世界史論述問題の自動評価のための時間的・地理的情報の利用

渋木 英潔 †1　藤田 彬 †1 阪本 浩太郎 †1　石下 円香 †2
狩野 芳伸 †3　三田村 照子 †4　森 辰則 †1　神門 典子 †2†5

†1 横浜国立大学　†2 国立情報学研究所　†3 静岡大学
†4 カーネギーメロン大学　†5 総合研究大学院大学

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(コミュニティQA の解答の自動生成)[1] や NTCIR の QA Lab[2, 3] など，現実世界における，特に解答が複数の文を含む文章とな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．QA Lab では，世界史の大学入試問題を対象としており，特に論述問題というチャレンジングな課題も設定している．ここでタスクオーガナイザの視点から問題となるのが，システムが出力した解答をどのように評価するかである．世界史の専門家による採点は一法であるが，評価基準がわかりにくい，採点者によるゆれがある，再利用できない，時間がかかるなどの問題がある．そのため，QA システム内で生成された複数の解候補を評価して最善解を選ぶために使用することができない．こういった背景から，自動的に評価する手法を確立したい．QA Lab では，専門家による採点に加えて，模範解答を参照要約とみなした ROUGE[4] や Pyramid 法 [5, 6] による評価を行っている．これらの評価尺度は，DUC[1] や TAC[2] などの要約タスクでも用いられる一般的なものであるが，世界史論述問題においては，専門家による採点と弱〜中程度の正の相関にとどまり，十分な評価尺度とはいえず，より適切な評価手法が望まれる．

ROUGE や Pyramid 法は参照要約との内容的な一致を測るものであり，記述の順序や構造の適切性などを考慮しない．DUC や TAC では言語学的適切性や可読性などに関する quality questions を用いて，内容一致以外の観点からの評価を行っている．しかしながら，世界史論述問題においては，そういった一般的な要約文章の適切性に加えて，解答となる論述文章全体が一貫した論理構成を有する必要がある．中でも，歴史分野では，時間的・空間的な一貫性は多くの問題において，必要となる文章構成要件のひとつである．時間情報を考慮した研究には Barzilay et al.[7] や Okazaki et al.[8]，地理情報を考慮した研究には Buscaldi et al.[9]，GeoTime 情報を考慮した研究には Gey et al.[10] などが存在するが，いずれも世界史論述問題という専門性を考慮したものではない．それゆえ，本稿では世界史論述問題における適切性を，時間的・地理的情報の観点からどのように評価するべきかを分析し，その分析結果に基づいた評価手法が専門家による採点とどの程度相関があるかを報告する．

[1]http://duc.nist.gov/
[2]http://tac.nist.gov/

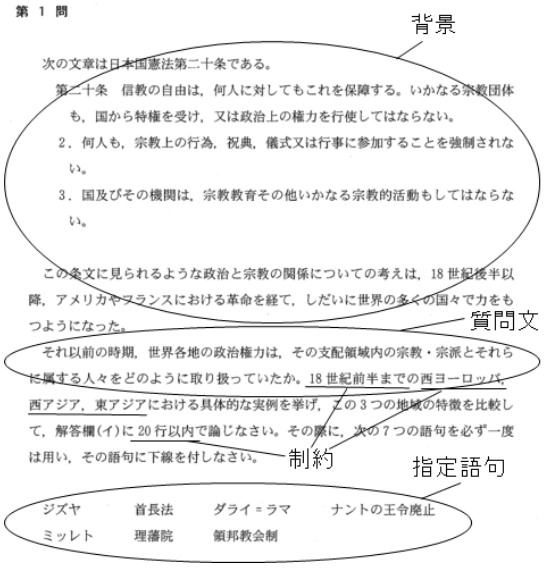

図 1: 世界史論述問題の例

2　世界史論述問題

図 1 は世界史論述問題の例である．問題には核となる質問文の他に多くの情報が含まれている．最初の段落は質問文に至る背景であり，質問文に続いて様々な制約が書かれている．制約には，「20 行以内」といった制限字数，「西ヨーロッパ，西アジア，東アジア」という地理的制約，「18 世紀前半まで」という時間的制約，使用すべき指定語句などが存在する．時間的制約と地理的制約は時間的・地理的な適切性を判断する上で重要である．

3　時間的・地理的観点からの適切性

3.1　構造に関する適切性

一般に，論述問題の解答は歴史的イベント (HE) の記述が連続しており，それぞれの HE には時間情報と地理情報の両方が存在する．ここで，論述問題に解答する際に時間情報と地理情報がどのように書かれるかを考える．時間情報が時系列順に容易に並べて書けるのに対し，空間的広がりをもつ地理情報を記述する順序を決定するのは容易ではない．いくつかの模範解答を調査した結果，一般的に以下の 2 通りのどちらかで書かれていることが分かった．(a) 地理情報を無視して，全ての HE を時系列順に記述する．(b) いくつか

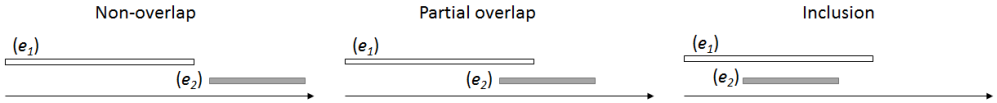

図 2: HE 間における時間的関係

の地理情報ごとに HE をグループ化して，グループごとに時系列順に記述する．ここで (a) の場合を，全ての HE が「全世界」というグループにまとめられていると仮定した場合，(a) と (b) どちらも HE をある地理情報単位でまとめて時系列順に記述しているとみなすことができる．我々は，解答中にまとめられている同じ地理情報をもつ HE の記述単位を地理的段落 (GS) と定義する．GS は階層的な入れ子構造をもっていることがあり，例えば，ヨーロッパの GS 中にイギリス，フランス，ドイツの GS が含まれ，イギリスの GS 中にロンドン，バーミンガム，マンチェスターの GS が含まれているといったようにである．

以上から，我々は適切な論述構造に関して以下の仮説を立てる．

(H1) 世界史の論述は 1 つの GS として記述される．

(H2) ある GS は，その地理情報内の地理情報をもつ複数の GS で構成されうる．

(H3) ある GS 内の HE は時系列順に記述される．

3.2 均一性に関する適切性

GS 中に含まれる GS 群の均一性 (uniformity) について考える．もしも，ヨーロッパの GS 中にイングランド東部，パリ，ドイツの GS が同じレベルで記述されていたとすると，それらは確かにヨーロッパの一部でありながらも不適切さを感じるであろう．これは，国，地域，都市という異なったレベルの地理情報を同じレベルで記述しているからである．したがって，同じレベルの地理情報で記述することが適切性の一つであると考えられる．また，イギリスについて何百字も費やして記述する一方で，フランスやドイツの記述が数十字しかなかった場合も，同じレベルの地理情報でありながら不適切さを感じるであろう．これは記述量のバランスを欠いているからである．したがって，記述量の均一性も適切性に影響すると考えられる．

以上から，GS の均一性による適切性について以下の仮説を立てる．

(H4) ある GS を構成する GS 群は，同じレベルの地理情報をもつ．

(H5) ある GS を構成する GS 群は，同程度の記述量をもつ．

3.3 記述順序に関する適切性

GS 中に含まれる HE の記述順序について考える．適切な論述において，HE は一般的に時系列順に記述されているが，必ずしも HE の起きた時系列と記述の順序が一致するとは限らない．HE の時間情報は開始時と終了時の範囲で示されるので，図 2 に示すように，2 つの HE 間には，「重ならない」，「一部重なる」，「包含される」の 3 通りの関係が存在する．どの関係においても，図中の e_1 の開始時は e_2 の開始時に先行しているが，包含される場合には，「ポツダム宣言の受諾により太平洋戦争が終結した」のように e_2 の後に e_1 が書かれる場合がある．それゆえ，包含される場合には時系列順と関係なく記述されうると仮定した．次に，GS 中に含まれる GS の記述順序について考える．GS の記述順序は HE と比較して自由であると考えられる．しかしながら，例えば，「アテネ，ローマ，カイロ，バグダッド，北京，上海」という記述順序の方が「アテネ，バグダッド，北京，カイロ，ローマ，上海」という順序よりも適切であると感じられるだろう．これは，地理的に近い距離にある順番で記述する方が自然であるからだと思われる．

以上から，GS 中の記述順序について以下の仮説を立てる．

(H6) 仮説 (H3) の例外として，包含関係にある HE は時系列と無関係に記述することができる．

(H7) GS 中の GS 群は地理的に近い距離にある順番で記述される．

3.4 制約に関する適切性

2 章で述べた時間的・地理的制約に関する適切性について考える．例えば，「18 世紀前半まで」という時間的制約において，紀元前の内容しか書かなかったとすると，確かに制約には違反していないが，出題の意図を汲んでいるとはいえない．適切に解答するためには，18 世紀前半の HE を少なくとも一つは記述する必要があるだろう．地理的制約に関しても同様で，「西ヨーロッパ，西アジア，東アジア」という地理的制約において，各地域の HE を少なくとも一つは記述しなければ不適切となろう．

以上から，GS の時間情報と地理情報を以下のように定義し，時間的・地理的制約に関する適切性について以下の仮説を立てる．GS 中の HE において最も早い開始時から最も遅い終了時までの期間を，その GS の時間情報と定義する．また，GS 中の HE の地理情報を全て含む最小の範囲を，その GS の地理情報と定義する．

(H8) GS の時間情報は時間的制約の範囲を可能な限り被覆している．

(H9) 論述全体に対応する GS の地理情報は地理的制約の範囲を可能な限り被覆している．

4 評価手法

4.1 全体の流れ

提案手法の流れを図 3 に示す．まず，入力された解答文書を句点で分割し，各断片を HE とみなす．各断

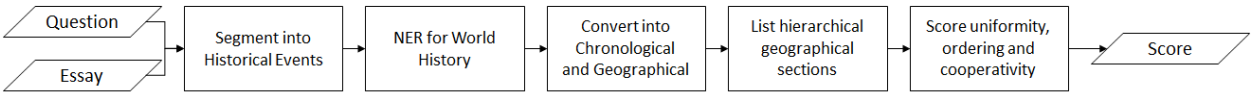

図 3: 提案手法の流れ

表 1: スコアリングの式一覧

$$sc(E,SS,CC,GC)=\begin{cases} sc_T(E,CC) & \text{(TGS の場合)} \\ sc_N(E,SS,CC)sc_{GC}(E,GC) & \text{(RGS の場合)} \\ sc_N(E,SS,CC) & \text{(その他の場合)} \end{cases} \quad (1)$$

$$sc_T(E,CC)=sc_{CO}(E)sc_{GO}(E)sc_{CC}(E,CC) \quad (2)$$

$$sc_N(E,SS,CC)=\frac{1}{|SS|}sc_{GU}(SS)sc_{QU}(SS)\sum_{i=1}^{|SS|}sc(events(s_i),sections(s_i),CC,GC) \quad (3)$$

$$sc_{CO}(E)=\frac{K-L}{K+L} \quad (4)$$

$$sc_{GO}(E)=\frac{1}{geochange(E)+1} \quad (5)$$

$$geochange(E)=\frac{1}{|E|-1}\sum_{i=1}^{|E|-1}distance(range(e_i),range(e_{i+1})) \quad (6)$$

$$sc_{CC}(E,CC)=\frac{overlap(period(E),CC)}{extend(period(E),CC)} \quad (7)$$

$$sc_{GC}(E,GC)=\frac{2P(E,GC)R(E,GC)}{P(E,GC)+R(E,GC)} \quad (8)$$

$$P(E,GC)=\frac{subsumed(geoentities(E),GC)}{|geoentities(E)|} \quad (9)$$

$$R(E,GC)=\frac{subsuming(geoentities(E),GC)}{|GC|} \quad (10)$$

$$sc_{GU}(SS)=1-\frac{sd_{GU}(S)}{am_{GU}(SS)} \quad (11)$$

$$sd_{GU}(SS)=\sqrt{\frac{1}{|SS|}\sum_{i=1}^{|SS|}(depth(s_i)-am_{GU}(SS))^2} \quad (12)$$

$$am_{GU}(SS)=\frac{1}{|SS|}\sum_{i=1}^{|SS|}depth(s_i) \quad (13)$$

$$sc_{QU}(SS)=\frac{-\sum_{i=1}^{|SS|}p(s_i,SS)\log_2 p(s_i,SS)}{\log_2|SS|} \quad (14)$$

$$p(s,SS)=\frac{length(s)}{\sum_{i=1}^{|SS|}length(s_i)} \quad (15)$$

片から固有表現抽出を行い，世界史用語の集合として HE を表現する．世界史用語の中には時間情報や地理情報を想起させるものがある．例えば,「ナポレオン・ボナパルト」であれば時間情報「1769 年 8 月 15 日生–1821 年 5 月 5 日没」や地理情報「フランス」といったようにである．我々は世界史用語集を用いて世界史用語の時間・地理情報をデータベース化しており，これにしたがって抽出された世界史用語の時間・地理情報を獲得し，3.4 で述べた GS の時間情報や地理情報と同様の方法で各断片の時間情報と地理情報を設定する．それから，GS の構造として考えられうる全てのパターンをリストアップし，各パターンのスコアを 4.2 に述べる方法で計算する．最終的に，最大のスコアとなったパターンをその解答の GS 構造として評価する．

4.2 スコアリング

GS は階層的な入れ子構造をもつため，子となる GS をもたない GS を終端 GS(TGS)，そうでない GS を非終端 GS(NGS) と定義する．また解答全体に対応する GS をルート GS(RGS) と定義する．ある GS は，HE 列 $E=(e_1,e_2,\cdots,e_m)$ と子となる GS 列 $SS=(s_1,s_2,\cdots,s_n)$ のペアで定義する．TGS の場合，SS は空となる．問題における時間的制約 CC は開始時 bt と終了時 et のペアで定義し，地理的制約 GC は地理情報の集合 $\{g_1,g_2,\cdots,g_k\}$ で定義する．

仮説 (H2) に基づき，GS のスコア sc は表 1 中の式 (1)–(3) で計算される．$events(s)$ と $sections(s)$ は，GSs の HE 列と GS 列をそれぞれ返す関数である．また，表中の式は [0, 1] の範囲で値を返すように設計している．

式 (4) は，仮説 (H3) に対応した時間的記述順序の適切性を計算する式であり，Kendall の順位相関係数を拡張したものである．時系列と記述順序が一致している HE ペアの数を K，一致しない数を L としてカウントするが，仮説 (H6) に基づいて，包含関係にある HE ペアに関しては除外する．E 中の HE の記述順序が全て時系列順になっている場合，$sc_{CO}(E)$ は 1 を返す．

式 (5)–(6) は，仮説 (H7) に対応した地理的記述順序の適切性を計算する式である．我々は，大陸，国，都市のように地理情報を階層的に配置したシソーラスを作成しており，これを用いて地理的な距離を計算する．$range(e)$ は，HEe の地理情報に対応するシソーラス上のノードであり，$distance(n_i,n_j)$ はシソーラス上の 2 つのノード間の距離を返す関数である．E 中の HE の地理情報が全て同一であった場合，$sc_{GO}(E)$ は 1 を返す．

式 (7) は，仮説 (H8) に対応した時間的制約の適切性を計算する式である．$period(E)$ は，E 中の HE において最も早い開始時と最も遅い終了時のペアを返す関数である．また，$overlap(P_1,P_2)$ は 2 つの開始時と終了時のペアを比較して重なっている期間を返す関数であり，$extend(P_1,P_2)$ は 2 つのペアを比較して早い方の開始時から遅い方の終了時までの期間を返す関数である．E の期間が時間的制約 CC の期間と一致する時，$sc_{CC}(E,CC)$ は 1 を返す．

式 (8)–(10) は，仮説 (H9) に対応した地理的制約の適切性を計算する式であり，地理的制約 GC に対する E の地理情報の適合率と再現率の調和平均である．

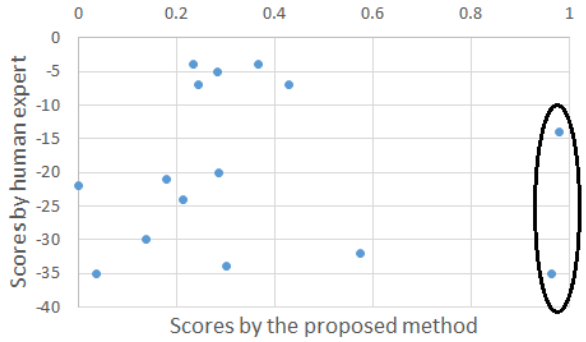

図 4: 提案手法のスコアと専門家による採点の散布図

$geoentities(E)$ は E 中の全ての地理情報の集合を返す関数である．また，$subsumed(G_1, G_2)$ はシソーラス上で G_2 の地理情報の下位に配置される G_1 の地理情報の数であり，$subsuming(G_1, G_2)$ は G_1 の地理情報を下位にもつ G_2 の地理情報の数である．E の地理情報が全て GC の範囲内に存在し，かつ，GC の全ての地理情報が E の地理情報を少なくとも一つは範囲内に収めている場合，$sc_{GC}(E, GC)$ は 1 を返す．

式 (11)–(13) は，仮説 (H4) に対応した地理情報の均一性を計算する式である．$depth(s)$ は s の地理情報に対応するシソーラス上のノードのルートノードからの距離（深さ）を返す関数である．SS 中の GS の地理情報が全て同じ深さである場合，$sc_{GU}(SS)$ は 1 を返す．

式 (14)–(15) は，仮説 (H5) に対応した記述量の均一性を計算する式であり，平均情報量の式を拡張したものである．$length(s)$ は s の文字数を返す関数である．SS 中の GS の文字数が全て等しい場合，$sc_{QU}(SS)$ は 1 を返す．

5 実験

QA Lab-2[3] の Phase-1 と-3 で提出された解答を用いて，提案手法によるスコアと専門家による採点結果とを比較する．提出された解答数は質問数 8 に対して 15 と決して十分な量とは言えないが，専門家には解答全体の点数だけではなく，加点と減点の箇所も指摘してもらった．基本的に，加点は内容の正しさによるものであり，減点は書き方の不適切さによるものであった．それゆえ，提案手法のスコアは減点の結果と比較するのが妥当と思われる．

図 4 に，提案手法によるスコアと専門家による減点の結果の散布図を示す．図中の丸で囲った 2 つの点が全体から大きく離れている．この 2 つの解答は同じ質問に対するものであり，文明発祥以降のエジプトの歴史を概説せよというものであった．したがって，時間的制約が殆ど存在しない一方で，地理的制約はエジプトという比較的小さい範囲に限定された質問といえる．結果として，エジプトに関する HE を時系列順に記述さえすれば，提案手法では高いスコアが得られることとなってしまった．この点に関しては，今後の課題である．この 2 つの解答を除いた場合の相関係数は 0.21 であり，弱い正の相関関係が観察された．この値は決して高いものではないが，最初の試みであることや，減点が必ずしも時間的・地理的観点からによるものでないことを考慮するとまずまずのスタートであると考えられる．現在，開催されている QA Lab-3 に提出されるであろう解答などを用いて，今後さらに調査を続けていきたい．

6 まとめ

本稿では，大学入試の世界史論述問題における時間的・地理的な観点からの適切性を，構造，均一性，記述順序，質問の制約の 4 点で分析し，その分析結果に基づいた評価手法を提案した．QA Lab-2 に提出された 15 の解答を用いて実験した結果，専門家の採点結果に対して 0.21 という弱い正の相関が得られた．今後，対象とする解答数を増やして調査し，より適切な論述問題の自動評価手法を目指していきたいと考えている．

参考文献

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

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

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

[4] C.-Y. Lin . 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of Workshop on Text Summarization Branches Out*, Post-Conference Workshop of ACL 2004, 74–81.

[5] A. Nenkova and R. J. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 145-152.

[6] R. J. Passonneau, E. Chen, W. Guo, and D. Perin. 2013. Automated Pyramid Scoring of Summaries using Distributional Semantics. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 143-147.

[7] R. Barzilay, N. Elhadad, and K. R. McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*, 17(1):35–55.

[8] N. Okazaki, Y. Matsuo, and M. Ishizuka. 2004. Improving Chronological Sentence Ordering by Precedence Relation. *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, 81–88.

[9] D. Buscaldi, J. J. G. Flores, J. L. Roux, and N. Tomeh. 2014. LIPN: Introducing a new Geographical Context Similarity Measure and a Statistical Similarity Measure Based on the Bhattacharyya Coefficient. *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval 2014, 400-405,

[10] F. Gey, R. Larson, N. Kando, J. Machado, and T. Sakai. 2010. NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search. *Proceedings of NTCIR-8 Workshop Meeting*.