.11 と表 5.12 と表 5.13 において，重要項目ごとではなく全ての重要項目に対しての正解率を求めている．「Wikipedia 内に正解がないもの」とは Wikipedia からの 3.1 節の手法に基づく情報抽出が 100%の正解率で行えており，正しく記載欠落箇所を過不足なく検出できた場合を想定した実験であり，すべての正しい記載欠落箇所での実験である．「表の記載欠落箇所かつ Wikipedia 内に正解がないもの」とは 3.1 節の手法に基づく情報抽出の失敗を考慮した場合の実験であり，3.1 節の手法で正しく特定できた記載欠落箇所だけでの実験である．

また，表 5.14 と表 5.15 と 5.16 は，評価実験の条件として 5.2.6 節で述べた条件で行った評価実験の結果である．システムによって正解候補が記事頻度上位 1 位から 5 位までの単語にないと判定された場合は，評価実験のときは，記事頻度上位 6 位から 20 位までにも正解候補がないかを取得した Web 文書から人手で確認する．さらに 6 位から 20 位までの単語にも正解候補となるものが Web 文書内になかった場合は，Web を利用して本当に正解候補が存在しないものなのかを確かめる．Web 上にも正解候補となるものが本当に存在しないのであれば，その箇所は正解がないままでよいものとして，省いて評価実験を行った．

3.1 節の情報抽出について，先行研究である藤原 [1]，赤野 [2] の論文を参照して記載欠落箇所の検出を行った．固有表現抽出に基づく手法で作成した表では，本研究で用いた 30 件の城データにおいて 20 個の正しい記載欠落箇所のうち 10 個の記載欠落箇所が正しく検出され，同様に上位下位知識に基づく手法で作成した表では，本研究で用いた 30 件の城データにおいて 37 個の正しい記載欠落箇所のうち 33 個の記載欠落箇所が正しく検出され，クラスタリングに基づく手法で作成した表では，本研究で用いた 30 件の城データにおいて 67 個の正しい記載欠落箇所のうち 53 個の記載欠落箇所が正しく検出された．「Wikipedia 内に正解がないもの」の実験では，5 位正解率の値は，固有表現抽出に基づく手法が 0.45，上位下位知識に基づく手法が 0.45 を検出できた．クラスタリング

に基づく手法が 0.44 であった．また，評価条件として 5.2.6 節で述べた条件で行った評価実験では，5 位正解率の値は，固有表現抽出に基づく手法が 0.50，上位下位知識に基づく手法が 0.56 を検出できた．クラスタリングに基づく手法が 0.63 であった．

表 5.11: 固有表現抽出の表の記載欠落箇所のみでの正解率

評価方法	表の記載欠落箇所かつ Wikipedia 内に正解がないもの	Wikipedia 内に正解がないもの
1 位正解率	0.10(1/10)	0.15(3/20)
5 位正解率	0.40(4/10)	0.45(9/20)
MRR	0.20	0.25

表 5.12: 上位下位知識の表の記載欠落箇所のみでの正解率

評価方法	表の記載欠落箇所かつ Wikipedia 内に正解がないもの	Wikipedia 内に正解がないもの
1 位正解率	0.21(7/33)	0.18(7/37)
5 位正解率	0.45(15/33)	0.45(17/37)
MRR	0.30	0.28

表 5.13: クラスタリングの表の記載欠落箇所のみでの正解率

評価方法	表の記載欠落箇所かつ Wikipedia 内に正解がないもの	Wikipedia 内に正解がないもの
1 位正解率	0.28(18/53)	0.38(26/67)
5 位正解率	0.50(27/53)	0.44(30/67)
MRR	0.40	0.40

表 5.14: 固有表現抽出の表の記載欠落箇所のみでの正解率 (Web にも正解がないものを除いた評価)

評価方法	表の記載欠落箇所かつ Wikipedia 内に正解がないもの	Wikipedia 内に正解がないもの
1 位正解率	0.12(1/ 8)	0.16(3/18)
5 位正解率	0.50(4/ 8)	0.50(9/18)
MRR	0.25	0.28

表 5.15: 上位下位知識の表の記載欠落箇所のみでの正解率 (Web にも正解がないものを除いた評価)

評価方法	表の記載欠落箇所かつ Wikipedia 内に正解がないもの	Wikipedia 内に正解がないもの
1 位正解率	0.25(7/27)	0.23(7/30)
5 位正解率	0.55(15/27)	0.56(17/30)
MRR	0.37	0.34

表 5.16: クラスタリングの表の記載欠落箇所のみでの正解率 (Web にも正解がないものを除いた評価)

評価方法	表の記載欠落箇所かつ Wikipedia 内に正解がないもの	Wikipedia 内に正解がないもの
1 位正解率	0.52(18/34)	0.55(26/47)
5 位正解率	0.79(27/34)	0.63(30/47)
MRR	0.63	0.57

5.3.3 情報抽出の比較

固有表現抽出に基づく手法により Web 文書から情報抽出を行った結果と，上位下位知識に基づく手法と，クラスタリングに基づく手法により Web 文書から情報抽出を行った結果において，3.1 節の手法により Wikipedia の文書から重要情報を抽出して作成した表の記載欠落箇所を検索エンジンによって補完した性能を 5 位正解率で求めたところ，固有表現抽出に基づく手法での実験では 0.45 で，上位下位知識に基づく手法での実験では 0.45 で，クラスタリングに基づく手法での実験では 0.44 であった．記載欠落箇所の補完の性能は，本研究で行った 3 手法の実験の中ではどの手法もほぼ同じ精度であることがわかった．

5.3.4　正解がないままでよい箇所の取り出し性能

本研究では 3.2 節の手法により作成した表には単語が出現した記事数の上位 1 位から 5 位までを出力する．システムにより出力された記事頻度上位 1 位から 5 位までの単語に正解となるものがなければ，全ウェブに正解がない (正解がないままでよい) と推定するものとする．この推定がどのくらいの性能でできるかを本節で確認する．上位 1 位から 5 位までの単語とその関連記事を見て，人手で上位 1 位から 5 位までの単語に正解がないかを調べることのみで，全ウェブに正解がないかを判定できれば，便利である．

システムにより出力された記事頻度上位 1 位から 5 位までの単語に正解候補がないと判定された箇所については，記事頻度上位 6 位から 20 位までにも正解候補がないかを取得した Web 文書から人手で確認する．さらに 6 位から 20 位までの単語にも正解候補となるものが Web 文書内になかった場合は，Web を利用して本当に正解候補が存在しないものなのかを確かめる．Web 上にも正解候補となるものが本当に存在しないのであれば，その箇所は正解がないままでよいものとする．本節では，そのような正解がないままでよいと判断した箇所を，正解がないままでよい箇所として正しく取り出せているかの評価実験を行った．

記事頻度上位 1 位から 5 位までに正解候補がない箇所について，正解がないままでよい箇所を正しく取り出すことができているかを F 値で求める．F 値の算出方法を以下に示す．

$$F = (\frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}) \tag{5.2}$$

$$\text{適合率} = \frac{\text{正しく正解がないものを特定した数}}{\text{システムが正解がないものとした数}} \tag{5.3}$$

$$\text{再現率} = \frac{\text{正しく正解がないものを特定した数}}{\text{正解がないものの数}} \tag{5.4}$$

本研究において，適合率はシステムにより正解がないとなったもの，すなわち 1 位から 5 位に正解がないもののうち，正しく正解がないものを抽出できた割合を表したものである．再現率は Web 文書内に正解の記載がなかったもののうち，正しく正解がないもの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 5.3，式 5.4 において「正解がないものの数」というのは表の正解候補がない部分のことである．また「システムが正解がないものとした数」というのは，システムにより表の 1 位から 5 位に正解候補がないものとして抽出されたもののことである．また「正しく正解がないものを特定した数」というのは，1 位から 5 位までに正解候補がなく，かつ 6 位から 20 位までにも正解候補が Web 文書内で見つからず，Web を利用して正解候補となるものがないかを探しても存在しなかったもののことである．F 値が大きいほど，抽出される情報が Web 上にも記載されていないことをシステムがより正しく抽出できたことを意味する．

正解がないままでよい箇所を正しく取り出せたかを F 値で求めた結果について，固有表現抽出に基づく手法での結果を表 5.17 に，上位下位知識に基づく手法での結果を表 5.18 に，クラスタリングに基づく手法での結果を表 5.19 に示す．

表 5.17 と表 5.18 と表 5.19 から，記事頻度上位 1 位から 5 位までに正解候補がない箇所が正解がないままでよいものとして正しく取り出すことができているかの F 値は，固有表現抽出に基づく手法では 0.21，上位下位知識に基づく手法では 0.36，クラスタリングに基づく手法では 0.82 であった．

表 5.17: 固有表現抽出での F 値

適合率	0.12(3/25)
再現率	1.00(3/3)
F 値	0.21

表 5.18: 上位下位知識での F 値

適合率	0.22(8/36)
再現率	1.00(8/8)
F 値	0.36

また，正解がないままでよいと判断したものの例を表 5.20 に示す．表 5.20 は,「省城」を検索エンジンに入力して取得した記事 50 件に対し固有表現抽出に基づく手法により

表 5.19: クラスタリングでの F 値

適合率	0.70(21/30)
再現率	1.00(21/21)
F 値	0.82

重要情報を抽出し，単語が出現した記事数の上位 1 位から 20 位の単語を表にまとめたものである．表において太字で記載してあるものは正解と判断したものである．重要項目「地名」について，正解候補が 1 位にあったため，これは正解として評価を行う．重要項目「人名」と「組織名」について，正解候補が上位 1 位から 5 位までにないため不正解であるが，正解候補となるものが上位 6 位から 20 位にもなく，Web で探しても正解が存在しないものであったため，正解がないままでもよいものとして評価実験を行った．

表 5.20: 正解がないままでよいものの例

城名	記事頻度順位	地名	人名	組織名
		正解候補が 1 位から 5 位にあったもの	Web 上にも正解がないもの	Web 上にも正解がないもの
省城	1	**中国**	建湖県	近畿運輸局
	2	台湾	台北	交通省
	3	台北	湖南	中医協
	4	日本	ジン	黒龍江省吉林省遼寧省内
	5	北京	大豊県	環境情報アーカイブズ事務所
	6	台	諸城	パクリ疑惑
	7	江蘇省	桂林	坊
	8	上海	関羽	絳県
	9	城県	永寧寺	産経
	10	茨城県	永寧	厚労省
	11	山西省	林豪泰	区市県年月日射陽県
	12	湖南省	明代	ＧＥ
	13	河北省	大豊	鹽城
	14	遼寧省	厚労省	漓江
	15	矢部川	ローソン	労働省医政局経済課
	16	日	李金早	労働省
	17	東台県	平遥	防衛大学校
	18	東京	武昌	防衛医科大学校医
	19	千葉県	泊元	不法投棄情報大川出張所
	20	城原川	梅田	売却･貸付）審議会

5.3.5　情報抽出の成功例

Web 文書からの情報抽出の成功例について説明する．

- 固有表現抽出に基づく手法での成功例

表 5.21 は，3.1 節の手法により Wikipedia の文書から固有表現抽出に基づく手法を用いて重要情報を抽出して作成した表であり，表 5.21 の記載欠落箇所は Wikipedia 内に正解の記載がなく，正しく空欄を抽出したものである．表 5.21 において太字で記載してあるものは正解と判断したものである．表 5.21 について，この表の記載欠落箇所を補完するために提案手法の 3.2 節の手法を用いる．実際に補完ができたものを表 5.22 に示す．括弧付きで記載されているものが補完した情報である．ただしこの例で補完されている情報は，出現した記事数の多さが 1 位のものを示しているのではなく，Web 文書から情報抽出をして得られた記事頻度上位 1 位から 5 位の単語のうち最初に正解と判断したものである．

Web 文書からの情報抽出の成功例として,「浦賀城」を検索エンジンに入力して得られた Web 文書に対し，固有表現抽出に基づく手法を用いて情報抽出を行い，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.23 に示す．表において太字で記載してあるものは正解と判断したものである．表 5.23 は，表 5.21 のうち重要項目「組織名」が記載欠落箇所となっていたため，表 5.22 で補完した重要項目「組織名」の「里見軍」が記事頻度何位であったかを，考察を行うときに分かりやすくするためのものである．ユーザには記事頻度上位 1 位から 5 位の単語をまとめた表と，5 個の単語がどのような文章に記載されていたかがわかるように，単語が記載されている文章を表とともに提示することにより文章の修正がしやすくなると考える．

「里見軍」が出現した文章の例を以下に示す.「里見軍」は太字で記載している．

「里見軍」が出現した文章の例

ここは、戦国時代の弘治年三浦半島が房総の**里見軍**に攻められたため、 北条氏康（後北条氏第代）が築城したらしい浦賀城があった所です

このように正しく記載欠落箇所を補完することができれば，文章修正支援に役立つ．

表 5.21: 固有表現抽出の表に記載欠落箇所がある例

城名	地名	人名	組織名
浦賀城	**浦賀**	**北条氏康**	○
溝口城	イギリス	ピーター・ランスリー	○

表 5.22: 固有表現抽出での文章修正支援の成功例

城名	県名	時代	地名
浦賀城	**浦賀**	**北条氏康**	(**里見軍**)
荊の城	イギリス	ピーター・ランスリー	(**荊城チカレプリ撮影会**)

表 5.23: 浦賀城の組織名 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	組織名	記事数
浦賀城	1	新井	11
	2	浦賀城跡	7
	3	三崎城	6
	4	**里見軍**	5
	5	静岡古城研究会	4
	6	浦賀奉行所	4
	7	浦賀船渠株式会社	4
	8	浦賀行政センター	4
	9	八丁堀日本考古学協会年度奈良大会	3
	10	日本城郭大系	3
	11	新人物往来社	3
	12	松坂城＆宮山城城友会	3
	13	曲輪跡	3
	14	浦賀定海賊衆	3
	15	浦賀城址	3
	16	ＥＸイン横須賀リサーチパーク	3
	17	問い合わせ市民部浦賀行政センター〒横須賀市浦賀丁目	2
	18	名所鎌倉研究部カテゴリー鎌倉遺構探索おすすめ寺社鎌倉	2
	19	名越坂古墳遺跡安国論寺妙法寺	2
	20	万福寺光明寺内藤家墓所小坪坂	2

- 上位下位知識に基づく手法での成功例

表 5.24 は，3.1 節の手法により Wikipedia の文書から上位下位知識に基づく手法を用いて重要情報を抽出して作成した表であり，表 5.22 の記載欠落箇所は Wikipedia 内に正解の記載がなく，正しく空欄を抽出したものである．表 5.24 において太字で記載してあるものは正解と判断したものである．表 5.24 について，この表の記載欠落箇所を補完するために提案手法の 3.2 節の手法を用いる．実際に補完ができたものを表 5.25 に示す．括弧付きで記載されているものが補完した情報である．ただしこの例で補完されている情報は，出現した記事数の多さが 1 位のものを示しているのではなく，Web 文書から情報抽出をして得られた記事頻度上位 1 位から 5 位の単語のうち最初に正解と判断したものである．

Web 文書からの情報抽出の成功例として,「小田原城」を検索エンジンに入力して得られた Web 文書に対し，上位下位知識に基づく手法を用いて情報抽出を行い，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.26 に示す．表において太字で記載してあるものは正解と判断したものである．本研究では取得記事が 50 件だったためか，この例では抽出できた単語が 5 つしかなかったため，6 位から 20 位は無記入の状態としている．表 5.26 は，表 5.24 のうち重要項目「時代」が記載欠落箇所となっていたため，表 5.25 で補完した重要項目「時代」の「戦国時代」が記事頻度何位であったかを，考察を行うときに分かりやすくするためのものである．ユーザには記事頻度上位 1 位から 5 位の単語をまとめた表と，5 個の単語がどのような文章に記載されていたかがわかるように，単語が記載されている文章を表とともに提示することにより文章の修正がしやすくなると考える．

「戦国時代」が出現した文章の例を以下に示す.「戦国時代」は太字で記載している．

「戦国時代」が出現した文章の例

【小田原城】小田原市にある城
鎌倉時代初め、土肥氏が築城
戦国時代、北条早雲が入城して後、北条氏の本城となり、関東の中心となった

このように正しく記載欠落箇所を補完することができれば，文章修正支援に役立つ．

表 5.24: 上位下位知識の表に記載欠落箇所がある例

城名	県名	時代	地名	元号
小田原城	○	○	○	○
溝口城	**愛知県**	○	**愛知**	**天正**

表 5.25: 上位下位知識での文章修正支援の成功例

城名	県名	時代	地名	元号
小田原城	(**神奈川県**)	(**江戸時代**)	(**箱根**)	(**天正**)
溝口城	**愛知県**	(**戦国時代**)	**愛知**	**天正**

表 5.26: 小田原城の時代 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	時代	記事数
小田原城	1	江戸時代	22
	2	**戦国時代**	16
	3	現代	6
	4	室町時代	5
	5	安土桃山時代	1
	6		
	7		
	⋮	⋮	⋮
	20		

- クラスタリングに基づく手法での成功例

表 5.27 は，3.1 節の手法により Wikipedia の文書からクラスタリングに基づく手法を用いて重要情報を抽出して作成した表であり，表 5.27 の記載欠落箇所は Wikipedia 内に正解の記載がなく，正しく空欄を抽出したものである．表 5.27 において太字で記載してあるものは正解と判断したものである．表 5.27 について，この表の記載欠落箇所を補完するために提案手法の 3.2 節の手法を用いる．実際に補完ができたものを表 5.28 に示す．括弧付きで記載されているものが補完した情報である．ただしこの例で補完されている情報は，出現した記事数の多さが 1 位のものを示しているのではなく，Web 文書から情報抽出をして得られた記事頻度上位 1 位から 5 位の単語のうち最初に正解と判断したものである．

Web 文書からの情報抽出の成功例として,「門司城」を検索エンジンに入力して得られた Web 文書に対し，クラスタリングに基づく手法を用いて情報抽出を行い，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.29 に示す．表において太字で記載してあるものは正解と判断したものである．表 5.29 は，表 5.27 のうち重要項目「クラスタ 765」が記載欠落箇所となっていたため，表 5.28 で補完した重要項目「クラスタ 765」の「交通」が記事頻度何位であったかを，考察を行うときに分かりやすくするためのものである．ユーザには記事頻度上位 1 位から 5 位の単語をまとめた表と，5 個の単語がどのような文章に記載されていたかがわかるように，単語が記載されている文章を表とともに提示することにより文章の修正がしやすくなると考える．

「交通」が出現した文章の例を以下に示す.「交通」は太字で記載している．

「交通」が出現した文章の例

大友義鎮と毛利元就との合戦、「門司城の戦い」の舞台となった城として知られています
関門海峡を見下ろす**交通**の要衝であったため、大友氏と大内氏、大内氏滅亡後は毛利氏の間でが争奪戦が繰り広げられました

このように正しく記載欠落箇所を補完することができれば，文章修正支援に役立つ．

表 5.27: クラスタリングの表に記載欠落箇所がある例

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)
安濃津城	○	○	○
門司城	**敗戦**	**本丸**	○

表 5.28: クラスタリングでの文章修正支援の成功例

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)
安濃津城	(**開城**)	(**本丸**)	(**街道**)
門司城	**敗戦**	**本丸**	(**交通**)

表 5.29: 門司城のクラスタ 765 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	クラスタ 765(交通関係)	記事数
門司城	1	**交通**	10
	2	海道	8
	3	便利	7
	4	要衝	5
	5	幹線	5
	6	北陸	4
	7	瀬戸内	3
	8	水上	3
	9	山陽	3
	10	押さえ	3
	11	連絡	2
	12	要所	2
	13	繋がる	2
	14	街道	2
	15	伊勢湾	2
	16	抑える	1
	17	生野	1
	18	山陰	1
	19	国境	1
	20	交差	1

5.3.6 情報抽出の失敗例

次に Web 文書からの情報抽出の失敗例について説明する．

- 固有表現抽出に基づく手法での失敗例 1

正解候補が記事頻度上位 1 位から 20 位になかった場合の例を以下に示す．

表 5.30 は，3.2 節の手法により「高橋城」を検索エンジンに入力して得られた Web の文書に対し，固有表現抽出に基づく手法を用いて情報抽出を行い，単語が出現した記事数の上位 1 位から 5 位までをまとめたものである．表において太字で記載してあるものは正解と判断したものである．表 5.30 には正解候補が記事頻度上位 1 位から 5 位にない重要項目「地名」がある．正解候補が上位 1 位から 5 位までにない場合は，記事頻度上位 6 位から 20 位までに正解候補がないかを取得した Web 文書から人手で確認する．「高橋城」の重要項目「地名」について，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.31 に示す．表 5.31 より，取得した Web 文書内に正解候補がないかを確認したところ，記事頻度上位 6 位から 20 位には正解候補となるものがなかった．「高橋城」の「地名」として正解になるものとしては，Web で調べたところ「京都府」が挙げられるが，システムにより正解かを判定するのは記事頻度上位 1 位から 5 位の単語であり，記事頻度上位 1 位から 5 位までに正解になる単語がなかったため，情報抽出としては失敗になった．

表 5.30: 固有表現抽出での文章修正支援の失敗例

城名	地名	人名	組織名
高橋城	日本	**高橋**	島津軍
	九州	島津	**宝塚歌劇団**
	筑前	大友	備中松山城
	熊本	立花宗茂	高橋紹運
	宝満山	立花道雪	紹運

表 5.31: 高橋城の地名 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	地名	記事数
高橋城	1	日本	44
	2	九州	42
	3	筑前	30
	4	熊本	22
	5	宝満山	20
	6	大友	20
	7	土橋	18
	8	日	17
	9	天正	17
	10	太宰府	14
	11	中国	12
	12	山城	12
	13	耳川	11
	14	龍造寺	10
	15	博多	10
	16	筑後	9
	17	肥前	8
	18	栃木県	8
	19	高梁市	8
	20	阿須那	8

- 固有表現抽出に基づく手法での失敗例 2

正解候補が記事頻度上位 1 位から 20 位にあった場合の例を以下に示す.

表 5.32 は，3.2 節の手法により「溝口城」を検索エンジンに入力して得られた Web の文書に対し，固有表現抽出に基づく手法を用いて情報抽出を行い，単語が出現した記事数の上位 1 位から 5 位までをまとめたものである．表において太字で記載してあるものは正解と判断したものである．表 5.32 には正解候補が記事頻度上位 1 位から 5 位にない重要項目「組織名」がある．正解候補が上位 1 位から 5 位までにない場合は，記事頻度上位 6 位から 20 位までに正解候補がないかを取得した Web 文書から人手で確認する．「溝口城」の重要項目「組織名」について，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.33 に示す．表 5.33 より，取得した Web 文書内に正解候補がないかを確認したところ，記事頻度上位 6 位から 20 位には正解候補となるものが，記事頻度 19 位にあった．表において太字で記載してあるものが正解と判断した箇所である．Web 文書内の，単語が正解と判断できる文章の例を以下に記載する．正解に相当する箇所を太字で記載している．

「新発田藩」が出現した文章の例

溝口城
新発田藩祖となった溝口秀勝ゆかりのお城

この場合,「組織名」の項目に対する正解候補である「新発田藩」が記事頻度上位 19 位に出現していたが，システムにより正解かを判定するのは記事頻度上位 1 位から 5 位までであるため，情報抽出としては失敗になった.

表 5.32: 固有表現抽出での文章修正支援の失敗例 2

城名	地名	人名	組織名
溝口城	**尾張**	**溝口**	岩倉
	新発田	溝口秀勝	別名豊場城城郭構造平城築城
	愛知県稲沢市	溝口城	足羽将監重成
	日本	秀勝	溝口メッキ電気亜鉛メッキ
	福岡県	溝口勝政	ささら屋福光本店

表 5.33: 溝口城の組織名 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	組織名	記事数
溝口城	1	岩倉	5
	2	別名豊場城城郭構造平城築城	3
	3	足羽将監重成	3
	4	溝口メッキ電気亜鉛メッキ	3
	5	ささら屋福光本店	3
	6	陸田市左衛門	2
	7	別名溝口城城郭構造平城築城	2
	8	美山の遺跡－伊那市教育委員会	2
	9	尾張溝口城尾張溝口	2
	10	南朝	2
	11	田長盛	2
	12	中日	2
	13	築城年代：応永	2
	14	祖父江大膳	2
	15	川崎	2
	16	生放送記事単語記事動画記事	2
	17	清水	2
	18	神戸電鉄公園都市線道	2
	19	**新発田藩**	2
	20	城溝口城新発田藩	2

- 上位下位知識に基づく手法での失敗例 1

正解候補が記事頻度上位 1 位から 20 位になかった場合の例を以下に示す．

表 5.34 は，3.2 節の手法により「三田城」を検索エンジンに入力して得られた Web の文書に対し，上位下位知識に基づく手法を用いて情報抽出を行い，単語が出現した記事数の上位 1 位から 5 位までをまとめたものである．表において太字で記載してあるものは正解と判断したものである．表 5.34 には正解候補が記事頻度上位 1 位から 5 位にない重要項目「県名」がある．正解候補が上位 1 位から 5 位までにない場合は，記事頻度上位 6 位から 20 位までに正解候補がないかを取得した Web 文書から人手で確認する．「三田城」の重要項目「県名」について，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.35 に示す．表 5.35 より，取得した Web 文書内に正解候補がないかを確認したところ，記事頻度上位 6 位から 20 位には正解候補となるものがなかった．「三田城」の「県名」として正解になるものとしては，Web で調べたところ「兵庫県」が挙げられるが，システムにより正解かを判定するのは記事頻度上位 1 位から 5 位の単語であり，記事頻度上位 1 位から 5 位までに正解になる単語がなかったため，情報抽出としては失敗になった．

表 5.34: 上位下位知識での文章修正支援の失敗例 1

城名	県名	時代	地名	元号
三田城	葉県	**戦国時代**	**三田**	明治
	大分県	江戸時代	山崎	**天正**
	千葉県	南北朝時代	神戸	寛永
	神奈川県	現代	東京	文化
	新潟県	室町時代	千石	慶長

表 5.35: 三田城の県名 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	県名	記事数
三田城	1	葉県	4
	2	大分県	4
	3	千葉県	4
	4	神奈川県	4
	5	新潟県	3
	6	宮城県	3
	7	愛知県	3
	8	鳥取県	2
	9	山形県	2
	10	岐阜県	2
	11	福岡県	2
	12	福井県	1
	13	富山県	1
	14	徳島県	1
	15	島根県	1
	16	長野県	1
	17	長崎県	1
	18	静岡県	1
	19	青森県	1
	20	秋田県	1

- 上位下位知識に基づく手法での失敗例 2

正解候補が記事頻度上位 1 位から 20 位にあった場合の例を以下に示す.

表 5.36 は，3.2 節の手法により「打吹城」を検索エンジンに入力して得られた Web の文書に対し，上位下位知識に基づく手法を用いて情報抽出を行い，単語が出現した記事数の上位 1 位から 5 位までをまとめたものである．表において太字で記載してあるものは正解と判断したものである．表 5.36 には正解候補が記事頻度上位 1 位から 5 位にない重要項目「地名」がある．正解候補が上位 1 位から 5 位までにない場合は，記事頻度上位 6 位から 20 位までに正解候補がないかを取得した Web 文書から人手で確認する.「打吹城」の重要項目「地名」について，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.37 に示す．表 5.37 より，取得した Web 文書内に正解候補がないかを確認したところ，記事頻度上位 6 位から 20 位には正解候補となるものが，記事頻度 6 位にあった．表において太字で記載してあるものが正解と判断した箇所である．Web 文書内の，単語が正解と判断できる文章の例を以下に記載する．正解に相当する箇所を太字で記載している.

「白壁」が出現した文章の例

白壁土蔵群のある打吹玉川は外堀と伝えられている。

この場合,「地名」の項目に対する正解候補である「白壁」が記事頻度上位 6 位に出現していたが，システムにより正解かを判定するのは記事頻度上位 1 位から 5 位までであるため，情報抽出としては失敗になった.

表 5.36: 上位下位知識での文章修正支援の失敗例 2

城名	県名	時代	地名	元号
打吹城	**鳥取県**	**戦国時代**	石垣	**元和**
	富山県	江戸時代	中村	明治
	島根県	室町時代	城下	慶長
	新潟県	南北朝時代	毛利	延文
	山口県	現代	池田	天正

表 5.37: 打吹城の地名 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	地名	記事数
打吹城	1	石垣	26
	2	中村	24
	3	城下	23
	4	毛利	21
	5	池田	19
	6	**白壁**	15
	7	室町	12
	8	山田	12
	9	吉川	11
	10	天皇	10
	11	天神	9
	12	東京	8
	13	中央	8
	14	愛知	7
	15	福井	7
	16	奈良	7
	17	東郷	7
	18	千石	7
	19	四国	6
	20	平成	6

- クラスタリングに基づく手法での失敗例 1

正解候補が記事頻度上位 1 位から 20 位になかった場合の例を以下に示す．

表 5.38 は，3.2 節の手法により「鏡島城」を検索エンジンに入力して得られた Web の文書に対し，クラスタリングに基づく手法を用いて情報抽出を行い，単語が出現した記事数の上位 1 位から 5 位までをまとめたものである．表において太字で記載してあるものは正解と判断したものである．表 5.38 には正解候補が記事頻度上位 1 位から 5 位にない重要項目「クラスタ 407(城の造り)」がある．正解候補が上位 1 位から 5 位までにない場合は，記事頻度上位 6 位から 20 位までに正解候補がないかを取得した Web 文書から人手で確認する.「鏡島城」の重要項目「クラスタ 407(城の造り)」について，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.39 に示す．表 5.39 より，取得した Web 文書内に正解候補がないかを確認したところ，記事頻度上位 6 位から 20 位には正解候補となるものがなかった.「鏡島城」の「クラスタ 407(城の造り)」として正解になるものとしては，Web で調べたところ「南門」が挙げられるが，システムにより正解かを判定するのは記事頻度上位 1 位から 5 位の単語であり，記事頻度上位 1 位から 5 位までに正解になる単語がなかったため，情報抽出としては失敗になった．

表 5.38: クラスタリングでの文章修正支援の失敗例

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)
鏡島城	**落城**	本丸	海道
	防戦	役所	**交通**
	出陣	大門	中山道
	戻り	移築	便利
	奮戦	門跡	街道

表 5.39: 鏡島城のクラスタ 407 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	クラスタ 407(城の造り)	記事数
鏡島城	1	本丸	6
	2	役所	4
	3	大門	3
	4	移築	3
	5	門跡	2
	6	表門	2
	7	二之	2
	8	二ノ	2
	9	二の丸	2
	10	土蔵	2
	11	長屋門	2
	12	大手門	2
	13	西丸	2
	14	御殿	2
	15	蓮池	1
	16	裏門	1
	17	門扉	1
	18	番所	1
	19	東門	1
	20	東丸	1

- クラスタリングに基づく手法での失敗例 2

正解候補が記事頻度上位 1 位から 20 位にあった場合の例を以下に示す.

表 5.38 は，3.2 節の手法により「田幡城」を検索エンジンに入力して得られた Web の文書に対し，クラスタリングに基づく手法を用いて情報抽出を行い，単語が出現した記事数の上位 1 位から 5 位までをまとめたものである．表において太字で記載してあるものは正解と判断したものである．表 5.38 には正解候補が記事頻度上位 1 位から 5 位にない重要項目「クラスタ 401(戦い状況)」がある．正解候補が上位 1 位から 5 位までにない場合は，記事頻度上位 6 位から 20 位までに正解候補がないかを取得した Web 文書から人手で確認する.「田幡城」の重要項目「クラスタ 401(戦い状況)」について，記事頻度上位 1 位から 20 位までの単語を表にまとめたものを表 5.41 に示す．表 5.41 より，取得した Web 文書内に正解候補がないかを確認したところ，記事頻度上位 6 位から 20 位には正解候補となるものが，記事頻度 16 位にあった．表において太字で記載してあるものが正解と判断した箇所である．Web 文書内の，単語が正解と判断できる文章の例を以下に記載する．正解に相当する箇所を太字で記載している．

「大敗」が出現した文章の例

愛知の城　田幡城尾張...
... 信長軍に**大敗**
その後、廃城となり、許されて信長に仕えた

この場合,「「クラスタ 401(戦い状況)」の項目に対する正解候補である「大敗」が記事頻度上位 16 位に出現していたが，システムにより正解かを判定するのは記事頻度上位 1 位から 5 位までであるため，情報抽出としては失敗になった.

表 5.40: クラスタリングでの文章修正支援の失敗例 2

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)
田幡城	出陣	役所	海道
	援軍	御殿	**交通**
	落城	本丸	交差
	奮戦	**正門**	便利
	直ぐ	二之	街道

表 5.41: 田幡城のクラスタ 401 記事頻度上位 1 位から 20 位まで

城名	記事頻度順位	クラスタ 401(戦い状況)	記事数
田幡城	1	出陣	4
	2	援軍	3
	3	落城	2
	4	奮戦	2
	5	直ぐ	2
	6	退却	2
	7	加わっ	2
	8	明け渡し	1
	9	防戦	1
	10	放火	1
	11	兵糧	1
	12	派兵	1
	13	転戦	1
	14	駐留	1
	15	着陣	1
	16	**大敗**	1
	17	全滅	1
	18	焼き討ち	1
	19	出撃	1
	20	向かわ	1

第6章　今後の課題

6.1　城データ以外の抽出内容

本研究では先行研究のデータを用いて実験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，文章修正支援に役立てられるかを確かめる必要がある．そのために現段階では，Wikipedia の「国」,「日本の内閣総理大臣」,「観光地百選」,「大学」のページを用いて記載欠落箇所のある表の作成を行うことを検討している．

- 国
- 日本の内閣総理大臣
- 観光地百選
- 大学

6.2　情報抽出の性能向上

本研究で行った 3.2 節の手法による情報抽出の実験では，Web 記事 50 件中で出現した記事の数が多い上位 5 つの単語を表の出力としているが，記事頻度以外のパラメータを用いていない．そのため頻度だけではなく，城名と重要項目との単語間の距離を求め頻度に足し込んでから評価を行うなど，新たに頻度以外のパラメータを増やすことによってさらに良い正解率が見込めると考える．

6.3　スコアを利用して表の作成

6.2 節と同様に正解率の向上のために，村田ら [6] [7] の行った逓減加点法を利用して評価を行うことを検討している．現在のところ記事頻度の上位 5 位までを答えとして抽

出しているが，スコアを用いて，スコアが低ければ答えを取り出さないようにして，一つも取り出さなければ，正解がないものと推定するという方法である．

第7章　おわりに

本研究では，文章の修正支援を行うことを目的に，Wikipedia の文書から重要情報を抽出し，その結果から書き漏れのある文章の検出と，検索エンジンを利用して書き漏れのある文章を修正するための情報の提示を行った．

表の全ての箇所に対して検索エンジンを用いた情報抽出の実験では，固有表現抽出に基づく手法と上位下位知識に基づく手法とクラスタリングに基づく手法の 3 手法による情報抽出の実験の，情報抽出の性能を 5 位正解率で求めたところ，すべての重要項目での正解率が 0.69 の性能で検出できた．5 位正解率で評価する理由として，表に出力した 5 個の単語をユーザに見せただけではどれが正解かわからないが，5 個の単語を取り出した記事も見せることによってどれが正解かわかるようになり，ユーザも単語を 5 個見せられても困らないので文章の修正支援に役立つ．5 個の中に正解があれば役に立つため，5 位正解率によって評価を行った．また 5 位正解率ならば 1 位正解率で精度が低くても，1 位正解率よりも高い精度を出すことができる．

表の記載欠落箇所のみに対して検索エンジンを用いた情報抽出の実験では，固有表現抽出に基づく手法と上位下位知識に基づく手法とクラスタリングに基づく手法の 3 手法それぞれに対して表の記載欠落箇所のみでの正解率を求めた結果，5 位正解率の値は，固有表現抽出に基づく手法が 0.45，上位下位知識に基づく手法が 0.45，クラスタリングに基づく手法が 0.44 であった．

今後の課題として，城名以外での実験を行う予定である．城データ以外でも適切に情報抽出ができ，文章修正支援に役立てられるかを確かめる必要があると考えている．また，情報抽出の正解率向上のために記事頻度以外のパラメータを足し込んで評価を行うことで，より良い正解率になると考えている．

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．

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”, 鳥取大学工学部卒業論文, 2015.

[2] 赤野北斗: “Wikipedia からの情報抽出における重要項目の選定と改良”, 鳥取大学工学部卒業論文, 2016.

[3] CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer `http://code.google.com/p/cabocha/`

[4] 上位下位関係抽出ツール: Hyponymy extraction tool `http://alaginrc.nict.go.jp/hyponymy/`

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Dis-tributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems, Vol.26, pp.3111-3119, 2013.

[6] 村田真樹, 井佐原均: “質問応答システムにおける再検索を用いた回答候補の抽出手法”，情報処理学会自然言語処理研究会，2004-NL-160，pp.115-122，2004.

[7] M. Murata， M. Utiyama， and H. Isahara: “Use of Multiple Documents as Evidence with Decreased Adding in a Japanese Question-answering System”，Journal of Natural Language Processing，Vol.12，No.2，pp.209-247，2005.

[8] 北斗修哉, 村田真樹, 馬青: “Web を情報源とする日本語質問応答システムに関する研究”, 言語処理学会, 第 12 回年次大会, pp.939-942, 2006.

[9] 岡田拓真, 村田真樹, 徳久雅人, 馬青: “論文における記載不備の自動検出と自動修正に向けた分析”, 言語処理学会, 第 22 回年次大会, pp.176-179, 2016.

[10] 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.

[11] Bing Search API :Microsoft Azure Marketplace http://datamarket.azure.com/dataset/bing/search