センター試験『世界史 B』文の正誤判定問題ソルバー

宮下 洋　石井 愛　小林 実央　星野 力
日本ユニシス株式会社 総合技術研究所
{Hiroshi.Miyashita, Ai.Ishii, Mio.Kobayashi, Chikara.Hoshino}@unisys.co.jp

1　はじめに

現在, 国立学情報研究所が中心となり, 細分化された人工知能 (AI) の分野の再統合を目的として, 大学入試問題を解く計算機プログラムの開発プロジェクトが進められている [新井 12]. 我々は 2015 年に本プロジェクトの世界史 B に参加し,「進研模試 総合学力マーク模試 (2015 年 6 月) 世界史 B」において受験者平均を 30 点上回る 76 点 (偏差値 66.5) を達成した.

センター試験世界史 B の問題は自然言語で記述されたデータを情報源とし, 設問の意図に沿った選択肢を選ぶ問題である. 質問のタイプとしては画像や表の理解が必要な問題を除くと「文の正誤」,「穴埋め」,「ファクトイド型」,「年代並び替え」の 4 タイプに分類される. その中で「文の正誤」の比率が毎年約 70% を占め, 高得点を目指すにはこのタイプの質問が重要となる. このことから, 本論文では「文の正誤」に注目し, 我々が開発した解答器を説明し, 実験結果と考察について述べる.

また,2015 年の世界史 B は NTCIR が主催する QA-Lab[NTCIR-11][NTCIR-12] と共同で行われており, 本論文で使用する試験データは NTCIR より提供されたものを使用している.

2　関連研究

センター試験の歴史の文の正誤問題に対するアプローチとして, 主に以下の 3 つがある.

1. 単語の分布情報で解く [狩野 14]
2. 質問応答に変換して解く [金山 13]
3. 文を抽象表現に変換して解く [田 14]

これらアプローチの特徴として, 単語の分布情報を使った方法はカバー率が高いが厳密性が低くなる傾向にあり, 質問応答に変換する方法は誤文の検知には強いが全ての文の正誤問題を質問応答へ変換できない, 抽象表現を使用する方法では厳密であるが問題のカバー率が低くなると言った傾向がある. そこで本研究では, これらのアプローチを相互で補完することを目指し, 上記のそれぞれのアプローチによる解答器を実装し, これらの結果を組み合わせた.

3　センター試験の分析

事前の観察により, 情報源での局所的な記述により正しい文が選択可能であること, 時間表現については問題文では「世紀」や「年代」など情報源に比べて広い時間範囲を表す表現に置き換わっていること, 場所表現については問題文では「ヨーロッパ」や「アジア」など情報源に比べて広い範囲を表す表現に置き換わっていることが予想された. そのため, 正文となる選択肢の文と問題文に対して以下の観点で調査を実施した.

- 正文に含まれる全ての固有表現が情報源の単一のパラグラフ内に出現するか
- 問題での地理/時間表現が情報源に比べ広い範囲に変換されているか

分析対象となる問題として,2007 年度,2009 年年度のセンター試験世界史 B,2014 年 9 月,2015 年 6 月の進研模試世界史 B の文の正誤問題を用い, 情報源としては以下を使用した.

1. 株式会社山川出版社・詳説世界史 B(世 B 016)
2. 東京書籍株式会社・世界史 A(平成 20 年度発行)
3. 東京書籍株式会社・世界史 B(平成 19 年度発行)
4. 東京書籍株式会社・新選世界史 B(平成 19 年度発行)
5. 世界史オントロジー[*1]
6. フリー百科事典　ウィキペディア日本語版
7. 年表データ (ウィキペディアのページより生成)

その結果, 正文となる選択肢に含まれる固有表現は情報源において 99.3% で同一パラグラフ内に存在していた. このことから, センターの世界史 B の文の正誤問題は情報源の局所的な情報によりほぼ判定可能であることが判明した. また, 質問及び選択肢の文中の時間表現

[*1] http://researchmap.jp/zoeai/event-ontology-EVT/

や地理表現については, その範囲が情報源と比べて広くなっているものがそれぞれ 39/135 件と 16/135 件存在した. このことから, 情報源と選択肢の文を比較する際に特に時間や地理を表す単語については包含関係を判定する仕組みが必要となることが判明した.

4 辞書

上記の分析を加味し, 以下の辞書及び共通モジュールを作成した.

1. 固有表現辞書
2. 同義/類義語辞書
3. 上位/下位語辞書
4. 語尾辞書
5. 国, イベントから年号への変換辞書
6. 時間表現の包含関係判定モジュール

固有表現辞書には, 付加情報として単語が属するクラス (Time,Person 等) の情報が付与されている.

また, これらに加え WordNet, 日本語彙体系も利用した.

4.1 単語間のマッチング

単語間のマッチングでは, 質問文内の単語が情報源内の単語の同義/類義/上位である場合に一致と判断した. また, いくつかのクラスに属する単語に関しては, 同一のクラスの異なる単語同士は排他一致と判断した. また, 単語の語尾は取り除いて比較した.

5 戦略

ここでは, 我々が作成した「単語の分布情報で解く」,「質問応答に変換して解く」,「文を抽象表現に変換して解く」それぞのアプローチに基づく解答器について説明する.

5.1 単語の分布情報で解く

正文に含まれる単語群は情報源の中で共起する確率が高く, 誤文に含まれる単語群には共起する確率の低い単語が含まれるという仮説から, 以下の項で示す単語ペアの共起確率を使ったスコアを計算し, 文の正誤問題に応用した. しかし, 単語ペアの共起確率では 2 単語の組合せしか考慮しないが,3 つ以上の単語の組み合わせが鍵となる場合がある. そのため, 文中の固有表現と, その他の全単語との関連を考慮するため, 検索ランクの情報をスコアに取り入れた. 検索エンジンは Apache Solr[*2]を使用し, 教科書および Wikipedia 中の 1 文を 1 文書として登録した. なお, この手法では固有表現辞書内のクラス情報を用いないため, 辞書の誤りによる影響を受けにくい.

5.1.1 PMI によるスコア

選択肢の文が複数の話題によって構成されることを考慮し, まず文から固有表現および内容語を抽出し, 隣り合う固有表現の間の単語郡において単語のペアを作成する. 次に, 全単語ペアに対して PMI(Pointwise Mutual Information) を計算し, その平均値を文のスコアとした.

$$PMIScore = \frac{1}{|S|} \sum_{(w_i,w_j)\in S} \log \frac{p(w_i,w_j)}{p(w_i)p(w_j)} \quad (1)$$

ここで,S は単語のペア集合,$|S|$ は S の要素数,$p(w_i,w_j)$ は単語 w_i,w_j の共起確率,$p(w_i)$ は単語 w_i の出現確率を表す. 確率の推定では,$p(w_i,w_j)$ は”AND 検索ヒット数/全文書数”,$p(w_i)$ は”検索ヒット数/全文書数”とした. 但し, クエリーとしては単語 w_i 及びその同義/類義語を OR で連結した.

5.1.2 検索ランクによるスコア

まず, 選択肢の文の固有表現 ne_i とその固有表現を取り除いた文 q_i のペアを作る. これにより, 固有表現の数だけペアができる. 次に, それぞれのペアについて,q_i を検索クエリーとして検索を行い,ne_i が出現する文書の最小順位を取得し, その平均をランクによるスコアとした.

$$RankScore = \frac{1}{|Q|} \sum_{(ne_i,q_i)\in Q} -Rank(ne_i,q_i) \quad (2)$$

ここで,Q は ne_i と q_i のペアの集合,$|Q|$ は Q の要素数,$Rank(ne_i,q_i)$ は閾値 N に ne_i 以下の順位に出現する場合は”順位-1”それ以外は $2N$ とした. また, 検索クエリに一番近い文書を検索の上位とするため, 単語の網羅率を重視するようにスコア計算を変更した.

5.1.3 文の正誤判定

PMI によるスコアと, 検索ランクによるスコアを合算し, スコアが高い選択肢が正文である確率が高いとして選択肢の文を選択した.

5.2 質問応答に変換して解く

情報源の中で, 相互に関係のある語は近くに共起する確率が高く, 関係のない語は近くに共起する確率が低くなるという仮説から, 以下の項で示す質問応答スコアを計算し文の正誤問題に応用した.

5.2.1 スコア

質問応答システムは, 与えられた情報源 D とクエリ (質問)Q から, 適切な単語 A を返答する. まず, 情

[*2] http://lucene.apache.org/solr/

報源および質問を形態素解析を用いて, それぞれ,$D = d_1, d_2, \ldots, d_n$, $Q = q_1, \ldots, q_d$ の単語列に分解する. またその際, 助詞, 助動詞などの付属語は取り除く. 以下ではこの表現を使用する. このとき, 情報源 D の中で解答候補 A が質問 Q の解答である確率を, 条件付確率を用いて,

$$p(A|Q, D) \equiv p(a|q_1, \ldots, q_d, d_1, \ldots, d_n) \quad (3)$$

と定義する. この右辺は,$p(a, d_1, \ldots, d_n)$ が a の値によらず一定と仮定し, ベイズの定理を適用して a に依存する項のみを考えると,

$$p(a|q_1, \ldots, q_d) \propto p(q_1, \ldots, q_d|a, d_1 \ldots, d_n) \quad (4)$$

となる. これより, もとの問題は, 解答候補の単語 a を条件として, クエリの単語列が起こる確率が高いものを見つける問題となる. 次に,

$$p(q_1, \ldots, q_d|a, d_1 \ldots, d_n) \quad (5)$$

を設計する. 局所的に正誤が判定できると仮定する場合, 従来では文内共起やパラグラフ内共起などが用いられているが, 質問が情報源の中でどの粒度によって記述されているかには, ばらつきがあるため, より直接的な距離を定義するほうが望ましい. 本研究ではそれを,

$$p(q_1 \ldots, q_d|a = d_k, d_1 \ldots, d_n) \propto \sum_{i=1}^{d} \alpha_{q_i, d_k} \exp(-\gamma l(q_i, d_k)^\beta) \quad (6)$$

で定義する. 但し, 解答候補 a は情報源の単語列 $d_1, \ldots, d_n$ から選ぶものとし, 上の式は, その中で k 番目に出現する単語 d_k のスコアを表す. また,$l(q_i, d_k)^\beta$ は, 情報源の中で q_i と一致する単語のうち d_k にもっとも近いもの $\min_{d_j = q_i} |k - j|$ であり,$\alpha_{q_i, d_k}, \gamma, \beta$ はハイパーパラメータである. システムの最終的な解答としては、上記のスコアの最も高い単語, もしくは与えられたタイプ (Time, Person 等) の中でスコアの最も高い単語をスコアとともに出力する.

5.2.2 文の正誤判定

上記の質問応答システムを文の正誤問題に応用することを考える. 選択肢の文には複数の固有表現が含まれるが, それぞれを解答として隠し, それ以外の語をクエリと見たてた擬似的な質問応答問題を作成する. クエリが解答を十分同定できれば, 隠した単語を高いスコアで返答することが期待できる. 具体的には, 誤文の単語が同一タイプ内で入れ替えられることが多いことを考慮し, 隠した単語と同クラスの単語の中で, 最も高いスコアを基準とし, 隠した単語のコストをその基準からのスコアの差分として定義した. また, 文のコストとしては, 隠した単語それぞれのコストの平均を用いた. 文のコストが小さいものが正文である確率が高いとし, 選択肢から解答を選択した.

5.3 文を抽象表現に変換して解く

文の正誤問題では, 選択肢の文と情報源の文を何らかの抽象表現に変換し, 選択肢の文の抽象表現と類似した抽象表現が情報源に見つかれば正文である確率が高くなると考えられる. そこで本研究では, 文を以下の節で説明する構文木と言う抽象表現に変換し, それらの類似度を定義しその類似度を利用して問題を解くことを試みた.

5.3.1 構文木の定義

本研究における構文木は, 述語を頂点とするツリーで表現され, それぞれの単語には,obj(ヲ格),sbj(ガ格),time(時間),loc(場所),loc-to(場所-へ),other(その他) のいずれかの格が割り振られている. この構文木は KNP[*3]が出力する述語項構造及び係り受けの情報を使用して簡単なルールで作成した. また, 選択肢や情報源の文では sbj,time,loc の省略が多く存在するため, 対象となる構文木よりも前に出現する構文木の情報を使用して補完した. また,「指示詞 + 語尾」という形式での参照については語尾情報を元に参照を解決した. その他, 受動表現についてはルールを用いて平常文の形式に変換した.

5.3.2 スコア

以下のスコアを定義した. ここで,T_t は選択肢の文を構文木に変換したものであり,T_h は情報源の文を構文木に変換したものである.

1. 構文木類似度スコア ($f_T(T_t, T_h)$)
   述語が不一致もしくは,sbj,obj 格をもつ単語について同一格の単語同士を比較し一致する単語がない場合は 0. それ以外の場合, T_t と T_h の同一格同士の単語を比較し, T_t 側の単語の中で T_h に一致する単語が見つかる比率.
2. 単語一致スコア ($f_W(T_t, T_h)$))
   構文木を構成する述語を除く単語についての一致の比率. 但し, 構文木の類似スコアが 0.5 以上場合, 重みとして 2.0 を掛ける. この重みを掛けることに

[*3] http://nlp.ist.i.kyoto-u.ac.jp/?KNP

表 1　正答率 (単位:%; 小数点第 2 位以下四捨五入)

データ	分布情報	質問応答	抽象表現	合議
学習 (97 題)	71.9	81.3	56.3	79.2
検証 (75 題)	70.7	72.0	64.0	73.3

より, 木の類似度を考慮しつつ単語の一致度を評価する.

3. 排他一致スコア ($f_{-W}(T_t, T_h)$)
   T_t の単語の中で T_h の単語と一致しない単語を比較し, 排他一致が存在する場合は 1.0 を, それ以外の場合は 0 とした.

5.3.3　文の正誤判定

情報源の中で最大の単語一致スコア $\max_{T_h}\{f_W(T_t, T_h)\}$ を情報源に対する選択肢のスコアとし, スコアが高い選択肢が正文である確率が高いとして選択肢から解答となる文を選択した. 但し, 単語一致スコアを最大化する T_h について排他一致スコア $f_{-W}(T_t, T_h)$ が 1.0 以上となる場合, その文は誤文とした.

6　結果

表 1 に文の正誤問題に対するそれぞの解答器の正答率と, それぞれの解答器が出力する各選択肢の順位の合計値を使用して回答を選択した場合 (合議) の正答率を示す. 学習用データとしては 2007,2009 年度センター試験と 2014 年 6 月,11 月の進研模試のデータを使用し, 検証データとしては 2011 年度センター試験と 2014 年 9 月,2015 年 6 月の進研模試を使用した. 学習データの問題数は 97 題であり, 検証データの問題数は 75 題であった. 学習データを用いてハイパーパラメータの調整及び語彙や同義/類義語のを追加を行い, 調整済みのハイパーパラメータ及び辞書を使用して学習データ及び検証データに対して正答率を評価した.

回答の選択肢は 4 つであるため, ランダムで選択した場合の期待値は 25.0% となるが, それぞれの解答器についてそれを大幅に上回る結果となった. また, それぞれのデータにおける最も良い解答器と合議での正答率を比較すると, 学習データについては合議での回答が 1.9% 低く, 一方検証データでは 1.3% 高くなった. このことから, 複数の解答器の出力を組み合わせて合議で選択することにより, 汎化誤差のバリアンスに由来する部分を減らす効果があったものと考えられる.

7　考察

全解答器において間違った問題 (10 問) を対象に分析を行った結果, 以下の原因が観察された (重複有).

1. 固有表現の同定力不足 (情報源上頻出の為)(4 件)
2. 文脈上での単語の同義語認識ができない (3 件)
3. 問題部からの重要単語の抽出ミス (2 件)
4. 原因や理由などの因果関係を理解できない (2 件)
5. 固有表現が辞書にない (1 件)
6. 主客逆転を認識できない (1 件)
7. 時間認識に失敗 (1 件)

このことから, さらに正答率を高めるには, 上記のエラーを解消するために個々のモジュールの精度向上や新しい仕組みが必要になるものと考えられる.

8　まとめ

センター試験世界史の文の正誤問題を対象とし,「単語の分布情報で解く」「質問応答に変換して解く」「文を抽象表現に変換して解く」という 3 つのアプローチで解答器を作成した. それぞれの解答器を組み合わせることで最終的に, 学習データで 79.2%, 検証データで 73.3% という, 過去の手法を上回る正答率を達成した.

参考文献

[新井 12] 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)

[NTCIR-11] Hideyuki Shibuki el al:Overview of the NTCIR-11 QA-Lab Task, in Proceedings of the 11th NTCIR Conference(2014)

[NTCIR-12] NTCIR 12 QALab-2 Task: http://research.nii.ac.jp/qalab/

[狩野 14] 狩野芳伸:Solving History Exam by Keyword Distribution: KJP, in Proceedings of the 11th NTCIR Conference(2014)

[金山 13] 金山博, 宮尾祐介:ファクトイド型質問応答を用いた正誤判定問題の解決, 言語処理学会第 19 回年次大会 (2013)

[田 14] 田然, 宮尾祐介:テキスト推論でセンター試験の歴史問題を解く, The 28th Annual Conference of the Japanese Society for Artiial Intelligence(2014)