概要

文章を作成する際に内容が欠落してしまうことがある．情報の欠落した文章はとても読み難いものである．そこで文書から重要情報の欠落を抽出しユーザに指摘する技術が求められている．

そこで本研究では，城に関する重要情報を Wikipedia から抽出し，抽出した情報をもとに文章の欠落箇所を抽出し文章作成支援をすることを目的とする．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て表の形に整理する．表において空欄になっている箇所は，Wikipedia 内で情報が欠けておりその情報を埋めるように文章を書くとよく，そのように文章作成支援をする．またその有効性を確認するための実験も行う．

実験の結果，重要情報の抽出実験においては，固有表現抽出に基づく手法では 0.6 から 0.8 の正解率で，上位下位知識に基づく手法では約 8 割の正解率であり，2 手法間にあまり性能の差は見られなかったが，文章作成支援の結果においては，固有表現抽出に基づく手法では 0.53 の F 値で，上位下位知識に基づく手法では 0.85 の F 値であった．さらに，提案手法と比較手法の F 値を比較したところ，固有表現抽出に基づく手法，上位下位知識に基づく手法ともに比較手法より性能が良かった．

目 次

表 目 次

図 目 次

第1章　はじめに

文章を作成する際に内容が欠落してしまうことがある．情報の欠落した文章はとても読み難いものである．そこで文書から重要情報の欠落を抽出しユーザに指摘する技術が求められている．

本研究では，城に関する重要情報を Wikipedia から抽出し，抽出した情報をもとに文章の欠落箇所を抽出し文章作成支援をすることを目的とする．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て表 1.1 のような形に整理する．表において空欄になっている箇所は，Wikipedia 内で情報が欠けておりその情報を埋めるように文章を書くとよく，そのように文章作成支援をする．またその有効性を確認するための実験も行う．

以下，第 2 章で関連研究の紹介をする．第 3 章では Wikipedia からの重要情報抽出の手法と文章作成支援の手法を提案する．第 4 章では本研究における実験環境を説明する．第 5 章で本研究の重要情報抽出の実験結果と，文章作成支援の性能の評価，また比較手法との性能の差を報告する．第 6 章で今後の課題について述べる．最後に第 7 章で本稿をまとめる．

表 1.1: 城の重要情報の表の例

城名	構築年	別名	構築者	...
大阪城	1583 年	錦城	豊臣秀吉	...
姫路城	1346 年	白鷺城	赤松貞範	...
熊本城	1600 年	銀杏城	加藤清正	...
名古屋城	1612 年	金鯱城	徳川家康	...
...	...	...	...	...

本研究の特徴を，重要情報の抽出と文章作成支援の二つに分けて以下に整理する．

- 重要情報の抽出
  - 重要情報の抽出に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を用いる．
  - 抽出した重要情報を表の形に可視化する．
  - 固有表現抽出に基づく手法では 0.6 から 0.8 の正解率で重要情報の抽出ができた．上位下位知識に基づく手法では,「地名」を除く項目で約 8 割の正解率であった．

- 文章作成支援
  - 重要情報の抽出のみならず文章作成支援も行えるという新規性がある．
  - 文章作成支援の性能は固有表現抽出に基づく手法では 0.53 の F 値であり，上位下位知識に基づく手法では 0.85 の F 値であった．
  - 提案手法と比較手法とを比較した結果，固有表現抽出に基づく手法，上位下位知識に基づく手法ともに比較手法より性能が良かった．

第2章　関連研究

村田ら [1] 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. 取り出した表現は，関連する論文を検索するためのキーワードとして利用できる．また，自然言語処理のサーベイを自動的に構築すためにも利用できる．この研究では情報の抽出に YamCha と教師あり機械学習を用いているが，本研究では固有表現抽出と上位下位知識を用いて情報の抽出を行う．

村田ら [2][3] の研究では，近年，アンケート分析やマーケティングリサーチに利用されているテキストマイニングの高度化を目的とし，半自動で大規模記事群から数値，固有表現情報を取り出して表やグラフを生成するテキストマイニング可視化システムを構築した．この研究では，抽出した情報を表やグラフとして可視化する，という点では本研究と類似しているが，本研究ではさらに文章作成支援を行うという点で違いがある．

樫本ら [4] の研究では，論文のサーベイを効率良く行うために，学術論文からルール及び機械学習である SVM を用いて，図や表，参考文献欄の書誌情報や脚注などの論文構成要素を抽出する手法を提案した．この研究では論文のサーベイを効率よく行うことを目的としているが，本研究では内容が欠落した文を抽出しユーザに知らせることを目的としている．

中渡瀬ら [5] の研究では，論文アブストラクトの中から特に重要な内容である「主旨」を表現している文を抽出するために,「本論文では」「本研究」などの主旨を誘発するキーワードを手ががりとして，対象となる文を獲得する．その処理で獲得できなかった文は,「主旨」を表現する文のサンプル集合を用意する，という二段階の手法を提案した．二段階目は，一段階目で獲得した文に含まれる述語を抽出して作成した述語リストを使って対象となる文を獲得する．この研究では「文」を抽出しているが，本研究では「文」ではなく対象となる語句を抽出するという違いがある．

村田ら [6][7] の研究では，質問応答システムの精度向上のために，得点を減らしながら複数の記事の得点を利用する新しい方法を提案した．ただ単純に得点を加算するだ

けではシステムの性能を下げる場合がある，そこでこの研究では単純に加算する際に生じる問題に対処するために，得点の加算の際に得点を減らしながら加算する手法を用いている．質問応答システムとは，与えられた質問に対してその答えを出力するシステムのことである．

以上 5 つの先行研究を紹介した．どれも情報抽出の研究という点では類似しているが，どれも本研究とは手法が違っていた．さらに多くの先行研究では重要情報の抽出を大きな目的としているが，本研究では情報の欠落した文章を指摘しユーザに知らせることを目的としている点で新規性がある．

第3章　提案手法

本研究の手法は文章内における重要情報の抽出と，文章作成支援の二つの段階からなる．

3.1　重要情報の抽出

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[8](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[9] の上位下位知識に基づく手法の 2 手法で抽出する．抽出は城のページ単位で行う．表 3.1 のように最初に出現した重要情報のみをまとめた表と，表 3.2 のように出現した全ての重要情報をまとめた表の 2 つを作成する．

表 3.1: 最初に出現した重要情報の表の例

城名	県	時代	地名	元号
川田城	岐阜県	室町時代	原	康正
宇和島城		江戸時代	石垣	慶長

表 3.2: 出現した全ての重要情報の表の例

城名	県	時代	地名	元号
川田城	岐阜県, 愛知県	室町時代, 戦国時代	原, 田町, 室町, 愛知, 一宮, 加, 一方, 城内, 関, 松原, 館	康正, 長久
宇和島城		江戸時代, 安土桃山時代, 現代	石垣, 四国, 中, 北, 兵衛, 東, 海, 名城, 原, 岸, 海岸, 小屋, 藤原, 早川, 大洲, 戸田, 関, 富田, 台, 宝, 城山, 平成, 三浦, 館, 千鳥, 楚, 谷	慶長, 寛文, 天慶, 明治, 文化, 太平, 昭和, 承平, 嘉禎, 天文, 天正, 文禄, 元和, 平成

3.1.1 固有表現に基づく手法

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．

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する．対象データで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手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．固有表現抽出に基づく手法と同様に．

3.2 文章作成支援

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，そのことをユーザに知らせ記載の追加を促すことで文章作成支援をする．表 3.3 に文章作成支援に用いられる表の例を示す．この表において空欄になっている箇所が情報抽出の結果 Wikipedia 内に正解がないと判定された箇所である．本研究の文章作成支援の研究では，このような表の空欄箇所を情報の欠落としてユーザに知らせることを目的とする．

表 3.3: 文章作成支援に用いられる表の例

城名	構築年	別名	構築者	…
大阪城	1583 年		豊臣秀吉	…
姫路城		白鷺城	赤松貞範	…
熊本城	1600 年	銀杏城	加藤清正	…
名古屋城	1612 年	金鯱城		…
…	…	…	…	…

第4章　実験環境

4.1　実験データ

本研究では Wikipedia(2014 年 11 月現在) のうち，記事タイトルが城で終っているページ (2665 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Wikipediaの城に関する記事から，固有表現を抽出するためにCaboChaを用いる．以下の図4.2が具体例である．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する．

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

4.3.1 頻度分析

上位下位知識を用いて頻度分析を行い，下位語の出現記事数が 100 件を超えている上位語を取り出した．その結果において出現記事数が多かったものと，少かったものの例をそれぞれ 15 件ずつを表 4.2 に示す．その取り出したものの中から重要項目になりうると思われるものを人手で選んだ．その結果「県」「時代」「地名」「元号」という 4 つの上位語を重要項目とした．その 4 つの上位語の下位語が出現した記事数をまとめたものを表 4.3 に示す.

表 4.2: 上位下位知識を用いた頻度分析の結果

上位語	下位語の出現記事数
城	2665
う	1912
日	1865
よう	1786
一	1706
年	1679
県	1655
関	1579
ト	1573
山	1568
市	1557
連	1508
す	1477
項	1419
名	1374
...	...
平氏	104
飯	104
ラー	104
藩庁	103
南朝	103
貫	103
サー	103
理由	102
大和	102
政権	102
文禄	101
極	101
平安時代	100
自動車	100
鬼	100

表 4.3: 上位下位知識を用いた頻度分析の結果

上位語	下位語の出現記事数
県	1665
時代	1061
地名	301
元号	238

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月現在)を用いる．Wikipediaからのデータの抽出は，記事単位で行う．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．

実験1 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を用いて，Wikipediaの城に関するページの情報抽出を行い，表にまとめる．さらに，固有表現抽出に基づく手法で抽出された重要情報の正解率を「地名」「人名」「組織名」でそれぞれ求め，上位下位知識に基づく手法で抽出された重要情報の正解率を「県」「時代」「地名」「元号」でそれぞれ求める．

実験2 重要情報抽出の実験において作成された表の空欄が，正しく抽出されているかどうかの性能評価を行う．その後，比較手法と提案手法の性能の比較を行う．

5.2 表の評価方法

5.2.1 固有表現抽出に基づく手法

ランダムに選択した 30 件を用いて評価を行う.「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.「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．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.

5.2.2 上位下位知識に基づく手法

ランダムに選択した 30 件を用いて評価を行う.「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．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.

5.2.3 比較手法

文章作成支援の実験において，有効性確認のため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で作成した表を，全て空欄と仮定して F 値を求める.

5.3 F 値の算出式

文章作成支援の評価実験では以下の算出式を用いて F 値を求める．

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

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

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

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，正解がいくつあるかの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．

5.4 実験結果

5.4.1 実験１ 固有表現抽出を用いた情報抽出の結果

固有表現抽出を用いて抜き出した重要情報のうち最初に出現したものだけをまとめたものを表 5.1 に，出現した全ての重要情報をまとめたものを表 5.2 に示す．その 2 つの表を評価したものを表 5.3 に示す．抽出した結果の正解率を求めると「地名」は 0.83,「人名」は 0.83,「組織名」は 0.63 という正解率であった．表 5.1 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，〇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．表 5.3 で，括弧で記載してあるものは 30 件を評価したうち正解と判断したものの数である．

表 5.1: 評価した最初に出現した重要情報の表

城名	地名	人名	組織名
宇和島城	**宇和島**	**藤堂高虎**	JPG
筑後十五城	**筑後**	大友	center
岡崎城	**岡崎**	**西郷**	”’岡崎城”’（おかざきじょう
桜尾城	日本	**友田興藤**	img
リンダーホーフ城	**独**		
小峯城	小峯		
高橋城	高橋城	**高橋城**	**同志社大学**
川田城	**川田**	康正	川島入道川田雅楽助
長森城	長森	長森	**切通陣屋**
石神井城	日本	**豊嶋**	img
鴨山城	倉敷	加茂山城	name=鴨山城
安濃津城			安濃津城
省城		○	○
打吹城	打吹城	**山名時**	倉吉
バルモラル城	**スコットランド**	**サー・ウィリアム・ドラモンド**	Estate
道本城	**長野県**		○
荊の城	**イギリス**	**ピーター・ランスリー**	○
白雲の城	氷川	**松井由利夫**	**日本レコード協会**
三田城	**三田**	車瀬城	img
門司城	日本	司城	JPG
下大留城	大留城	**谷口友之**	○
作山城	山城	**香西佳清**	○
溝口城	溝口城	溝口城	”’溝口城”’（みぞぐちじょう
新屋城	屋城	**新屋**	name=新屋城
浦賀城	**浦賀**	**北条氏康**	○
幻想水滸伝Ⅴ黎明の城	○	○	○
寒河江城	**寒河江**	寒河江城古	**薬師堂**
鏡島城	**鏡島**	**斉藤帯刀左衛門**	[[安藤守就—安藤守就
河渡城	**河渡**	稲葉良	[[安藤守就
田幡城	**田幡**	越智信	JPG

表 5.2: 評価した出現した全ての重要情報の表の一例

城名	地名	人名	組織名
宇和島城	宇和島, 日本, 宇和島城, 愛媛県, 鶴島城, 板島城, 丸串城, 宇和島市, 江戸, 四国, 丸之内, 板島丸串城, 二ノ丸, 三ノ丸, 平山城, 水堀, 築城, 築城術, 太平洋戦争, 昭和, 持田右京, 筑前, 戸田与左衛門, 宇和郡, 関ヶ原, 今治市, 移, 伊勢, 大阪, 三重, 安土桃山, 竪三つ引, 予讃線, 宇和島駅, 宇和島城天守, 宇和島藩	藤堂高虎, 伊達宗利, 藤堂, 伊達, 藤兵衛, 代右衛門丸, 長門丸, 徳川, 高虎, 藤原純友, 大友, 長宗我部, 西園寺宣久, 小早川隆景, 隆景, 戸田勝隆, 富田信, 藤堂良勝, 伊達政宗, 伊達秀宗, 大手門, 三浦正幸, 寛文, 児島惟謙, 井上宗和, 橘, 西園寺, 小早川, 戸田, 富田	JPG, 安土桃山, 広島大学, 宇和島市立城山郷土館, 宇和島市教育委員会文化課, 宇和島市観光情報センター作成パンフレット
筑後十五城	筑後, 下筑後, 龍造寺, 九州, 肥前, 龍造寺隆信, 蒲池, 柳川市, 本城町, 山門郡, 三潴郡, 下妻郡, 吉井町, 福岡県, 富永妙見山, 生葉郡, 竹野郡, 八女市, 黒木町, 北木屋, 上妻郡, 久留米市, 草野町, 山本郡, 高良山, 三井郡, 大刀洗町]] 下高橋）,,[[御原郡, 本郷, 三井郡大刀洗町本郷）,, 御原郡, 三原, 城島, 城島町, 周, 尾城, 筑後国), 鷹尾城, 大和町, 鷹尾, 矢部村, 矢部, 筑後市, 溝口, 大牟田市, 今山, 三池郡, 日本	大友, 筑後, 筑後十五城, 蒲池, 蒲池鑑久, 蒲池親, 島津, 蒲池鑑盛, 蒲池鑑, 龍造寺隆信, 隆信, 龍造寺, 田尻鑑, 黒木家永, 黒木, 星野, 星野吉実, 河崎, 草野, 丹波, 高橋, 高橋鑑種, 三原, 西牟田, 田尻, 田尻鑑種, 溝口, 三池	center
岡崎城	岡崎, 阿波, 撫養, 相模, 日本, 岡崎城, 愛知県, 岡崎市, 康生町, 江戸, 菅生川, 矢作川, 龍頭山, 曲輪, 北曲輪, 三ノ丸, 東曲輪, 東海, 岡崎城絵図, 愛媛県, 松山市, 松山, 伊予, 今川, 浜松, 石川, 関東, 移, 東海道, 上野, 白井, 岡崎県, 額田県, 名古屋市, 額田郡額田町, 西尾市西浅井町, 下青野町慈光寺, 東名高速道路, 愛知県岡崎市康生町 561　岡崎公園, 名古屋, 名古屋本線]][[東岡崎駅, 岡崎藩	西郷, 頼嗣, 三河国守護仁木, 松平清康, 田中吉政, 松平, 田中, 本多, 水野, 徳川家康, 徳川, 仁木, 西郷信貞, 松平昌, 平山, 本多康重, 三浦正幸, 大手門, 稗田, 稗田曲輪, 城内, 岡崎, 山田景, 今川義元, 松平元康, 家康, 今川, 松平信康, 信康, 本多重次, 豊臣, 吉政, 本多康, 白井, 本多忠直, 本多康紀, 本多忠利, 本多利長, 水野忠善, 水野忠春, 水野忠盈, 水野忠之, 水野忠輝, 水野忠辰, 水野忠任, 松平康福, 本多忠粛, 本多忠典, 本多忠顕, 本多忠考, 大林寺郭堀跡, 豊臣秀吉, 岡竒城, 三河, 菅生郷なり, 大草城, 祥松平	"' 岡崎城"'（おかざきじょう, 安土桃山,"' 龍燈山城"'（りゅうとうざんじょう, 新人物往来社, 備前曲輪, 坂谷曲輪, 白山曲輪, 菅生曲輪, 北曲輪門, 広島大大学院, 岡崎市教育委員会, 夏目, 名鉄]][[名鉄名古屋本線,Keep,Okazakijyo, 岡崎公園カラクリ人形,Statue

表 5.3: 固有表現抽出を用いて作成した表の評価結果

	地名	人名	組織名
全ての重要情報を抽出した表	0.83(25/30)	0.83(25/30)	0.63(19/30)
最初に出現した重要情報のみ抽出した表	0.50 (15/30)	0.53(16/30)	0.36(11/30)

5.4.2　実験 1 上位下位知識を用いた情報抽出の結果

頻度分析によって得られた上位語を用いて抜き出した重要情報のうち，最初に出現したものだけをまとめたものを表 5.4 に，出現した全ての重要情報をまとめたものを表 5.4 に示す．その 2 つの表を評価したものを表 5.6 に示す．抽出した結果の正解率を求めると「県」は 0.83,「時代」は 0.93,「地名」は 0.26,「元号」は 0.83 という正解率であった．表 5.4 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．表 5.6 で，括弧で記載してあるもののうちの分子の値は 30 件を評価したうち正解と判断したものの数であり，分母は評価した数の 30 である．

表 5.4: 評価した最初に出現した重要情報の表

城名	県	時代	地名	元号
宇和島城		**江戸時代**	石垣	**慶長**
筑後十五城	**福岡県**	**戦国時代**	中	大和
岡崎城	**愛知県**	**戦国時代**	**愛知**	昭和
桜尾城	○	**江戸時代**	桜	**天文**
リンダーホーフ城	○	○	南	○
小峯城				
高橋城	○	戦前	中	○
川田城	**岐阜県**	**室町時代**	原	**康正**
長森城	**岐阜県**	**戦国時代**	鎌倉	**文治**
石神井城	神奈川県	**室町時代**	鎌倉	**文明**
鴨山城		**戦国時代**	関	**応永**
安濃津城	○	○	○	○
省城	○	○		○
打吹城	**鳥取県**	**室町時代**	加	**応安**
バルモラル城	○	○	関	○
道本城	**長野県**	○	中	○
荊の城	○	○	原	○
白雲の城	○	○	前田	
三田城	大分県	**南北朝時代**	**三田**	**慶安**
門司城	**福岡県**	**南北朝時代**	関	**元暦**
下大留城	**愛知県**	○	**愛知**	昭和
作山城	**香川県**	○	南町	○
溝口城	**愛知県**	○	**愛知**	**天正**
新屋城	**青森県**	**戦国時代**	館	**慶長**
浦賀城	**神奈川県**	**戦国時代**	北条	○
幻想水滸伝Ⅴ 黎明の城	○	○	○	○
寒河江城	**山形県**	**南北朝時代**	南	**元和**
鏡島城	**岐阜県**	**戦国時代**	鏡	承久
河渡城	**岐阜県**	**戦国時代**	中山	**天正**
田幡城	**愛知県**	○	**田幡**	**天文**

表 5.5: 出現した全ての重要情報の表の一例

城名	県	時代	地名	元号
宇和島城		江戸時代, 安土桃山時代, 現代	石垣, 四国, 中, 北, 兵衛, 東, 海, 名城, 原, 岸, 海岸, 小屋, 藤原, 早川, 大洲, 戸田, 関, 富田, 台, 宝, 城山, 平成, 三浦, 館, 千鳥, 楚, 谷	慶長, 寛文, 天慶, 明治, 文化, 太平, 昭和, 承平, 嘉禎, 天文, 天正, 文禄, 元和, 平成
筑後十五城	福岡県	戦国時代	中, 山下, 田尻, 種, 木, 城町, 千石, 新川, 北, 城山, 河崎, 原, 本郷, 大和, 矢部	大和
岡崎城	愛知県, 額田県	戦国時代, 安土桃山時代, 江戸時代, 現代	愛知, 木, 中, 石垣, 菅, 北, 北方, 三浦, 東, 谷, 白山, 南, 川沿, 海, 城内, 平成, 原, 関, 今川, 山田, 石川, 城下, 上野, 名城, 田町, 桜, 館	昭和, 享徳, 明治, 文化, 慶長, 正保, 元和, 平成, 天明, 徳元, 康正, 正元, 享禄, 天文, 永禄, 元康, 元亀, 天正
桜尾城		江戸時代, 室町時代, 戦国時代	桜, 加, 海, 藤原, 原, 岸, 中, 中原, 室町, 毛利, 北, 関, 東, 草津	天文, 承久, 慶長, 永享, 嘉吉, 天正
リンダーホーフ城			南, 加, 館, ポンパドゥール, 関	
小峯城				
高橋城		戦前	中, 宝, 南, 関, 加	
川田城	岐阜県, 愛知県	室町時代, 戦国時代	原, 田町, 室町, 愛知, 一宮, 加, 一方, 城内, 関, 松原, 館	康正, 長久
長森城	岐阜県, 笠松県	戦国時代, 南北朝時代	鎌倉, 谷, 渋谷, 南, 北, 宝, 加, 木, 中, 千石, 中山, 関	文治, 暦応, 文和, 享和, 明治
石神井城	神奈川県	室町時代	鎌倉, 東, 東京, 台, 室町, 中, 宮城, 館, 関, 城内, 宝, 扇, 谷, 江古田, 原, 沼, 沼袋, 愛宕, 早稲田, 稲田, 足立, 川崎, 加, 北, 世田谷, 中央, 南, 大門, 木, 皿	文明, 貞和, 応安, 安元, 応永, 明治, 昭和

表 5.6: 上位下位知識を用いて作成した表の評価結果

	県	時代	地名	元号
全ての重要情報を抽出した表	0.83(25/30)	0.93(28/30)	0.26(8/30)	0.83(25/30)
最初に出現した重要情報のみ抽出した表	0.83(25/30)	0.93(28/30)	0.23(7/30)	0.80(24/30)

5.4.3 実験２ 文章作成支援の性能評価

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．この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 5.7 に示す．固有表現抽出に基づく手法では 0.53 の F 値であり．上位下位知識に基づく手法では 0.85 の F 値であった．上位下位知識に基づく手法の性能の方が良かった．

表 5.7: 文章作成支援の結果の評価

手法	再現率	適合率	F 値
固有表現抽出	0.50(10/20)	0.56(10/18)	0.53
上位下位知識	0.89(33/37)	0.83(33/40)	0.85

5.4.4 比較実験

固有表現抽出に基づく手法と比較手法との比較結果を表 5.8，上位下位知識に基づく手法との比較結果表 5.9 に示す．比較実験の結果どちらの手法とも比較手法より性能が良かった．

表 5.8: 固有表現抽出に基づく手法との比較結果

手法	再現率	適合率	F 値
固有表現抽出	0.50(10/20)	0.56(10/18)	0.53
比較手法	1.00(20/20)	0.22(90/20)	0.36

表 5.9: 上位下位知識に基づく手法との比較結果

手法	再現率	適合率	F 値
上位下位知識	0.89(33/37)	0.83(33/40)	0.85
比較手法	1.00(37/37)	0.30(120/37)	0.46

5.4.5 文章作成支援の成功例

文章作成支援の成功例について説明する．表5.10では，情報抽出した結果Wikipedia内に正解の記載が無く，空欄を抽出したことになっている．そこで実際にWikipedia内を確認したところ，実際に正解の記載が無かった．空欄が正しく抽出できていたものについてはウェブの他のページを用いて正解を書き込んだ．表5.10について，この表では空欄を正しく抽出できており，Wikipedia内に正解の記載がなかったので，他のウェブページを参考に正解の情報を書き込んだ．実際に書き込んだものを表5.11に示す．このように正しく空欄を抽出でき，かつ，空欄の内容は他のページを参考にすれば記載可能であるため，表5.10は文章作成支援に役立つ例となっている．

表 5.10: 空欄の抽出の成功例

城名	県	時代	地名	元号
作山城	**香川県**	○	南町	○
溝口城	**愛知県**	○	**愛知**	**天正**

表 5.11: 文章作成支援を行った例

城名	県	時代	地名	元号
作山城	**香川県**	(鎌倉時代)	南町	(承久)
溝口城	**愛知県**	(安土桃山時代)	**愛知**	**天正**

5.4.6　文章作成支援の失敗例

次に文章作成支援の失敗例について説明する．下の表 5.12 では，情報抽出した結果 Wikipedia 内に正解の記載が無く，空欄を抽出したことになっているが，人手で評価を行ったところ，Wikipedia の宇和島城のページの冒頭に，以下のように記載されており，県の項目に対する正解が出現していたが，上手く抽出できていなかった．

“宇和島城（うわじまじょう）は，四国の 愛媛県 宇和島市丸之内にあった城．”

同様に，鴨山城のページの冒頭にも以下のように記載されていたが，正解の「岡山県」が上手く抽出できていなかった．

“鴨山城（かもやまじょう）は日本の城．所在地は 岡山県 浅口市鴨方町鴨方．”

このように表が空欄となっているにも関わらず Wikipedia 内に正解の記載が発見された場合，文章作成支援が上手くできなかったと判定した．文章作成支援の失敗の傾向としては，上記の愛媛県宇和島市や岡山県浅口市のように，名詞が連続で続いた場合情報抽出が上手くできないと考える．

表 5.12: 文章作成支援の失敗例

城名	県	時代	地名	元号
宇和島城		**江戸時代**	石垣	**慶長**
鴨山城		**戦国時代**	関	**応永**

第6章　今後の課題

6.1　情報抽出

重要情報の実験では，固有表現抽出に基づく手法では 0.6 から 0.8 の正解率で重要情報の抽出ができた．上位下位知識に基づく手法では,「地名」を除く項目で約 8 割の正解率であった．この数値はそれなりに実用可能な数値であると思われるが，他の手法を用いればよりよい数値が得られる可能性がある．そこで今後の課題として，重要情報の抽出に先行研究の手法などを利用する方法をあげる．

本研究では Wikipedia の城ページのみを用いて重要情報の抽出を行ったが，今後は城以外のページにも本研究の手法が有効かどうかの研究も行うとよいと思われる．

上位下位知識に基づく手法では「地名」の項目の正解率が 0.23 となり，低い精度であった.「地名」を抽出する場合他の上位語を用いると，上手く抽出できる可能性があるので重要項目の決定を再度検討したい.

6.2　文章作成支援

本研究では，評価を全て一人で行ったが，今後の課題として複数の被検者を用いた文章作成支援の実験を行い，文章作成支援が実際に役に立つのかを考察したい．また，Wikipedia 以外のウェブページでも本研究の文章作成支援が使えるかの実験も行いたい.

第7章　おわりに

本研究では文章中の重要情報の記載欠落を指摘するために，2段階の手法を提案した．その手法とはWikipediaからの重要情報抽出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の2つである．また，重要情報の抽出と同時に文章作成支援をする実験を行った．Wikipediaからの重要情報の抽出実験の結果，固有表現抽出に基づく手法では0.6から0.8の正解率で重要情報の抽出ができた．上位下位知識に基づく手法では,「地名」を除く項目で約8割の正解率であった．文章作成支援の性能は，固有表現抽出に基づく手法では0.53のF値であり，上位下位知識に基づく手法で0.85のF値であった．上位下位知識に基づく手法の性能の方が良かった．また，重要情報の抽出で作成した表の項目を全て空欄とみなす比較手法と比較した結果，固有表現抽出に基づく手法，上位下位知識に基づく手法ともに比較手法より性能が良かった．

参考文献

[1] 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, 3F2-NFC3-9, 2009.

[2] 村田真樹, 岩立将和, 一井康二, 馬青, 白土保, 金丸敏幸, 塚脇幸代, 井佐原均, “ 大規模記事群からの数値固有表現情報のテキストマイニング可視化システム ”, 情報処理学会自然言語処理研究会　研究報告, 2008-NL-184, pp.25-32, 2008.

[3] Masaki Murata, Masakazu Iwatate, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru and Kentaro Torisawa ”Extraction and Visualization of Numerical and Named Entity Information from a Large Number of Documents” IEEE NLPKE-08, pp.122-139, 2008.

[4] 樫本達矢, 太田学, 高須淳宏, ”学術論文からの構成要素抽出の一手法”, 第 6 回データ工学と情報マネジメントに関するフォーラム, C5-2, 2014.

[5] 中渡瀬秀一, 大山敬三“ 論文アブストラクトからの主旨抽出法 ”, 人工知能学会研究会資料, 情報編纂研究会第 6 回, pp.13-16, 2011.

[6] 村田 真樹，井佐原 均“ 質問応答システムにおける逓減加点法に基づく複数記事情報の利用 ” 情報処理学会自然言語処理研究会 2004-NL-160，pp115-122，2004.

[7] Masaki Murata, Masao Utiyama, and Hitoshi 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] CaboCha/南 瓜: Yet Another Japanese Dependency Structure Analyzer http://code.google.com/p/cabocha/

[9] 上 位 下 位 関 係 抽 出 ツ ー ル Version1.0 : Hyponymy extraction tool http://alaginrc.nict.go.jp/hyponymy/

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，徳久雅人講師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．