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

石下円香[†1] 阪本浩太郎[†1†2†3] 中山周[†3†4] 渋木英潔[†2] 森辰則[†2] 神門典子[†1]

[†1]国立情報学研究所 [†2]横浜国立大学 [†3]Carnegie MellonUniversity [†4]東北大学
E-mail: {ishioroshi,kando}@nii.ac.jp,
{sakamoto,shib,mori}@forest.eis.ynu.ac.jp, nakayamas@ecei.tohoku.ac.jp

1 はじめに

現在，国立情報学研究所では，大学入試問題を解く計算機プログラムを開発することを目的とした、人工知能プロジェクト (以下，東ロボプロジェクトとする) を進めている [3]．このプロジェクトでは，2016 年までに大学入試センター試験 (以下，センター試験とする) において高得点をマークし，2021 年までに東大二次試験に合格することを目指している．また，NTCIR では東ロボプロジェクトと連携した QALab-2 タスクが開催されている．QALab-2 タスク [1] では，世界史の大学入試問題 (センター試験，東大を含む 5 大学の二次試験，センター試験模試，東大模試) を対象とした課題が設定されている．

我々は，QALab-2 タスクにおいてはすべての試験について解答を提出したが，特に東大二次試験 (及び東大模試) を対象としたシステム構築を行った．東大二次試験の問題は，3 問の大問から成っており，第 1 問が指定文字数 500 文字程度の大論述問題，第 2 問が指定文字数数百文字程度の小論述問題，第 3 問が一問一答型の語句問題となっている．本稿では，第 2 問の小論述問題と第 3 問の語句問題に焦点を当てる [2]．東大入試第 2 問，第 3 問の特徴及び，それらに解答するために構築したシステムについて述べる．

2 関連研究

歴史問題などの知識を問う問題に対するアプローチとして，含意関係認識を使ったアプローチと，イベントオントロジー [2] を用いるアプローチなどがある [6]。これらのアプローチは，センター試験問題，特に言明の真偽を問うタイプの問題に焦点を当てており，二次試験の解答作成においてはそのままでは利用できないが，イベントオントロジーは世界史の知識が詰まった重要な知識源といえる．本手法では，イベントオントロジーに含まれるインスタンスデータを固有表現辞書の基として利用した．

語句問題に解答するシステムとしては，アメリカのクイズ番組 Jeopardy!のクイズに答えるシステムの Watson[1] が挙げられる．Watson では，Jeopardy!のクイズに対応するため、多数の質問タイプが用意されている．本手法では，イベントオントロジーのインスタンスデータを基に，5 節に示すように，世界史に特化した質問タイプを用意している．

3 大学入試問題

東大二次試験では，3 問の大問から成っており，大問毎に，大論述，小論述，語句と解答の形式がやや異なっている．本稿では，小論述問題と語句問題を対象とする．

東大第 2 問では，図 1 上のような，1 行～数行 [3] で解答する小論述問題が出題される．解答に入れるべき語句 (指定語句) はなく，数十～数百文字での解答が求められる．

東大第 3 問では，図 1 下のような，語句を答える問題が出題される．既存の質問応答システムにおいて用意されている代表的な質問タイプは，人名，地名，組織名，数量といったものだが，図 1 の例では,「記録・伝達手段」が解答として求められており，一般的な質問とは異なる範囲の固有表現が解答として用いられていることが分かる．

4 小論述問題解答システム

本節では，東大第 2 問の小論述問題に解答するシステムについて，提案するシステム構成と評価実験について述べる．小論述問題解答システムは，阪本ら [4] に示す東大第 1 問の大論述問題に対する手法を利用した．

[1] http://research.nii.ac.jp/qalab/
[2] 第 1 問に対する取り組みは，阪本ら [4] を参照されたい．
[3] 1 行は 30 文字

小論述問題の例

中国王朝の首都を考えると，華北においては，唐代までは…その様子は張択端の「清明上河図」に描かれている。
（b）明代の長江流域の農業・工業について，2 行以内で説明しなさい。

語句問題の例

アメリカ大陸では，中米のメソアメリカ文明が独自の文字を使用していたのに対し，南米のアンデス地方を支配したインカ帝国には文字がなく，ほかの手段を用いて数量などの情報を記録・伝達していた。インカ帝国で用いられた記録・伝達手段の名称を記しなさい。

図 1: 小論述問題と語句問題の例

解候補のスコア付けなどの詳細に関しては，阪本ら [4] を参照されたい．

4.1 小論述問題解答システムの構成

小論述問題解答システムでは，図 1 上の (b) の文を質問文として入力し，処理を行う．

まず，質問文解析を行い，質問文キーワードの抽出と指定語句に相当する語 (以下，仮想指定語句) の決定を行う．阪本ら [4] の手法では，文書検索に指定語句を用いているが，小論述問題では指定語句が示されていないため，仮想指定語句を用意する必要がある．小論述問題の質問文には，解答にいれるべき固有表現が直接現れていない場合が多いため，教科書などの知識源から問と関係が深そうな固有表現を抽出し，それを仮想指定語句とした．固有表現の抽出には，5 節で述べる語句問題解答システムの手法を流用する．問の文から質問焦点に相当する語を抽出し，抽出した質問焦点を使って語句問題解答システムと同様の流れで仮想指定語句を抽出する．1 の例では,「農業」「工業」が質問焦点として抽出され，それぞれ「稲作」「家内制手工業」が仮想指定語句として抽出される．

次に，仮想指定語句を用いて知識源の検索を行う．知識源として，東京書籍の教科書 (世界史 A, 世界史 B, 新選世界史 B)，山川出版社の教科書 (諸説世界史) 及び用語集を用いた．教科書は一段落，用語集は 1 つの用語説明を一文書とした．文書検索エンジンは，indri[4] を用いた．

仮想指定語句ごと関連文書を抽出し，関連文書から仮想指定語句を含む文を句点区切りで抽出する．そして，それぞれの文に解候補らしさのスコアをつける．解答らしさの判定には，質問文キーワードが含まれる度合い (包含度) や時間情報が一致しているかどうかの情報が考慮される．

仮想指定語句ごとの文集合から一文ずつ選択して組み合わせることで解候補を生成する．すべての解候補が指定文字数に収まらない場合は，仮想指定語句を減らして再度解候補を生成する．各文の解候補らしさのスコアの和が最大になるような解候補を，最終的な解答として出力する．

4.2 評価実験及び考察

小論述問題解答システムの評価実験を行った．比較手法として，仮想指定語句を質問文から直接抽出する手法を用いた．質問文の焦点になっている名詞以外の名詞を仮想指定語句として抽出した．図 1 の例では,「長江」「明代」が抽出された．

使用した試験問題は，駿台東大模試 (2013 年度第 2 回，2015 年度第 1 回) の第 2 問 (小論述問題) の計 10 問である．解候補の評価の際には，駿台が公表している正解の加点ポイントを利用した．図 1 の問に対する正解例と加点ポイントを図 2 に示す．出力した解候補に一つでも加点ポイントが含まれていれば正解とし，精度を算出した．加点ポイントが含まれているかどうかの判定は，自然言語処理を研究する大学院生 4 名が行った．判定者毎の精度を求めた後，その平均値を全体の精度とした．

(正解例) 下流域で綿織物など家内制手工業や綿花などの原料栽培が広がり，中流域が穀倉地帯となり「湖広熟すれば天下足る」と称された。
・加点ポイント (4 点上限)
1. 下流域で、家内制手工業が広がった (工業化した) こと
2. 下流域で原料 (商品作物・換金作物) 栽培が広がったこと
3. 家内制手工業の内容 (絹織物，綿織物・生糸)or 原料の内容 (綿花，桑)
4. 中流域が穀倉地帯 (米の栽培) の中心となったこと
5. 湖広熟すれば天下足る」という表現

図 2: 小論述問題の正解例と加点ポイントの例

表 1: 小論述問題解答システムの精度

仮想指定語句の抽出元	質問文	知識源
精度	0.18	0.23

結果を表 1 に示す．表 1 より，知識源から仮想指定語句を抽出した方が精度が良いことが分かる．一方で，どちらの場合でも精度の値はおおむね 0.2 と低く，加点ポイントが含まれる解候補が出力されることが少ないことが分かる．実験では，仮想指定語句の抽出に失敗しているために解候補が出力されない場合があった．また，本手法では論述問題で重要な,「問われている内容」に関する尺度が実装されていない．さらなる改良が必要である．

[4] http://www.lemurproject.org/indri.php

5　語句問題解答システム

本節では，東大第 3 問の語句問題に解答するシステムについて，提案するシステムの構成及び実験結果について述べる．

5.1　語句問題解答システムの構成

語句問題解答システムを図 3 に示す．

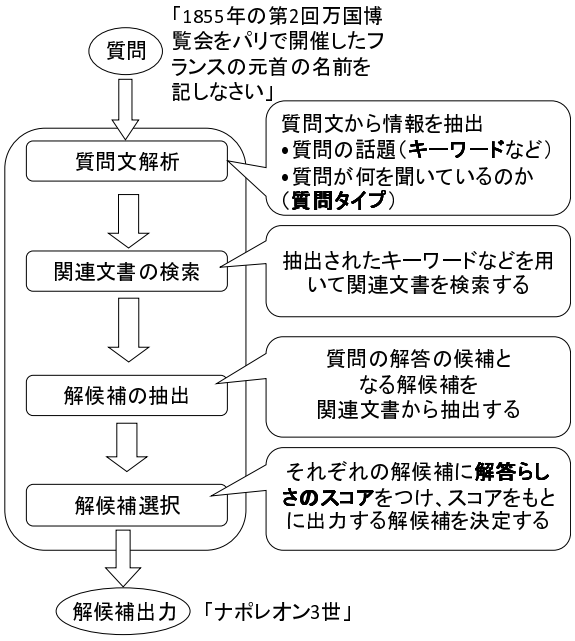

図 3: 語句問題解答システムの構成図

語句問題解答システムでは，複数文にまたがる問の文すべてを質問文として使用する．

質問文解析部分では，質問文から質問文キーワードや質問タイプなどの情報を抽出する．質問文キーワードは，形態素解析の結果から，内容語を抽出した．質問タイプは，質問が対象としている語 (質問焦点) を抽出し，質問焦点を基に判定する．判定の際には，事前に作成した質問焦点と質問タイプを紐付ける正規表現のリストを利用する．用意した質問タイプを以下に示す．カッコ内は代表的なサブクラスである．図 1 の例では，「記録・伝達手段」が質問焦点として抽出され，質問タイプは「技術」と判定される．

人物，**場所** (国，都市，地域)，**出来事** (革命，会議，戦争)，**文明**，**言語**，**技術** (発明，文字，道具)，**時代**，**建造物** (宮殿，道路，寺院)，**民族**，**作品** (小説，詩，絵画)，**制度** (政策，法令，思想)，**組織** (同盟，共同体，結社)，**社会概念** (権利，通貨)，**宗教**，**宗教概念** (神)，**様式**，**数値**

文書検索部分では，知識源から，質問文キーワードを用いて質問に関連する文書を検索する．知識源及び検索エンジンは，4 節の小論述問題解答システムとほぼ同様のものを用いたが，教科書では節で区切って一文書としている点で異なる．

解候補抽出部分では，検索された文書から解候補を抽出する．解候補として，名詞及び名詞連続の複合語を用いた．

解候補選択部分では，解候補にスコアをつけ，スコアが高い解候補を出力する．解候補らしさのスコアには，以下の 2 つの尺度を用いた．

キーワードスコア 解候補が抽出された文にキーワードが含まれる度合い

質問タイプスコア 解候補と質問タイプの合致の度合い

キーワードスコアは，解候補が抽出された文に含まれるキーワード数から算出した．

質問タイプスコアの算出においては，世界史用の固有表現辞書を用意し，これを利用した．固有表現辞書は，イベントオントロジー [2] のインスタンスデータを利用した．インスタンスデータでは概念クラスごとに固有表現が分けられているが，一部のクラスをマージしたのち，それを固有表現クラスとして利用した．もともとの概念クラスはサブクラスとして保存した [5]．インスタンスデータには，概念クラスの他に，開始年，終了年，異表記，その他の情報 (国名クラスにおける首都など) が含まれるため，これらも利用した．また，辞書に不足があった場合には適宜人手で追加している．質問タイプスコアの算出には，以下の指標を用いた．

- (必須) 解候補の固有表現クラスが質問タイプと一致するかどうか
- 解候補の固有表現のサブクラスが質問焦点と一致するかどうか
- 質問文中に時間情報がある場合，解候補の固有表現の年情報に含まれているかどうか
- 解候補の固有表現の「その他の情報」が質問文キーワードとして現れているかどうか

解候補の最終的なスコアは，キーワードスコアと質問タイプスコアを積算して算出する．スコアの最も高いスコアを解答として出力する．

5.2　評価実験及び考察

5 節で述べたシステムの評価実験を行った．本研究では，世界史用の固有表現辞書を作成し，利用している．そのため，固有表現辞書を利用しない従来手法 (松井ら [5]) との比較を行った．

使用した試験問題は，東大二次試験の過去問 (2011 年度) 及び，駿台東大模試 (2013 年度第 1 回，2013 年度第 2 回，2015 年度第 1 回) の第 3 問 (語句問題) の計 41 問である．表 2 に正解数を示す．

表 2 より，提案手法の方が精度がいいことが分かる．提案手法で正解し，従来手法で正解できなかった問題

[5] 人物クラスについては，職業をサブクラスとした．

表 2: 語句問題解答システムの正解数

	提案手法	従来手法
正解数/問題数	19/41	11/41

を見ると，従来手法では問で聞かれている内容とは別の固有表現クラスを解答している例が多かった．このことから，質問タイプを用いることは精度向上に役立っていることが分かる．

また，正解できなかった 22 問について，失敗原因の分析を行った．失敗分析の結果を表 3 に示す．

表 3: 語句問題解答システムの失敗原因

失敗原因	問題数
質問文解析ミス [6]	1
質問タイプ判定誤り [7]	3
文書検索の失敗	5
解抽出失敗 (辞書不足以外)	10
固有表現辞書不足 [8]	3

「質問文解析ミス」は，一つの問文書中に (a)，(b) といった二つの質問が含まれているものである．入力する xml 文書では，質問 (a) の範囲と質問 (b) の範囲が明確には示されていないため，自動での解析に誤りがあった．このような問は全体に占める割合は低いものの，きちんと対応する必要がある．

質問タイプ判定では，未知の質問焦点が抽出された場合失敗する可能性が高い．また，正解が固有表現辞書に載っていなかったために失敗した問もあった．提案手法では，質問タイプに重きを置いているため，質問タイプ判定失敗や固有表現辞書不足のリスクが高い．従来手法では正解したが，提案手法では不正解だった問の多くは質問タイプ判定に関するものであった．これらの問では，質問焦点のリストや固有表現辞書を更新すると正解できる場合が多かった．東大第 3 問の問内容は多岐にわたるため，未知の質問焦点や固有表現をなるべく減らすとともに，質問タイプの判定がうまくいかない場合でも正解が出力できるようなスコアリングを検討する必要がある．

また，正解を含む文書が検索できなかったものや，固有表現辞書不足以外での解抽出の失敗もかなりあった．提案手法では，質問文キーワードはすべて同じ重要度で扱っているが，あまり重要ではないキーワードが文書検索や解抽出において邪魔になっていることが多くあった．複数文にまたがる問の文章のうち，後半に現れるキーワードの方が重要である場合が多いため，後半に現れるキーワードを重要視するなど，キーワードに重要度をつけることで解決できる可能性がある．

[6] 人手で正しい問に直した場合，正解できた．
[7] タイプ判定のための正規表現の修正で正解できた．
[8] 固有表現辞書への追加で正解できた．

6 おわりに

本稿では，東大二次試験 (及び東大模試) の第 2 問の小論述問題と第 3 問の語句問題に解答する質問応答システムについて述べた．

第 2 問の小論述問題に解答するシステムでは，指定語句に相当する語句を抽出する際には，問の文から直接抽出するよりも，一段処理を加えて知識源から抽出する方が有効であることが分かった．しかし，さらに改良を加え，精度を向上させる必要がある．

第 3 問の語句問題に解答するシステムでは，約半数の問で正解することができた．しかし，東大試験においては第 3 問は約 8 割以上の精度が望ましいとされている．そのため，さらなる精度向上が必要である．今後は固有表現辞書の拡充や複数文にまたがる問の解析方法の再検討の他，既存の質問応答システムで有効だった手法の検討をする予定である．

謝辞

本研究の実施にあたっては，東京書籍の教科書，山川出版社の教科書及び用語集のデータを使用した．また，試験データとして，駿台予備校の東大実践模試を使用した．東京書籍株式会社，株式会社山川出版社，駿台予備校には深く感謝いたします．

参考文献

[1] David A Ferrucci. Introduction to "This is Watson". *IBM Journal of Reserch and Cevelopment*, Vol. 56, No. 3.4, pp. 1:1–1:15, 2012.

[2] Ai Kawazoe, Yusuke Miyao, Takuya Matsuzakiand Hikaru Yokono, and Noriko Arai. World history ontology for reasoning truth/falsehood of sentences: Event classification to fill in the gaps be-tween knowledge resources and natural language texts. In *Proceedings of LENLS 10. 2013.*, 2013.

[3] 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, 9 2012.

[4] 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 22 回年次大会発表論文集, 3 2016.

[5] 松井兵庫, 阪本浩太郎, 松永詠介, 神貴久, 渋木英潔, 石下円香, 森辰則, 神門典子. 大学入試の穴埋め問題を解く質問応答システムの検討. 言語処理学会 21 回年次大会発表論文集, 3 2015.

[6] 宮尾祐介, 川添愛. 「大学入試問題を解く」ことから見える言語，知識，世界理解に関する研究課題. 人工知能学会論文誌, 9 2012.