1K2-4

表層類似性と含意認識結果の組み合わせによる英語長文問題解答の試み

Solving Entrance Examination of English
based on Surface Similarity and Automatically Recognized Textual Entailment Relations

菊井 玄一郎 *1*2 Genichiro KIKUI
三村 奈央 *1 Nao MIMURA da
藤川 哲志 *2 Tetsushi FUJIKAWA
但馬 康宏 *1*2 Yasuhiro TAJIMA

*1岡山県立大学 情報工学部
Faculty of Computer Science and Systems Engineering, Okayama Prefectural University

*2岡山県立大学大学院 情報系工学研究科
Graduate School of Computer Science and Systems Engineering, Okayama Prefectural University

This paper proposes a method of solving English examination problems in the National Center Test for University Admission. We focus on so-called *long text questions*, where applicants are requested to read a text of 600-800 word length and to select an appropriate answer option from four alternative English sentences or phrases. The proposed method chooses the best ranked option by using ranking svm (support vector machine) trained with two kinds of features: surface similarity and strength of entailment relation between an option and the text. The former kind of features are simply calculated by counting overlapping words and the latter are obtained from TIFMO, an textual entailment recognition program. The proposed method achieved 42% correct rate.

1. はじめに

大学入試問題を計算機に解かせる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どれ程度近づくことができるかを明らかにしようという目的で「ロボットは東大に入れるか」と呼ばれるプロジェクトが国立情報学研究所を中心に実施されている [新井 2012]．このプロジェクトの中間目標は 2016 年までに大学入試センター試験（以降「センター試験」と略す）において最 難関大学に合格できるレベルの高得点を取ることにある．

我々はセンター試験の「英語」の解答手法の開発に取り組み，予備校が大学受験生向けに実施している模擬試験において受験生の平均を上回る性能を達成している [東中 2015]．しかしながら，この時の得点源は発音問題や単語補完問題，対話文補完問題，語順整序問題などであり，長文問題については配点が高いものの，有効な手法は見つかっておらず，選択肢をランダムに選んだレベルの正答率にとどまっている．そこで本研究では，長文問題のうちイラストやグラフなどを含まない，テキストのみで構成される「第 6 問 (A)」を対象として正答率の向上を目指す．

以下，第 2 章では本研究で対象とする「第 6 問 (A)」について説明し，3 章では関連研究を踏まえて回答の手掛かりについて検討する．そのあと 4 章で提案手法について説明し，5 章で評価実験について述べ，6 章でまとめる．

2. 対象とする長文問題 (第 6 問 (A))

センター試験英語第 6 問 (A) は，まず 650 から 750 語程度の英語長文が置かれ，そのあと原則として 5 つの「小問」が並ぶ．1 つの小問は「問題文」と 4 つの「選択肢」から構成される (図 1)．

長文は 2008 年以前は小説，日記，随筆などだったが，2008 年からは論説文に変わり，現在に至る．

問題文は平叙文の場合と疑問文の場合の 2 種類がある．

連絡先: 菊井 玄一郎，岡山県立大学情報工学部，岡山県総社市窪木 111， kikui'at'cse.oka-pu.ac.jp

問題文: According to paragraph (3), the Mbuti Pygmies [47].
選択肢：
① disciplined careless hunters through dance
② handed down customs and traditions through dance
③ made lazy members dance after a day's hunt
④ performed culturally desirable behavior by dance

図 1: 2013 年度 本試験, 第 6 問 A, 問 2

平叙文の場合は先の図 1 に示すように，一部が空欄になっており，長文を前提として（メタ的な意味も含めて）この文が真となるように，選択肢のいずれかを空欄に埋めることが求められる．空欄に入る英語表現はこの例のように句 (phrase) の場合もあれば節 (clause) や単語の場合がある．なお，問題文にはほとんどの場合'According to paragraph (3),' のような「長文の範囲を限定するような表現」 (メタ表現) が含まれる．

「長文の範囲を限定する表現」を除くとこの問題は典型的な含意関係認識の問題と言える．すなわち，問題文の空欄に各選択肢を代入してできた文の中で長文（の指定された段落）の内容から最も良く含意されるものを選び *1，その選択肢の番号を回答すればよい．

一方，問題文が疑問文の場合，この疑問文は長文全体，あるいは，特定の段落の内容を問う質問であり，この質問に対する最も適切な答（英語表現）を選択肢から選ぶことが求められる．選択肢で与えられる表現はフルセンテンスの場合と句や節 (e.g.,'to save money') の場合がある．

選択肢がフルセンテンス (long answer) の場合は長文と各選択肢（＝文）との間で含意関係が成立するかどうか調べ，最も強い含意関係にあるものを選べばよい．但し，選択肢は疑問文に対する自然な回答文になるように代名詞化などが行われて

*1 本来は選択肢の 1 つのみに含意関係があり，そのほかの場合は含意関係が成立しないはずである

問題文: What first attracted Beth to the study of sleep? 46
選択肢：
① A book on sleep which she got as a gift.
② ...

図 2: 2007 年度 追試験, 第 6 問, 問 2(一部)

いるため，選択肢のみでは意味をなさない場合がある．その場合は問題文との間で共参照解析等を行なって情報を補う必要がある．

選択肢が句や節の場合は問題文と統合して長文との間で含意関係のチェックが行えるようなフルセンテンスの long answer を生成する必要がある．

例えば図 2 のような問題文と選択肢が与えらえたとき，選択肢① については例えば次のような文を作る．

> A book on sleep which Beth got as a gift first attracted her to the study of sleep.

これを行うには疑問文を平叙文に変換して，疑問詞部分に選択肢の英語表現を代入する処理が必要である．

また，平叙文の場合と同様，回答の前提となる長文の範囲を限定する表現が問題文に含まれることが多い．

3. 回答の手掛かり

上述の通り，英語長文問題の「第 6 問」は本質的には長文から含意される選択肢を選ぶ「含意関係認識」の問題に帰着する．従って，最も直接的な解法は各選択肢から含意関係認識の対象となる「仮説（テキスト）」を生成し，含意関係認識を使って最も良く長文から含意されるものを選ぶ，という方法である．小松ら [*2] は TIFMO[Tian2014] と呼ばれる含意認識プログラムを用いて回答を試みている．我々も同じプログラムを第 6 問に適用して正解率を計算した．含意関係認識の入力である「前提」と「仮説」うち「前提」は長文全体とした．「仮説」は問題文が平叙文の場合，問題文から “According to paragraph (2)” のような長文の範囲を指定する表現をパターンマッチで取り除いたあと，空欄に選択肢を埋めたものとし，問題文が疑問文の場合は選択肢のみとした．この入力に対して含意認識を行い，出力値が最大でかつ値が 0.4 以上であるような選択肢を選んだところ，全体の正答率は 25%となり偶然の一致レベルであった．しかしながら，問題文が平叙文の場合のみに限定すると正答率は 35%となり，かなり向上した．これは，問題分が疑問文の場合，選択肢が文でないなど，含意関係認識の入力（仮説）として不適当な場合があるのに対して，平叙文の場合は必ず文になっていて内容的にも妥当なことが多いためと考えられる．

以上から，適切な含意関係認識の入力を生成することで，正解率の向上が図れる一方，現状の含意関係認識だけでは限界があることが示唆される．

第二の方法として，表層的な類似性の利用が考えられる．佐藤ら [佐藤 2014] は，本研究で対象としている英語第 6 問と言語は異なるものの内容的に類似している，センター試験の国語の「傍線部問題」を対象として，長文との間で最も表層的に類似した選択肢を選ぶという方法を試み半分以上の問題に正解できることを示した．当該研究では表層的類似性として文字列の一致数をベースにした尺度を用いている．

この方法が英語でも効果的かどうかを明らかにするために，各問題について，長文との間で表層的類似性の高い選択肢から順に並べ，類似性の順位と正解率との関係を調査した．長文と選択肢の間の表層的類似性の尺度としては文字列の一致度ではなく，選択肢中の単語（異なり語）のうち長文にも出現する単語の割合（相対頻度）を用いた．なお対象単語は stanford coreNLP[Manning 2014] の品詞 tagger による品詞が FW, JJ*, NN*, RB*, VB* (*は 2 個以下の文字列) であるもの [*3] に限定した．単語の一致条件は同 tagger の出力の品詞と原形が一致するものとした．

さらに，選択肢の単語数が少ない場合など一つの問題の選択肢の間で長文との重複率が一致する場合がしばしば存在する（今回 73 問中，正解選択肢と同じ重複率の別の選択肢が存在した数は 27 問あった）．そこで，もし同じ重複率の選択肢が複数存在した場合は各順位の出現数を同率の選択肢の個数で割ることにした．たとえば，重複率の降順に並べた場合の 1 位と 2 位が同じ重複率でこれらの片方が正解であった場合，1 位と 2 位の場合それぞれについて 1/2 回とカウントする．

結果を表 1 に示す．たとえば，単語の重複順位が 1 位の選択肢を選ぶと平均して 16.3%の割合で正解することが分かる．この表を見ると重複率の高い候補を選ぶことは平均正解率としてはあまり得策でないことが分かる．むしろ，最も重複率の低い候補を選んだ方が良い可能性がある．実際この表によれば重複率の低い単語を選ぶことで偶然以上の正解率が得られることが示唆される．受験参考書等で指摘されているように，英語を「本当に理解していること」と単に表層を見ていることを区別するために，敢えて「表層的には近いが意味的には異なる」言語表現を選択肢に含めている可能性がある．

表 1: 長文との単語重複率による順位と正解率との関係

単語重複率の順位	1 位	2 位	3 位	4 位
正解率	16.3%	24.5%	25.9%	33.2%

4. 提案手法

以上より，含意認識結果のスコア（含意認識スコア），および，表層的類似性はいずれも正解選択肢を選ぶための手がかりを与えているものの，少なくともそれぞれ単独での効果は限定的であることが分かる．

そこで，本研究では TIFMO から得られる含意関係認識スコアと表層類似性を素性（特徴量）として組み合わせ，最適な選択肢を選ぶ方法を試みる．試験問題から適切な含意関係認識の入力を自動生成することが難しい場合があることから，近似的な仮説テキストを複数通り生成し，それぞれに対する TIFMO のスコアを全て素性として利用することとする．また，この方法の限界を探る目的で，人手によって含意認識の入力を作成した場合の含意認識スコアも利用する．表層的類似性についても上述の「重複率」に加えて各選択肢の単語数も素性とする．

これらの素性値の最適な組み合わせ（関数）を求めるために，本研究では ranking svm(support vector

*2 ” 代ゼミセンター模試タスク”，「ロボットは東大に入れるか」成果報告会, 2013 年 11 月

*3 形　　容詞類，名詞類，副詞類, 動詞類 (be 動詞を除く)

machine)[Joachims2002] を用いる．すなわち，あらかじめ正解が分かっている試験問題において，各選択肢に対して表層的類似性，および，含意認識結果から素性ベクトルを作成する．次に ranking svm を用いて，各小問ごとに正解の選択肢が他の選択肢より高いスコアになるように素性からスコアを計算する関数のパラメータを学習する．

以下では，本手法で用いた素性について，表層類似性に関する素性，含意認識から得られる素性の順で説明する．

4.1 表層類似性に関する素性

4.1.1 S1:長文との重複単語数 (頻度)

長文と当該選択肢の双方を Stanford coreNLP[Manning 2014] で解析し，品詞 tagger の出力が FW, JJ*, NN*, RB*, VB* [*は 2 個以下の文字列] である単語について，長文と選択肢の間で品詞と単語原形が一致した語の異なり数とする．これは 3 章の分析で用いたものと同じ条件である．

4.1.2 S2:選択肢の単語数

当該選択肢において S1 と同じ品詞条件を満たす単語の異なり数である．これは選択肢の「内容語の量」を表すパラメータである．この素性単独では表層類似性に直接対応しないが，次に述べる相対頻度の分母にもなることと，選択肢の長さが解答の手掛かりになるという仮説に基づいて表層類似性の素性に含める．

4.1.3 S1r:長文との重複単語数 (相対頻度)

S1 を S2 で割った値である．

4.2 含意関係に関する素性

含意関係に関しては 3 つの素性 (R1,R2,R3) を用いる．これらは 3 通りの前提・仮説ペアそれぞれに対する TIFMO の出力（含意関係の強さを表す値）である．3 通りの前提・仮説ペアの全てで「前提」の部分は長文全体で共通であり，違いは「仮説」の部分にある．以下では R1 から R3 の「仮説」の生成方法について述べる．

4.2.1 R1:代入と並置

問題文が平叙文の場合と疑問文の場合で異なる．平叙文の場合 (基本的に 2009 年試験以降) は一部が空欄になっているので，そこに当該選択肢を入れたものを「仮説」とする．疑問文の場合，疑問文の直後に空欄が置かれている場合と空欄がない場合があるが，いずれの場合でも問題文の直後に選択肢を単純に並置する．

(2007 年度本試験, 問 1)
問題文: Why did Valerie and Grandpa laugh? 46
選択肢: Valerie had not finished her preparation.
仮説テキスト: Why did Valerie and Grandpa laugh? Valerie had not finished her preparation.

この場合，生成される仮説には疑問文が含まれるため，含意認識の対象として不適当になるが，最小限の処理という観点からこの形式を用いる．また，選択肢がフルセンテンスでない場合，文でないものが並置されるが，これについてもそのままで含意認識を適用する．

4.2.2 R2:選択肢のみ

当該選択肢のみを仮説とする．

問題文: Why did Valerie and Grandpa laugh? 46
選択肢: Valerie had not finished her preparation.
仮説テキスト: Valerie had not finished her preparation.

但し，問題文が平叙文の場合については R1 と同様， 空欄に選択肢を埋めたものとする．

4.2.3 R3:疑問文の変形と代入（人手作成）

問題文が平叙文の場合は空欄に選択肢を埋めたものを仮説とする (R1, R2 と同じ)．そうでない場合は，フルセンテンス形式の回答文を作成して仮説とする．回答文の生成の基本的な方法は，問題文（疑問文）に対して統語的変形を行って平叙文化し，疑問詞部分に選択肢を置き換えることである．

問題文: Why did Valerie and Grandpa laugh? 46
選択肢：Valerie had not finished her preparation.
仮説テキスト: Valerie and Grandpa laughed because Valerie had not finished her preparation.

なお，選択肢部分に存在する代名詞は参照先名詞に置き換えて疑問文なしで解釈可能な回答文となるようにする．

(2005 年度本試験, 問 2)
問題文: What did Kevin do to help Tommy? 47
選択肢: He taught him how to become a good dancer.
仮説テキスト: Kevin taught Tommy how to become a good dancer.

以上の処理はある程度自動処理が可能であるが，正確な構文解析や共参照解析が必要となることから，今回は人手で行った．

5. 実験

上記手法を評価するため，過去のセンター試験本試験と追試験 14 回の計 73 問を 5 分割して交差検定を行った．

5.1 実験条件

5.1.1 実験データ

下記の大学入試センター試験英語筆記試験の第 6 問 (A) で問題文の回答として選択肢のいずれかを選ぶ形式の問題を使用した．

種別	対象年	大問の数	小問総数
本試験	1997-2013 の奇数年 (2001 を除く)	8	41
追試験	1997-2009 の奇数年 (2005 を除く)	6	32

これらのデータを 1 回分 (ある年の本試験あるいは追試験の小問 5-7 からなる) を最小単位としてランダムに 5 つのサブセットに分けて 4 つを訓練用，残り 1 つを評価用として交差検定を行った．

5.1.2 ランキング SVM のプログラム

Cornell 大学で開発された SVM^{rank} [Joachims2002] を使用した．

5.1.3 カーネル関数

ranking svm のカーネル関数は線形カーネル，多項式カーネル，RBF カーネル (ガウスカーネル) を試した．

5.1.4 パラメータ

SVM においては学習時のマージン制約違反に対するペナルティー係数であるソフトマージンパラメータの設定が必要である．さらに，線形カーネルにおいては次数，RBF カーネルにおいてはガンマの値を設定する必要がある．これらは訓練

データをさらに分割して交差検定とグリッドサーチによって決定した．

5.1.5 素性値の正規化

素性のうち TIFMO から得られる値は [0,1] であるのに対して，表層類似性に関するパラメータの一部は整数値であり，値域が大きく異なる．そこで，各素性について訓練データの最大値で除すことにより正規化を行った [*4]．

6. 実験結果

提案手法を用いて交差検定により ranking svm のパラメータを学習して評価セットに対する正答率を算出した．

比較のため，表層に関する素性のみ (S1,S2,S1r)，含意関係認識から得られる素性のみ (R1, R2, R3)，全素性の 3 通りに加え，含意関係認識から得られる素性については人手が介在した素性 R3 を除く R1+R2 のみの場合を試した．また，単語重複率については相対頻度と一致単語数の両方を試した．結果を表 2 に示す．

全素性を利用した場合，正答率は 42%となり，ランダムのベースラインより大幅に向上する．表層類似性に関する素性のみ，または，含意関係認識に関する素性のみの場合に比べて高い正答率になる．含意認識に関する素性のみの場合，3 つの素性を全て用れば 36%となり全素性を使った場合には及ばないものの，ベースラインよりかなり向上する．しかしながら，含意関係のみの場合でも，問題を最も適切に含意関係認識の入力に変換したと思われる R3 を除いた場合には 32%程度に低下する．一方，表層類似度に関する素性も含めた場合は R3 を除いても正答率は低下しない．このことから，現状の含意認識を前提とするなら，選択肢と疑問文を単に並置する程度の処理で十分であることが分かった．

表 2: 提案手法による英語第 6 問の正答率

	カーネル関数		
素性	linear	Poly	Rbf
S1+S2	.26	.30	.33
R1+R2+R3	.33	.32	.36
R1+R2	.30	.32	.31
S1+S2+R1+R2+R3	.40	.40	.38
S1r+S2+R1+R2+R3	.32	.30	.42
S1+S2+R1+R2	.37	.42	.38
S1r+S2+R1+R2	.32	.30	.32

7. まとめ

センター試験英語の長文問題（第 6 問）の解答法を提案した．提案手法は選択肢と長文の間の表層的類似性に関する素性（特徴量），および，選択肢と問題文から構成されるいくつかパターンのテキストと長文との間の含意関係を素性として，これらを ranking svm で組み合わせることにより，選択肢のスコアを計算し，最大のものを出力する．この手法を過去のセンター試験 14 回分に含まれる合計 73 問について評価したところ，過去の手法を上回る正答率 42%を達成した．

*4 交差検定においては各分割に対してこの操作を行った

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」を利用した．データをご提供頂いた「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」，プロジェクト「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とする国立情報学研究所の方々に謝意を表する。

また，本研究の一部は以下のメンバー（組織）との共同研究として行われた．東中竜一郎, 杉山弘晃 (以上 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）．記して感謝する．

参考文献

[新井 2012] 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 -”, 人工知能学会誌, Vol.27, No.5, 463-469 (2012).

[東中 2015] 東中竜一郎，杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩, “センター試験における英語問題の回答手法”, 言語処理学会第 21 回年次大会, pp.187-190 (2015).

[佐藤 2014] 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.465-482 (2014).

[Manning 2014] C. Manning, M. Surdeanu, et al. “The Stanford CoreNLP Natural Language Processing Toolkit”, ACL-2014: System Demonstrations, pp. 55-60 (2014).

[Tian2014] Ran Tian, Yusuke Miyao, Takuya Matsuzaki, “Logical Inference on Dependency-based Compositional Semantics”,In Proc. of ACL 2014, pp.79-89 (2014).

[Joachims2002] T. Joachims, “Optimizing Search Engines Using Clickthrough Data”, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.