大学入試の穴埋め型問題に対する語順を考慮した自動解答手法

田上　諒† 木村　輔†† 宮森　恒††

† 京都産業大学コンピュータ理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都産業大学大学院先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g1344775,i1658047}@cse.kyoto-su.ac.jp, ††miya@cc.kyoto-su.ac.jp

あらまし　近年，ユーザからの多様な情報要求を満たす技術として，質問応答などの自動解答技術が注目されている．しかし，それらの技術は，大学入試をはじめとする現実に即した多様で複雑な質問に対して，現状では十分に対応できているとは言い難い．例えば，大学入試等における文書中の空欄部分の単語を解答するような穴埋め型問題に対して，従来手法では，主に語順を考慮しない検索ベースのファクトイド型解答技術が用いられているため，十分な正答率を得られていない．本稿では，大学入試二次試験の世界史 B 穴埋め型問題を対象とし，語順を考慮した自動解答手法を提案する．具体的には，まず，解答カテゴリの推定の際に，語順を考慮した分散表現を用いる手法を導入する．また，穴埋め語を選択する際に，従来手法にはない，いくつかの条件を用いたスコアリング手法に改良する．実験では，ベースライン手法と提案手法を比較し，語順を考慮したうえで周辺単語から中心単語を予測するモデルを，解答カテゴリ推定に取り入れることで，正答率にどのような変化があるかを明らかにする．

キーワード　自然言語処理, factoid 型質問応答，自動解答，大学入試問題, 分散表現，語順

1. はじめに

近年，ユーザからの多様な情報要求を満たす技術として，質問応答などの自動解答技術が注目されている．大量に存在する情報源の中から自身が必要な情報を得る手段としては，関連するキーワードをクエリとして文書検索を行い，検索結果となる複数文書から得たい情報を探し出す方法が一般的である．しかし，この方法は，クエリ生成の過程や，複数文書の中から要求を満たす情報を選択する過程をユーザ自身で行わなければならない．それに対し質問応答は，ユーザ自身の情報要求を自然言語で入力し，情報源から 1 つの解答を出力する技術である．自分の得たい情報を身近な言語で伝えることができ，かつ，複数の情報を比較する必要が無い点が特徴といえる．

質問応答で扱う質問は，人名や地域名など，短い語句による事実を解答すればよいファクトイド型と，ある語句の定義や手順などを説明する必要があるノンファクトイド型に大別することができる．ファクトイド型の解答技術に関する研究はこれまでに多く行われているが，NTCIR-13(注1) QA Lab-3 タスクが目的としているような，大学入試をはじめとする現実に即した多様で複雑な質問に対して，現状では十分に対応できているとは言い難い．

大学入試問題におけるファクトイド型の問題例を図 1 に示す．解答形式を記述式問題に限定すると，主に穴埋め（Slot Filling）型と事実（Factoid）型に分けることができる．一般的なファクトイド型質問応答の場合，図 1（b）のような事実型の質問を対象とする場合が多い．このような質問では，質問文の疑問詞に着目することによって「人の名前を聞いているのか」「地域名を聞いているのか」などの解答カテゴリを推定でき，さらに，疑問詞直前の単語を確認することによって「王の名前を聞いているのか」「神の名前を聞いているのか」などの質問の焦点を推定できる．解答カテゴリや質問の焦点を推定することにより，解答候補を効率よく絞り込むことができると考えられる．しかし，図 1（a）のような穴埋め型問題の場合，基本的には疑問詞が問題文中に出現することはないため，上述のような推定手法を用いることができない．

(a)穴埋め型問題の例

大問2 （抜粋）

次の短文(1～8)は，19世紀後半から20世紀初頭までのヨーロッパ各国で起きた出来事について述べたものである。空欄の(A)～(H)に適切な語句を入れ，また下記の【設問】に答えなさい。

1. ボスニアの(A)を訪問中の帝位継承者夫妻が暗殺されたため，その一か月後にセルビアに対して宣戦を布告した。
2. 革命運動が高まる中，皇帝は(B)の起草した十月宣言を発して国会(ドゥーマ)の開設を約束し，彼を首相に登用した。

解答：(A)サライェヴォ (B)ウィッテ

(b)事実型問題の例

大問1 (抜粋)

(2) 前13世紀前半にシリア北部のカデシュでヒッタイトと戦い，戦いの後ヒッタイト王と講和条約を結んだ新王国時代の王は誰か。

(3) アメンホテプ4世が唯一の神としたこの太陽神を何というか。

解答：(2)ラメス2世[ラメセス2世] (3)アトン

図 1　大学入試問題におけるファクトイド型問題の例
（2015 年度中央大学文学部世界史）

(注1)：NTCIR-13 : http://research.nii.ac.jp/ntcir/ntcir-13/

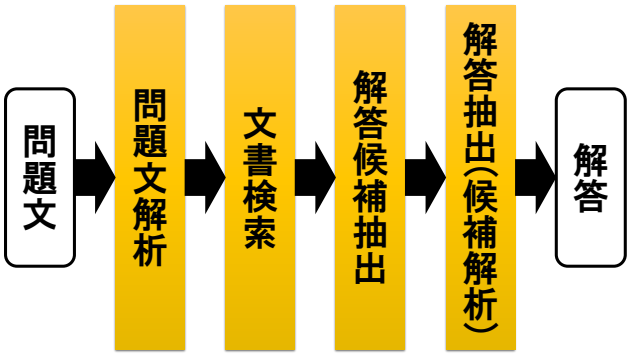


図 2　ファクトイド型質問応答システムの基本的な処理手順

ファクトイド型質問応答システムの基本的な処理手順を図 2 に示す．入力データとして 1 つの問題文が与えられ，問題文解析，文書検索，解答候補抽出，解答抽出 (候補解析) のモジュールの処理が順に実行され，解答が出力される．例えば，問題文解析モジュール内の処理の 1 つとして，解答カテゴリを推定するが，先述の通り，穴埋め型問題では同様の手法を用いることが困難なため，穴埋め型問題にも対応できる推定手法を取り入れる必要がある．解答抽出モジュールでは，解答候補ごとに，より正解に近いかどうかのスコアリングを行う．このモジュールにおいても，穴埋め型問題に対応した解析手法を取り入れる必要がある．

本稿では，大学入試二次試験の世界史 B における記述式の穴埋め型問題を対象とし，語順を考慮した自動解答の手法を提案する．なお，処理の基本的な流れは図 2 に沿ったものとする．提案手法のポイントは次の 2 点である．1 点目は，語順を考慮した単語の分散表現を用いる点である．あらかじめ，教科書から得られる文書群を用いて，周辺単語とその出現順序を入力することによって，中心単語を予測できるモデルを生成する．実際の解答処理では，問題の穴埋め部分の周辺単語とその出現順序をモデルに入力し，予測単語候補にはどのようなカテゴリの単語が含まれているかによって，問題文解析におけるカテゴリ推定の精度を向上させる．2 点目は，穴埋め型問題に適したスコアリングを導入する点である．穴埋め語を選択する際に，問題文非既出単語判定や後方一致判定など，従来手法にはないいくつかの条件を用いてスコアリングを行う．

実験では，語順を考慮しないベースライン手法と提案手法を比較し，正答率とどのような関係があるかを明らかにする．

本稿の構成は以下の通りである．2 章では本稿と関連する研究について述べ，3 章では提案手法について詳しく説明する．4 章では正答率に関する実験を実施し，5 章では実験結果を踏まえた考察を記す．最後に 6 章では，まとめと今後の課題について整理する．

2. 関 連 研 究

質問応答システムに関する研究は，これまでに数多く行われている．

David ら [1] は IBM において，オープンドメインのファクトイド型質問応答システムである Watson を開発した．質問応答の仕組みとして，情報源と統計情報から，仮説の生成と根拠の探索を行う DeepQA フレームワークを設計している．このフレームワークの処理の流れは，図 2 で示した，現在におけるファクトイド型質問応答システムの基本的な処理手順のベースとなっている．なお，このシステムは，米国のクイズ番組「Jeopardy!」において，実際に人間 2 名と対戦を行い，両者を突き放して勝利した．Iyyer ら [2] は，再帰ニューラルネットワーク（Recursive Neural Network）を用いたオープンドメインのファクトイド型質問応答システムを開発している．同ネットワークを使って問題文を構造木で表現し，解答を分類している．従来の質問応答システムの精度を上回った結果が示されている．

大学入試問題を対象とした質問応答システムに関する研究も，近年活発に行われている．

Takada ら [3] は，大学入試問題の自動解答について，特に論述問題に焦点を当てたシステムを開発している．文書検索部分では，知識源として用意した参考書内の各一文ごとに，対象問題文との類似度を，文中に出現する名詞から計算し，その類似度をスコアとして取り扱っている．その際，各名詞ごとに，世界史の単語集に掲載されている単語かどうかや，Wikipedia の記事になっているかどうかなどで，重要語句かどうかの度合いを推定したり，問題文中のどこにその名詞が出現するかを判定することによって，スコアリング時の重みを変えている．その後，時間関係が一致しているかどうかのラベリングや，文要約などの過程を経て，論述問題の解答を生成している．事実型問題においても，初めに解答単語のカテゴリ解析を行っているが，文書検索部分の処理については論述問題と同じ手法を用いている．なお，このシステムにおける手法は，事実型問題に特化したスコアリング手法を採用しているため，穴埋め型問題にそのまま適用することはできない．

Sakamoto ら [4] は，事実型問題と穴埋め型問題を，同一の単語回答問題とみなして処理を行っている．おおまかな処理の流れは図 2 に沿っており，問題文解析モジュールでは問題の焦点や解答タイプの推定が行われている．解答抽出モジュールでは，解答候補ごとに，抽出元の文に当該候補が含まれる度合いや解答タイプの一致度によってスコアリングし，スコアが最も高かったものを解答として出力している．この手法についても，基本的に事実型問題の特徴を考慮した処理となっているため，穴埋め型問題の特徴が生かされていない．

また，単語の分散表現についても様々な研究が行われている．

単語の分散表現学習ツールとして代表的なものに，Mikolov ら [5] が開発した word2vec がある．学習モデルとして幾つか用意されているが，その 1 つとして Continuous Bag-of-Words（CBOW）モデルがある．この学習モデルは，周辺単語から中心単語を予測する構造となっており，本稿が取り扱っている穴埋め型問題に，人間が回答する際の考え方と類似している．しかし，CBOW モデルは周辺単語の語順までは考慮されておらず，そのままでは文脈に適した中心単語を予測することは難しい．また word2vec 自体は，生成されたモデルを使用して周辺単語から中心単語を予測するようなタスクを必ずしも意図しているわけではない．

有賀ら [6] は，word2vec の CBOW モデルに語順情報を付加した新しい学習モデルを提案している．具体的には，中心単語の前側（左側）と後ろ側（右側）の周辺単語群を区別する Left and Right（LR）モデル，及び，周辺単語の各出現位置を全て区別する Word Order（WO）モデルの 2 つを挙げている．提案モデルによって中心単語の予測精度が向上したという結果が示されている．CBOW モデルとは違い，語順情報が付与されているため，文脈に適した中心単語を予測しやすくなると考えられる．

以上の関連研究を踏まえて，我々は穴埋め型問題に解答することができる自動解答手法を提案するにあたり，まず従来解答手法のスコアリング等を改良する．また，周辺単語から語順情報を保ったまま中心単語を予測する Word Order モデルが，人間が穴埋め型問題を解く際の考え方と類似しているため，同手法で構築されたモデルをシステム内における解答カテゴリの推定手段として活用する．

3. 提 案 手 法

3.1 提案手法の概要

大学入試の世界史問題における記述式穴埋め型問題を対象とし，図 2 のようなファクトイド型質問応答の処理手順をベースとした手法とする．主に，単語の語順を考慮したカテゴリ推定と，穴埋め型問題に対応した解答候補スコア計算を導入した手法を提案する．問題文解析モジュールでは，語順を考慮した単語の分散表現による単語予測モデルに，穴埋め部分の周辺単語とその出現位置を入力することで得られる予測単語候補から，解答となる穴埋め部分のカテゴリを推定する．また，解答抽出モジュールでは，問題文非既出単語判定や後方一致判定など，穴埋め型問題に対応したいくつかの条件を用いスコアを計算する．

3.2 節では，本稿で取り扱う国公立大学二次試験の世界史 B 問題について述べ，3.3 節及び 3.4 では事前に準備した知識源や辞書，語順を考慮した単語予測モデルを用いたカテゴリ推定について記述する．3.5 節以降では，各モジュールについて詳述する．

3.2 対象とする世界史の穴埋め型問題の概要

大学入試の世界史問題における記述式穴埋め型問題の例を図 3 に示す．

世界史の穴埋め型問題は，大学や出題年度によって多少異なるものの，基本的に，問題指示部と問題文脈部から構成される．問題指示部は，解答方法を指示する 1 つ以上の文からなり，問題文脈部は，複数の穴埋め部分が埋め込まれた複数の文から構成される．問題文脈部は，原則として同じテーマについて記述されており，異なるテーマが混在する例はほとんど存在しない．

3.3 世界史に関する辞書および知識源の構築

世界史問題の自動解答を行うにあたり，同科目に関する辞書や知識源を用意した．

まず，自動解答システムの処理過程において，形態素解析エンジンである MeCab(注2) を使用する．基本辞書として，Sato [7] が開発した，mecab-ipadic-NEologd を採用した．また，世界史に関する単語を適切に形態素解析できるようにするため，世界史分野に特化した固有名詞が含まれる辞書 [8] を独自に作成し，ユーザ辞書として用いた．これにより，「16 世紀初頭」「1945 年」などの時間に関する表現についても固有名詞として扱われる．

次に，本稿では以下の書籍を，主に文書検索モジュールで使用する知識源とした．

教科書　詳説 世界史（山川出版社）[9]

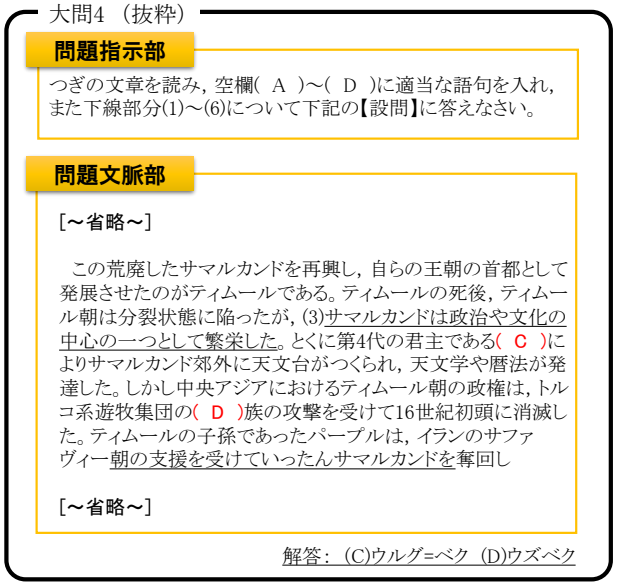
教科書　世界史 B（東京書籍）[10]

教科書　新選世界史 B（東京書籍）[11]

教科書　世界史 A（東京書籍）[12]

参考書　山川一問一答世界史（山川出版社）[13]

上記書籍を知識源として登録する際には，書籍中の本文 1 文ごとに 1 文書となるように登録した．


大問4 （抜粋）

問題指示部

つぎの文章を読み，空欄(A)～(D)に適当な語句を入れ，また下線部分(1)～(6)について下記の【設問】に答えなさい。

問題文脈部

［～省略～］

この荒廃したサマルカンドを再興し，自らの王朝の首都として発展させたのがティムールである。ティムールの死後，ティムール朝は分裂状態に陥ったが，(3)サマルカンドは政治や文化の中心の一つとして繁栄した。とくに第4代の君主である(C)によりサマルカンド郊外に天文台がつくられ，天文学や暦法が発達した。しかし中央アジアにおけるティムール朝の政権は，トルコ系遊牧集団の(D)族の攻撃を受けて16世紀初頭に消滅した。ティムールの子孫であったバーブルは，イランのサファヴィー朝の支援を受けていったんサマルカンドを奪回し

［～省略～］

解答：(C)ウルグ=ベク (D)ウズベク

図 3　大学入試世界史問題におけ記述式穴埋め型問題の例
（2014 年度中央大学文学部世界史）

3.4 語順を考慮した単語予測モデル

単語の語順を考慮したカテゴリ推定を行うため，word2vec の CBOW モデルをもとに有賀らが提案した，WO（Word Order）モデルを導入する．モデルの構築には，深層学習フレームワークである Chainer [14] を使用した．

ここで，CBOW モデルと，WO モデルについて説明する．それぞれの学習モデルの入力層から出力層の概要を図 4 に示す．ここで t は予測したい中心単語であり，$t \pm x$ は t の前後 x 番目の周辺単語であることを表す．図 4(a) の CBOW モデルの場合，中間層のベクトル H は周辺に存在する単語の位置関係を保たないまま生成されるため，語順が考慮されていない．(b) の WO モデルの場合，H は単語の位置関係を保ったまま生成されるため，語順を考慮して単語 t を予測することができる．

この学習モデルをもとに，単語の分散表現を用いて，周辺単

(注2)：MeCab : http://taku910.github.io/mecab/

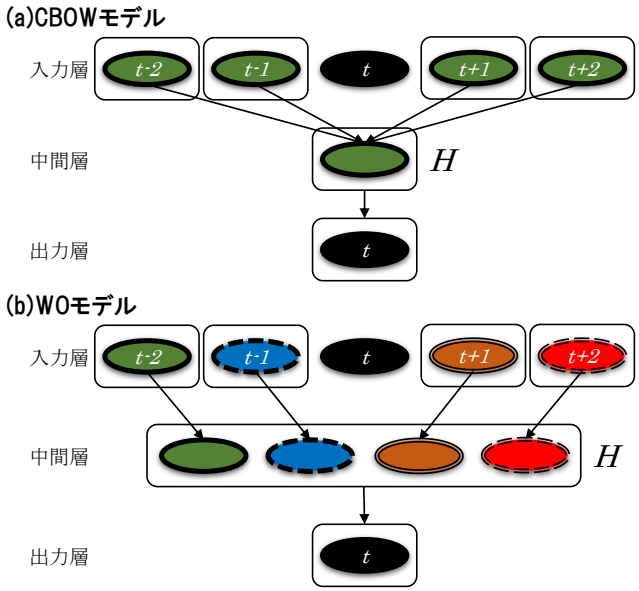


図 4　word2vec の学習モデルの概要図

語から中心単語を予測するモデルを構築する．このモデルを用いることにより，周辺単語が入力されると，中心単語に入り得ると思われる候補が，可能性が高い順に並べられて出力される．なお，周辺単語に入力される単語数は，モデル構築時のウィンドウサイズ x に依存する．本稿では，この単語予測モデルを，問題文解析モジュールにおける解答カテゴリ推定で使用する．

学習に用いるデータは，3.3 節で紹介した知識源のうち，教科書 4 冊分を採用する．ただし，学習データを作成する際には，対象とする単語の品詞の種別を考慮する必要がある．例えば，名詞のみを対象として作成する場合と，自立語（名詞・動詞・形容詞など）を対象として作成する場合では，予測精度が変化する可能性がある．どのような作成方法が適しているかについては，4.1 節の実験で明らかにする．

3.5　問題文解析モジュール

このモジュールでは，問題文脈部を入力し，穴埋め部分を含む文ごとに，文書検索用のクエリ生成や問題文のカテゴリ推定を行う．文中の穴埋め部分には，あらかじめ特殊文字列を記述しておき，システムが穴埋め部分を認識できるようにする．

まず初めに，当該問題の解答カテゴリ推定を行う．3.3 節で説明したユーザ辞書には，世界史分野の固有名詞に対し，「人物」「場所」「民族」などの全 18 種類のカテゴリが付与されているため，これを利用する．例えば，図 3(D) の解答である「ウズベク族」には，「民族」というカテゴリが付与されている．各問題の穴埋め部分にはどのようなカテゴリの単語が入り得るかを本モジュールで推定し，後の解答抽出モジュールにおいて，その結果を解答候補単語のスコアリングの一部として使用する．解答カテゴリ推定には，3.4 節で説明した単語予測モデルを使用する．例として，図 3(D) の解答カテゴリを推定する手順を，図 5 に示す．あらかじめ構築した単語予測モデルに，穴埋め部分の周辺単語を入力すると，中心単語の予測候補となる単語群が出力される．この際，名詞以外の品詞の単語は無視され，候補単語としてはカウントされない．候補群の各単語に付与されているカテゴリが照合され，照合されたカテゴリ全てを当該問題の解答カテゴリとする．なお，一般的な名詞が候補として挙げられた場合など，カテゴリが付与されていない場合は無視する．

単語予測モデルが出力する候補単語の件数は，任意に設定することができる．例えば，出力件数を 5 件と設定すると，カテゴリ推定の際には，予測上位 5 件の単語から推定処理を行う．ここで，出力件数を複数件とした場合．推定されるカテゴリも図 5 のように複数件となる場合がある．出力件数を少なくすると，推定カテゴリ群に列挙されるカテゴリは少なくなるが，もしその中に正解単語のカテゴリが含まれており，それ以外に推定されたカテゴリが少なければ少ないほど，高い精度でカテゴリを推定できたと言うことができる．逆に，参照件数を多くすると，推定カテゴリ群に正解単語のカテゴリが含まれる確率は高くなるが，それ以外のカテゴリが多く含まれる確率も高くなるため，正解単語のカテゴリが含まれていても，それ以外に推定されたカテゴリが多ければ多いほど，カテゴリ推定の精度は低くなる．出力件数は何件が適しているかについては，4.1 節の実験で明らかにする．

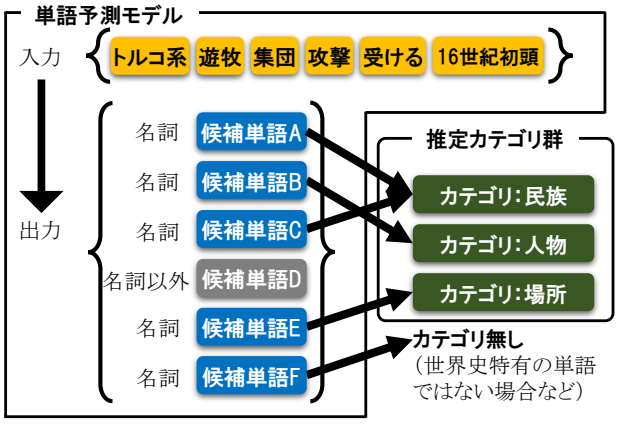


図 5　解答カテゴリ推定の概要図

次に，問題ごとにクエリ q の生成を行う．当該穴埋め部分が含まれる 1 文から，形態素解析を行って名詞をすべて抽出する．抽出された全ての名詞を使って OR 検索を行うようなクエリを生成する．例えば，図 3 の (D) の穴埋め部分を解答するための文書検索を行うクエリ q は図 6 に示すとおりである．

```
中央アジア OR ティムール朝 OR 政権 OR トルコ系 OR 遊牧
OR 集団 OR 族 OR 攻撃 OR 16世紀初頭 OR 消滅
```

図 6　文書検索用クエリ q の例

3.6　文書検索モジュール

このモジュールでは，3.3 節で説明した知識源に対して，3.5 節で生成したクエリ q による文書検索を行い，解答候補を含む文書群を取得する．検索エンジンには，オープンソースの全文検索システムである Apache Solr(注3) を使用する．また，検索時の文書の重み付けの手法には Okapi BM25 [15] を採用した．

(注3)：Apache Solr : http://lucene.apache.org/solr/

クエリ q による検索結果としてヒットした文書 d の集合を，BM25 のスコアが高い順に並び替えたものを *RankingResult*(q) とし，その上位 50 件の文書 d の集合を，解答候補抽出モジュールの入力とする．

3.7 解答候補抽出モジュール

このモジュールでは，3.6 節で得られた文書 d の集合から解答候補 w を抽出する．世界史の穴埋め型問題の解答となる単語は，原則として固有名詞になると考えられるため，解答候補としては，各文書 d に含まれている固有名詞を採用する．この条件により得られる解答候補 w の集合を C とすると，C は *RankingResult*(q) の上位 50 件の文書 d に含まれる全固有名詞の集合ということができる．

3.8 解答抽出（解答候補解析）モジュール

このモジュールでは，3.7 節で得られた解答候補について，各候補単語 w ごとにスコアを計算し，最終的な解答となる単語を決定する．各単語のスコア $Score(w)$ の計算式は式 (1) のとおりである．

$$Score(w) = \max_{w \in d} Score_{BM25}(q,d) + f_{existence}(w) + f_{backward}(w) + f_{category}(w), \forall w \in C \quad (1)$$

式 (1) におけるスコアリングに用いた指標は以下のとおりである．

$\max_{w \in d} Score_{BM25}(q,d)$：$w$ が解答となる潜在的可能性を表す指標

候補単語 w が当該問題の解答となる潜在的な確からしさとして，候補単語 w を含む文書 d のうち，クエリ q に対する BM25 のスコアが最大のスコアを加算する．ただし，クエリ q は 3.6 節の文書検索モジュールで用いたものである．このスコアが，候補単語 w の基準スコアとなり，このスコアから後の 3 つの判定による指標を考慮して，w の最終的なスコアが決定する．

$f_{existence}(w)$：問題文非既出単語判定による指標

w がすでに問題文中に含まれていないかを判定し，含まれていない場合は正の値 a を加算する．

$$f_{existence}(w) = \begin{cases} a\ (w \text{ が問題文中に含まれていない}) \\ 0\ (\text{それ以外}) \end{cases} \quad (2)$$

このような判定を行う理由は，問題文中に既に出現した単語が穴埋め部分の解答となる可能性は極めて低いため，そのような単語を除外するためである．

$f_{backward}(w)$：後方一致判定による指標

w の後方部分が，穴埋め部分の次の単語と一致している場合は正の値 b を加算する．

$$f_{backward}(w) = \begin{cases} b\ (w \text{ の後方部分が穴埋め部分の次の単語と一致}) \\ 0\ (\text{それ以外}) \end{cases} \quad (3)$$

例えば，図 3 の (D) では，穴埋め部分の次に「族」という単語があるため，候補単語の後方が「族」となった場合にスコアを加算する．

$f_{category}(w)$：カテゴリ不一致判定による指標

w のカテゴリを照合し．そのカテゴリが 3.5 節で推定された解答カテゴリ群に含まれていなければ正の値 c を減算する．

$$f_{category}(w) = \begin{cases} -c\ (w \text{ のカテゴリが推定済みのカテゴリと一致しない}) \\ 0\ (\text{それ以外}) \end{cases} \quad (4)$$

全ての候補単語 w に対してスコアを計算し，最も $Score(w)$ の値が大きい w を最終的な解答として出力する．

4. 実　　験

4.1 実験 1:単語予測モデルの違いによる解答カテゴリ推定の精度

4.1.1 目　　的

単語予測モデルの違いにより，解答カテゴリ推定精度がどのように変化するかについて明らかにする．3.8 節の解答候補抽出モジュールで行われる**カテゴリ不一致判定による指標**に基づくスコアリングは，あらかじめ問題文解析モジュールで推定されている解答カテゴリの結果に依存するが，3.4 節や 3.5 節で述べたように，解答カテゴリの推定精度は，単語予測モデル構築時の学習データやパラメータ，及び単語予測モデルの出力単語件数の設定によって変化すると考えられる．本実験では，これらの複数の異なる条件のうち，どれがより適切な条件であるかを明らかにする．

4.1.2 方　　法

まず，単語予測モデルを構築する際の条件の一覧を表 1 に示す．**モデル名**は，単語予測モデル構築時に使用した学習モデル名を接頭にし，学習データ及びウィンドウサイズ (x) の組み合わせを通し番号として付与した．**学習データ作成方法**は，どの品詞を対象として学習データを作成したかで区別される．**ウィンドウサイズ (x)** は，学習時に周辺の何単語から中心単語を予測しているかの数値である．例えば，$x = 4$ の場合，中心単語の前後各 4 単語から中心単語を予測することになる．なお WO モデルについては，学習データ作成時の対象品詞を全品詞とした場合のモデルのみを作成した．

次に，それぞれのモデルを使用してカテゴリ推定を行う際に，モデルの予測候補出力件数を 1 件，5 件，10 件とした場合を考え，それぞれの場合（条件）で，カテゴリ推定の精度を算出する．

精度を調べる問題文のデータセットとしては，NTCIR12 QA Lab-2 タスクで提供された世界史問題のうち，穴埋め型問題のみを使用する．具体的には，同タスクの Phase1 で提供された下記のデータセット（全 57 問）となる．

- 2003 年度 北海道大学（全 9 問）
- 2003 年度 東京大学（全 4 問）

表 1　実験で用意した各単語予測モデルの構築条件

モデル名	学習データ作成方法	ウィンドウサイズ (x)
CBOW-1	名詞のみ	4
CBOW-2	名詞のみ	5
CBOW-3	名詞のみ	6
CBOW-4	自立語のみ	4
CBOW-5	自立語のみ	5
CBOW-6	自立語のみ	6
CBOW-7	全品詞（記号除く）	4
CBOW-8	全品詞（記号除く）	5
CBOW-9	全品詞（記号除く）	6
WO-1	全品詞（記号除く）	4
WO-2	全品詞（記号除く）	5
WO-3	全品詞（記号除く）	6

- 2003 年度 中央大学（全 15 問）
- 2003 年度 早稲田大学（全 15 問）
- 2003 年度 京都大学（全 14 問）

最後に，各条件のカテゴリ推定の精度を示す指標として，問題ごとの MAP（Mean Average Precision）を用いる．MAP を指標とすることで，本来正解となる単語のカテゴリが，単語予測モデルの出力結果となる候補単語群の上位にどれだけ出現しているかが分かる．

4.1.3　結　　果

各条件におけるカテゴリ推定の精度を測定した結果を表 2 及び図 7 に示す．

表 2　各条件におけるカテゴリ推定の精度

モデル名＼参照件数	上位 **1** 件	上位 **5** 件	上位 **10** 件
CBOW-1	0.088	0.234	0.236
CBOW-2	0.228	0.269	0.277
CBOW-3	0.158	0.281	0.291
CBOW-4	0.246	0.305	0.306
CBOW-5	0.070	0.196	0.217
CBOW-6	0.175	0.250	0.249
CBOW-7	0.246	0.356	0.347
CBOW-8	0.246	0.314	0.315
CBOW-9	0.228	0.373	0.362
WO-1	0.211	0.332	0.318
WO-2	0.404	0.380	0.344
WO-3	0.070	0.183	0.221

CBOW モデルにおいては，学習データ作成時に全品詞の単語を対象としたモデルが，他と比べて精度が高い傾向であることが分かる．また，モデルの出力候補件数が 1 件の場合と 5 件の場合では，5 件のほうが精度が高いが，5 件の場合と 10 件の場合では，その差は僅かである．以上の結果を踏まえて，4.2 節の実験における，CBOW の予測モデルを用いたカテゴリ推定には，CBOW-9 モデルの出力件数 5 件の結果を使用する．

次に，WO モデルにおいては，$x = 5$ のモデルが他と比べて精度が高いことが分かる．また，$x = 5$ の場合，モデルの出力候補件数が 1 件，5 件,10 件となるにつれて，精度が低下していることが分かる．この結果を踏まえて，4.2 節の実験における，WO の予測モデルを用いたカテゴリ推定には，WO-2 モデルの出力件数 1 件の結果を使用する．

4.2　実験 2:手法の違いによる自動解答の正答率

4.2.1　目　　的

本実験では，3.8 節で述べた解答抽出モジュールにおけるスコアリングで，カテゴリ不一致判定を行うか否かで，正答率及び誤答率がどの程度変化するかを明らかにする．また，カテゴリ不一致判定を行う場合，4.1 節の実験 1 で調査した，解答カテゴリの推定精度が高かった各単語予測モデルを用いることが，カテゴリ不一致判定にどの程度影響を及ぼすか明らかにする．

4.2.2　方　　法

まず，問題文のデータセットとしては，4.1 節と同様に，NTCIR12 QA Lab-2 タスクで提供された世界史問題のうち，穴埋め型問題のみを使用する．システム開発時のトレーニングデータは，4.1 節と同様に同タスクの Phase1 で提供されたものを使用する．評価用データとして，同タスクの Phase3 で提供された下記のデータセット（全 54 問）を使用し，正答率を調べた．

- 2011 年度 北海道大学（全 9 問）
- 2011 年度 中央大学（全 15 問）
- 2011 年度 早稲田大学（全 8 問）
- 2011 年度 京都大学（全 22 問）

なお，以上のデータセットには，本来，多肢選択式であった問題も含まれているが，本実験ではそのような問題においては選択肢を無視し，記述式の穴埋め型問題として取り扱った．

次に，比較するシステムの解答処理方法として，下記の 3 つの方法を用意した．

方法 1

解答抽出モジュールにおいて，基準スコアに加え，非既出単語判定・後方一致判定のみを用いて解答を決定する

方法 2

解答抽出モジュールにおいて，**方法 1** に，4.1 節における CBOW-9 モデルの上位 5 件の出力によるカテゴリ推定に基づいたカテゴリ不一致判定を加えて，解答を決定する

方法 3

解答抽出モジュールにおいて，**方法 1** に，4.1 節における WO-2 モデルの上位 1 件の出力によるカテゴリ推定に基づいたカテゴリ不一致判定を加えて，解答を決定する

最後に，本実験における正答・誤答の定義を表 3 に示す．解答抽出モジュールにおけるスコアリングの結果と，その問題の本来の正解単語との関係により，表中に示す小分類のいずれかに分類される．なお，正答の定義については，**単独**と**同率**に分かれているが，実質的にシステムが正解できた問題数は**単独**に分類されたもののみである．

4.2.3　結　　果

実験で得られた解答結果の内訳を表 4 に示す．方法ごとに，正答率，誤答率及びそれぞれの問題数を記している．

単独の正答に注目すると，方法 2 の正答率が，すべての方法

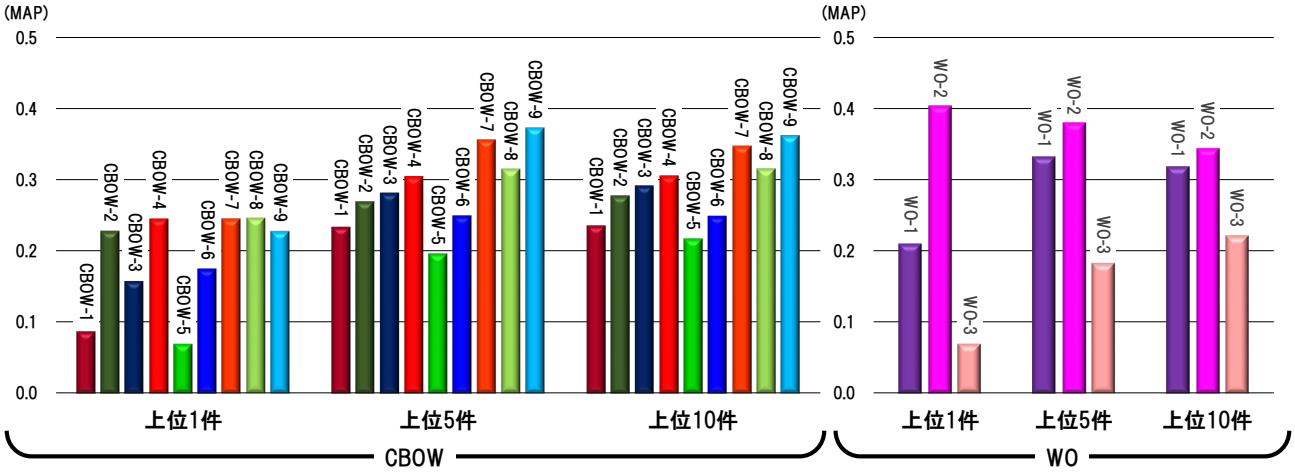


図 7　各条件におけるカテゴリ推定の精度

表 3　本実験における正答・誤答の定義

大分類	小分類	定義
正答	単独	本来の正解単語が唯一スコア 1 位となった
	同率	本来の正解単語がスコア 1 位であったが，他にも同一のスコアの単語が存在した
誤答	候補存在	本来の正解単語が解答候補には含まれていたが，スコア 1 位では無かった
	候補不在	本来の正解単語が解答候補に含まれていなかった

表 4　解答処理方法による解答結果の内訳
（カッコ内は該当問題数/全問題数）

処理方法	正答		誤答	
	単独	同率	候補存在	候補不在
方法 1	0.22　(12/54)	0.22　(12/54)	0.39　(21/54)	0.17　(9/54)
方法 2	0.26　(14/54)	0.13　(7/54)	0.44　(24/54)	0.17　(9/54)
方法 3	0.18　(10/54)	0.13　(7/54)	0.52　(28/54)	0.17　(9/54)

において最も高いことが分かる．また，方法 2 と同じくカテゴリ不一致判定を導入した方法 3 の正答率が低いことが分かる．しかし，単独と同率を合わせたすべての正答においては，方法 1 が該当問題数の合計が最も多くなっている．

5. 考　　察

4.1 節及び 4.2 節の各実験について考察する．

5.1　考察 1:解答カテゴリ推定の精度

実験 1 では，単語予測モデルの構築方法及び出力件数の違いにより，カテゴリ推定の精度に大きく差が出ることが分かった．

CBOW の各モデルにおいて，記号を除く全品詞を対象とした学習データで構築したモデルは，名詞や自立語のみを対象としたものと比べて，全体的に高い精度となっている．中心単語を予測するためには，周辺の自立語だけではなく，助詞などの機能語も考慮する必要があるためと考えられる．

CBOW の各モデルの中で，最も精度が高かったのは，CBOW-9 モデルの出力件数 5 件の場合で，MAP は 0.373 である．一方，WO の各モデルの中で，最も精度が高かったのは，WO-2 モデルの出力件数 1 件の場合で，MAP は 0.404 である．よって，語順を考慮した場合の方が，考慮しない場合より推定精度が改善することが分かった．また，WO のモデルは CBOW のモデルと異なり，出力件数が 1 件の場合での精度が高い．このことから，WO による単語予測モデル構築における学習をさらに強化することで，カテゴリ推定精度のさらなる向上が見込まれる．

5.2　考察 2:自動解答の正答率

実験 2 では，解答抽出モジュールの**カテゴリ不一致判定**の有無による正答率の変化を見た．

まず，CBOW の単語予測モデルによるカテゴリ推定結果を用いて判定を行うと，判定を行わない場合に比べて単独正答率が僅かに向上しているが，WO のモデルによる同様の判定を行うと，単独正答率は低下している．これは，現状のカテゴリ推定の精度自体が低く，問題文解析モジュールで誤ったカテゴリを推定してしまい，カテゴリ不一致判定によるスコアリングに悪影響を与えているためだと考えられる．なお，**方法 2** 及び**方法 3** において，仮に単語予測モデルによるカテゴリ推定精度が完全であった場合，単独正答率は 0.44 となり，**方法 1** の約 2 倍の精度となることが分かった．単語予測モデルの予測性能を向上させることが，カテゴリ不一致判定の効果の向上にもつながると思われる．

次に，解答抽出モジュールの**非既出単語判定**及び**後方一致判定**は，正答率の向上に大きく貢献していることが分かった．実験 2 の方法 1 による自動解答では，同判定のスコアリングのみを行い，単独正答したのは 54 問中 12 問であったが，このうち 11 問が，同判定のスコアリングによって，スコア単独 1 位となっている．これらの判定は，穴埋め型問題に適した判定ということができる．

6. ま　と　め

本稿では，大学入試の世界史問題における，穴埋め型問題を適切に解答する手法を提案した．まず，語順を考慮して中心単語を予測するモデルを構築し，そのモデルを用いて解答カテゴリを推定した．次に，穴埋め型問題に対応した幾つかのスコアリング手法を導入し，正答率の向上を目指した．

実験では，単語予測モデルを用いたカテゴリ推定の精度，及

び提案手法による自動解答の正答率を調べた．その結果，単語予測モデルの構築方法の違いによって，カテゴリ推定の精度が大きく変わることが分かったが，語順を考慮したモデルの場合の方が，考慮しない場合より推定精度が改善することを確認した．また，問題文非既出単語判定や後方一致判定は，穴埋め型問題の解答に大きく貢献していることが分かった．

今後は，カテゴリ推定の精度を向上させるために，単語予測モデルの構築方法の見直しや，場合によっては新たな手法を提案する予定である．また，穴埋め型問題の特徴を生かした，新たな判定の導入等も検討する．

謝　　辞

本研究は京都産業大学総合学術研究所の研究活動によるものです．

文　　献

[1] Ferrucci David, Brown Eric, Chu-Carroll Jennifer, Fan James, Gondek David, Kalyanpur Aditya A., Lally Adam, Murdock J. William, Nyberg Eric, Prager John, Schlaefer Nico, and Welty Chris. Building Watson: An Overview of the DeepQA Project. *AI MAGAZINE*, Vol. 31, pp. 59–79, 2010.

[2] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daum III. A Neural Network for Factoid Question Answering over Paragraphs. In *Empirical Methods in Natural Language Processing*, pp. 633–644, 2014.

[3] Takada Takuma, Imagawa Takuya, Matsuzaki Takuya, and Sato Satoshi. SML Question-Answering System for World History Essay and Multiple-choice Exams at NTCIR-12 QA Lab-2. In *12th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 421–424, 2016.

[4] Sakamoto Kotaro, Ishioroshi Madoka, Matsui Hyogo, Jin Takahisa, Wada Fuyuki, Nakayama Shu, Shibuki Hideyuki, Mori Tatsunori, and Kando Noriko. Forst : Question Answering System for Second-stage Examinations at NTCIR-12 QA Lab-2 Task. In *12th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 467–472, 2016.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] 有賀竣, 鶴岡慶雅. 単語のベクトル表現による文脈に応じた単語の同義語拡張. 言語処理学会 第 21 回年次大会 発表論文集, pp. 752–755, 2015.

[7] Sato Toshinori. Neologism dictionary based on the language resources on the Web for Mecab, 2015.

[8] Kimura Tasuku, Nakata Ryosuke, and Miyamori Hisashi. KSU Team's Multiple Choice QA System at the NTCIR-12 QA Lab-2 Task. In *12th NTCIR Conference on Evaluation of Information Access Technologiesnference on Evaluation of Information Access Technologies*, pp. 437–444, 2016.

[9] 佐藤次高, 木村靖二, 岸本美緒. 詳説世界史. 山川出版社, 2008.

[10] 尾形勇. 世界史 B. 東京書籍, 2007.

[11] 相良匡. 新選世界史 B. 東京書籍, 2007.

[12] 加藤晴康. 世界史 A. 東京書籍, 2008.

[13] 今泉博. 山川 一問一答世界史. 山川出版社, 2015.

[14] Tokui Seiya, Oono Kenta, Hido Shohei, and Clayton Justin. Chainer: a Next-Generation Open Source Framework for Deep Learning. In *Proceedings of Workshop on Machine Learning Systems in NIPS*, 2015.

[15] Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2010.