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

福原 優太 †1　阪本 浩太郎 †2　渋木 英潔 †3　森 辰則 †3

†1 横浜国立大学 理工学部　†2 横浜国立大学 大学院 環境情報学府
†3 横浜国立大学 大学院 環境情報研究院
E-mail: {yuta_f,sakamoto,shib,mori}@forest.eis.ynu.ac.jp

1　はじめに

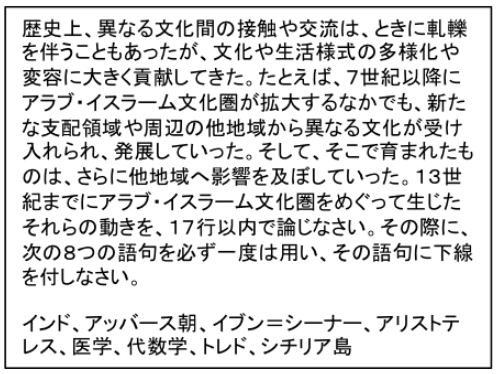
近年，文書情報に対する情報要求は複雑化，高度化しており，そのような要求を満たすアクセス技術として質問応答が注目されており，TREC の Live QA[1] や NTCIR の QA Lab[2, 3] など，現実世界における質問応答を目的とした取り組みも盛んに行われている．質問応答とは，利用者の自然言語による質問に対して情報源となる文書集合から回答そのものを抽出する技術であり，複雑高度な情報要求を自然言語で表現できる点に特徴がある．QA Lab では，世界史の大学入試問題を対象としており，特に論述問題への自動解答がチャレンジングな課題として設定されている．

我々はこれまでに，世界史の教科書や参考書などを知識源とし，知識源から文単位で抽出して論述問題に解答する QA システムを提案しており，QA Lab や東ロボプロジェクト [1] に参加して一定の成果を収めている [4]．しかしながら，模範解答と比較すると，文抽出による解答では内容的に過不足が生じることが多く，より適切な解答を生成するためには，不要な記述を削除したり，必要な記述を別の文から追加したりする必要がある．また，意味的に同一の記述であっても異なる表現であったり，前後の文脈に依存して省略されていたりと，表層だけで対応する記述を自動的に推測するのが困難な場合も多い．それゆえ，我々は模範解答の各記述が知識源中のどの記述と対応しているかを調査し，文抽出を超えた QA システムの開発を目指している．

以上の背景から，開発の基礎となるコーパスを必要としているが，世界史論述問題という専門的な知識や推論を要する分野のアノテーションは自明ではない．従来研究 [5, 6] では専門的な分野を対象としたアノテーションを行っており，その分野に特有の課題に対応するためのアノテーションを提案している．それゆえ，本稿ではプロトタイプ的に世界史論述問題のためのアノテーションを行い，模範解答と知識源との対応関係をアノテーションする際の問題点等を考察する．なお，模範解答として赤本 [2] を，知識源として山川出版社 [9] と東京書籍 [10, 11, 12] の教科書，山川出版社の世界史用語集 [13] を用いている．

2　模範解答と知識源の対応関係

世界史の論述問題と模範解答の例を図 1 と図 2 にそれぞれ示す．また，模範解答の (A1) と (A2) に対応する記述を含む知識源（教科書）の例を図 3 に示す．図 3 中の (a1) と (a2) と (a3) が (A1) と (A2) に対応する記述である．例で示したように，模範解答の記述が知識源の記述と表層的に一致することは殆どなく，文単位で一対一に対応しているとも限らない．それゆえ，文よりも小さい単位での対応付けを可能とするアノテーションが必要がある．


歴史上、異なる文化間の接触や交流は、ときに軋轢を伴うこともあったが、文化や生活様式の多様化や変容に大きく貢献してきた。たとえば、7世紀以降にアラブ・イスラーム文化圏が拡大するなかでも、新たな支配領域や周辺の他地域から異なる文化が受け入れられ、発展していった。そして、そこで育まれたものは、さらに他地域へ影響を及ぼしていった。13世紀までにアラブ・イスラーム文化圏をめぐって生じたそれらの動きを、17行以内で論じなさい。その際に、次の8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。

インド、アッバース朝、イブン＝シーナー、アリストテレス、医学、代数学、トレド、シチリア島

図 1: 2011 年度東京大学の問 1

我々は，1 章で述べたように QA システムの開発を目的としており，内部で文書検索モジュールを利用している．現段階では，教科書であれば教科書の段落単位で，用語集であれば見出し語の単位でインデキシングを行っていることから，知識源側の単位として上記の基準を用いている．以降，本稿ではこの知識源側の単位を検索パッセージと呼ぶ．

3　タグセット

本稿のアノテーションは XML タグで記述する．表 1 にタグの一覧を示す．

大きく模範解答側の情報を保持する <gs> タグと知識源側の情報を保持する <doc_set> タグの 2 部で構成される．<doc_set> の中にある <doc> の単位は検索パッセージと一致する．また模範解答の記述と知識源の記述を，文よりも小さい単位で対応付けるために，start と length という属性により，文字単位で模範解答と知識源から必要な記述をそれぞれ抜き出し，その集合によって構成された記述により両者を対応させる．<gs> タグの構造下にある <gs_index> は模範解答側の文字情報を保持するためのタグであり，<doc_set> タグの構造下にある <doc_index> は知識源側の文字情報を保持したタグである．また <gs> タグ側の <gs_sentence> タグには <gs_index> の組み合わせによる模範解答側の記述を持ち，<doc_set> タグ側の <extraction_text> には <doc_index> の組み合わせによる知識源側の記述を持っている．

4　アノテーション作業

アノテーション作業の過程は図 4 に示す五工程で行った．

最初に，作業者には検索パッセージの集合が与えられている．検索パッセージの内容は XML 構造で記述されており，それぞれ <DOCNO> という固有の ID

[1] http://21robot.org/
[2] https://akahon.net/

表 1: タグ一覧

タグ名	属性	内容
<data>		管理情報．<gs> と <doc_set> に対応する要素を保持するタグ
	version	タグ付きコーパスのヴァージョン情報
	exam_id	論述問題の年度情報
<gs>		<content>,<gs_index_set>,<gs_sentence_set> に対応する要素を保持するタグ
<content>		模範解答全文
<gs_index_set>		<gs_index> の集合に対応する要素を保持するタグ
<gs_index>		模範解答のインデックス情報
	expression_id	模範解答の抽出記述の ID
	start	模範解答の抽出記述の開始位置
	length	模範解答の抽出記述の長さ
<gs_sentence_set>		<gs_sentence> の集合に対応する要素を保持するタグ
<gs_sentence>		<gs_index> により抽出した文集合
	id	模範解答の文の ID
<doc_set>		模範解答と対応する知識源のパッセージ集合
<doc>		模範解答と対応する知識源のパッセージ
	docno	知識源のパッセージ ID
<doc_index_set>		<doc_index> の集合
<doc_index>		知識源のパッセージのインデックス情報
	gs_sentence_id	模範解答の文の ID
	start	知識源の抽出記述の開始位置
	length	知識源の抽出記述の長さ
<extraction_text_set>		<extraction_text> の集合
<extraction_text>		<doc_index> により抽出した文集合
	gs_sentence_id	模範解答の文の ID
<title>		知識源のパッセージ名称
<text>		知識源のパッセージのテキスト

アラブ人は、正統カリフ時代にビザンツ帝国からエジプトを奪い、ササン朝を滅ぼしてイランを支配し、ウマイヤ朝時代には(A1)イベリア半島やインダス川流域まで領域を広げた。(A2)こうした領域の拡大過程で、ビザンツ帝国に保存されていた古代ギリシア・ローマ文化が吸収され、アッバース朝の首都バグダードに建てられた知恵の館ではアリストテレスの哲学やヒポクラテスの医学などの文献がアラビア語に翻訳され、さらに研究が深められた。インドからはゼロの概念や十進法などが伝わり、フワーリズミーによる代数学の確立に大きく寄与した。また、タラス河畔の戦いによって中国から製紙法が伝えられた。これらは、十字軍やレコンキスタの過程でヨーロッパに流入し、トレドやシチリア島においてアラビア語文献がラテン語に翻訳され、「12世紀ルネサンス」に大きな影響を与えた。医学ではイブン=シーナーの『医学典範』が中世ヨーロッパの大学の講義にも使われ、イブン=ルシュドによるアリストテレスの注釈はトマス=アクィナスによるスコラ哲学体系化に寄与した。また、イスラームの暦学に基づいて元の郭守敬が「授時暦」を作成するなどヨーロッパだけではなく中国にも影響を及ぼしている。

図 2: 東京大学 2011 年度「赤本」模範解答

661年，ムアーウィヤがダマスカスを都としてウマイヤ朝をひらき，カリフを世襲制にした。
このとき，ウマイヤ朝の正統性をめぐり，ムスリムはスンナ派とシーア派に分裂した。
ウマイヤ朝は(a1)，正統カリフ時代からの征服をつづけ，東は中央アジア(a2)，西は北アフリカからイベリア半島までの広大な地域を支配し(a3)，アラビア語を公用語として，貨幣の統一も進めた。

図 3: 模範解答に対応する知識源 (教科書) の記述

第一工程　命題を単位とした模範解答の断片化
第二工程　命題に該当する検索パッセージの収集
第三工程　模範解答内の命題の記述の抽出
第四工程　命題に該当する検索パッセージ内の記述の抽出
第五工程　第一工程～第四工程を繰り返す

図 4: アノテーション作業工程の例

とその検索パッセージの内容を端的に表す <TITLE> が振られている．<TITLE> は教科書であれば章のタイトル，用語集であれば見出し語である．

まず第一工程は，知識源から模範解答に対応する記述を収集しやすくするために，模範解答を命題ごとに切り分ける．尚，本稿では単文を命題として扱う．単文は概ね節に対応する．また，一つあるいは複数の単文で構成されているものを文と呼ぶ．第一工程の例として，知識源に「ササン朝を滅ぼしてイランを支配し」とあれば「ササン朝を滅ぼして」と「イランを支配し」に分ける．

第二工程は，まずある一つの命題に該当する記述を見つける．次に知識源からそれを含む検索パッセージごと収集する．この際，再検索の手がかりとして <DOCNO> を <doc> タグの属性である docno に残す．また冗長であっても多くの情報を残すことで考察を進めやすいように <TITLE> を <title> に同時に保存しておく．

第三工程は，知識源の記述に対応する模範解答内の命題中の記述を抽出するために，<gs_index> タグの属性である start と length を設定する．このとき，命題中の記述と知識源の記述の表層が一致していればそのまま，一致していなければ知識源の記述に沿うように命題の記述を対応させる．例えば，知識源に「ウマイヤ朝は中央アジアまでの広大な地域を支配し」というような記述があり，模範解答の記述が「ウマイヤ朝時代にはインダス川流域やイベリア半島まで領域を広げた」となっていれば，start と length を二箇所とることで「ウマイヤ朝時代にはインダス川流域」と「まで領域を広げた」という二つの文字列をつくりあげ，それぞれ <gs_index> タグの属性である expression_id に数字で ID を振る．尚，これは模範解答の命題について知識源に現れる表現に対応する部分だけを抜粋していることに相当する．また，この例では「中央アジア」と「インダス川流域」の違いがあるが，地図上では粒度が異なるだけで同一と見なせるので同じ意味のものとして扱う．その後，同じ expression_id の集合を組み合わせ「ウマイヤ朝時代にはインダス川流域まで領域

```
<data version="1.0.0" exam_id="tokyo_2011">
 <gs>
  <content><![CDATA[ 図2と同様のため省略 ]]></content>
  <gs_index_set>
   <gs_index expression_id="1,14,19,20,22,23,30,33,34,42,44,46" start="1" length="5">アラブ人は</gs_index>
    ……(省略)……
   <gs_index expression_id="2,3,44" start="48" length="9">ウマイヤ朝時代には</gs_index>
   <gs_index expression_id="2,34,42" start="64" length="7">インダス川流域</gs_index>
   <gs_index expression_id="42" start="63" length="1">や</gs_index>
   <gs_index expression_id="3,20,42,44" start="57" length="6">イベリア半島</gs_index>
   <gs_index expression_id="2,3,20,34,42,44" start="71" length="9">まで領域を広げた。</gs_index>
   ……(省略)……
 </gs_index_set>
  <gs_sentence_set>
   <gs_sentence id="1">アラブ人はササン朝を滅ぼし</gs_sentence>
   <gs_sentence id="2">ウマイヤ朝時代にはインダス川流域まで領域を広げた。</gs_sentence>
   <gs_sentence id="3">ウマイヤ朝時代にはイベリア半島まで領域を広げた。</gs_sentence>
   ……(省略)……
  </gs_sentence_set>
 </gs>
 <doc_set>
 <doc docno="T-WH-A-1-1-4-13">
 <doc_index_set>
      <doc_index gs_sentence_id="2,3" start="83" length="6">ウマイヤ朝は</doc_index>
      <doc_index gs_sentence_id="2" start="109" length="5">中央アジア</doc_index>
      <doc_index gs_sentence_id="3" start="124" length="6">イベリア半島</doc_index>
      <doc_index gs_sentence_id="2,3" start="130" length="12">までの広大な地域を支配し</doc_index>
  </doc_index_set>
  <extraction_text_set>
      <extraction_text gs_sentence_id="2" >ウマイヤ朝は中央アジアまでの広大な地域を支配し</extraction_text>
      <extraction_text gs_sentence_id="3" >ウマイヤ朝はイベリア半島までの広大な地域を支配し</extraction_text>
  </extraction_text_set>
  <title>近・現代世界史の背景～諸地域世界とその交流-ユーラシアの諸地域世界-西アジア世界(イスラーム世界)-ウマイヤ朝の成立</title>
  <text><![CDATA[ 図3と同様のため省略 ]]></text>
 </doc>
 ……(省略)……
 </doc_set>
</data>
```

図 5: アノテーションの例

を広げた」という新たな命題を作成する．抽出，あるいは新たに作成した命題は <gs_sentence> タグ内に書き込み，その属性である id にその命題の <gs_index> タグで設定した expression_id を記述する．

第四工程では，模範解答の命題に対応する知識源の記述の範囲を注釈づけるために，<doc_index> タグの属性である start と length を設定する．また，同じタグの属性である gs_sentence_id には，知識源に対応させる命題の <gs_index> タグで設定した expression_id を記述する．

その後、抽出した知識源の記述は <extraction_text> に書き込み、その属性である gs_sentence_id に <doc_index> タグの gs_sentence_id を設定する．

最後に第五工程では，以上の工程をすべての命題に対して行う．実際に，アノテーションを行った例を図 5 に示す．図では，属性である gs_sentence_id や expression_id に複数の数字が設定されているが，これは知識源や模範解答中の記述のゼロ化代名詞などの照応を取ったためである．同様の理由で，図では「ウマイヤ朝時代にはインダス川流域」を「ウマイヤ朝時代には」と「インダス川流域」に分けている．

また，今回の作業は第 1 著者が行い，作業中に生じた疑問や問題点についてはその対応とともに適宜メモに残した．

5 考察

本節では，アノテーションを行った際に現れた検討事項を列挙し，それぞれの対応を示す．尚，知識源には同じ意味に思えるが異なる意味を示す記述はほとんどなく，異なる意味を示すように思えるが同じ意味を示す記述が多かったため，ここでは後者について列挙する．

類義語による言い換え

模範解答) アラブ人は正統カリフ時代にエジプトを奪い
知識源）正統カリフ時代にムスリム軍はエジプトを征服した．

アラブ人とムスリム軍の差異はあるが，この例では「奪い」と「征服した」の違いに注目する．両者は本来の意味は異なるが，文脈上意味が一致するため同一と見なす．

専門知識による言い換え

模範解答）アラブ人はササン朝を滅ぼし
知識源）ササン朝はイスラームの拡大のなかで滅亡した

本来であれば，アラブ人とイスラームは一致しない．ただし，この時代に限っていえば同一と見なせる．

地理的知識による言い換え

模範解答）ウマイヤ朝時代にはインダス川流域まで領域を広げた．
知識源）ウマイヤ朝は中央アジアまでの広大な地域を支配し

インダス川と中央アジアは粒度が異なるだけで，地理的には同一と見なせる．

格要素が省略されている場合

模範解答）ゼロの概念などが伝わり
知識源）ゼロの概念などはイスラーム世界に伝えられ

この例では「イスラーム世界」の有無が異なるが，ゼロの概念を伝えられた対象が文脈上同一であれば同じ意味の文章と見なす．この場合は対象が同じだったので，同一と見なした．

詳細化

模範解答）タラス河畔の戦いによって中国から製紙法が伝えられた．
知識源）タラス河畔で中国の紙すき職人が製紙法をイスラーム世界に伝えた．

この例では，伝えた側である「中国」の詳細として「中国の紙すき職人」が挙げられているので，同一の意味を持つと見なす．

推論が必要である場合

模範解答）バグダードに建てられた知恵の館では文献がアラビア語に翻訳され
教科書）バグダードにギリシア語の文献を集めてアラビア語に翻訳する機関をつくり「知恵の館」とよばれる機関に発展した．

この例では，知識源側では「知恵の館」の前身のことを「ギリシア語の文献を集めてアラビア語に翻訳する機関」と言っており，正確には同一の意味の文章とは見なせない．しかしながら，この文章から「知恵の館」もまた「ギリシア語の文献を集めてアラビア語に翻訳する機関」であると推論することができるため同一と見なす．

文と名詞句の対応

模範解答）イブン=シーナーの『医学典範』
知識源）イブン＝シーナーは『医学典範』を著している．

「米英首脳」と「アメリカ合衆国の大統領とイギリスの首相」といったような名詞句同士の対応は，論述問題としての解答として利用したときに意味がないため行わない．しかしながら，上記の例のように文と名詞句の対応であれば論述問題の解答して利用できるため解答に含める．

例示がある場合

教科書）トレドにアラビア語文献がラテン語に翻訳され
模範解答）トレドでは医学，哲学，数学，天文学などのアラビア語文献のラテン語への大々的な翻訳が行われた．

この例では，知識源側の「医学，哲学，数学，天文学などの」という余計な要素を含んでいるが同一と見なす．ただし，東京大学の論述問題には文字数制限があり，また複数の話題を扱うためにその範囲に記述を収めるためには簡潔にまとめることが必要である．そのため，上記のように例示が述べられている場合は省略して「トレドではアラビア語文献のラテン語への大々的な翻訳が行われた」とする．

また，今回の仕様で上手くいかなかったものを例示する．

世界史用語集による主語の省略

模範解答）アラブ人は正統カリフ時代にササン朝を滅ぼし
知識源）7 世紀にイスラーム教勢力によって滅ぼされた．

知識源側は，世界史用語集の見出し語が「ササン朝」であるものの記述を抜き出しているものである．しかしながら，世界史用語集の特性上，その用語の説明箇所には主語である見出し語が抜けていることがほとんどである．これを補完するために，見出し語を構造の中に組み込む必要がある．

6 まとめ

本稿では，世界史論述問題のためのアノテーションに関する検討を行い，その際に生じた検討事項を列挙して対応を示した．今後は，対応できなかった検討事項を解消し，他の論述問題に対しても有効であるか否かの調査を行いたい．また，それらの分析を行うことで文抽出を超えた QA システムの開発に繋げたいと考えている．

参考文献

[1] E. Agichtein, D. Carmel, D. Harman, D. Pelleg, Y. Pinter, “Overview of the TREC 2015 LiveQA Track”, *Proceedings of The Twenty-Fourth Text REtrieval Conference*, 2015.

[2] H. Shibuki, K. Sakamoto, Y. Kano, T. Mitamura, M. Ishioroshi, K. Y. Itakura, D. Wang, T. Mori, N. Kando, “Overview of the NTCIR-11 QA-Lab Task”, *Proceedings of the 11th NTCIR Conference*, 2014.

[3] H. Shibuki, K. Sakamoto, M. Ishioroshi, A. Fujita, Y. Kano, T. Mitamura, T. Mori, N. Kando, “Overview of the NTCIR-12 QA Lab-2 Task”, *Proceedings of The NTCIR-12 Conference*, 2016.

[4] K. Sakamoto, M. Ishioroshi, H. Matsui, T. Jin, F. Wada, S. Nakayama, H. Shibuki, T. Mori, N. Kando, “Forst: Question Answering System for Second-stage Examinations at NTCIR-12 QA Lab-2 Task”, *Proceedings of The NTCIR-12 Conference*, 2016.

[5] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治, テキスト情報分析のための判断情報アノテーション, 電子情報通信学会論文誌 (D), **J93-D**(6), pp. 705–713, 2010.

[6] 渋木英潔, 中野正寛, 宮林太郎, 石下円香, 金子浩一, 永井隆広, 森辰則, 情報信憑性判断支援のための Web 文書向け要約生成タスクにおけるアノテーション, 自然言語処理, **21**(2), pp. 157–212, 2014.

[7] 狩野芳伸, 神門典子, 質問応答システムとセンター試験解答フロー: Kachako 対応による標準化・互換化, 2013 年度人工知能学会全国大会（JSAI2013）論文集, 2013.

[8] 石下円香, 狩野芳伸, 神門典子, 質問応答システムでの解答に向けた大学入試問題の分析, 2013 年度人工知能学会全国大会（JSAI2013）論文集, 2013.

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

[10] 東京書籍株式会社・世界史 A（平成２０年度発行

[11] 東京書籍株式会社・新選世界史 B（平成１９年度発行）

[12] 東京書籍株式会社・世界史 B（平成１９年度発行）

[13] 株式会社山川出版社・世界史 B 用語集　改訂版（平成２０年度発行）