世界史大論述問題解答の自動生成に向けた指定語句による検索結果の分析

飯塚章裕[†1]　福原優太[†1]　阪本浩太郎[†1†2]　渋木英潔[†1]　森辰則[†1]

[†1]横浜国立大学　[†2]国立情報学研究所
E-mail: {iizuka_a, yuta_f, sakamoto, shib, mori}@forest.eis.ynu.ac.jp

1871年から73年にかけて、岩倉具視を特命全権大使とする日本政府の使節団は、合衆国とヨーロッパ諸国を歴訪し、アジアの海港都市に寄航しながら帰国した。その記録『米欧回覧実記』のうち、イギリスにあてられた巻は、「この連邦王国の……形勢、位置、広狭、および人口はほとんどわが邦と相比較す。ゆえにこの国の人は、日本を東洋の英国と言う。しかれども営業力をもって論ずれば、隔たりもはなはだし」と述べている。その帰路、アジア各地の人々の状態をみた著者は、「ここに感慨すること少なからず」と記している。(引用は久米邦武『米欧回覧実記』による。現代的表記に改めた所もある。)

世界の諸地域はこのころ重要な転機にあった。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。

<指定語句>

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: 大論述問題の例（東京大学、2008 年）

1　はじめに

近年、大学入試問題をコンピュータに解かせる試みとして、「ロボットは東大に入れるか」プロジェクト [1](以下、東ロボプロジェクト[1]) や NTCIR[2]の QALab-3[2] がある。QALab では世界史の大学入試問題 (センター試験、二次試験) を対象とした課題が設定されており、図 1 に示すような大論述問題も出題されている。論述問題は解答の制限字数から、500 字前後で解答する大論述問題と、150 字以内で解答する小論述問題に分類できる。図 1 の下側で示されているように、大論述問題では「インド大反乱」や「クリミア戦争」といった指定語句が 7 語から 9 語程度与えられ、それらの指定語句を解答に必ず含めなければならない。そのため、指定語句に関連する文章を知識源等から抽出することが必須である。

我々は文献 [3] などにおいて、大論述問題を解答するシステムの開発を行っている。このシステムでは、大学入試の世界史問題を教科書や用語集などを知識源とした情報要求課題と捉えるとともに、論述問題を制限字数以内に収める必要があることから、情報要求の存在する抽出型の複数文書要約と位置付けて解答している。つまり、システムは、要求された情報を知識源から探してくる前半の検索部と、探した情報を適切にまとめる後半の要約部の 2 段階で構成されている。このシステムは一定の成果を収めているが、解答に含めるべき記述が検索できないなど改善の余地がいくつか残されている。

見出し語：万国博覧会
説明文：各国の産業・技術・製品を展示する国際的な博覧会。その開催国の国力誇示の場ともなった。第1回は1851年ロンドン，第3回は55年パリで開催された。日本（幕府・薩摩藩・佐賀藩）は67年のパリで開かれた万国博覧会に初参加（出品）した。

図 2: 模範解答に必要だが指定語句の表現を含んでいないパッセージの例 (用語集)

我々は文献 [4] において、模範解答[3]と知識源との対応関係を調査した結果、知識源と対応付けられる模範解答中の記述の割合が文字単位で 85.1%であることが分かった。したがって、システムの検索部における問題は、知識源の不足よりも、知識源中の適切な記述 (以下、単に「適切な記述」と記す) を見つけられないことが大きいといえる。適切な記述を見つけられない原因としては、適切な記述中の表現が問題文の表現と一致しないことと、適切な記述中の語句が問題文に書かれていないことの 2 通りが考えられる。前者に焦点を当てると、検索クエリとなる問題文の表現を知識源の表現に対応させるクエリ拡張の枠組みが必要である。例えば、人間であれば「クリミア戦争」という表現と「クリミア半島を中心に起きた紛争」という表現を対応付けることは可能である。さらに、世界史を勉強した人間であれば、「第 1 回万国博覧会」と「ロンドン万国博覧会」を対応付けることも可能である。図 1 において人名である「ビスマルク」以外は「クリミア戦争」(「クリミア」＋「戦争」) のような複合語であり、指定語句の多くは複合語であるといってよい。本稿では、世界史分野において、問題文と知識源の間でどのような表現の言い換えが行われているかを分析し、指定語句をそのまま検索クエリとして用いた場合と比較して、指定語句をその構成語に分解した語の集合を検索クエリに用いた場合にどのような影響があるかを調査する。

[1]https://21robot.org/
[2]http://research.nii.ac.jp/ntcir/index-ja.html
[3]赤本の解答を模範解答とした。

表 1: 完全一致で検索できない要因の類型と該当数

要因	該当数
指定語句の構成語がパッセージ中に散在している（散在型）	85
指定語句の構成語が異表記である（異表記型）	6
指定語句の構成語が同義語により言い換えられている（同義型）	34
指定語句の構成語が上位概念語により言い換えられている（上位概念型）	50
指定語句の構成語がシリーズの番号で言い換えられている（シリーズ型）	3

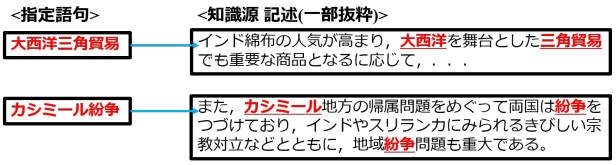

図 3: 指定語句の構成語がパッセージ中に散在している例

2 従来手法の問題点

従来手法 [3] の検索部では、問題文で示される指定語句（図 1 参照）を検索クエリとして受け取り、検索クエリを含む一文を出力とする検索モジュールを検索エンジン Indri[4]を基に作成している。知識源として、山川出版社[5]の教科書 1 冊 [5] と東京書籍[6]の教科書 3 冊 [6, 7, 8] と山川出版社の用語集 1 冊 [9] を用いている。Indri では、教科書は 1 つの段落、用語集は 1 つの用語とその説明文を 1 パッセージ（文書）として文字 unigram でインデキシングしている。しかしながら、従来手法では検索クエリと完全一致する記述を含むパッセージのみを検索しており、検索クエリと表現が異なるパッセージを検索することはできなかった。

例えば、図 1 で示す東大の過去問 2008 年の指定語句には「第 1 回万国博覧会」があり、この問題に対する赤本[7]の模範解答には「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」という記述が存在する。知識源の中には、この模範解答の記述と同等の内容を含むパッセージとして、図 2 に示すパッセージが存在するが、指定語句である「第 1 回万国博覧会」という表現が含まれていないため従来手法で検索することはできなかった。このパッセージを検索するためには、「万国博覧会」や「ロンドン[8]」などを検索クエリとして用いる必要があると考えられる。

[4]http://www.lemurproject.org/indri/
[5]https://www.yamakawa.co.jp/
[6]https://www.tokyo-shoseki.co.jp/
[7]https://akahon.net/
[8]第 1 回万国博覧会の開催地

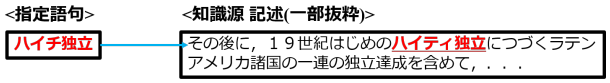

図 4: 指定語句の構成語が異表記である

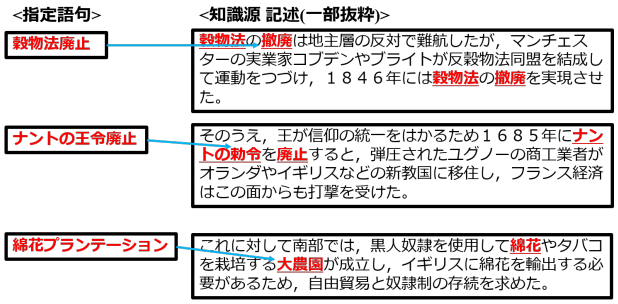

図 5: 指定語句の構成語が同義語により言い換えられている

3 完全一致で検索できない事例の分析

従来手法の検索モジュールでは、指定語句（検索クエリ）との完全一致による検索を行っているため、検索結果に解答に必要とされるパッセージが含まれない場合がある。そのため、解答に必要な記述を含むパッセージが指定語句の完全一致では検索できない事例を人手で調査した。調査対象としたデータは、2007 年から 2013 年までの東京大学入試における世界史科目第一問の 7 問である。完全一致で検索できなかった事例は全部で 121 件あった。これらの事例を分析すると、表 1 に示す 5 つの要因に類型化できることが分かった。表 1 に各要因に該当するパッセージ数を示す。ただし、複数の要因があるパッセージも存在するため、合計は 121 件にならない。

第 1 のタイプは「指定語句の構成語がパッセージ中に散在している」場合（以下、**散在型**）であり、その例を図 3 に示す。指定語句を構成する語（構成語）には、図中で下線を引いている。指定語句の「大西洋三角貿易」の構成語である「大西洋」と「三角貿易」がパッセージ中の別の場所にそれぞれ記述されている。

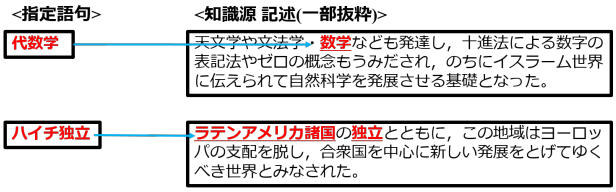

図 6: 指定語句の構成語が上位概念語により言い換えられている

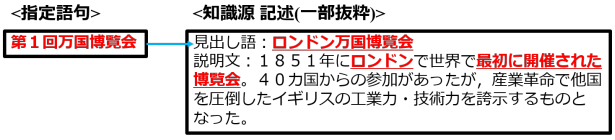

図 7: 指定語句の構成語がシリーズの番号で言い換えられている

これに該当する事例は 85 件あった。

第 2 のタイプは「指定語句の構成語が異表記である」場合（以下、**異表記型**）であり、その例を図 4 に示す。指定語句の「ハイチ独立」の構成語である「ハイチ」がパッセージでは異表記の「ハイティ」となっている。これに該当する事例は 6 件あった。

第 3 のタイプは「指定語句の構成語が同義語により言い換えられている」場合（以下、**同義型**）であり、その例を図 5 に示す。指定語句の構成語である「廃止」、「王令」、「プランテーション」がパッセージでは同義語の「撤廃」、「勅令」、「大農園」となっている。これに該当する事例は 34 件あった。

第 4 のタイプは「指定語句の構成語が上位概念語により言い換えられている」場合（以下、**上位概念型**）であり、その例を図 6 に示す。指定語句の構成語である「代数学」や「ハイチ」がパッセージでは上位概念の「数学」や「ラテンアメリカ諸国」となっている。これに該当する事例は 50 件あった。

第 5 のタイプは「指定語句の構成語がシリーズの番号で言い換えられている」場合（以下、**シリーズ型**）であり、その例を図 7 に示す。指定語句の構成語である「万国博覧会」は複数回開催されたシリーズの概念であり、番号の「第 1 回」によって「ロンドン万国博覧会」を示している。これに該当する事例は 3 件あった。

以上の事例を概観すると、指定語句の多くが複合語であり、完全一致で検索できなかったパッセージにおいても、それぞれの構成語が（同義語や上位概念など表現が違っても）基本的には全て含まれているといえる[9]。ただし、それらの構成語が連続して現れるとは限らない。したがって、指定語句を構成語に分解し、構成語を全て含むパッセージを検索することで、従来手法で検索できなかったパッセージがある程度検索できるようになるのではないかと考えた。なお、解答では指定語句の表現をそのまま用いる必要がある。そのため、指定語句を構成語に分解するなどして検索クエリの制約を緩めた場合に得られたテキストにおいて、対応する箇所を指定語句に言い換える必要がある。しかしながら、本稿ではその点について扱わない。今後の課題である。

[9]唯一の例外は、「第 1 回万国博覧会」の指定語句に関連する「世界各国の科学技術の成果を一堂に集めて展示する万国博覧会もひらかれるようになった」というパッセージで、「ひらかれるようになった」という記述から暗に「第 1 回」であることを示しているが、対応する構成語が含まれているわけではない。

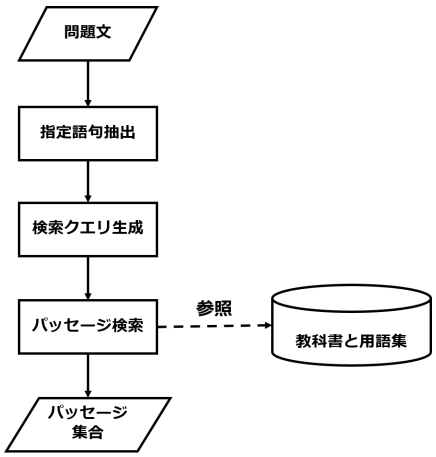

図 8: 調査の流れ

表 2: 調査の結果

検索クエリ	再現率	適合率
完全一致（従来手法）	0.649	0.348
構成語の集合	0.816	0.322

4 構成語に分解した検索クエリの調査

指定語句を構成語に分解した検索クエリを用いることで、パッセージの検索精度にどのような影響があるかを調査する。調査の流れを図 8 に示す。与えられた問題文から指定語句を抽出し、指定語句を形態素解析器 MeCab[10] を用いて解析する。解析結果の形態素の集合を構成語の集合として検索クエリを生成する。従来手法と同じ Indri による検索モジュールを用いて AND 検索し、パッセージ集合を求める。

調査対象は、3 節の分析対象と同じ、2007 年から 2013 年までの東京大学入試における世界史科目第一問の 7 問である。7 問で示された指定語句は全部で 56 語句であったが、構成語に分解するのが適切ではない人名と国名を除外した 49 語句について調査した。比較

[10]http://taku910.github.io/mecab/

表 3: 構成語の集合を用いることによる類型ごとの改善率

	散在型	異表記型	同義型	上位概念型	シリーズ型
改善率	0.624 (53/85)	0.000 (0/6)	0.588 (20/34)	0.180 (9/50)	0.333 (1/3)

対象として、従来手法と同じ、指定語句をそのまま検索クエリとして完全一致で検索した場合と比較した。検索精度の指標として再現率 R と適合率 P を用い、模範解答の記述と同等の内容を含むパッセージ（重要パッセージ）がどの程度検索できたかを以下の式で計算した。

$$R = \frac{\text{検索された重要パッセージ数}}{\text{重要パッセージ数}} \quad (1)$$

$$P = \frac{\text{検索された重要パッセージ数}}{\text{検索されたパッセージ数}} \quad (2)$$

調査結果を表 2 に示す。指定語句を検索クエリとしてそのまま用いた場合と比較して、構成語の集合の AND 検索を用いた場合は、適合率が 0.348 から 0.322 に僅かに低下したものの、再現率が 0.649 から 0.817 に上昇した。文献 [3] のシステムにおいて、検索部が要約部の前処理であることを考慮すると再現率を重視した方が良いと考えられ、適合率の減少も相対的に小さいことから、構成語の集合の AND 検索を前提としたクエリ拡張を行った方が良いと考えられる。

構成語の集合の AND 検索が、表 1 で示した類型にどのように効果があったかを考察する。表 3 に、類型ごとの改善率を示す。括弧内は、分母が指定語句の完全一致で検索できなかったパッセージ数（表 3 参照）、分子が構成語の集合で検索できるようになったパッセージ数である。散在型、同義型、シリーズ型、上位概念型、異表記型の順に改善率が高かった。しかしながら、構成語に分解する方法が最も適していると思われた散在型においても 6 割程度の改善率であった。3 節で述べたように、検索できないパッセージには複数の要因があるため、該当する全ての要因を解決しなければパッセージを検索できない。そのため、今後、異表記や同義語などによるクエリ拡張を行うことで改善していきたい。

5 まとめ

本稿では、世界史分野において、問題文と知識源の間でどのような表現の言い換えが行われているかを分析し、問題文の表現をそのまま検索クエリとして用いた場合と比較して、問題文の表現を形態素に分解した構成語の集合を用いた場合にどのような影響があるかを調査した。完全一致で検索できない要因を人手で調査した結果、散在型、異表記型、同義型、上位概念型、シリーズ型の 5 つの類型があることが分かった。また、指定語句を検索クエリとしてそのまま用いた場合と比較して、構成語の集合の AND 検索を用いた場合は、適合率が 0.348 から 0.322 に僅かに低下したものの、再現率が 0.649 から 0.817 に上昇した。今後、分析結果を基にしたクエリ拡張手法を開発したいと考えている。

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] 新井紀子, 松崎拓也. ロボットは東大に入れるか?-国立情報学研究所「人工頭脳」プロジェクト. 人工知能学会論文誌, (2012).

[2] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, Noriko Kando. Overview of the NTCIR-13 QA Lab-3 Task. Proceedings of the 13th NTCIR Conference, 2017.

[3] 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, pp. 553556, (2016).

[4] 福原優太, 阪本浩太郎, 渋木英潔, 森辰則. 世界史論述問題における模範解答－知識源の対応を表すアノテーションの検討. 言語処理学会第 23 回年次大会発表論文集, pp. 593–596, 2017.

[5] 株式会社山川出版社・世界史 B 詳説世界史 改訂版 (世 B 016)

[6] 東京書籍株式会社・世界史 A(平成 20 年度発行)

[7] 東京書籍株式会社･新撰世界史 B(平成 19 年度発行)

[8] 東京書籍株式会社･世界史 B(平成 19 年度発行)

[9] 株式会社山川出版社・世界史用語集 改訂版 (平成 20 年度発行)