世界史用語集の語釈文における見出し語に照応するゼロ代名詞の表層格の推定

大矢 康介 †1　阪本 浩太郎 †2　渋木 英潔 †3　森 辰則 †3

†1 横浜国立大学 理工学部　†2 横浜国立大学 大学院 環境情報学府
†3 横浜国立大学 大学院 環境情報研究院
E-mail: {kosuke-o,sakamoto,shib,mori}@forest.eis.ynu.ac.jp

1　はじめに

近年，文書情報に対する情報要求は複雑化，高度化しており，そのような要求を満たすアクセス技術として質問応答が注目されている．質問応答とは利用者の自然言語による質問に対して情報源から解答そのものを抽出する技術であり，現実世界における，特に解答が複数の文を含む文章となる質問応答を目的とした取り組みも盛んにおこなわれている．そのような質問の例として，大学入試の世界史論述問題がある．以上の背景から，我々は大学入試における世界史分野の論述問題を対象とした質問応答システムの構築を目指している．

阪本ら [2] は，情報要求の存在する抽出型の複数文書要約としてこの課題を位置づけ，知識源から句点区切りの単位でテキストを抽出・整列して論述問題に解答する質問応答システムを提案している．知識源には 4 冊の教科書と 1 冊の用語集のテキストデータを用いており，用語集は見出し語と語釈文に分かれて構成されている．

用語集の見出し語と語釈文の例

見出し語）シク戦争
語釈文）西北インドのシク教徒をイギリス軍が撃破した戦い。イギリスはパンジャーブ地方を併合した。

論述問題に解答する際に，この語釈文だけをそのまま解答文に含めてしまうと，何について述べているかわからない文になってしまう．また，論述問題において解答に含めなければならない指定語句が見出し語となっている場合，語釈文だけから解答を構成すると大きく減点されてしまう．このため阪本らは，用語集の語釈文を解答の材料として抽出した際には，見出し語を文頭に主題として付け加えた文を生成し，これを解答の一部とする手法を提案している．同手法により語釈文より生成された文の例を以下に示す．

阪本らの手法により生成された文の例 1

シク戦争は，西北インドのシク教徒をイギリス軍が撃破した戦い。

この手法は，例 1 のようなコピュラ文を対象とした場合等には文法的に問題ないが，一部語釈文によっては例 2 のように文法的に誤りのある文が生成され得るという問題点がある．

阪本らの手法により生成された文の例 2

シク戦争は，イギリスはパンジャーブ地方を併合した。

また，この手法によって得た文の主題は例 2 のようにかならず見出し語になるため，解答文に適していない場合があるという問題点も挙げられる．例えば図 1 に示す 2008 年度の問 1 は，各国の動きを述べよという問題であり，図 2 に示す模範解答[1]でもイギリスを主題として文章が展開されているが，用語集を用いた阪本らの手法の解答は「クリミア戦争」という出来事が主題となっており，解答に適している文であるとは言えない．

...(省略)...
世界史が大きなうねりをみせた1850年ころから70年代までの間に、日本をふくむ諸地域がどのようにパクス・ブリタニカに組み込まれ、また対抗したのかについて解答欄(イ)に18行（540字）以内で論述しなさい。その際に、以下の9つの語句を必ず一度は用い、その語句に下線を付しなさい。

インド大反乱　クリミア戦争　江華島事件　総理衙門　第1回万国博覧会
日米修好通商条約　ビスマルク　ミドハト憲法　綿花プランテーション

図 1: 東京大学 2008 年度の問 1

阪本らの手法により生成された文の例 3

クリミア戦争は，東地中海へのロシアの南下を阻止するためイギリスとフランスなどが，1854 年オスマン帝国側について参戦した。

これらの問題を解消し，可読性の高い文章を解答するためには，

[1] https://akahon.net/

イギリスは第1回万国博覧会を開催して圧倒的な国力を誇示し，ヨーロッパではクリミア戦争に介入してロシアの南下を阻止するなど外交の主導権を握ったが、これに対しロシアでは戦後農奴解放令を発布して近代化を模索し、ドイツは...

図 2: 2008 年度の問 1 に対する模範解答の一部

1. 見出し語を語釈文に埋め込むことができるか，否かを判定する．埋め込めることができるのであれば，見出し語の表層格を推定する．
2. 問題文ならびに論述文章の前後の文等から何を主題にするかを決定し，それに応じて格交替などを行い論述問題の解答の一部とする．

ことが必要であると考えられる．

そこで本研究では，先に示した問題を解消するための第一段階として，語釈文中の動詞に着目し，見出し語に照応するゼロ代名詞の表層格の推定を行う．

2 世界史用語集

株式会社山川出版社・世界史 B 用語集　改訂版 [9] について見出し語と語釈文との間の照応に関する分析を行った．分析対象には用語集の見出し語 7037 語の中から 70 語間隔で抽出した見出し語 100 語とその語釈文を用いた．見出し語と語釈文中の単文の関係を人手で分類した結果を表 1 に示す． また，表 1 の分類

表 1: 語釈文中の単文と見出し語との間の関係の分類ならびにその出現数

分類	出現数
1. 単文においてゼロ代名詞化された格要素の一つが，見出し語に照応する	207
2. 単文が，主題がゼロ代名詞化されたコピュラ文で，ゼロ代名詞が見出し語に照応する	89
3. 単文中に見出し語そのものが現れている	6
4. 単文に現れる代名詞が見出し語に照応する	3
5. 1〜4 のいずれでもない	156

1 には動詞・形容詞・形容動詞・名詞に関わる格要素がゼロ代名詞化されている場合があり，それらの出現数については表 2 の通りである． 表 2 より，見出し

表 2: 語釈文中の単文においてゼロ代名詞化された格要素と見出し語との間の分類ならびにその出現数

分類	出現数
A. 動詞のいずれかの格要素になっている	190
B. 形容詞・形容動詞のガ格になっている	7
C. 名詞のノ格になっている	11

語を語釈文に埋め込むことができる場合，その多くは語釈文中の動詞のいずれかの格要素になっている．本研究では，語釈文中の各動詞について，ゼロ代名詞化された格要素の一つが，見出し語に照応しているか否か，すなわち，表 2 の A であるか否かを判定し，表 2 の A である場合はさらにその表層格を推定する．埋め込める場合の表層格は，「ガ格」「ヲ格」「ニ格」「デ格」「ト格」「カラ格」「ガ 2 格[2]」となるケースがあったが，「ト格」「カラ格」「ガ 2 格」となるケースは非常に少なかった．そのため本研究では「埋め込めない」「ガ格」「デ格」「ヲ格」「ニ格」「それ以外の格」のうちのどれであるかを推定する．また，表 1 の分類 2 に属すコピュラ文については，体言止めの文でありその体言がサ変名詞以外であるか否か，という簡単な条件で判定できる[3]．その場合は，助詞「は」を添えて見出し語を文頭に置くだけで良いので，本研究の扱うタスクの対象としない．

[2]文献 [4]2 節，「二重主語構文」を参照されたい．

3 日本語ゼロ照応解析

日本語では格要素の省略が頻繁に起きることから，日本語ゼロ照応解析に関する研究が多く行われている．飯田ら [7] は，文内ゼロ照応を対象とし，文の構造を素性として取り入れた機械学習によるモデルを提案している．笹野ら [8] は文内，文間両方のゼロ照応を対象とし，構文的手掛かりおよび大規模格フレーム [4][5] による語彙的手掛かりを素性とした対数線形モデルに基づく日本語ゼロ照応解析モデルを提案している．

これらの関連研究はいずれも Web テキストや新聞記事などの文章中のゼロ照応解析を行っているのに対し，本研究では世界史用語集を対象とし，質問応答における解答の流暢性の向上を目的としている．タスクとしては，一般的なゼロ照応解析では先行詞とその表層格を同定しているのに対し，本研究では先行詞候補が用語集の見出し語に限られているという点で異なる．また，先行詞である見出し語が文中に存在するわけではないので，照応解析において有効な手がかりとなる先行詞の文脈情報がまったく使えないという点が特徴的である．

4 カスケード型分類器による見出し語に照応するゼロ代名詞の表層格の推定手法

4.1 全体の処理の流れ

本研究の扱うタスクにおいては，用語集の見出し語と語釈文を入力とし，語釈文中の各動詞句[4]に対して「埋め込めない」「ガ格」「デ格」「ヲ格」「ニ格」「それ以外の格」のいずれかを出力する．その際，対象となる動詞句が受身形，および使役形の場合であっても原型に戻さず，それぞれの態のままで表層格の推定を行う．入力と出力の例を図 3 に示す．このタスクを実現するにあたって，本研究では入力された語釈文に対して KNP[6] により格構造解析を行った後，各動詞句に対して「埋め込めない」「ガ格」「デ格」「ヲ格」「ニ格」

[3]一般的にコピュラ文は名詞＋判定詞（だ）で終わるが，世界史用語集においては判定詞（だ）が省略されている．例，見出し語）シュメール人，語釈文）メソポタミア南部で最古の文明を築いた民族。また，サ変名詞の場合はコピュラ文にならない．例，見出し語）ボンディシェリ，語釈文）1672〜74 年フランスが獲得

[4]KNP の解析の結果，< 用言:動 > タグがふられた句を動詞句としている．

【入力】
見出し語）聖職売買
語釈文）教会が腐敗した9〜10世紀にさかんに行われた。1075年，教皇グレゴリウス7世が禁止したが，以後もつづいた。
【出力】
1. 腐敗する＝＞埋め込めない
2. 行われる＝＞ガ格
3. 禁止する＝＞ヲ格
4. 続く＝＞ガ格

図 3: 本研究が扱うタスクにおける入出力例

「それ以外の格」の分類を，多クラス分類に応用したカスケード型分類器により段階的に行っていく手法を提案する．処理の流れを図 4 に示す．各 2 値分類器にお

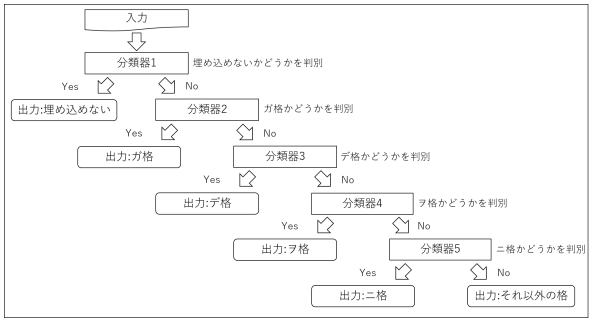

図 4: カスケード型分類器による表層格推定の流れ

いて「Yes」と判別された場合はその分類器が「Yes」とするラベル (例えば「ガ格」) を出力し，「No」と判別された場合は次の分類器で判別を行う．最後まで一貫して「No」となったものを「それ以外の格」として出力する．各 2 値分類器には，Support Vector Machine を使用する．

また，それぞれの分類器が学習する訓練事例は，それ以前の分類器によって「Yes」と判別されるべきデータに該当しないものである[5]．

4.2 使用する素性

飯田ら [7] によると，ゼロ代名詞照応解析に使用する素性は一般的に以下の 3 種類に分けられる．

1). 対象となるゼロ代名詞を持つ述語の語彙，統語情報に関する素性
2). 先行詞候補に関する語彙，統語，意味，位置情報に関する素性
3). 対象となるゼロ代名詞を持つ述語と先行詞候補の対から抽出可能な情報に関する素性

語釈文におけるゼロ代名詞照応の場合，先行詞である見出し語に文脈がないため，これらのうち 2) に含まれる先行詞候補に関する統語，位置情報に関する素性を扱うことはできない．本研究では 1) に分類される素性として「節の種類」「格の埋まりやすさ」「候補格かどうか」，2) に分類される素性として「見出し語の意味カテゴリ」，3) に分類される素性として「意味クラス PMI」「意味的な類似度」の合計 6 種類の素性を使用する．「格の埋まりやすさ」「候補格かどうか」「意味クラス PMI」「意味的な類似度」については，KNP の解析時に使用された格フレームをもとに，「ガ格」「デ格」「ヲ格」「ニ格」それぞれに設定するが，格解析によって既に対応付けられている格については値を 0 とする．また，カスケード型分類器において，すでに分類が終わった格の素性は除外する[6]．

節の種類

ゼロ代名詞を持つ述語の統語的選好を考慮するため，節の種類を素性として使用する．節の種類とは，注目する述語がどの種類の節に含まれているかであり，「主節」「連用節」「連体節」のいずれかになる．KNP の解析結果から得られる．

格の埋まりやすさ

格フレームのそれぞれの格の埋まりやすさは，その格が明示されていない場合に，ゼロ代名詞として省略されているのか，単にその格が考慮されていないのかの判断の手がかりになると考えられる．そこで，格の埋まりやすさの素性として格スロット生成確率[7]を用いる．

候補格かどうか

KNP の格解析に使用された格フレームでは必須格となっているが，文中に対応する格要素が存在しない格に対し 1，それ以外の格に対し 0 を設定する．

見出し語の意味カテゴリ

どのような意味カテゴリに属する見出し語に対する語釈文であるかによって，語釈文中のある格での省略されやすさに偏りがあると考えられる．世界史用の固有表現辞書[8]を登録した MeCab により見出し語を形態素解析し，固有表現辞書の 17 種類の固有表現クラスのうちどのクラスに属するかを素性として使用する．

意味クラス PMI

笹野ら [8] と同様に，見出し語の意味クラスと対象の格スロットの意味クラス情報との間の自己相互情報量 (PMI) を素性として使用する．ここで，意味クラスには日本語語彙大系 [1] において深さ 5 の位置にある意味属性を使用する．また，世界史用語集の見出し語が日本語語彙大系にそのまま登録されていることは稀であるため，見出し語の汎化を行う必要がある．見出し語の汎化には世界史用の固有表現辞書のサブクラスにあたる語を使用する．

意味的な類似度

[5] 例えば，「デ格」を判別する分類器 3 においては，「デ格」を「Yes」，「ヲ格」「ニ格」「それ以外の格」が「No」である訓練事例により学習させ，「埋め込めない」「ガ格」が「Yes」である訓練事例は利用しない．

[6] 例えば，「ヲ格」かどうかを分類する際には「ガ格」「デ格」に関する素性は使用しない．

[7] 文献 [8]，5 節の対象格の埋まりやすさ，$P(A(s_j)=1|cf_i,s_j)$ を使用．

[8] 文献 [3]，5 節を参照されたい．

ある格スロットにどのくらい先行詞候補が入りやすいかを考慮するため，世界史用の固有表現辞書により得られた見出し語のサブクラスにあたる語と，対象の格スロットの用例との意味的な類似度を日本語語彙大系により計算し[9]，最も値の大きいものを素性として使用する．意味クラス PMI と異なり，用例の出現頻度を考慮せず，1 つでも見出し語と意味が類似している語があれば大きな値をとる．

5 評価実験

5.1 使用するデータと実験設定

データセットは，世界史用語集中の見出し語 900 語[10]に対応する語釈文中のすべての動詞句に対し，見出し語がどの表層格に埋め込むことができるか，もしくは埋め込むことができないかの正解データを人手で作成したものを使用する．実験は，すべての分類器をカスケード型に連結したものによる多クラス分類と，分類器 1...N-1 がそれぞれ 100%の精度で分類できると仮定したときの分類器 N による 2 値分類をそれぞれ 10 分割交差検定により行った．

5.2 実験結果

カスケード型分類器による実験結果を表 3 に示す．全体の再現率と適合率算出時の分母が異なっているのは，ト格，カラ格，ガ 2 格を推定対象から除外しているためである．全体の F 値は 0.665 であり，改善の余地があると思われる． 次に，それぞれの分類器による

表 3: カスケード型分類器による表層格推定結果

格	再現率	適合率	F 値
埋め込めない	0.700(967/1382)	0.663(967/1459)	0.681
ガ格	0.722(1017/1409)	0.748(1017/1360)	0.735
デ格	0.080(10/125)	0.137(10/73)	0.101
ヲ格	0.241(19/79)	0.200(19/95)	0.218
ニ格	0.182(8/44)	0.151(8/53)	0.165
全体	0.665(2021/3039)	0.665(2021/3040)	0.665

実験結果を表 4 に示す．カスケード型に連結した際には低かった「デ格」「ヲ格」「ニ格」はいずれも単体の分類器では F 値が 0.7 以上となっている．「埋め込めない」を分類する分類器の F 値が低いことから，「埋め込めない」を分類する分類器の精度を上げる必要があると思われる．

表 4: 各分類器による表層格推定結果

格	再現率	適合率	F 値
埋め込めない	0.700(967/1382)	0.663(967/1459)	0.681
ガ格	0.840(1183/1409)	0.949(1183/1247)	0.891
デ格	0.760(95/125)	0.785(95/121)	0.772
ヲ格	0.886(70/79)	0.854(70/82)	0.870
ニ格	0.886(39/44)	0.812(39/48)	0.848

5.3 素性の有効性

使用した 6 つの素性の有効性を考察するため，素性を 1 つずつ除いて各分類器による実験を行った．実験結果を表 5 に示す．

[9]文献 [4] 付録，式 (1) を使用.
[10]分析に使用した 100 語と，無作為に抽出した 800 語を用いた.

表 5: 各分類器において素性を 1 つずつ除いた場合の F 値

使用しない素性タイプ	埋め込めない	ガ格	デ格	ヲ格	ニ格
全ての素性を使用	0.681	0.891	0.772	0.870	0.848
節の種類	0.651	0.882	0.727	**0.848**	0.842
格の埋まりやすさ	0.669	0.884	0.750	0.871	0.813
候補格かどうか	**0.638**	0.881	0.743	0.859	0.848
見出し語の意味カテゴリ	0.660	**0.854**	**0.710**	0.885	**0.791**
意味クラス PMI	0.679	0.890	0.748	0.875	0.848
意味的な類似度	0.675	0.879	0.779	0.882	0.870

全ての素性を使用した場合の F 値と，ある 1 つの素性を除いた場合の F 値を比較することで，各分類器毎にどの素性が有効であるかを考察することができる．全体としては「見出し語の意味カテゴリ」の素性が有効であり，見出し語の意味カテゴリによって，語釈文の書かれ方に傾向があるためだと考えられる．また，「格の埋まりやすさ」「意味クラス PMI」「意味的な類似度」の素性はそれぞれ除いてもあまり F 値に変化は見られなかった．

6 まとめ

本研究では，文章中のゼロ照応解析と異なり，見出し語と語釈文に分かれている文書データに対し，語釈文における見出し語に照応するゼロ代名詞の表層格の推定を行った．

今後の課題としては，「埋め込めない」を判別する分類器の精度の向上や，見出し語が形容詞・形容動詞・名詞に関わる格要素ゼロ代名詞化しているか否かの判定を行うシステムの構築などが挙げられる．

参考文献

[1] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系　岩波書店，1997

[2] 阪本浩太郎，中山周，渋木英潔，石下円香，森辰則，神門典子．東大入試世界史第 1 問 （大論述問題）を解く質問応答システムの検討，言語処理学会　第 22 回年次大会　発表論文集，2016.

[3] 石下円香, 阪本浩太郎, 中山周, 渋木英潔, 森辰則, 神門典子．東大入試世界史第 2 問 (小論述問題) 及び第 3 問 (語句問題) を解く質問応答システムの検討　言語処理学会　第 22 回年次大会　発表論文集，2016

[4] 河原大輔, 黒橋禎夫, 格フレーム辞書の漸次的自動構築, 格フレーム辞書の漸次的自動構築 自然言語処理, Vol.12，No.2，pp.109-131, 2005.

[5] 河原大輔，黒橋禎夫，高性能計算環境を用いた Web からの大規模格フレーム構築，情報処理学会 自然言語処理研究会 171-12，pp.67-73，2006.

[6] 河原大輔，黒橋禎夫，自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル，自然言語処理, Vol.14，No.4，pp.67-81，2007.

[7] 飯田龍，乾健太郎，松本裕治，文の構造を利用した文内ゼロ照応解析 言語処理学会第 12 回年次大会，pp.488-491，2006.

[8] 笹野遼平，黒橋禎夫，大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析，情報処理学会論文誌 Vol.52，No. 12，pp.3328-3337 2011.

[9] 株式会社山川出版社・世界史 B 用語集　改訂版　2008.