二段階選抜による選択式評論読解問題の自動解法

木村 遼 † 佐藤 理史 ‡ 松崎 拓也 ‡
† 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

2011 年に、国立情報学研究所で「ロボットは東大に入れるか」というプロジェクトが開始された。このプロジェクトは大学入試問題を解く人工知能の開発に挑戦するものである。我々はこのプロジェクトに参加し、大学入試センター試験「国語」評論の「傍線部問題」と呼ばれる読解問題の自動解法を研究している [1, 2]。

大学入試センター試験の「国語」では大問 4 題が出題され、大問の第 1 問が評論の問題である。評論の問題では、何らかの評論から抜き出された文章 (本文) とそれに対する設問 6 問から構成される。本研究では、そのうちの問 2 から問 5 の読解問題を対象とする。これらの読解問題は、本文中の一部 (傍線部) に対する 1 つの設問文と 5 つの選択肢から構成される。設問文で傍線部の内容に対する説明や理由を問われ、その解答として最も適当なものを選択肢から 1 つ選ぶ選択式の問題である。

本稿では、この読解問題の自動解法として二段階選抜による解法 (二段階選抜法) を提案する。この解法では、一段目の選抜で選択肢 5 個から解答候補を 2 個に絞り込み、二段目で最終的な解答 1 個を決定する。

このような方法を採用する理由は、人が読解問題を解くとき、選択肢 5 個のうち 2 個にまで絞り込んでから解答に迷う場合が多いという経験に基づく。これは、4 個の間違い選択肢のうちの 3 個は、比較的簡単に間違いであることがわかるのに対し、残りの 1 個はその判定が難しいことを意味する。すなわち、2 個に絞り込むための判定と、最終的な解答を決めるための判定は、質的に異なる可能性がある。二段階選抜法では、そのそれぞれの判定を、専用の判定器で行う。

2 二段階選抜法

本稿で提案する二段階選抜法は、すでに提案されている Binary Classifier-Based Method(BCBM)[2] を拡張した解法で、以下の手順で読解問題を解く。

1. 一段目: 解答候補を 2 個に絞り込む
   (a) 特徴ベクトル作成: 5 個の選択肢を特徴ベクトルに変換する。
   (b) 順位付け: 5 個の選択肢 (特徴ベクトル) から総当たり 10 ペアを作成する。各ペアを 2 値分類モジュールに入力し、「より本文に合致する」と判定された選択肢に 1 点を与える。点数の大きい選択肢から順に順位を付ける。
   (c) 出力: (b) の順位付けで 1 位と 2 位となった 2 個の選択肢を一段目の結果として出力する。
2. 二段目: 最終解答を決定する
   (a) 特徴ベクトル作成: 出力された 2 個の選択肢を一段目とは異なる特徴ベクトルに変換する。
   (b) 判定: 2 個の特徴ベクトルを 2 値分類モジュールに入力し、「より本文に合致する」と判定された選択肢を解答として出力する。

2.1 2 値分類モジュール

各段階で使用する 2 値分類モジュールは、2 個の選択肢 (特徴ベクトル) を入力とし、「より本文に合致する」選択肢の番号を出力する。具体的には、選択肢に対応する 2 つの特徴ベクトルの差分 $\boldsymbol{x} = \boldsymbol{f}^{(i)} - \boldsymbol{f}^{(j)}$ を入力とし、i または j を出力する。

2 値分類器には、サポートベクターマシン (SVM) を用いる。SVM の設定は、BCBM と同様の設定を用いる。すなわち、カーネルには線形カーネルを用い、他のパラメータは全てデフォルト値を用いる。SVM の学習例には、正解選択肢と間違い選択肢のペアを用いる。

一段目と二段目の 2 値分類モジュールの機構は同一である。異なる点は、特徴ベクトルを構成する特徴量集合、および、学習に使用する実例集合である。

表 1: 新たに導入する特徴量

式	説明
$f_{18}(c_i, C) = \begin{cases} 1 & (c_i = \underset{c_j \in C}{\mathrm{argmin}}(\sum_{k \neq j} \mathrm{overlap}(c_j, c_k))) \\ -1 & (\mathrm{otherwise}) \end{cases}$	他の選択肢との文字オーバーラップ率が最も低いなら 1、それ以外なら-1
$f_{19}(c_i, E) = \sum_{(t,t') \in P(c_i,E)} \{\mathrm{td}(t, t'; E) + \mathrm{td}(t', t; E)\}$	名詞・動詞の総当たりでの出現依存度の和
$f_{20}(c_i, E) = \frac{1}{2\lvert P(c_i,E)\rvert} f_{19}(c_i, E)$	名詞・動詞の総当たりでの出現依存度の平均
$f_{21}(c_i, E) = \sum_{(t,t') \in P(c_i,E)} \max\{\mathrm{td}(t, t'; E), \mathrm{td}(t', t; E)\}$	名詞・動詞の出現依存度が大きい組み合わせでの総和
$f_{22}(c_i, E) = \frac{1}{\lvert P(c_i,E)\rvert} f_{21}(c_i, E)$	名詞・動詞の出現依存度が大きい組み合わせでの平均
$f_{23}(c_i, E) = \sum_{(t,t') \in P(c_i,E)} \min\{\mathrm{td}(t, t'; E), \mathrm{td}(t', t; E)\}$	名詞・動詞の出現依存度が小さい組み合わせでの総和
$f_{24}(c_i, E) = \frac{1}{\lvert P(c_i,E)\rvert} f_{23}(c_i, E)$	名詞・動詞の出現依存度が小さい組み合わせでの平均
$f_{25}(c_i, E) = \sum_{(t,t') \in P(c_i,E)} \mathrm{zero}(\mathrm{td}(t, t'; E))$	選択肢中の名詞・動詞の出現依存度が 0 となる数 ($\mathrm{zero}(x)$ は、$x = 0$ で 1 、$x \neq 1$ で 0 とする)
$f_{26}(c_i, E) = \frac{1}{\lvert P(c_i,E)\rvert} f_{25}(c_i, E)$	選択肢中の名詞・動詞の出現依存度が 0 となる割合
$f_{27}(c_i, E) = \sum_{t \in N(c_i)} \mathrm{attr}(t; E)$	選択肢に出現する名詞の吸引力の総和
$f_{28}(c_i, E) = \frac{1}{\lvert N(c_i)\rvert} f_{27}(c_i, E)$	選択肢に出現する名詞の吸引力の平均
$f_{29}(c_i, E) = \sum_{t \in NV(c_i)} \mathrm{attr}(t; E)$	選択肢に出現する名詞・動詞の吸引力の総和
$f_{30}(c_i, E) = \frac{1}{\lvert NV(c_i)\rvert} f_{29}(c_i, E)$	選択肢に出現する名詞・動詞の吸引力の平均
$f_{31}(c_i, E) = \lvert NV(c_i) \cap NV(E)\rvert$	選択肢、照合領域両方に出現する名詞・動詞の数
$f_{32}(c_i, E) = \lvert NV(c_i) \cap \overline{NV(E)}\rvert$	選択肢 c_i に出現し、照合領域に出現しない名詞・動詞の数
$f_{33}(c_i, E) = \frac{\lvert NV(c_i) \cap \overline{NV(E)}\rvert}{\lvert NV(c_i)\rvert}$	選択肢 c_i に出現する名詞・動詞のうち、照合領域には出現しない割合

3 特徴量

二段階選抜法では、BCBM で使用された 17 種類の特徴量 (文献 [2] の表 2) に加え、新たに 16 種類の特徴量を導入する。

新たに導入する特徴量の概要を表 1 に示す。ある選択肢 c_i の 1 つの特徴量は、関数 $f_j(c_i, X)$ によって計算される。X は選択肢 c_i の比較対象であり、選択肢集合 C または照合領域 E である。照合領域 E は本文中の傍線部を中心とした領域であり、どの領域を抽出するかをパラメータで指定する。パラメータでは、段落単位または文単位で傍線部から前後どこまでを領域とするかを指定する。

f_{18} の特徴量は、選択肢 c_i が他の選択肢と比べて表層的に最も離れているか否かを示す。表層的な類似度は、文字オーバーラップ率 [3] によって計算する。

f_{19} から f_{30} の特徴量は、選択肢と照合領域の比較において、語句の重要度を考慮する特徴量である。語句の重要度の計算には、語の出現依存度と吸引力 [4] を利用する。語の出現依存度 $\mathrm{td}(t, t'; E)$ は、文章 (本研究では照合領域)E に出現する異なる 2 つの語 t と t' について、語 t が出現した同じ文中に語 t' が出現する確率として定義される。語 t の吸引力 $\mathrm{attr}(t; E)$ は、他の語 t'' から語 t に対する出現依存度 $\mathrm{td}(t'', t; E)$ の総和として定義される。f_{19} から f_{26} は出現依存度を利用した特徴量で、f_{27} から f_{30} は吸引力を利用した特徴量である。

表 1 中の $N(T)$、$NV(T)$ はそれぞれ文章 T 中に出現する名詞の集合、名詞と動詞の集合を示す。$P(c_i, E)$ は選択肢 c_i と照合領域 E 両方に出現する名詞または動詞 2 個の組み合わせの集合を示す。

f_{31} から f_{33} の特徴量は、選択肢 c_i と照合領域 E に共通して出現する単語の数や比率である。

33 種類の特徴量のうち、24 種類は、特徴量の計算に照合領域を使用する。これらの特徴量の計算において、最適な照合領域はそれぞれ異なる可能性がある。そこで、はじめに複数の照合領域を用意して各照合領域ごとに特徴量を計算し、その後、有用な特徴量を取捨選択する手法 (特徴選択) を適用する。具体的には、10 通りの照合領域を定め、それぞれの照合領域に対して 24 種類の特徴量を計算する。すなわち、最初の段階では、照合領域が不要な特徴量 9 種類とあわせて合計 249 個の特徴量を作成する。

4 性能評価

4.1 データセット

性能評価には、センター試験の過去問題および各種予備校のセンター試験対策問題集と模擬試験の問題を使用した。使用したデータセットの構成を表 2 に示す。データセットには全 284 問の読解問題が含まれる。

表 2: データセット内訳

問題	年代 (20xx)	大問数	設問数
センター試験過去問	01–15	14	56
代ゼミ問題集+模試	05,14,15	19	76
駿台問題集	06,14	14	56
河合塾問題集	14,15	10	40
旺文社問題集	14	4	16
ベネッセ模試	14,15	10	40
計		71	284

4.2 評価手順

性能評価は、一段目、二段目の順に行う。それぞれの段では、まずは特徴選択を実施し、2 値分類モジュールの分類精度が高くなる特徴ベクトルの構成を定める。249 個の特徴量集合の全ての部分集合に対して評価を行うことは困難であるため、一部の部分集合に対して性能を評価する。なお、特徴ベクトルの次元（使用する特徴量の数）は、予備実験により定める。

4.3 一段目の評価

3 節で述べた 249 個の特徴量から、60 個をランダムに選んで構成した 5 万通りの特徴ベクトルに対して、2 値分類モジュールの分類精度を測定した。測定には、データセットの各設問につき正解選択肢と間違い選択肢のペアを 8 個 (4 個× 2 方向) 作成し、これらに対して 10-fold cross validation を適用した。

次に、分類精度が最も高かった 10 種類の 2 値分類モジュールのそれぞれに対し、それを用いた場合の一段目の性能を測定した。この測定は、大問単位の leave-one-out cross validation で行った。すなわち、大問 1 題 (設問 4 問) に対し、2 値分類モジュールの学習に使用する実例集合は、データセット中のその大問を除いた 70 題に含まれる設問の選択肢から作成したペア 2240 個 (70 題× 4 問× 4 個× 2 方向) である。

評価結果を表 3 に示す。一段目の目的は、選択肢を 2 個に絞ることであるため、正解選択肢を 1 位とするよりも、2 位までに正解選択肢を含むことがより重要である。表 3 には、分類精度が高かった 10 種類の 2 値分類モジュールのうち、正解選択肢が 2 位までに含まれた設問数が多かった 5 種類に対し、その設問数 (2 位まで)、正解選択肢を 1 位として出力した設問数 (1 位)、使用した 2 値分類モジュールの分類精度を示した。

表 3: 一段目の性能

ID	2 位まで	1 位	分類精度 [%]
1-1	**201**	107	76.40
1-2	195	122	76.18
1-3	195	118	**76.58**
1-4	194	**124**	76.54
1-5	194	123	76.49

表 4: 二段目およびソルバー全体の性能

ID	正解数	分類精度 [%]
2-1	**139**	70.89
2-2	137	71.14
2-3	136	**73.13**
2-4	134	70.64
2-5	133	71.14

4.4 二段目の評価

一段目の性能が最も良かった表 3 の ID1-1 では、284 問中 201 問で正解選択肢が出力の 2 個に含まれていた。二段目の 2 値分類モジュールの性能評価には、その 201 問で出力された正解選択肢と間違い選択肢によるペア合計 402 個 (201 組× 2 方向、以下、学習セット A と記す) を用いる。

まず、3 節で述べた 249 個の特徴量から、65 個をランダムに選んで構成した 5 万通りの特徴ベクトルに対して、2 値分類モジュールの分類精度を測定した。測定には、学習セット A に対して 10-fold cross validation を適用した。

この測定で分類精度が最も高かった 50 種類の 2 値分類モジュールのそれぞれに対し、一段目 (ID1-1) と組み合わせた場合のソルバー全体の性能を評価した。この評価には、大問単位の leave-one-out cross validation を用いた。すなわち、一段目の学習には、大問 1 題を除いた 70 題から作成した実例集合を用いる。二段目の学習では、この実例集合の中の、学習セット A に含まれる実例のみを用いる。

評価結果を表 4 に示す。表 4 には、全体の性能評価を行った 50 種類のうち、正解数が多かった 5 種類に対し、その正解数と、二段目で使用した 2 値分類モジュールの分類精度を示した。この表に示すように、正解数の順位と分類精度の順位は必ずしも一致しない。これは、正解数を測定した評価 (大問単位の leave-one-out cross validation) と、分類精度を測定した評価 (10-fold cross validation) では、二段目の 2 値分類モジュールの学習に使用した実例集合が異なるためである。

表 5: 二段階選抜法と BCBM の比較

		BCBM 正解	BCBM 不正解	計
二段階選抜法	正解	92	47	139
	不正解	28	117	145
	計	120	164	284

表 6: 二段階選抜法と一段目のみの比較

		一段目 ID1-4 正解	一段目 ID1-4 不正解	計
二段階選抜法	正解	94	45	139
	不正解	30	115	145
	計	124	160	284

表 7: 一段目のみと BCBM の比較

		BCBM 正解	BCBM 不正解	計
一段目 ID1-4	正解	84	40	124
	不正解	36	124	160
	計	120	164	284

4.5 検証

二段階選抜法の効果を検証するため、これまで提案された解法の中で最も成績が良かった BCBM と比較する。4.1 節で示したデータセットに対するそれぞれの解法の正解数は、二段階選抜法が 139 問、BCBM が 120 問である。この結果に McNemar 検定を適用したところ (表 5)、有意水準 5%で成績に差が見られた。

二段階選抜法の、BCBM からの変更点は、(1) 特徴量を増やす、(2) 選抜を二段階で行う、の 2 つである。特徴量を増やしたことの効果は、一段目のみで正解数が最も多い場合 (表 3 の ID1-4、124 問正解) と BCBM(120 問正解) の差である。一方、選抜を二段階で行う効果は、二段階選抜法の最高成績 (表 4 の ID2-1、139 問正解) と一段目のみで正解数が最も多い場合 (表 3 の ID1-4、124 問正解) の差である。

これらに対して、それぞれ McNemar 検定を適用した (表 6、表 7) が、どちらにも有意な差は見られなかった。つまり、性能の向上は、特徴量を増やすことと選抜を二段階で行うことの両方の効果によってもたらされている。

ID2-1 では、一段目として ID1-1 を用いている。一段目で正解選択肢が出力に含まれた 201 問を対象に、一段目の時点で解答を確定する場合と二段階選抜法の成績を比較した (表 8)。表 8 から、一段目と二段目で異なる判定が下された設問が 70 問あるとことが分かる。このうち 51 問は、正解が 2 位から 1 位に上がり、残りの 19 問は正解が 1 位から 2 位に下がった。この結果より、2 つの判定器を利用する手法は、有効に機能していることが分かる。なお、ID1-1 と ID2-1 の特徴ベクトルに共通する特徴量は、11 個である。

表 8: 一段目と二段目の比較

		一段目 正解	一段目 不正解	計
二段目	正解	88	51	139
	不正解	19	43	62
	計	107	94	201

5 まとめ

本研究では、選択式の評論読解問題の自動解法として、二段階選抜法を提案し実装した。この解法は今回対象とした読解問題に対して正解率 48.9%を達成するとともに、これまでの解法 (BCBM) よりも有意に性能が高い。

この解法では、選択肢選抜の判断材料として、選択肢と照合領域の文字の一致率や共通して出現する語句の数など、表層的な特徴を主として使用している。性能をさらに向上させるため、論理の展開や文間関係の推定など、表層的には表れにくい特徴量を取り入れていくことが必要であると考えられる。

謝辞 本研究は、JSPS 科学研究費基盤研究 (B) 「文章の読解と産出のための言語処理技術」(課題番号 15H02748) の助成を受けている。本研究では、プロジェクト「ロボットは東大に入れるか」から提供された試験問題データを使用した。

参考文献

[1] 佐藤理史, 加納隼人, 西村翔平, 駒谷和範. 表層類似度に基づくセンター試験『国語』現代文傍線部問題ソルバー. 自然言語処理 Vol.21 No.3 pp.465-483, 言語処理学会, 2014.

[2] 加納隼人, 佐藤理史, 松崎拓也. 表層的特徴を用いたセンター試験『国語』評論読解問題の自動解法. 人工知能学会論文誌, Vol. 32, No. 1, 2017.

[3] 服部昇平, 佐藤理史. 多段階戦略に基づくテキストの意味関係認識: RITE2 タスクへの適用. 情報処理学会研究報告 2013-NL-211 No.4/2013-SLP-96 No.4, 情報処理学会, 2013.

[4] 赤石美奈. 文書群に対する物語構造の動的分解・再構成フレームワーク. 人工知能学会論文誌 Vol. 21, No. 5, pp. 428-438, 2006.