

表層類似度を用いたセンター試験「世界史」解答システムの改善

今川 拓哉†

高田 拓真‡

松崎 拓也‡

佐藤 理史‡

† 名古屋大学工学部 電気電子情報工学科 ‡ 名古屋大学大学院 工学研究科

1 はじめに

国立情報学研究所主催のグランドチャレンジ「ロボットは東大に入れるか」プロジェクト [1] や NTCIR の QA Lab タスク [2] など、大学入試問題を計算機で解く試みが進められている。我々もこれらの試みに参加し、センター試験「世界史」の自動解答に取り組んでいる。センター試験「世界史」の問題の多くは選択肢の内容が史実に一致するかどうかを判定する問題（以下、FV 形式と呼ぶ）であり、選択肢と教科書のテキストとの間の含意関係認識によって解答を導きだすことができる。本研究では、表層類似度によってテキスト間の含意関係を判定する服部らの手法 [3] を拡張し、選択肢と教科書の表記の違いを吸収するための同義語辞書や世界史用語とそうではない語を区別するための名詞の分類といった要素を加えることでより精度の高い解答システムを実現することを試みた。

2 センター試験「世界史 B」の概要

センター試験「世界史 B」の問題は他の教科と同様にマークシート形式である。例えば 2003 年度の本試験は、4 つの大問によって構成され、各大問に 9 つの設問がある。その中でも特に多い設問は図 1 のような FV 形式であり、これらは 2013 年度の本試験では全問題のおよそ 55% を占めている。それ以外の設問の形式には、文章中の空欄に語句を当てはめる空欄問題、2 つの文の正誤の組み合わせを解答する正誤問題、3 つの文を年代順に並び替える年代並び替え問題、地図や写真を見て解答する地図・写真問題、年表を利用して解答する年表問題などが存在する。本研究では主に FV 形式の問題を対象とした。

3 世界史解答システムの構成

3.1 ベースラインシステムの概要

本研究では、服部らの表層類似度に基づく日本語テキスト含意関係認識システム [3] をベースにしている。

... ベストを拠点として⑤ 交易を行い、ドイツ、オーストリア、バルカン半島へと広がる商業圏を築いた。

問 6 下線部⑤の歴史について述べた文として正しいものを、次の①～④のうちから一つ選べ。

- ① 16 世紀に、ポルトガルは元と交易を行った。
- ② オーストリアを中心に、ドイツ関税同盟が発足した。
- ③ ジョン=ヘイは、中国の門戸開放を主張した。
- ④ スウェーデンは、ネルチンスク条約によって清と通商を始めた。

図 1: センター試験「世界史 B」 FV 形式問題の例 (2013 年 センター試験 第 2 問の問 6)

システムの構成を図 2 に示す。服部らのシステムは、主に図 2 の左側の部分に対応している。図 2 の点線部分は本研究の改善点に対応している。服部らの手法では、まず各選択肢と教科書の各段落の表層類似度を計算し、それぞれの選択肢に対し最も類似度の高い段落を決定する。類似度の計算については後述する。

次に、選択肢と教科書の最も表層類似度の高い段落から固有名詞と年代表現（「1946 年」など）を抜き出し、選択肢に含まれるすべての固有名詞が、教科書から抜き出した段落にも含まれるかどうかを調べる。年代表現については、選択肢に含まれるすべての年代表現が、それぞれ教科書の段落から抜き出した年代表現のうちどれか 1 つとでも合致していた場合、その選択肢は年代表現が合致しているものとする。例えば、選択肢に「1946 年」と「1950 年代」の 2 つの年代表現が含まれ、教科書には「1894 年」と「19 世紀」が含まれていたとすると、選択肢の「1946 年」と「1950 年代」がそれぞれ「19 世紀」と合致するので、この選択肢は年代表現が合致していると判断する。また、選択肢には年代表現が含まれるが教科書には含まれない場合は年代表現が合致していないと判断する。

最後に、表層類似度が閾値を上回り、かつ上記の固有名詞と年代表現が合致している選択肢を真と判断し、それ以外の選択肢を偽と判断する。正しい選択肢を選ぶ問題の場合は、真の選択肢の中から最も表層類似度の高いものを解答として出力する。誤っている選択肢を選ぶ問題の場合は、偽の選択肢の中から最も表層類似度の低いものを解答として出力する。

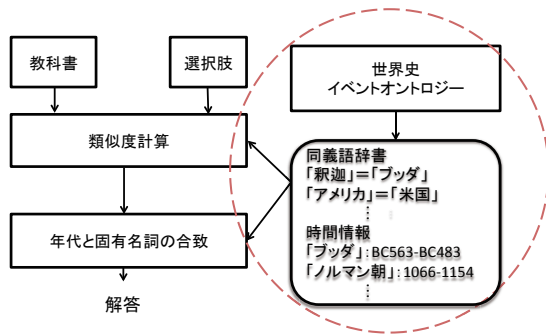


図 2: 世界史解答システムの構成

本稿の実験では、教科書として東京書籍出版世界史 A(2008 年発行)、東京書籍出版世界史 B(2007 年発行)、東京書籍出版新選世界史 B(2007 年発行)、山川出版詳説世界史 B(2007 年発行) を使用した。

3.2 改善点

服部らのシステムのエラー分析の結果に基づき、類似度スコアの改良 (3.2.1 節, 3.2.2 節) と年代の合致チェックの改良 (3.2.3 節) を行った。また、教科書との類似度と年代・固有名詞の合致チェックを総合的に扱う手段として機械学習を用いることを試みた (3.2.4 節)。さらに、FV 形式以外の問題タイプである年代並び替え問題に対する解答モジュールを作成した (3.2.5 節)。

3.2.1 名詞の分類

服部らのシステムにおける類似度計算では教科書や選択肢から抽出した名詞の分類を行わないため、「下線部」や「文」などの類似度計算をする上で重要ではない名詞もその他の名詞と同じように類似度に影響してしまうという問題があった。そこで、類似度計算の際、名詞の分類に基づく重み付けを加えることで改善を行った。服部らのシステムにおける類似度を式 (1)、本研究で用いた類似度を式 (2) に示す。

$$score_h = \frac{\sum_{x \in N^1} \{\min(f(x, t), f(x, q))\}}{\sum_{x \in N^1} f(x, q)} \quad (1)$$

$$score = \frac{\sum_{x \in N^1} \{\min(f(x, t), f(x, q)) \cdot a(x)\}}{\sum_{x \in N^1} f(x, q)} \quad (2)$$

ここで t は教科書の段落、 q は選択肢であり、 $f(x, t)$ は集合 N^1 の要素 x が、 t 中に出現する回数を表す。

N^1 は名詞 unigram の集合を表す。また、 $a(x)$ は以下に示す重み付けの関数である。

$$a(x) = \begin{cases} 1 & \text{if } x \in \text{山川出版世界史用語集の見出し語} \\ 0.9 & \text{else if } x \in \text{教科書の固有名詞} \\ 0.7 & \text{else if } x \in \text{WikipediaPageTitle} \\ 0.1 & \text{otherwise} \end{cases}$$

3.2.2 同義語辞書

世界史イベントオントロジー [4] から世界史用語の同義語情報を抽出し、同義語辞書を作成した。この辞書には、「釈迦」と「ガウタマ=シッダールタ」は同義語であるといった情報が含まれる。この辞書は、類似度計算の際に使用し、同義関係にある名詞は代表表記に書き換えたのちに式 (2) の類似度計算を行った。

3.2.3 世界史用語からの時間情報の抽出

服部らのシステムにおいては文中に「1923 年」、「18 世紀」などの直接の年代表現がある場合のみに選択肢と教科書の段落の間での年代表現の合致をチェックしていた。しかし、文中に直接、時間表現が書かれている場合は多くない。そこで、世界史イベントオントロジーを参照し「第 1 次世界大戦」などのイベント名や人物名などの時間情報を参照することで世界史用語からも時間情報を抽出できるように改善した。

3.2.4 Support Vector Machine を用いた解答選択システム

選択肢と教科書との表層類似度および固有名詞と年代表現の合致を総合的に扱うために SVM を用いた解答システムを作成した。表 1 に用いた特徴量を示す。学習器としては、Support Vector Machine for Ranking [5] を利用した。

まず、教科書 (山川、東京書籍)、用語集、4 つの選択肢を入力とし、それぞれの選択肢について表 1 の特徴量を計算する。表 1 で、「教科書 (山川)」、「教科書 (東京書籍)」、「用語集 (山川)」はそれぞれ、選択肢との表層類似度の最も大きい山川出版と東京書籍の教科書の段落および用語集の項目を意味する。東京書籍の教科書に関しては前述の 3 冊を 1 つにまとめたものを使用した。表層類似度の計算は式 (2) を使用している。

次に、正しい選択肢を選ぶ問題と誤った選択肢を選ぶ問題のそれぞれで、各選択肢の特徴量ベクトルと重みベクトルの内積を計算し、解答を決定する。

表 1: SVM の特徴量一覧

教科書 (山川) との表層類似度
教科書 (東京書籍) との表層類似度
用語集との表層類似度
教科書 (山川) と選択肢の年代が一致するかどうか
教科書 (東京書籍) と選択肢の年代が一致するかどうか
教科書 (山川) と選択肢の固有名詞が一致するかどうか
教科書 (東京書籍) と選択肢の固有名詞が一致するかどうか

問 3 下線部②に関連して、中国の文化について述べた次の文 a~c が、年代の古い順に正しく配列されているものを、下の①~⑥のうちから一つ選べ。
a 欧陽脩らの名文家が、活躍した。
b 顧炎武が、考証学の基礎を築いた。
c 『永楽大典』が編纂された。
① a → b → c ② a → c → b ③ b → a → c
④ b → c → a ⑤ c → a → b ⑥ c → b → a

図 3: センター試験「世界史 B」年代並び替え問題の例 (2013 年度 センター試験 第 2 問の問 3)

3.2.5 年代並び替え問題の解答モジュールの追加

センター試験「世界史 B」には図 3 のような年代並び替え問題が出題される。このような形式の問題に解答するためのモジュールを追加した。

このシステムは「欧陽脩」のような世界史用語を並べ替える文中から抜き出し、世界史イベントオントロジーからその用語に関する時間情報を抽出する。世界史オントロジーから時間情報が抽出できなかった場合、文と教科書の各段落との表層類似度を計算し、最も表層類似度の高い段落から年代表現を抜き出す。最後に、抜き出した時間情報を比較して文を並び替える。

4 実験

4.1 評価データ

実験には、2001 年度センター試験本試験と 2003 年度追試験の FV 形式の問題を使用した。FV 形式の問題は、2001 年度は全 41 問のうち 33 問、2003 年度は全 41 問のうち 30 問であった。SVM の学習データは 2003、2005、2007、2009、2011、2013 年度の本試験の FV 形式 163 問を利用した。年代並び替え問題については、2001 年度本試験と 2003 年度追試験に年代並び替え問題が存在しなかったため、開発データの 2003、2007、2011、2013 年度の本試験、2013 年度代ゼミセンター模試第 1 回～第 4 回の問題を使用した。

4.2 結果

服部らのシステム (ベースライン)、ベースラインにそれぞれ同義語辞書、名詞の分類、時間情報の抽出、

3 つ全てを加えたシステムおよび SVM を用いたシステムの結果を表 2 に示す。「○ → ×」と「× → ○」はそれぞれベースラインでは正解した問題が不正解になった数と、不正解だった問題を正解した数である。また表 3 は年代並び替え問題に対する結果である。

5 結果の検討

5.1 同義語辞書の結果

2001 年度本試験では、同義語辞書を用いることで正解数が増加している。同義語辞書の追加によって正解になった問題の例として、「ビスマルクは、普仏戦争でフランスを破った。」という選択肢を含む問題がある。選択肢では「普仏戦争」となっているが教科書では「プロイセン＝フランス戦争」と呼称されている。同様に、同義語辞書によって吸収できた表記の差異として、「ソ連」と「ソビエト連邦」、「訓民正音」と「ハングル」などがあった。

5.2 名詞の分類の結果

名詞の分類によって正解できるようになった問題には、「『四書大全』の編纂が行われた。」という正しい選択肢を含むものがあつた。ベースラインでは、この選択肢との表層類似度が最も高い教科書の段落は「... 法学者を集めて編纂させた『ローマ法大全』がその集大成である。...」だった。一方、名詞の分類を加えたシステムでは「...『四書大全』や『五経大全』などの編集と出版が行われた。...」という段落が最も類似度が大きくなった。この選択肢で重要なのは一般語の「編纂」よりも世界史用語の「四書大全」であり、これらに対する重み付けを変えることで、選択肢と内容が一致するような段落を検索することができるようになった。

5.3 時間情報の抽出の結果

時間情報の抽出を用いた手法では大きな改善は見られなかった。時間情報の抽出が悪影響を与えた問題には、「1848 年、フィヒテが、『ドイツ国民に告ぐ』を発表した。」という誤った選択肢を含む問題があつた。この選択肢との表層類似度が最も高い教科書の段落は「フランスによる... フィヒテは連続講演『ドイツ国民に告ぐ』で...」というものだった。時間情報の抽出をした結果、「フランス」からも時間情報が抽出されたため、選択肢と教科書との年代が合致してしまい真と

表 2: FV 形式問題の評価結果

	2001 年本試験			2003 年追試験		
	正解数	○ → ×	× → ○	正解数	○ → ×	× → ○
ベースライン	8/33	-	-	6/30	-	-
+同義語辞書	10/33	1	3	6/30	1	1
+名詞の分類	11/33	2	5	8/30	1	3
+時間情報の抽出	8/33	2	2	6/30	0	0
+全て	11/33	2	5	8/30	2	4
SVM	9/33	2	3	9/30	3	6

表 3: 年代並び替え問題の評価結果

	正解数
2003 年度本試験	0/1
2007 年度本試験	1/1
2011 年度本試験	1/3
2013 年度本試験	0/2
2013 年度代ゼミ模試 1	1/3
2013 年度代ゼミ模試 2	0/3
2013 年度代ゼミ模試 3	2/2
2013 年度代ゼミ模試 4	0/2
合計	5/17

判断してしまった。このようなことを防ぐためには、時間情報を抽出する対象を限定する必要がある。

5.4 SVM による解答システムの結果

2001 年度本試験では、SVM を用いた手法では大きな改善が見られなかった。2003 年度追試験では、ベースラインに比べ正解数は増加している。しかし、いずれの場合も SVM とベースライン+全てと比較すると、2001 年度では正解数が 2 問減少し、2003 年度では 1 問増加しただけだった。正解数をさらに増やすためには、表層類似度を用いたシステムでは正解できないような問題を正解できるようになるような特徴量を考え追加する必要がある。

5.5 年代並び替えの結果

年代並び替え問題は 6 通りの解答が存在する。よって、システムの正解率 5/17 (約 30%) はランダムに解答した場合よりは良い結果である。誤りの多くは、世界史オントロジーから時間情報の抽出ができなかった場合に、選択肢に合致するような教科書の段落を検索することができなかったことと、選択肢と合致するような段落を検索することはできたが段落中に年代表現が存在しないことによるものだった。これを解決するためには、正確に選択肢と合致する教科書の段落を抜き出せるようにする必要がある。

5.6 課題

今後の課題として、「禁止した」と「許可した」などのような知識を用いることがあげられる。例えば、2011 年の大問 3 の問 5 は「ルイ=フィリップは、七月革命によって廃位された。」という選択肢を含む。この選択肢は偽であり、「廃位」ではなく「即位」であれば史実と一致し真の選択肢である。現在のシステムでは教科書の段落と「ルイ=フィリップ」、「七月革命」

が一致し、表層類似度が高くなるためこの選択肢を真であると判断してしまう。これを防ぐためには、選択肢と教科書で対義語の関係にある語が出てきた場合はペナルティをかけるなどの方法が考えられる。

6 結論

本研究では、服部らのシステムに同義語辞書の追加と名詞の分類による重み付け、イベントの時間情報の抽出と年代並び替えシステムの追加、機械学習の利用、などによる改良を試みた。同義語辞書、名詞の分類による重み付け、機械学習の利用では、ベースラインを上回る正解数になった。一方で、イベントの時間情報の抽出は改善につながらなかった。

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? 人工知能学会誌, Vol. 27, No. 5, pp. 463-469, 2012.
- [2] Hideyuki Shibuki, Kotaro Sakamoto, Yoshionobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. Overview of the NTCIR-11 QA-Lab Task. In *Proceedings of the 11th NTCIR Conference*, 2014.
- [3] 服部昇平, 佐藤理史, 駒谷和範. 表層類似度に基づく日本語テキスト含意認識. 人工知能学会論文誌, Vol. 29, No. 4, pp. 416-426, 2014.
- [4] Ai Kawazoe, Yusuke Miyao, Takuya Matsuzaki, Hikaru Yokono, and Noriko Arai. World history ontology for reasoning truth/falsehood of sentences: Event classification to fill in the gaps between knowledge resources and natural language texts. In *New Frontiers in Artificial Intelligence*, pp. 42-50. Springer, 2014.
- [5] Thorsten Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 217-226. ACM, 2006.