# 世界史用語集のゼロ代名詞の表層格推定における自動生成された擬似訓練データの利用

大矢 康介 †1　阪本 浩太郎 †1　渋木 英潔 †2　森 辰則 †2

†1 横浜国立大学 大学院 環境情報学府
†2 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 [1] 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，教科書や用語集等の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．知識源に使用される用語集は見出し語と語釈文に分かれて構成されており，語釈文には見出し語が明示されていないため，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．

そこで，用語集を利用し，可読性の高い文章を解答するためには，

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．我々はこの課題を解消するための第一段階として，図 1 に示すタスクを設定し，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った [2]．

【入力】
見出し語）聖職売買
語釈文）教会が腐敗した9～10世紀に盛んに行われた。
【出力】
1. 腐敗する＝＞埋め込めない
2. 行われる＝＞ガ格

図 1: 大矢らのタスクにおける入出力例

しかしこの研究では，表層格推定の手がかりとなる素性を抽出する際に，格助詞を付与した見出し語を語釈文に埋め込んで作られる文 (以下，「見出し語を埋め込んだ文」という．) 全体の適格性を考慮せず，そのままの語釈文のみを解析し，素性抽出を行なっている．すなわち，図 2 の語釈文を見出し語とは独立に解析して素性抽出を行い，埋め込まれる見出し語の持つ素性と組み合わせることにより間接的に埋め込み文の適格性を判断していたが，図 3 のように，見出し語を埋め込んだ状態の文の適格性評価に直接つながる素性は考慮していなかった．これにより，3 節で述べる「格フレーム誤り」などの問題が起こり，精度低下の原因の 1 つとなっていると考えられる．

教会が腐敗した9～10世紀にさかんに行われた.

図 2: 「聖職売買」のそのままの語釈文の例

教会が腐敗した9～10世紀にさかんに聖職売買が行われた.
教会が腐敗した9～10世紀にさかんに聖職売買を行われた.
教会が腐敗した9～10世紀にさかんに聖職売買に行われた.
教会が腐敗した9～10世紀にさかんに聖職売買で行われた.

図 3: 格助詞を付与した見出し語を語釈文に埋め込んで作られた文の例

また，「ヲ格」「ニ格」「デ格」に関して精度が非常に低い値となっているが，この原因の 1 つとして訓練データがかたよっており，用語集において見出し語を「ヲ格」「ニ格」「デ格」に埋め込める事例数が少ないことが挙げられる．訓練データを増やすことで「ヲ格」「ニ格」「デ格」の精度向上が見込めるが，用語集を用いて人手により訓練データを作成するのには大きなコストがかかってしまう．

そこで本研究では，見出し語を埋め込んだ文を考慮した素性を導入することで全体の精度向上を目指す．さらに「ヲ格」「ニ格」「デ格」の精度向上のために，世界史教科書から擬似的な訓練データを自動生成する手法を提案する．

## 2　関連研究

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．笹野ら [3] は文内，文間両方のゼロ照応を対象とし，構



Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

文的手掛かりおよび大規模格フレーム [4] 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．

笹野らの手法では，注目する用言と直接係り受け関係にある語を格スロットに対応づけた後，対応づけられなかった格スロットに対し直接係り受け関係にない談話要素の対応づけを行ない，ゼロ照応を考慮した対応づけ候補を生成している．本研究においても，見出し語を用言の格スロットに対応づけた際の素性を考慮する．

また，機械学習手法を用いた研究において正解情報を人手で付与するのには膨大なコストがかかるため，擬似訓練データを利用した研究も行われている．大森ら [5] は Q&A サイトの質問分類において，Q&A に対して擬似的な分類ラベルを付与したデータを利用して特徴表現の収集を行なっており，訓練データ作成のコストを抑えている．本研究においては，用語集の語釈文が，元の文から「見出し語が抜けた文」であることを利用し，教科書中の文から用語集の語釈文に近い文を生成することで擬似訓練データを自動生成することで，訓練データ不足の課題を解消することを目指す．

# 3　大矢ら [2] における課題

大矢ら [2] による機械学習手法を用いた表層格推定の評価実験における失敗分析を行う．見出し語 90 語分のデータセットにおける調査を行なった結果，失敗事例のパターンおよび出現数は表 1 の通りであった．

表 1: 失敗事例のパターン別出現数

| 失敗事例のパターン | 出現数 |
|---|---|
| 格フレーム誤り | 33 |
| 固有表現辞書に登録されていない | 13 |
| 見出し語に照応する代名詞が注目する動詞の格要素になっている | 11 |
| 注目する動詞が複合格助詞の一部となっている | 7 |
| 名詞のノ格がゼロ代名詞化されていて見出し語に照応する | 2 |

このうち最も多いのが格フレーム誤りである．格フレーム誤りとは，語釈文の格構造解析に用いた KNP の格解析結果として提示された格フレームの必須格の集合に，正解となる格が含まれない場合である．例えば，図 4 において KNP の格解析に使用された格フレームは「独立」の格フレーム 4 であり必須格はデ格のみであるが，正解となる格はガ格となっている．このような事例では，正解の格に関する「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」の素性値が 0 となってしまうため，上手く分類できなくなってしまうと考えられる．これは，見出し語を埋め込んだ文を考慮せず，そのままの語釈文を解析した際に適用された格フレームを素性抽出に利用していることが問題である．

さらに，大矢らのモデルでは事例数の多い「埋め込めない」や「ガ格」の精度を向上させることで全体の精度向上を目指しているが，阪本らの手法[1]を適用した時に特に意味がわかりにくくなってしまう事例はむしろ，見出し語が「ヲ格」「ニ格」「デ格」など「ガ格」以外の格に埋め込むことができる場合である．例えば，図 5 は「聖職売買」が「行われた」の「ガ格」に埋め込める場合であるが，この場合阪本らの手法においても，生成された文単体としては，適格な文となっている[2]．しかし，図 6 のように「クリミア戦争」が「参戦した」の「ニ格」に埋め込める場合は，阪本らの手法を適用すると文単体として意味がわかりにくい文になってしまう．

**【用語集の記述】**
**見出し語）チャンパー**
**語釈文）初め中国の統治下にあったが，192年頃後漢より独立した**
**【期待される出力】**
**独立する＝＞ガ格**
**【選択された「独立」の格フレーム4】**
**｛会社+法：243｝で　独立する**

図 4: 格フレーム誤りの例

**聖職売買は，教会が腐敗した9〜10世紀にさかんに行われた．**

図 5: 見出し語が「ガ格」に埋め込める時に阪本らの手法を適用した場合

**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．**

図 6: 見出し語が「ニ格」に埋め込める時に阪本らの手法を適用した場合

そのため「ヲ格」「ニ格」「デ格」の精度向上は重要な課題であるが，大矢らのモデルでは「ヲ格」「ニ格」「デ格」の精度は非常に低い．その原因の 1 つとして，表 2 に示すように，学習に用いることのできるデータ量が「埋め込めない」「ガ格」と比較して少ないことが挙げられる．相対的に割合の低い事例の数を増やすためには，非常に多くの語釈文を正解データとして人手で解析しておく必要があり[3]，非常に大きなコストがかかってしまう．

表 2: 見出し語 900 語文の語釈文における正解ラベル毎の事例数

| 正解ラベル | 埋め込めない | ガ格 | ヲ格 | ニ格 | デ格 | その他の格 |
|---|---|---|---|---|---|---|
| 事例数 | 1399 | 1432 | 82 | 44 | 134 | 31 |

これらのことから本研究では，見出し語を埋め込んだ文を考慮して格フレームの再選定を行い，再選定された格フレームを利用した素性を用いることにより全体の精度向上を目指す．また，学習データ不足の問題を解消し，「ヲ格」「ニ格」「デ格」の精度を向上させるために，教科書から擬似訓練データを自動生成する手法を提案する．

# 4　ベースライン手法

本研究の提案手法は大矢らの手法をベースラインとしている．本節ではこのベースライン手法について簡単に説明する．

[1]「見出し語＋は、」を文頭に付け加える

[2]ただし，格の推定を行っていないので，主題の変更などは行えない．

[3]例えば，「ニ格」の正解データを 500 個作成したい場合，「埋め込めない」「ガ格」の正解データをおおよそ 3 万個作成しなければならない



Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

### 4.1 全体の処理の流れ

見出し語に照応するゼロ代名詞の表層格の推定を行うにあたって，入力された語釈文に対して KNP[4] により格構造解析を行った後，各動詞句に対して KNP の解析結果や，KNP を使用した際の京大格フレーム [4]，日本語語彙大系 [7] などの言語資源を用いて素性抽出を行い，「埋め込めない」「ガ格」「ヲ格」「ニ格」「デ格」「それ以外の格」の多クラス分類を行う．また，多クラス分類には，Support Vector Machine を one-versus-the-rest 法により多クラス分類に拡張したモデルを使用する．

### 4.2 使用する素性

「節の種類」「格の埋まりやすさ」「見出し語の意味カテゴリ」「意味クラス PMI」「意味的な類似度」の 5 つの素性に加え，一般的にゼロ代名詞照応解析に使用される 2 値素性である「直前格」「埋まっているか」「複合格助詞」「態」の 4 つの素性を使用する．表 3 に使用する素性の一覧と，それらの抽出に用いる言語資源を示す．

表 3: 使用する素性

| 素性 | 使用する言語資源 |
|---|---|
| 節の種類 | KNP |
| 格の埋まりやすさ | 京大格フレーム |
| 見出し語の意味カテゴリ | 世界史辞書[5] |
| 意味クラス PMI | 京大格フレーム，日本語語彙体系，世界史辞書 |
| 意味的な類似度 | 京大格フレーム，日本語語彙体系，世界史辞書 |
| 直前格 | 京大格フレーム |
| 埋まっているか | KNP |
| 複合格助詞 | KNP |
| 態 | KNP |

## 5 提案手法

本研究では，全体の正答率の向上のために，「KNP スコア差分」素性および「再選定された格フレーム」に関する素性を導入する．また，「ヲ格」「ニ格」「デ格」の精度向上のために，世界史教科書から擬似訓練データを生成する手法を提案する．

### 5.1 見出し語を埋め込んだ文の適格性を考慮した素性

**KNP スコア差分**

見出し語を埋め込んだ文を KNP で解析した時のスコアから，元の語釈文を KNP で解析した時のスコアを引いたものを「KNP スコア差分」とする．KNP スコアは KNP による構文・格解析結果のもっともらしさであり，このスコアが高いほど日本語として出現しやすい文であると言える．そのため，埋め込みの前後でのスコアの差は，埋め込みの良さを表すと考える，表 4 に KNP スコア差分の算出の例を示す．この例では，KNP スコア差分の値が最も大きい (絶対値が一番小さい) ガ格が適切であると判断される．

表 4: KNP スコア差分算出例

| 見出し語を埋め込んだ文 | KNP スコア | KNP スコア差分 (素性) |
|---|---|---|
| 教会が...聖職売買が行われた | -58.5 | -58.5-(-43.2)=-15.3 |
| 教会が...聖職売買を行われた | -61.4 | -61.4-(-43.2)=-18.2 |
| 教会が...聖職売買に行われた | -61.1 | -61.1-(-43.2)=-17.9 |
| 教会が...聖職売買で行われた | -59.9 | -59.9-(-43.2)=-16.7 |

**再選定格フレームに関する素性**

見出し語を埋め込んだ文を KNP により解析し，その時に使用された格フレームを使用し，表 3 における京大格フレームを使用して抽出する素性「格の埋まりやすさ」「意味クラス PMI」「意味的な類似度」「直前格」を再計算・再抽出したものを「再選定格フレーム」素性とする．

見出し語を埋め込んだ文を考慮したこれら 2 つの素性を使用することにより，格フレーム誤りの問題を解消し，精度が向上することが期待される．

### 5.2 擬似訓練データの使用

東京書籍株式会社・世界史 B[9] を使用して擬似訓練データを自動生成する．生成手順としては，教科書中の文を句点区切りで分割し，分割されたテキストごとに以下の処理を行う．

1. 「を」「に」「で」のいずれかでマークされた世界史固有表現に注目し，2，3 の処理を行う．
2. その世界史固有表現を文中から取り除き，見出し語とする．
3. 直後の「を」「に」「で」のいずれかの格助詞を取り除き，正解ラベルとする．

図 7 に擬似訓練データの自動生成の例を示す．

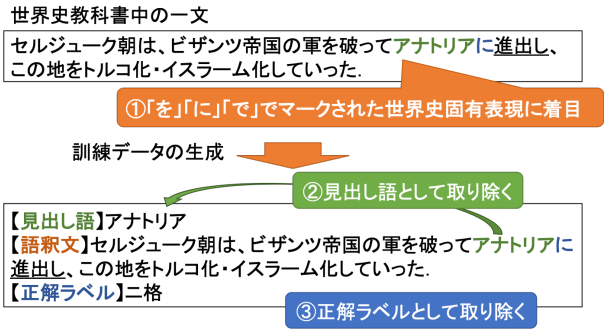


図 7: 擬似訓練データの自動生成の例

この手法によって生成された擬似訓練データは，大きく分けて 3 つのパターンに分けられる．わかりやすさのために，図 7 同様に着目する世界史固有表現を緑，正解ラベルとして取り除く格助詞を青とし，着目した固有表現が 1 つの項となっている述語に下線を引いて図 8，9，10 に 3 つのパターンの例を示す．

セルジューク朝は、ビザンツ帝国の軍を破って＜アナトリアに＞進出し、この地をトルコ化・イスラーム化していった.

図 8: パターン A

[4] http://nlp.ist.i.kyoto-u.ac.jp/?KNP



Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

<モンゴルに>蹂躙されたバグダードにかわって、カイロがイスラーム世界の内外を結ぶ国際交易の中心となった

図 9: パターン B

シーア派に対し、ムハンマドのスンナを尊重し、正統カリフ、ウマイヤ朝、アッバース朝と続いた<カリフを>正統な指導者と肯定する党派をスンナ派という

図 10: パターン C

パターン A は着目している文節を取り除いても特に問題がないものであるが，パターン B の場合には取り除きたい文節の係先 (図 9 黄マーカ部) が連体節となっている．この場合，「ガ格」の項が存在しないために不自然な文となってしまう．また，パターン C の場合には取り除きたい文節にかかる文節 (図 10 赤マーカ部) がかかっているため，「カリフを」を取り除くと非文になってしまう．そこで本研究では自動生成された擬似訓練データのうち，パターン C を使用せず，パターン A，B のみを使用する．

# 6 評価実験

## 6.1 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[6] 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．また，実験は 10 分割交差検定により行い，擬似訓練事例は訓練事例のみに含まれており評価事例には含まれていない．

## 6.2 実験結果

提案手法を適用して評価実験を行なった結果を表 5 に示す．ベースラインに「KNP スコア差分」の素性を追加しても全体の F 値に変化はなかったが，再選定格フレームによる素性を使用することで全体の正答率が向上した．これにより，見出し語を埋め込んだ文を考慮した格フレームの再選定が精度向上に有効であることがわかる．

表 5: 使用した素性ごとの F 値

| | ベースライン | ベースライン +KNP スコア差分 | ベースライン +KNP スコア差分 +再選定格フレーム |
|---|---|---|---|
| 埋め込めない | 0.778 | 0.783 | 0.788 |
| ガ格 | 0.842 | 0.840 | 0.847 |
| ヲ格 | 0.358 | 0.346 | 0.400 |
| ニ格 | 0 | 0 | 0 |
| デ格 | 0 | 0 | 0 |
| その他の格 | 0.708 | 0.638 | 0.681 |
| Micro F | 0.777 | 0.777 | 0.784 |

次に，擬似訓練データを使用して評価実験を行なった結果を表 6 に示す．また，今回は事例数の少ない「ヲ格」「ニ格」「デ格」の精度も重視したいため，全体結果を「Micro F」「Macro F」で評価する．パターン A のみを使用した場合には Macro F が改善されたが，パターン B を合わせると全体的に精度が下がっている．このことから，質の高い擬似訓練データを使用することで，Micro F を犠牲に Macro F を改善することができるということがわかる．

表 6: 擬似訓練データのパターンごとの F 値

| | 擬似訓練データなし | パターン A のみ | | パターン A+B | |
|---|---|---|---|---|---|
| | F 値 | 追加した擬似訓練事例数 | F 値 | 追加した擬似訓練事例数 | F 値 |
| 埋め込めない | 0.788 | | 0.780 | | 0.774 |
| ガ格 | 0.847 | | 0.844 | | 0.842 |
| ヲ格 | 0.400 | 201 | 0.483 | 315 | 0.452 |
| ニ格 | 0 | 186 | 0.215 | 256 | 0.150 |
| デ格 | 0 | 40 | 0.043 | 59 | 0.041 |
| その他の格 | 0.681 | | 0.717 | | 0.708 |
| Micro F | 0.784 | | 0.773 | | 0.766 |
| Macro F | 0.453 | | 0.514 | | 0.494 |

# 7 まとめ

本研究では，見出し語を埋め込んだ文を考慮した素性を利用することにより，精度が向上することが確認された．また，擬似訓練データを使用することで「ヲ格」「ニ格」「デ格」の精度が向上することが確認されたが，質の低い擬似訓練データを追加した際には精度が下がったため，今後はさらに質の高い擬似訓練データに絞る方法を検討する必要があると考えられる．

## 謝辞

本研究の一部は，JSPS 科研費 16K00296 の助成を受けたものである．

# 参考文献

[1] 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.

[2] 大矢康介，阪本浩太郎，渋木英潔，森辰則，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 言語処理学会第 24 回年次大会，pp.492-495，2018.

[3] 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.

[4] 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.

[5] 大森勇輔，森田和宏，泓田正雄，青江順一，擬似訓練データを用いた Q&A サイトの質問分類，言語処理学会　第 21 会年次大会　発表論文集，2015.

[6] 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

[7] 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

[8] 株式会社山川出版社・世界史 B 用語集　改訂版　2008.

[9] 東京書籍株式会社・世界史 B　2007.

[6] 大矢らの研究と同じ 900 語



Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.