表現の類似性と文書分類を併用した センター試験英語段落タイトル付与問題の解答手法

井内　健人*　，菊井　玄一郎*，杉山　弘晃+，但馬　康宏*
*:岡山県立大学大学院 情報系工学研究科
+:ＮＴＴコミュニケーション科学基礎研究所

Choosing a Title for a Paragraph in ‘Center Test’ of English

Kento Iuchi*, Genichiro Kikui*, Hiroaki Sugiyama+ and Yasuhiro Tajima*
*:Okayama Prefectual University Graduate School of Computer Science and Systems Engineering
+: NTT Communication Science Laboratories

1. はじめに

大学入試問題を解く計算機プログラムを開発することにより，計算機が人間の知的能力にどれほど近づくことができるかを明らかにする目的で「ロボットは東大に入れるか」というプロジェクト[1]が行われている．我々はこのプロジェクトに参画し英語の入試問題の解答手法を検討している．本研究では近年のセンター試験英語第 6 問 B で出題されている，長文の各段落にタイトルを付与する問題（「以下，タイトル付与問題」と呼ぶ）に取り組む．この問題は配点が高く高得点を取る上で重要である．本研究では word2vec による段落とタイトルの類似度計算に教師あり文書分類を組み合わせた手法を提案し，模擬試験を含むセンター試験問題を対象に評価を行う．

2. 段落タイトル付与問題

センター試験英語第 6 問では最初に 6 つの段落から成る 650 から 750 語程度の論説文（本文）が提示される．第 6 問 B はこの本文の指定された段落に対してタイトルとして適切なものを選択肢から選ぶ問題であり，表の空欄を埋める形式になっている．図 1 の上側の枠で囲まれた部分に本文の一部，その下に空欄付きの表，さらにその下に選択肢を示す．表の Paragraph の列は本文の段落番号であり，Content の列には各段落に対応した 5〜10 単語ほどのタイトルが入る．Content 欄のうち 4〜5 個には解答欄の番号(例[51])のみが書かれた空欄になっており，受験生は空欄に入れるべきタイトルを選択肢から重複が無いように選ぶ．本問は完答した場合，すなわち，選んだ段落-選択肢の組み合わせ全てが正解した場合にのみ得点となることから，ランダムに選択した場合に得点できる確率は 1/4!≒4%から 1/5!≒0．8%と非常に低い．

この問題においてタイトルの内容は主に次の 3 種類に分けられる．

①段落の文章の内容を短く要約したもの
　例：Using dance to point out unfavorable actions
②当該段落が文章全体の主題のどのような属性について述べているかを表したもの
　例：The value of giving music your full attention
③文章全体の構成における当該段落の役割を示したもの
　例：Introduction,　Conclusion

このうち③が空欄（選択肢）となることはほとんど無い．①についてはテキストの内容が選択肢の内容を包含しているかどうかを判定することにより正解を求めることができると考えられる．②については狭い意味で①のような意味的な包含関係はない．

Dance is one of the oldest forms of art, and it is seen in every culture and performed for a variety of purposes. In modern society, dance is widely recognized as a form of entertainment: many people enjoy dancing for fun or watching their favorite artists dance on stage or screen. It can also be a form of sport: there are dance competitions of various types. In addition to these obvious functions, however, there are other more complex roles dance can play in a society.

Paragraph	Content
(1)	Typical roles of dance today
(2)	[51]
(3)	[52]
(4)	[53]
(5)	[54]
(6)	[55]

① Dance for passing down appropriate cultural behavior
② How dance improves a group's status
③ The common function of dance and its significance
④ The demonstration of group force through dance
⑤ Using dance to point out unfavorable actions

図 1．2013 年実施センター試験（本試験）における 段落タイトル付与問題　上は本文の一部

3. 関連手法

段落タイトル付与問題に直接適用できる可能性がある手法として，「文章からタイトルを自動生成する手法」が挙げられる．文献[3]は，ニュース放送を書き起こした文章と，その文章に人手で付与したタイトルのデータを約 40,

000 個用意し，文章とタイトルの(単語間の）対応関係を EM アルゴリズムによって学習するというものである．文章とタイトルの対応関係を学習することは段落タイトル付与問題の解答手法としても有効性が期待できる．ここで問題は学習データをどのように準備するかである．

センター試験の段落タイトル付与問題のように，多岐にわたる分野を扱った文章の段落に対してタイトルを付与したデータは存在しないと思われる．近いものとして Wikipedia の説明本文が考えられる．Wikipedia における一つの見出しに対する説明本文はそれぞれにタイトルが付いた幾つかの項目（サブテキスト）に分かれている．しかしながら，これらのタイトルはその殆どが 1~4 語程度と短く，2 章で述べた①,②のようなタイトルとは異なっており不適切であると考える．

4. 提案手法

提案手法の全体の流れを図 2 に示す．

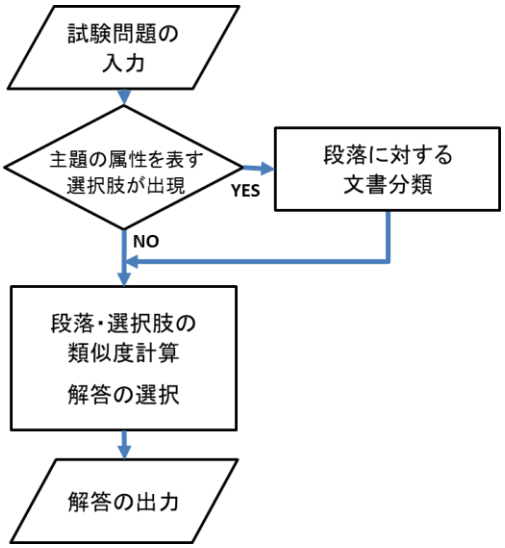

図 2. 全体の流れ

まず，選択肢の中に 2 章の②に属するもの，すなわち，文章の主題の属性名を含む選択肢(例：”History of xxx”)，があるかどうか判定する．ここでは属性として「歴史」と「利点や長所，効能」の 2 種類のみを考え，属性名を表す単語として前者については “history“，後者については “effect“，“benefit“，“merit“として，これらが選択肢に現れる場合に「属性名を含む選択肢（タイトル)」であると判定した．

上述の処理で属性名を表す選択肢が存在すると判定された場合，4.1 で説明する文書分類を用いて当該属性について述べている段落を一つ選び，この選択肢と対応づける．なお，複数の属性名に対してそれぞれ選択肢が存在した場合は，各属性に対して文書分類の処理を適用する．

「主題の属性名」を含む選択肢（およびこれに対応すると判断された段落）以外の選択肢と段落について，4.2 で述べる，段落と選択肢の意味的類似性による回答選択を行う．

4.1. 教師あり文書分類による解答の選択

本研究では，ある段落が事物の特定の属性を述べているか否かはこの段落の出現単語を手がかりとしてある程度推測できるものと考えた．例を挙げると，段落が歴史（“history “）について述べている場合，“1800 “などの年代を表す数字や，“ancient “，“since “といった単語の出現が見られる．“history “と “since “の間の関係は共起関係をベースにした word2vec では捉えにくいと考え，ここでは教師付きの文書分類手法を使って，段落が「歴史」のような特定の属性を表しているかどうかを判定することにした．なお，学習データ及び実験の条件については 5 章で述べる．

4.2. 表現の類似性による解答の選択

2 章で述べた①については，段落とタイトルの意味が包含関係にあるか，類似関係にあると考えられる．本研究では word2vec による類似性の尺度を用いることにした．

処理全体は 2 つのステップからなる．最初のステップで，段落と選択肢の全ての組みに対してこれらの間の類似度を求める．次のステップではこうして得られた類似度をもとに最も解答として適当と思われる組み合わせを選ぶ．

4.2.1. ステップ 1：類似度の計算

最初のステップの類似性は，段落と選択肢の文を word2vec によってベクトル化したものの，コサイン距離とする．段落，および，選択肢のベクトル化については，これらを単純に bag-of-words と考えて，含まれる全ての単語を word2vec でベクトル化し IDF(Inverse Document Frequency)を重みとして足し合わせたものをベースとする．IDF の重みは，より段落内容の特徴を表現したベクトルを作るためであり，イディオムの意味をベクトルで扱う先行研究に基づいている[4]．すなわち，段落あるいは選択肢のベクトル $\boldsymbol{A}$ は次の式で表される．

$$\boldsymbol{A} = \boldsymbol{w_1} \times idf(w_1) + \boldsymbol{w_2} \times idf(w_2) + \cdots + \boldsymbol{w_n} \times idf(w_n)$$

ここで，段落あるいは選択肢に含まれる各単語を $w_1, w_2, \ldots, w_n$,とし，それぞれの word2vec によるベクトルを $\boldsymbol{w_1}$，$\boldsymbol{w_2}$，$\cdots \boldsymbol{w_n}$ で表す．また，$idf(w)$ を単語 w の IDF 値とする．

なお，段落をベクトル化する際に段落全体を使うのではなく，その一部を使うことも試みた．選び方は段落のほかに 4 通りとした．具体的には i)段落の最初の 1 文，ii)段落の最後の 1 文，iii)段落の最初の 1 文と最後の 1 文，iv) 段落中で最も選択肢との類似度が高い文である．なお，最初と最後の 1 文に注目したのはこれらが段落における重要文であることが多いとされるためである．

また，選択肢をベクトル化する際に，選択肢同士で共通する単語を削除する方法も試した．これは，共通する単語を削除することで，正解の組み合わせとそうでない組み合

わせの類似度差を大きく出来るのではないかと考えたからである.

こうして求められた類似度を, 図 3 のようなマトリクス形式にまとめる. 段落の行と選択肢の列の交点が類似度となっている. 例えば段落 2 と選択肢 2 の類似度は 0.7 である.

	選択肢1	選択肢2	選択肢3
段落1	0.2	0.9	0.3
段落2	0.6	0.7	0.4
段落3	0.8	0.5	0.1

図 3. 段落・選択肢の類似度をまとめたマトリクスの例

4.2.2. ステップ２：解答の選択

解答の作成については 3 通りの方法を試した.

①段落と選択肢の全組み合わせのうち類似度最大のものから段落や選択肢の重複が無いよう選ぶ（貪欲法),

②全ての解答候補 (段落に対して選択肢を重複なく割り当てたもの）の中から類似度の平均値が最大の候補を選ぶ,

③同じく全ての解答候補の中から類似度の最小値が最大の候補を選ぶ.

5. 実験

以上の手法を，過去のセンター試験（本試験と追試験）および予備校が実施したセンター試験用の模擬試験の段落タイトル付与問題，計 42 問分に適用し，全ての段落-選択肢の組み合わせが一致した問題数 (完答数) で評価した. なお，より細かい粒度で比較するため段落（解答欄）単位で集計した「平均正解率」も求めた（後述）. また，比較のため 4.1 節，4.2 節で述べた手法を単独で適用した場合も試した.

5.1. 学習データ

提案手法においては word2vec，idf 計算，および文書分類において学習データが必要である. まず word2vec では googlenews 記事から学習したとされるモデル[5]を使用した. また idf は New York Times の 3 年分の記事から，一つの記事を一つの文書として計算した.

文書分類用の学習データとして英語版 wikipedia を使った. wikipedia の記事はそれぞれにタイトルがついた幾つかの項目に分かれている. “history“の場合，“history“をタイトルに含む項目の文章を正例, そうでない項目の文章を負例とする. ある記事から正例を抽出したとき，同一記事内から同じ数の負例を抽出する. よって正例 : 負例の比は 1 : 1 となる. データ数を表 1 に示す.

表 1. SVM の学習データ数

属性（属性名を表す単語）	データ数
歴史（history）	195388
利点・長所（effect，benefit，merit）	21306

5.2. パラメータ

文書分類に用いる素性として, 正例の出現単語のうち出現頻度上位 1000 単語を使用した. 素性の値は 1 または 0 である. 今回は線形 SVM を使用した. ソフトマージン係数 C は，0.1， 1.0，10，100 の 4 通りで学習データに対する 5 分割交差検定を行い決定した. 結果を表 2 に載せる.

表 2. SVM のパラメータ

属性	C(ソフトマージン係数）	交差検定の正解率
歴史	1. 0	78. 70%
利点 長所	1. 0	72. 11%

6. 結果と考察

6.1. 類似度の計算と解答選択の結果

類似度を求める手法のみを使って全 42 問のタイトル付与問題を解き，実際の正答と比較することで評価した. 表 1 は完答数が多い順に上位 5 つの結果とその条件についてまとめたものである. この表で「段落」とは段落のどの部分の単語を利用してベクトルを作成したかを表す. また平均正解率とは個々の解答欄単位での正解率である (1 問中, 4 つの解答欄のうち 2 つ正解した場，平均正解率は 50%, 完答率は 0 となる).

表 3. 類似度を計算する手法での実験結果

IDF	段落	選択肢の共通単語	解答の選び方	完答数 (42 問中)	平均正解率 (%)
あり	類似度最大文	削除	②	19	67
あり	全体	削除	②	17	64
あり	全体	残す	②	17	62
あり	類似度最大文	残す	②	16	63
あり	最初と最後の 1 文	削除	②	16	61

最もよい結果となった方法は「IDF 重みあり，段落内で最も選択肢と類似している文の単語を利用, 選択肢間の共通単語を削除してベクトル化」して類似度を計算し，類似度平均が最大の組み合わせを解答とする方法だった. 42 問中 19 問に完答し，各問題の 4〜5 つの空欄の内，平均して 67%を正解した. 全体の傾向としては IDF を重みとして掛けたほうが良い結果を出しており, 有効性が認められる. 段落は全体をベクトル化し，選択肢は共通する単語を削除したほうが完答数・正解率ともに高い傾向にある.

6.2. SVM を使った段落分類の結果

次に，SVM を使った段落の分類結果を表 4 に示す.

表 4. 段落の文書分類結果

属性(属性名を示す語)	正解数
歴史（history）	4/7
利点・長所（effect，benefit，merit）	10/14

ここで，42 問の試験問題中に，“history“を含む選択肢（タイトル候補）は 7 個，“effect“，“benefit“，“merit“のいずれかを含む選択肢は 14 個存在する．

6.3. 類似度の計算と段落分類の統合結果

提案手法全体の評価結果を表 5 に示す．なお，類似度を計算する条件として，6.1 で最良だった「IDF 重みあり，段落内で最も選択肢と類似している文の単語を利用，選択肢間の共通単語を削除」を用いた．分類処理については今回作成した 2 つの属性について 1 つずつ適用した場合(上段と中段)と，両方とも適用した場合（下段）を求めた．

表 5．2 つの手法の統合結果

分類処理を適用した属性	正解数
「歴史」のみ	21/42
「利点・長所」のみ	17/42
両方	19/42

分類を使用しない場合と比較して，「歴史」の分類モデルを適用した場合，完答できた問題数は 2 問増えた．「利点・長所」の場合は，完答できた問題数が 1 問増えたものの，類似度のみによる手法で完答できていた問題 3 問が不正解になってしまった．

6.4. 考察

6.4.1. 類似度計算での誤り分析

類似度を計算する手法で正解できなかった例として，「同調圧力」が引き起こすよくない出来事について，身近な例をいくつか述べた段落に対する“negative aspects of peer pressure“というタイトルや，産業用ロボットの特徴を列挙した段落に対する“features of industrial robots“といったものがある．こうした例のような，段落中で述べられた事例を簡潔に言い換えたタイトルの場合，類似性を word2vec（の加重和）で捉えるのは難しいと思われる．

6.4.2. 文書分類の効果

提案手法ではタイトルが「歴史」および「利点・長所」に関するものについて，教師あり文書分類の手法を用いて対応する段落を選んだ．文書分類の精度は後者の方がやや高い（表４）．しかし，今回の問題に対しては，前者が若干有効であったのに対して，後者は逆効果となった．これは歴史について述べている文章中に「歴史(history)」という単語あるいはこれに類似した単語が出現しなかったため，他の手がかり語を使う教師あり分類の効果があったのに対して，「利点」についてはこれに類する単語が出現しがちであり，word2vec による類似性判定でも対応できたためであると推測される．

なお，”history”の分類に失敗した例として，複数の段落にわたって「歴史」に関する話題を扱っている問題がある．この問題では，正解の段落のみを”history”と判断させることができなかった．

また”effect”，”benefit”，”merit”の分類に失敗した段落には，「長所」や「効能」を述べている段落と判定するにあたって手がかりとなりそうな，”helpful”，”improve”，”can”といった単語が出現する．しかし，今回作成したモデルではこれらの単語が素性となっていないことが多かった．素性の数を増やすことや，学習データの抽出方法を改善することが必要と思われる．

7. まとめ

本研究ではセンター英語段落タイトル付与問題の解答手法を提案した．段落とタイトルをベクトル化してコサイン類似度を求め，解答を出力することで 42 問中 19 問を完答した．これはランダムに解答した場合のベースラインより明らかに高く有効であると言える．一方，教師あり文書分類によって，特定の属性を表すタイトルと段落とを対応づける処理を追加したが，改善には至らなかった．

今後の方向性としては，段落と選択肢との類似性を計算するときに段落全体ではなく中心的な部分を使うことや言い換えの処理が挙げられる．また，属性を表す段落の処理について，今回は２つの属性のみ考えたが更に拡張することも検討したい．

謝辞

本研究を遂行するにあたり「『ロボットは東大に入れるか』大学入試センター試験関連オンラインタスクデータ」を利用しました．ご提供下さった「独立法人大学入試センター」および「株式会社ジェイシー教育研究所」に感謝いたします．また，模擬試験データをご提供下さった学校法人高宮学園，株式会社ベネッセコーポレーション，「ロボットは東大に入れるか」を推進している新井紀子教授をはじめ，国立情報学研究所の方々に深く感謝いたします。また，本研究の一部は以下の各氏（組織）との共同研究として行われました．東中竜一郎（NTT），磯崎秀樹（岡山県立大），堂坂浩二（秋田県立大），平博順（大阪工業大），南泰浩（電気通信大）．熱心な議論に感謝いたします．

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？-国立情報科学研究所「人工頭脳プロジェクト-」『人工知能学会誌』27 巻第 5 号，pp.463-469,2012

[2] Tomas Mikolov, Kai Chen,Greg Corrado,and Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space. “arXiv preprint arXi 1301.3781

[3] Paul E.Kennedy, Alexander G, Hauptmann, “Automatic title generation for EM“, Proceedings of the 5th ACM Conference on Digital Libraries, 2000,pp.230-231

[4] Maria Pershina, Yifan He, Ralph Grishman, “Idiom Paraphrases: Seventh Heaven vs Cloud Nine“ Proceed-ings of the EMNLP Workshop on Linking Models of Lexical,Sentential and Discourse-level Semantics, Association for Computational Linguistics(2015) pp.76-82.

[5] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing