

# 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<sup>1,†1,a)</sup> 木村 輔<sup>2,b)</sup> 宮森 恒<sup>2,c)</sup>

受付日 2017年3月9日, 採録日 2017年7月7日

**概要:** 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いえない. たとえば, 大学入試など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, 問題文解析時に穴埋め部分の周辺単語から解答カテゴリを推定し, 解答候補抽出に利用するとともに, 解答候補評価時に, カテゴリとの一致性や周辺単語の既出状況などを用いた指標を導入することで解答候補を評価する. 特に, 解答カテゴリを推定する際には, 語順を考慮した分散表現による単語予測モデルを導入する. 実験では, まず, 単語予測モデルの精度を比較する. また, ベースライン手法と提案手法を比較し, 提案手法を解答処理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:** ファクトイド型質問応答, 自動解答, 大学入試問題, 分散表現, 語順

## Automatic Answering Method Considering Word Order for Slot Filling Question of University Entrance Examinations

RYO TAGAMI<sup>1,†1,a)</sup> TASUKU KIMURA<sup>2,b)</sup> HISASHI MIYAMORI<sup>2,c)</sup>

Received: March 9, 2017, Accepted: July 7, 2017

**Abstract:** Recently, automatic answering technologies such as question answering have attracted attention as a technology to satisfy various information requests from users. However, it is difficult to say that these technologies can adequately respond to the diverse and complicated questions in realistic situations including university entrance examinations. For example, conventional methods can not provide correct answers sufficiently for the slot filling questions in the university entrance examinations, because retrieval-based factoid-type answering technologies are mainly used, which do not consider word order. In this paper, we propose an automatic answering method considering word order for the slot filling questions in the university entrance examination world history problems. In particular, when in analyzing the question sentence, the answer category is estimated from the surrounding words of the filling slot and used for extracting the answer candidates, and these candidates are evaluated by introducing the indicator using the consistency with the category, and the occurrence situation of the surrounding words. Especially, we introduce a word prediction model by distributed expression considering word order in estimating the answer category. In the experiment, we first compare the accuracy of the word prediction models. In addition, we compare the proposed method with the baseline method and clarify what kind of change is observed in the correct answer rate by incorporating proposed method.

**Keywords:** factoid question answering, automatic answering, university entrance examination, distributed representation, word order

<sup>1</sup> 京都産業大学コンピュータ理工学部  
Faculty of Computer Science and Engineering, Kyoto Sangyo University, Kita, Kyoto 603-8555, Japan

<sup>2</sup> 京都産業大学大学院先端情報学研究科  
Division of Frontier Informatics, Graduate School of Kyoto Sangyo University, Kita, Kyoto 603-8555, Japan

<sup>†1</sup> 現在, 京都産業大学大学院先端情報学研究科  
Presently with Division of Frontier Informatics, Graduate School of Kyoto Sangyo University

a) i1788124@cse.kyoto-su.ac.jp

b) i1658047@cse.kyoto-su.ac.jp

c) miya@cc.kyoto-su.ac.jp

## 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ユーザ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。ユーザ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な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<sup>\*1</sup> 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いえない。

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。多くの従来研究では、このような問題を自動解答させる手法として、問題文から様々な手がかりを取得し、それを基に解答を抽出する方法がとられている。たとえば、当該問題文が「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する手法は、従来研究でも多く取り入れられている。実際の推定手法については様々な提案がなされているが、一般的に図1(a)のような穴埋め型問題を想定しておらず、図1(b)のような事実型問題の文の構造を前提としたものとなっているため、それらの手法を直接用いることはできない。なぜならば、穴埋め型問題と比較すると、事実型問題のほうが、表層的なレベルでカテゴリを判断しやすいからである。これは、解答カテゴリの推定に限ったことではなく、問題の焦点を推定する場合などについても同様のことがいえる。以上の点から、穴埋め型問題を自動解答させるためには、それに適した手法を、解答処理内に取り入れる必要がある。

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、解答候補評価のモ

### (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)ウイッテ

### (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よいか。

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)

Fig. 1 Example of factoid-type question in university entrance examination (World history, Faculty of Letters, Chuo University, 2015).



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  
Fig. 2 Basic processing steps of factoid QA system.

ジュールの処理が順に実行され、解答が出力される。そのうち、先に述べたような、問題文から様々な手がかりを取得する処理は、問題文解析モジュールにあたる。また、それらの手がかりを基に、最終的に解答を選択する処理は、解答候補評価モジュールにあたる。よって、穴埋め型問題にも対応したシステムとするには、これらのモジュールにおいて、適した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。穴埋め型問題は、その問題文の特徴から、穴埋め部分の周辺に出現する単語およびその語順が、解答候補

<sup>\*1</sup> NTCIR-13: <http://research.nii.ac.jp/ntcir/ntcir-13/>

評価のための重要な手がかりになると考えられる。提案手法としては、問題文解析モジュールおよび解答候補評価モジュールにおいて、先述の情報をもとに、解答カテゴリの推定や、単語の後方文字列の一致判定を行う。特に、解答カテゴリの推定においては、語順を考慮した単語予測モデルを活用する。

実験では、まず、単語予測モデルの生成条件の違いによる、カテゴリ推定の精度を比較する。また、使用するモデルや文書検索ドキュメントの構築方法の違いによる、自動解答の正答率についても比較する。さらに、解答候補評価モジュールにおけるスコアリングの各指標が、どの程度正答率向上に貢献しているかを明らかにする。

本稿の構成は以下のとおりである。2章では本稿と関連する研究について述べる。3章では自動解答システムの一連の処理について説明し、それをふまえたうえで、4章では提案手法について詳しく説明する。5章では精度・正答率に関する実験を実施し、6章では実験結果をふまえた考察を記す。最後に7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。

Ferrucci 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。この手法により、従来の質問応答システムの精度を上回ったという結果が示されている。

評価型ワークショップである NTCIR でも、質問応答に関するタスクがたびたび実施されている。Murata ら [3] が開発したシステムは、ファクトイド型質問応答を取り扱った NTCIR-5\*2 QAC-3 タスクにおいて、最も良い成績を残している。このシステムの特徴は、複数の文書を使用し、文書ごとに得られた解答候補の評価スコアを、最終的に足し合わせている点である。システム内では、解答カテゴリを推定する処理も取り入れられているが、ルールベースによる推定手法となっている。つまり、ルールベースによる手法でも、ファクトイド型質問応答では一定の正答率を期待することができる。しかし、4.1 節でも詳しく述べるが、

大学入試問題のような複雑ないい回しで、かつ、一般的な事実型問題とは文の構造が異なる穴埋め型問題では、ルールの定義が複雑になり、網羅できる範囲が限られてしまうと予測される。

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 [4] は、大学入試問題の自動解答において、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1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たりしている。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を解析しているが、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 [5] は、事実型問題と穴埋め型問題を、同一の単語回答問題と見なして処理を行っている。おおまかな処理の流れは図2に沿っている。問題文解析モジュールでは、問題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し、さらに、疑問詞直前の単語によって「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している。解答候補評価モジュールでは、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カテゴリ・焦点の一致性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。穴埋め型問題の自動解答においても、問題の解答カテゴリを推定することができれば、正答率向上に大きく貢献できると考えられるが、Sakamoto らの手法は本稿における事実型問題を対象としており、穴埋め型問題とは文の構造が異なるため、まったく同じ手法を適用することはできない。

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 [6] が開発した word2vec がある。学習モデルとしていく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

\*2 NTCIR-5: <http://research.nii.ac.jp/ntcir/ntcir-ws/5/>

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 [7] 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、および、周辺単語の各出現位置をすべて区別する Word Order (WO) モデルの 2 つをあげている。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。

以上の関連研究をふまえて、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、従来の解答カテゴリ推定手法を見直し、さらに、穴埋め型問題に適した解答候補を評価できる新しい手法を取り入れる。特に、解答カテゴリ推定においては、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活用する。

### 3. 自動解答手法

本稿で取り扱う自動解答システムは、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ものとする。この章では、システムの一連の処理の流れを説明する。それをふまえ、4 章では、システム内に取り入れる我々の提案手法について述べる。

#### 3.1 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。

#### 3.2 世界史に関する辞書および知識源

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、自動解答の処理過程において、形態素解析エンジンである MeCab<sup>\*3</sup>を使用する。システム辞書として、

<sup>\*3</sup> MeCab: <http://taku910.github.io/mecab/>

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る記述式穴埋め型問題の例 (2014 年度中央大学文学部世界史)

Fig. 3 Example of descriptive slot filling question in university entrance examination world history problem (World history, Faculty of Letters, Chuo University, 2014).

表 1 用意した文書集合  
Table 1 Prepared document set.

文書集合の名称	文書登録方法	文書数
文書集合	書籍内の 1 文を 1 文書とする	18209
段落集合	書籍内の 1 段落を 1 文書とする	3642

Sato [8] 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 [9] 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、文書検索モジュールに格納する知識源として以下の書籍を用いた。教科書 4 冊については、QA Lab-3 のオーガナイザから提供されたデータであり、参考書 1 冊については、我々が独自に用意したデータである。

- 教科書 詳説 世界史 (山川出版社) [10]
- 教科書 世界史 B (東京書籍) [11]
- 教科書 新選世界史 B (東京書籍) [12]
- 教科書 世界史 A (東京書籍) [13]
- 参考書 山川一問一答世界史 (山川出版社) [14]

なお、知識源へは表 1 に示す 2 通りの方法で文書を登録した<sup>\*4</sup>。各文書集合は、1 文書の登録方法が異なるだけで、同じ知識源から構築される。それぞれの文書集合を使用する際の利点および欠点については、3.4 節で述べる。ただし、文書集合においては、1 文が 10 文字以下の文は除外し

<sup>\*4</sup> なお、参考書「山川一問一答世界史」については、書籍の構成上、1 節分を 1 段落として扱う。

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 4 文書検索用クエリ  $q$  の例

Fig. 4 Example of query  $q$  for document retrieval.

た。文書検索モジュールで検索する際には、いずれかの文書集合を選択したうえで実行する必要がある。

### 3.3 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と解答カテゴリの推定を行う。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。

クエリ  $q$  の生成については、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。抽出されたすべての名詞を使って OR 検索を行うようなクエリを生成する。たとえば、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 4 に示すとおりである。

本稿における解答カテゴリの推定とは、当該穴埋め部分に入るべき単語のカテゴリを推定することである。たとえば、人の名前を問う問題文であれば、解答カテゴリは「人名」となるべきである。我々が提案した、カテゴリの推定手法については、4.1.2 項で述べる。

### 3.4 文書検索モジュール

このモジュールでは、3.2 節で説明した知識源に対して、3.3 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<sup>\*5</sup>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。

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べ替えたものを  $RankingResult(q)$  とする。3.2 節で述べたとおり、検索の際には表 1 に示す文書集合のうち、どの集合に対して検索を行うかを指定する必要がある。

文書集合を被検索文書集合とした場合、穴埋め部分を含む問題文と類似する 1 文が文書集合に存在した場合、検索結果としてその文書がヒットしやすくなり、かつ、BM25 のスコアが高くなりやすい。また、詳細は 4.4 節で説明するが、検索結果から解答候補単語  $w$  を抽出した後、各候補  $w$  の基準スコアとして抽出元文書の BM25 のスコアを割り当てる。その際、各文書  $d$  (1 文) から抽出される単語は数件程度であるため、基準スコアが同率である候補  $w$  が比較

的少なくなるという利点がある。しかし、穴埋め部分の正解単語について、その事象が知識源内で複数文にわたって詳しく記されていた場合、文書を被検索文書集合とすると、正解単語を含む文書の BM25 のスコアが、他の文書のスコアに劣ってしまう可能性がある。

一方、段落集合を被検索文書集合とした場合、問題文に関連する事象について述べられた文書を段落単位で検索することとなるため、正解単語を含む文書が、BM25 のスコアが高い状態でヒットしやすくなる。しかし、各文書  $d$  (1 段落) から抽出される単語は数十件以上となるため、基準スコアが同率である候補  $w$  が比較的多くなるという欠点がある。基準スコアが同率である候補単語  $w$  が多すぎると、3.6 節で述べる穴埋め型問題に適したスコアリングの効果が低くなると懸念される。

文書集合を被検索文書集合とした場合、 $RankingResult(q)$  の上位 50 件、段落集合を被検索文書集合とした場合、 $RankingResult(q)$  の上位 5 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。

### 3.5 解答候補抽出モジュール

このモジュールでは、3.4 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 (文書集合) または上位 5 件 (段落集合) の文書  $d$  に含まれる全固有名詞の集合となる。

### 3.6 解答候補評価モジュール

このモジュールでは、3.5 節で得られた解答候補について、候補単語  $w$  ごとに解答らしさを評価し、最終的な解答となる単語を決定する。本稿のシステムでは、 $w$  ごとに様々な指標を用いてスコア  $Score(w)$  を計算し、スコアの順位をもとに、最終的なシステムの解答を出力する。

## 4. 提案手法

3 章で説明した自動解答システムをもとに、問題文解析モジュールおよび解答候補評価モジュールにおいて、問題文の出現単語および語順の情報を考慮した、問題文の解析および解答候補評価手法を提案する。

### 4.1 解答カテゴリ推定およびカテゴリ不一致判定

2 章でも述べたように、穴埋め型問題においても、問題文の解答カテゴリを正確に推定することができれば、正答率の向上が期待できる。

まず、解答カテゴリの推定は、基本的な手法として、ルールベースによる手法を考えることができる。たとえば、

\*5 Apache Solr: <http://lucene.apache.org/solr/>

図 3(C) の問題であれば、穴埋め部分の直前にある「君主である」という文脈から、穴埋め部分に入るべき単語のカテゴリは「人名」とであると推定できるかもしれない。しかし、ルールベースによる手法は、過去の問題集などを参考にしながらルールを列挙する必要がある、新しい問題文などの、あらゆるいい回しを網羅したルールを作成することは現実的とはいえない。また、ルールの定義が非常に複雑になる可能性もあり、複雑ないい回しなども存在する大学入試問題では網羅できる範囲に限られることが予想される。よって、我々は、どのような問題文のいい回しに対してもより柔軟に適用できる解答カテゴリを推定する手法として、単語予測モデルを用いた手法を提案する。この手法では、穴埋め部分の周辺の単語群から、穴埋め部分に入るべき正解単語のカテゴリを推定する。なお、人間の解答者が穴埋め部分の周辺単語群からどの程度穴埋め単語のカテゴリを推定できるかを、4名の被験者に対する予備実験で確認したところ、穴埋め部分の前後それぞれ7単語が与えられれば、世界史の穴埋め型問題の平均55%で、正しいカテゴリを推定できることが分かった。

自動解答処理を実行する前に、あらかじめ、周辺の出現単語とその語順から中心単語を予測するモデル（単語予測モデル）を構築しておく。自動解答処理内では、まず、問題文解析モジュールにおいて、単語予測モデルを用いた問題文の解答カテゴリを推定する。その結果をもとに、解答候補評価モジュールでは、各解答候補のスコアリングの指標として、カテゴリ不一致判定を導入する。

#### 4.1.1 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 [16] 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5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5(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、周辺単語から中心単語を予測する単語予測モデルを構築する。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 $x$ に依存するが、この値によってモデルの精度が左右される可能性がある。どのようなパラメータが適しているかについては、

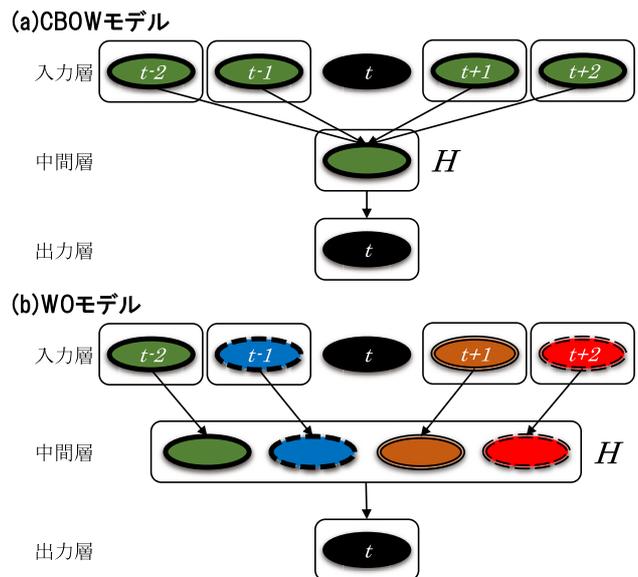


図 5 単語の分散表現獲得に用いた2つの学習モデル  
Fig. 5 Two learning models used to acquire distributed representation of words.

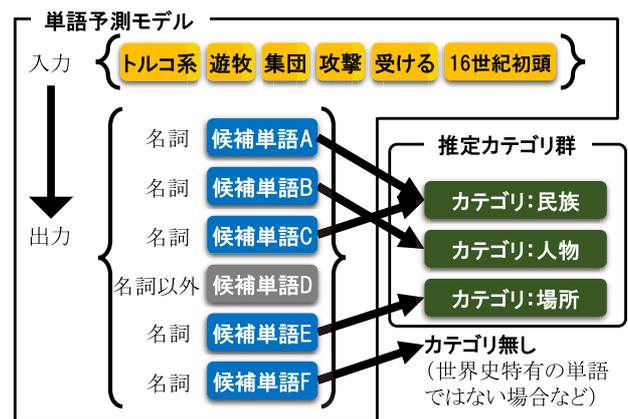


図 6 解答カテゴリの推定  
Fig. 6 Prediction of answer category.

5.1 節の実験で明らかにする。

#### 4.1.2 解答カテゴリ推定

4.1.1 項で構築された単語予測モデルを用いて、入力された問題文から、解答カテゴリを推定する処理を、問題文解析モジュール内に取り入れる。3.2 節で説明したユーザー辞書には、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。例として、図3(D)の解答カテゴリを推定する手順を、図6に示す。なお、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。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すべてを当該問題

の解答カテゴリとする。なお、一般的な名詞が候補としてあげられた場合など、カテゴリが付与されていない場合は無視する。

もし、穴埋め部分の直後が名詞であった場合、例外的にその名詞は単語予測モデルへの入力へ含まない。たとえば、図 3(D) の場合、穴埋め部分の直後にある「族」という単語は、単語予測モデルへの入力へ含まない。これは、本来、穴埋め部分とその直後の名詞が複合されて、1つの名詞とみるべきであるためである。図 3(D) の模範解答は「ウズベク」となっているが、これはあくまで空欄部分に入る文字列であり、問題が意図している解答は「ウズベク族」である。3.2 節でも述べたように、本稿では、世界史特有の固有名詞に対応したユーザ辞書を形態素解析時に用いており、このような世界史特有の複合名詞は1つの単語として認識される。これは、単語予測モデル構築時と同様である。よって、単語予測モデルへ、「族」のような穴埋め部分の直後の名詞を入力させることは、周辺単語を入力させているのではなく、正解単語の断片を入力する行為になってしまうため、入力へは含まないこととした。なお、穴埋め部分の直後の名詞自体は、解答候補評価の大きな手がかりになると考えることができる。よって、4.2 節で述べる後方一致判定で、この情報を活用することとする。

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。たとえば、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 6 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いえる。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、5.1 節の実験で明らかにする。

#### 4.1.3 カテゴリ不一致判定

3.6 節で述べたスコア計算の指標の1つとして、カテゴリ不一致判定による指標を、式 (1) のとおり導入する。 $w$  のカテゴリを照合し、そのカテゴリが 4.1.2 項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  $a$  を減算する。

$$f_{category}(w) = \begin{cases} -a (w \text{ のカテゴリが推定済みの} \\ \text{カテゴリと一致しない}) \\ 0 (それ以外) \end{cases} \quad (1)$$

## 4.2 後方一致判定

4.1.2 項で述べたが、穴埋め部分の直後に名詞が存在している場合、その名詞の文字列は、正解単語の後方の文字列と一致する可能性がきわめて高い。よって、3.6 節で述べたスコア計算の指標の1つとして、後方一致判定による指標を、式 (2) のとおり導入する。 $w$  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。

$$f_{backward}(w) = \begin{cases} b (w \text{ の後方部分が穴埋め部分の} \\ \text{次の単語と一致}) \\ 0 (それ以外) \end{cases} \quad (2)$$

たとえば、図 3 の (D) では、穴埋め部分の次に「族」という名詞があるため、候補単語の後方が「族」となった場合にスコアを加算する。

## 4.3 問題文非既出単語判定

大学入試問題の場合、問題文にすでに出現している単語が穴埋め部分の正解単語となることは、非常に考えにくい。よって、3.6 節で述べたスコア計算の指標の1つとして、問題文非既出単語判定による指標を、式 (3) のとおり導入する。 $w$  がすでに問題指示部および問題文脈部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  $c$  を加算する。

$$f_{existence}(w) = \begin{cases} c (w \text{ が問題文中に} \\ \text{含まれていない}) \\ 0 (それ以外) \end{cases} \quad (3)$$

## 4.4 各判定指標を用いた解答候補評価

3.6 節で述べた各単語のスコア  $Score(w)$  の計算については、式 (4) の計算式のとおりとする。

$$Score(w) = \max_{w \in d} Score_{BM25}(q, d) + f_{category}(w) + f_{backward}(w) + f_{existence}(w), \forall w \in C \quad (4)$$

式 (4) における  $\max_{w \in d} Score_{BM25}(q, d)$  は、 $w$  が解答となる潜在的可能性を表す指標である。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 3.4 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、4.1 節から 4.3 節にかけて示した指標による、スコアの加算・減算を経たうえで、 $w$  の最終的なスコアが決定する。

## 5. 実験

### 5.1 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

#### 5.1.1 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。4.1.3 項のカテゴリ不一致判定による指標は、あらかじめ 4.1.2 項で推定されている解答カテゴリの結果に依存するが、4.1.1 項や 4.1.2 項で述べたように、解答カテゴリの推定精度は、単語予測モデル構築時のパラメータや、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。

#### 5.1.2 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 2 に示す。

モデル名は、学習モデル名と、ウィンドウサイズを組み合わせたものとなっている。

学習モデルは、単語予測モデル構築時に使用した学習モデル名であり、CBOW か WO のいずれかである。4.1.1 項で述べたように、CBOW は語順が考慮されていないモデル、WO は語順が考慮されているモデルに対応する。

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。たとえば、 $x = 4$  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。WO モデルを提案している有賀ら [7] は、予測精度を調査する実験において、 $x = 5$  と設定していたため、このことを参考に、本実験では、 $x = 3, 4, 5, 6, 7$  の各場合でモデルを構築することとした。

モデルを構築する際の学習データとしては、3.2 節で示した知識源のうち、教科書 4 冊分を用いる。なお、モデルを構築する際の穴埋め部分の周辺単語について、どの範囲の品詞を用いるかでモデルの予測精度が変化する可能性が

表 2 実験で用意した各単語予測モデルの構築条件

Table 2 Construction condition of word prediction model used by experiment.

モデル名	学習モデル	ウィンドウサイズ ( $x$ )
CBOW-3	CBOW	3
CBOW-4	CBOW	4
CBOW-5	CBOW	5
CBOW-6	CBOW	6
CBOW-7	CBOW	7
WO-3	WO	3
WO-4	WO	4
WO-5	WO	5
WO-6	WO	6
WO-7	WO	7

ある。よって、予備実験として、穴埋め部分の周辺単語として、名詞のみ、自立語のみ<sup>\*6</sup>、全品詞（記号除く）の合計 3 通りの場合で学習データを用意し<sup>\*7</sup>、それぞれの場合で単語予測モデルを構築して、精度を比較した。その結果、全品詞を用いた学習データで構築されたモデルが、最も精度が高いことが分かった。よって、以降の実験では、全品詞を対象とした学習データを用いて、単語予測モデルを構築することとした。

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、NTCIR-12<sup>\*8</sup> 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）
- 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。

#### 5.1.3 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 3 および図 7 に示す。

CBOW モデルでは  $x = 4$ 、WO モデルでは  $x = 5$  のモデルが、他と比べて精度が高いことが分かる。それぞれのモデルに注目すると、CBOW-4 は、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、その差はわずかである。WO-5 は、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。

### 5.2 実験2：手法の違いによる自動解答の正答率

#### 5.2.1 目的

本実験では、自動解答処理内で用いる手法を変えることで、正答率および誤答率がどの程度変化するかを明らかにする。具体的には、被検索文書集合にする文書集合の違い、およびカテゴリ推定の際に使用する単語予測モデルの違い

<sup>\*6</sup> 自立語のみとは、助詞および助動詞以外の全品詞を指す。

<sup>\*7</sup> 品詞分類は、3.2 節で述べた形態素解析エンジン MeCab の仕様に依存する。たとえば接尾辞は、MeCab の仕様上、名詞に分類される。

<sup>\*8</sup> NTCIR-12: <http://research.nii.ac.jp/ntcir/ntcir-12/>

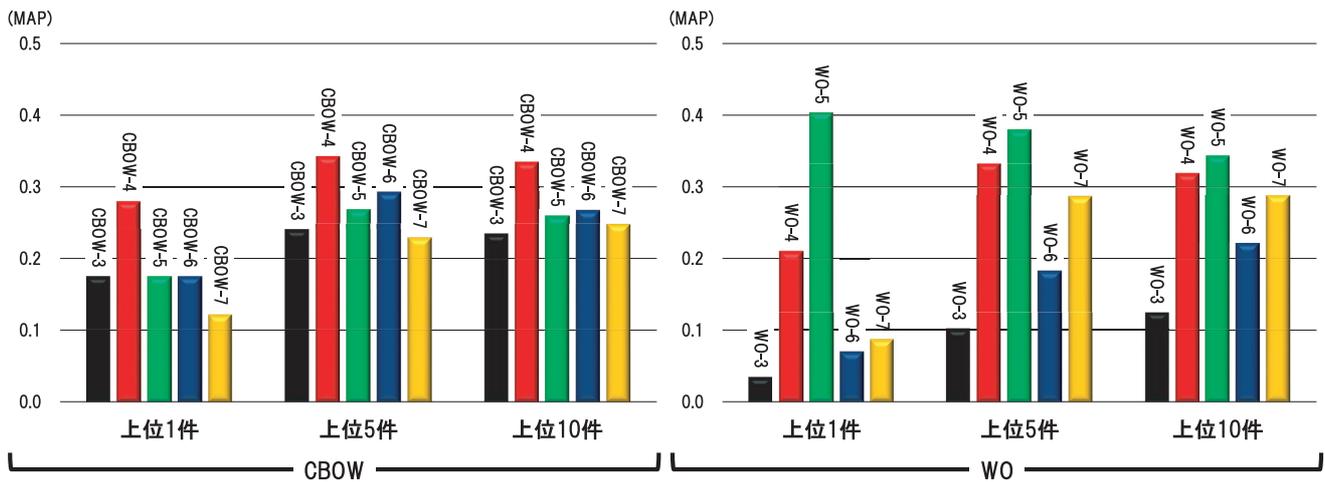


図 7 各条件におけるカテゴリ推定の精度 (MAP)

Fig. 7 Accuracy (MAP) of category prediction by each condition.

表 3 各条件におけるカテゴリ推定の精度 (MAP)

Table 3 Accuracy (MAP) of category prediction by each condition.

モデル名	参照件数	上位 1 件	上位 5 件	上位 10 件
CBOW-3		0.175	0.241	0.235
CBOW-4		0.281	0.343	0.335
CBOW-5		0.175	0.268	0.260
CBOW-6		0.175	0.293	0.268
CBOW-7		0.123	0.229	0.248
WO-3		0.035	0.102	0.124
WO-4		0.211	0.332	0.318
WO-5		0.404	0.380	0.344
WO-6		0.070	0.183	0.221
WO-7		0.088	0.287	0.288

表 4 比較する解答処理手法

Table 4 Answering methods to be compared.

名称	被検索文書集合	使用する単語予測モデル
方法 1	文集合	CBOW-4
方法 2	文集合	WO-5
方法 3	段落集合	CBOW-4
方法 4	段落集合	WO-5

による変化を調べる。

### 5.2.2 方法

まず、問題文のデータセットとしては、5.1 節と同様に、NTCIR-12 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、5.1 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、上記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、比較する解答処理手法として、表 4 に示す 4 つ

の方法を用意した。まず、文書検索モジュール内で被検索文書集合とする文書集合について、表 1 に示した 2 種類を比較する。また、問題文解析モジュールで使用する単語予測モデルについて、語順が考慮されているか否かで比較する。なお、各モデルの構築条件については、5.1.3 項の結果に照らし、語順が考慮されていないモデルについては CBOW-4 モデル、語順が考慮されているモデルについては WO-5 モデルを使用し、どちらも、モデルの予測候補出力件数を 1 件にする。

解答候補評価モジュールにおけるスコアリングの各指標について、各パラメータを設定する。本実験で設定すべきパラメータは、4 章で定義した、定数  $a$ ,  $b$ ,  $c$  となる。我々は、予備実験として、5.1 節で用いた問題文のデータセットを自動解答させた際に、正答率が最も高くなるようなパラメータの値を調べた。ただし、4.3 節でも述べたように、問題文にすでに出現している単語が正解単語となることは、非常に考えにくいので、問題文非既出単語判定による指標の定数は、 $c = 50$  とした。この値は、各解答候補  $w$  の基準スコアとなる BM25 のスコアのとりうる値を考慮し、決定したものである。これをふまえたうえで、カテゴリ不一致判定による指標の定数  $a$ 、および、後方一致判定による指標の定数  $b$  について、それぞれ 10~90 の範囲を 10 ずつ変化させたところ、 $a = 10$ ,  $b = 30$  としたとき、最も正答率が高くなった。よって、以降の実験でのパラメータは、以上の値を使用する。

最後に、本実験における正答・誤答の定義を表 5 に示す。解答候補評価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。

5.2.3 結果

実験で得られた解答結果の内訳を表 6 に示す。方法ごとに、正答率、誤答率およびそれぞれの問題数を記している。

まず、検索対象となる文書集合による違いについては、単独の正答率で見ると、文集合の場合のほうが高くなっているが、同率も含めた全体的な正答率は、段落集合の場合のほうが、全体的な正答率が向上している。また、使用する単語予測モデルによる違いについては、語順が考慮されていない CBOW-4 モデルを使用した場合のほうが、高い正答率となっている。

表 5 実験における正答・誤答の定義

Table 5 Definition of correct/wrong answer in experiment.

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 6 解答処理手法の違いによる解答結果の内訳  
(カッコ内は該当問題数/全問題数)

Table 6 Breakdown of answer results by difference in answering methods (Correct or incorrect number of questions/total number of questions in parentheses).

処理方法	正答		誤答	
	単独	同率	候補存在	候補不在
方法 1	0.24 (13/54)	0.19 (10/54)	0.41 (22/54)	0.17 (9/54)
方法 2	0.19 (10/54)	0.17 (9/54)	0.48 (26/54)	0.17 (9/54)
方法 3	0.22 (12/54)	0.32 (17/54)	0.28 (15/54)	0.19 (10/54)
方法 4	0.19 (10/54)	0.24 (13/54)	0.39 (21/54)	0.19 (10/54)

表 7 各指標の使用有無の違いによる解答結果の内訳  
(カッコ内は該当問題数/全問題数)

Table 7 Breakdown of answer results by usage of each scoring method (Correct or incorrect number of questions/total number of questions in parentheses).

解答候補評価モジュールの 状態	正答		誤答	
	単独	同率	候補存在	候補不在
方法 3 (基準)	0.22 (12/54)	0.32 (17/54)	0.28 (15/54)	0.19 (10/54)
カテゴリ不一致判定による指標を除外	0.19 (10/54)	0.37 (20/54)	0.26 (14/54)	0.19 (10/54)
後方一致判定による指標を除外	0.04 (2/54)	0.44 (24/54)	0.33 (18/54)	0.19 (10/54)
問題文非既出単語判定による指標を除外	0.19 (10/54)	0.33 (18/54)	0.30 (16/54)	0.19 (10/54)

5.3 実験 3：解答候補評価モジュールにおける各指標の効果

5.3.1 目的

本実験では、3.6 節で述べた解答候補評価モジュールの各指標について、正答率の向上にどの程度貢献しているのかを明らかにする。

5.3.2 方法

評価用の問題データは、5.2 節の実験 2 と同様のデータを使用する。自動解答手法については、同実験で同率を含めた全体の正答率の高かった表 4 の方法 3 を基準とする。提案手法で導入した 3 つの指標を 1 つずつ除いた場合の正答率および誤答率を調べる。正答率が大きく下がった場合、当該指標は正答率向上に大きく貢献しているといえることができる。また、カテゴリ不一致判定を行わない場合というのは、問題文解析モジュールで、カテゴリ推定を行わなかった場合と同義となる。

5.3.3 結果

実験で得られた解答結果の内訳を表 7 に示す。解答候補評価モジュールの状態ごとに、正答率、誤答率およびそれぞれの問題数を記している。結果より、後方一致判定による指標をスコアリングで使用しなかった場合の正答率が著しく低下していることが分かる。

6. 考察

5 章の各実験について考察する。

6.1 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および出力件数

の違いにより、カテゴリ推定の精度に大きく差が出るということが分かった。

まず、ウィンドウサイズの観点から結果を見ると、CBOWでは4、WOでは5をピークとして、それよりもサイズが小さい場合・大きい場合ともに、精度が下がる傾向があることが分かる。ウィンドウサイズが小さすぎる場合、手がかりとなる語数が少なくなるため、精度が下がることが考えられる。反対に、ウィンドウサイズが大きすぎる場合、手がかりとなる語数は増えるが、その単語の中には、穴埋め部分の正解単語とは無関係な単語も含まれやすくなる。また、穴埋め部分がより文頭または文末に近い場合、指定サイズ分の単語を確保することができず、確保できなかった部分は未知語の扱いとなる。図5に示す中間層のベクトル  $H$  を生成する際に、未知語が多く含まれていると、精度に影響が出ると考えられる。

CBOWの各モデルの中で、最も精度が高かったのは、CBOW-4モデルの出力件数5件の場合で、MAPは0.343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-5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されることが分かった。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。

## 6.2 考察2：自動解答の正答率

実験2では、文書検索モジュールにおいて被検索文書集合とする文書集合の違いや、問題文解析モジュールにおいて使用する単語予測モデルの違いによる自動解答の正答率の変化を見た。

まず、被検索文書集合とする文書集合については、知識源の1文を1文書としている文書集合よりも、1段落を1文書とした段落集合のほうが、全体的な正答率が高くなっている。具体的には、単独正答率の向上に現状ではつながらなかったものの、解答候補群の正解単語のスコアが同率1位となった割合は高くなっている。これは、3.4節でも述べたとおり、問題文に関連する文書が、検索結果としてBM25のスコアが高い状態でヒットしやすくなり、かつ、同率の基準スコアである解答候補  $w$  が多く存在するためだと考えられる。

次に、使用する単語予測モデルについては、語順が考慮されていないモデルのほうが、考慮されているモデルよりも正答率が高くなった。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、解答候補評価モジュールにおいて、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、仮に単語予測モデルによるカ

表8 従来手法との正答率の比較  
(カッコ内は正答問題数/全問題数)

Table 8 Comparison of correct answers with conventional methods (Correct or incorrect number of questions/total number of questions in parentheses).

手法	正答率
提案手法 (方法1)	0.28 (15/54)
提案手法 (方法2)	0.22 (12/54)
提案手法 (方法3)	0.28 (15/54)
提案手法 (方法4)	0.20 (11/54)
Takadaら [4]の手法	0.00 (0/54)
Sakamotoら [5]の手法	0.26 (14/54)

テゴリ推定精度が100%であった場合、方法1および方法2における単独正答率は0.40、方法3および方法4における単独正答率は0.33となることが分かった。よって、現状の自動解答手法のみでは、十分な正答率を達成することは困難といわざるをえない。今回我々が構築した単語予測モデルは、「解答カテゴリ」という粒度で穴埋め部分を推定している。今後、この予測モデルが改良され、「単語」の粒度でも高い精度で穴埋め部分を推定することができるようになれば、より直接的な解答導出による正答率向上につながる可能性が考えられる。

最後に、実験2の各解答方法におけるシステムの最終的な正答率と、同一の問題データセットを用いて実験している従来研究の最終的な正答率との比較を表8に示す。従来手法の正答率については、NTCIR-12 QA Lab-2タスクが発表している各チームの解答結果 [17] から、穴埋め型問題のみを抜粋して正答率を再計算したものである。なお、提案システムでの最終的な正答率は、単独正答した問題数と、同率正答のうち偶然正答できた問題数の合計の割合となる。たとえば、表8における提案手法(方法1)の正答数15問は、表6で示されている方法1の単独正答数13問と、同率正答10問のうち偶然正答できた2問との合計となる。この表から分かるのとおり、従来研究と比べても、提案手法による大きな正答率の向上は見られなかった。このことから、単語予測モデルの改良が課題であるといえる。

## 6.3 考察3：各指標の効果

実験3では、解答候補評価モジュールにおける各指標の効果について見た。結果として、後方一致判定による指標を無効とした場合、ベースとなる手法と比べて単独正答率が1/5程度と大幅に下がり、本指標が穴埋め型問題の正答率向上に大きく貢献していることが分かる。つまり、穴埋め部分の直後に、接尾辞などの名詞が存在した場合、当該名詞は問題解答のための大きな手がかりとなり、問題に正解する可能性が高くなると考えられる。

また、カテゴリ不一致判定による指標を無効とした場合は、正答率にあまり変化は見られなかったが、これは6.2節

でも述べたとおり、現状のカテゴリ推定の精度自体が低い  
ためであると考えられる。

## 7. 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。具体的には、従来研究で用いられているファクトイド型質問応答システムの処理をベースとし、問題文解析モジュールにおいて、穴埋め部分の周辺単語から語順を考慮した解答カテゴリ推定手法を導入し、解答候補評価モジュールにおいて、カテゴリとの一致性や周辺単語の既出状況などを用いたスコアリング指標を取り入れた。

実験では、単語予測モデルを用いたカテゴリ推定の精度、手法の違いによる自動解答の正答率、解答候補評価モジュールにおける各指標の効果を調べた。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ほうが、考慮しない場合より推定精度が改善されることを確認した。また、後方一致判定による指標は、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、新たなスコアリング指標の導入なども検討する。

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。

## 参考文献

[1] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J. et al.: Building Watson: An overview of the DeepQA project, *AI magazine*, Vol.31, No.3, pp.59-79 (2010).

[2] Iyyer, M., Boyd-Graber, J.L., Claudino, L.M.B., Socher, R. and Daumé III, H.: A Neural Network for Factoid Question Answering over Paragraphs., *EMNLP*, pp.633-644 (2014).

[3] Murata, M., Utiyama, M. and Isahara, H.: Japanese Question-Answering System Using Decreased Adding with Multiple Answers at NTCIR 5, *NTCIR-5 Workshop Meeting* (2005).

[4] Takada, T., Imagawa, T., Matsuzaki, T. and Sato, S.: SML Question-Answering System for World History Essay and Multiple-choice Exams at NTCIR-12 QA Lab-2, *12th NTCIR Conference on Evaluation of Information Access Technologies*, pp.421-424 (2016).

[5] Sakamoto, K., Ishioroshi, M., Matsui, H., Jin, T., Wada, F., Nakayama, S., Shibuki, H., Mori, T. and Kando, N.: Forst: Question Answering System for Second-stage Examinations at NTCIR-12 QA Lab-2 Task, *12th NTCIR Conference on Evaluation of Information Access Technologies*, pp.467-472 (2016).

[6] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).

[7] 有賀竣哉, 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).

[8] Sato, T.: Neologism dictionary based on the language resources on the Web for Mecab (2015).

[9] Kimura, T., Nakata, R. and Miyamori, H.: KSU Team's Multiple Choice QA System at the NTCIR-12 QA Lab-2 Task, *12th NTCIR Conference on Evaluation of Information Access Technologies*, pp.437-444 (2016).

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

[11] 尾形 勇: 世界史 B, 東京書籍 (2007).

[12] 相良 匡: 新選世界史 B, 東京書籍 (2007).

[13] 加藤晴康: 世界史 A, 東京書籍 (2008).

[14] 今泉 博: 山川 一問一答世界史, 山川出版社 (2015).

[15] Robertson, S., Zaragoza, H. et al.: The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends® in Information Retrieval*, Vol.3, No.4, pp.333-389 (2009).

[16] Tokui, S., Oono, K., Hido, S. and Clayton, J.: Chainer: a Next-Generation Open Source Framework for Deep Learning, *Proc. Workshop on Machine Learning Systems in NIPS* (2015).

[17] Shibuki, H., Sakamoto, K., Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T. and Kando, N.: Overview of the NTCIR-12 QA Lab-2 Task, *12th NTCIR Conference on Evaluation of Information Access Technologies*, pp.392-408 (2016).



田上 諒

2017年京都産業大学コンピュータ理工学部ネットワークメディア学科卒業。現在、同大学大学院先端情報学研究科博士前期課程在学中。主に、情報検索、自然言語処理、質問応答に関する研究に従事。日本データベース学会

会員。



木村 輔

2013年京都産業大学コンピュータ理工学部インテリジェントシステム学科卒業。2016年同大学大学院先端情報学研究科博士前期課程修了。現在、同大学院先端情報学研究科博士後期課程在学中。主に、自然言語処理、情報抽出、パターン認識、質問応答の研究に従事。日本データベース学会

会員。



宮森 恒 (正会員)

1992年早稲田大学理工学部電子通信学科卒業。1994年同大学大学院理工学研究科修士課程修了，1997年同大学院理工学研究科後期博士課程修了。1996～1997年同大学理工学部助手。1997年郵政省通信総合研究所入所。

独立行政法人情報通信研究機構主任研究員サブグループリーダー兼務を経て，2008年京都産業大学コンピュータ理工学部准教授。2013年より同大学同学部教授。工学博士。主に，マルチメディアデータ工学，パターン認識，情報検索に関する研究に従事。2006年日本データベース学会平成17年度論文賞。ACM，電子情報通信学会，日本データベース学会，人工知能学会，自然言語処理学会，映像情報メディア学会各会員。

(担当編集委員 宮尾 祐介)