2K3-4

英語センター試験「未知語語義推定問題」解答のための評価尺度について

On Similarity Measures for Choosing Semantically Equevalent Expressions in Entrance Examination of English

菊井 玄一郎[*1] 湯藤 真大[*1] 但馬 康宏[*1]
Genichiro KIKUI Masahiro YUTO Yasuhiro TAJIMA

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

This paper investigates word sense estimation (WSE) for solving English problems in the National Center Test for University Admission. The WSE problem requests an examinee to choose one English word/phrase out of four choice options which has the closest meaning to the given word/phrase in a English text. We applied three kinds of measures for choosing options: 1) phrase level similarity between the given word and an option, 2) language model likelihood of the text replacing the given word with an option, and 3) sentence level similarity calculated by a paraphrase recognizer. Experimental results have shown that the phrase level similarity using word2vec combined with an English dictionary and idf weights is most effective achieving 69% accuracy.

1. はじめに

国立情報学研究所が主導して「ロボットは東大に入れるか」と呼ばれるプロジェクト（グランドチャレンジ）が行われている [新井 12]．これは，最難関大学の入試問題を解くことができる計算機プログラムを開発することを通じて，自然言語処理を含む人工知能の各領域の進歩と統合を目指すものである．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 15]．しかしながら，難関大学に入学できるレベルとの間には依然大きな乖離があり，既に受験生の平均を上回る成績にある「単文問題」についても更なる検討が必要である．

本研究では「単文問題」として扱われている問題のうち，未知語の語義推定問題に取り組む．ここで、語義推定問題とは英語テキスト中の下線で指定された言語表現（以降、下線部と呼ぶ）に対して，文脈を考慮した上で意味的に最も近い言語表現を４つの選択肢から選ぶ問題である。なお，言語表現の多くは語または句であるが，文のこともある．

先行研究において我々は下線部と選択肢の言語表現の間の意味的な類似性を word2vec を用いて評価することにより，一部の模擬試験で 75%の平均正解率を達成している [東中 15]．これは十分高いようにも思われるが，後述するように，この手法をセンター試験や他の模試等広範な試験に適用すると平均正解率は 65%程度であり，このタイプの問題が一回のテストに２問しか出題されないことを考えると十分とはいえない。

本研究ではこの語義推定問題の解法について２つの方向から検討を行う．一つ目は我々の提案した手法の改良であり，二つ目は新たな同義性尺度の適用である．前者については英語辞書（英英辞典）の語義説明文を利用して元の表現をパラフレーズすることにより，同義性の認識漏れの低減を目指す．後者については，含意関係認識や翻訳評価などの分野で提案されている「言語表現間の意味的同値性・類似性を評価する尺度」などの適用を試みる．これらの尺度はそれぞれの分野でその有効性が示されているが，一定量 (約 100 問) の大学入試問題に適用してその有効性を評価することには大きな意味がある．

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

次の問いの英文を読み，下線部の語句の意味をそれぞれの文章から推測し，［　　］に入れるのに最も適当なものをそれぞれ下の① - ④ のうちから選べ．

問 2: In my high school years, my friend and I felt that Mr. Bell was the epitome of a good high school PE teacher. He was not tall or well-built, but he was able to teach sports which often required a lot of strength and endurance. Furthermore, he had the ability to make us do our best and never give up. Even today I believe I have never met a better PE teacher.
In this situation, the epitome of a good PE teacher is one who is the [28].
選択肢：
① athletic kind
② perfect example
③ practical sort
④ strict type

図 1: 2013 年度 本試験, 第 3 問 A 問 2

以下，第 2 章では本研究で対象とする「未知語語義推定問題」について説明し，3 章では本研究で試みた解答手法について説明し，4 章で評価実験について述べる。5 章はまとめである．

2. 未知語語選択定問題

未知語語義推定問題は，英語の文章（パラグラフ），英語による回答指示文 (“In this situation” で始まる文)，および，選択肢から構成される．例として，2013 年度センター試験（本試験）の大問 3A の問 2 を図 1 に示す．英語の文章中の一つの言語表現（単語あるいは連語）には下線が付与されている．下線部の英語表現は大学受験生の語彙レベルを大きく超える語句，すなわち，**未知語句** であることから「未知語語義推定問題」と呼ぶ．ほとんどの受験生は下線部の意味を知らないた

め，英文全体のコンテクスト（と選択肢）を利用して意味を推測することになる（問題冒頭の日本語指示文にも「下線部の語句の意味を文章から推測し」と書かれている）．

回答指示文は下線部とどういう関係にある選択肢を選ぶべきかを指示するもので，例は間接的ながら下線部と同じ意味の単語を選ぶことを求めている．実際は次のような直接的な表現になっている回答指示が多い．

expression means 00.

回答指示文は任意の英文であり得るため，問題に正しく解答するためには，問題ごとに回答指示文を解釈する必要がある．しかしながら，実際にはほぼすべての問題において下線部と広い意味での同義関係にある選択肢を選ぶことが求められており、本研究でもそのような「決め打ち」を行い回答指示文の解釈はスキップする．

なお，下線部と選ぶべき選択肢は（意味的には）置換可能になるものが多いが，置換すると統語的に不適格になるものが全体の 16.5%存在する．違いの軽微なものとして，統語的素性の微妙な違い (名詞句における specifier の有無や動詞句における to 不定詞形と finite 形の違いなど) があるが，より難しいものとして，以下の例のように，人物を表す名詞に対して，その人物を定義するような行動を表す動詞句が選択肢になっているようなものがある．これらを適切に扱うことは今後の課題である．

..Mr. Joseph Malunga is a maverick politician. ...
In this passage, maverick means someone who 28.
選択肢：④ thinks differently from most people
(2009 年度センター試験英語 追試験第 3 問 A の一部)

3. 解答手法

本研究で取り組む「未知語語義推定問題」は，下線部の言語表現に対して 4 つの選択肢で示される曖昧性があり，この曖昧性を次の二つの手掛かりを用いて解消する処理と考えられる．

1. 下線部の言語表現そのもの（**下線部** と呼ぶ）
2. 下線部の周りの文章（**文脈** と呼ぶ）

多くの受験生は下線部の意味を知らないため，文脈（2 番目の手掛かり）を頼りに曖昧性を解消せざるを得ないが，計算機の場合は膨大な語彙を保持させて，下線部を既知語にすることで，文脈を使わずに解ける可能性がある．実際，従来手法 [東中 15] は文脈を一切使わない．

以下では，本研究で適用を試みた手法について，下線部，および，文脈のいずれの情報を使用するかに着目し，３つに分けて説明する．

3.1 下線部のみを用いる方法

3.1.1 従来手法 (word2vec)

下線部と各選択肢の間の「意味的類似性」を word2vec[Mikolov 13] を用いて計算し，最も類似性の高いものを選ぶ方法である．文脈は一切参照しない．二つの言語表現間の意味的な類似性は各言語表現をベクトル化し，それらの間の角度のコサイン値とする．この手法を適用する上で解決すべき最大の問題は複数単語からなる表現，特に，全体の意味が構成単語の意味の「合成」になっていない慣用表現をどのようにベクトル化するかということである．我々は，英語辞書 (具体的には wiktionary [*1]) を参照し，当該連語が見出しに掲載されている連語についてはその語義記述を用いることとし，そうでない場合は連語を構成する単語のベクトルの和を用いることとした．なお，英語辞書の語義記述が複数単語から構成されている場合は各単語のベクトルの和とする．また，辞書の一つの見出しに対して複数の異なった語義記述が存在する場合は，これら各々の語義記述（から得られたベクトル）と選択肢との類似性を評価し，最大のものを使うこととした．

なお，下線部は受験生にとって未知語，すなわち，出現頻度が低い言語表現であることから，単語 (意味) ベクトルを作成するために必要なコーパスは相当大規模である必要がある．これについては超大規模なニュース記事コーパスから学習されたベクトルデータを利用することとした．

3.1.2 従来手法の改良

既存手法のエラー分析の結果，二つの言語表現が意味的に類似しているにもかかわらず，表層的な違いによってそのように判断されない例があり，これらの一部は英語辞書によるパラフレーズによって救済できることが分かった．そこで，本研究では下線部に加えて選択肢も含め，また，連語であるなしに限らず，全ての言語表現に対して英語辞書を引き，得られた語義記述文を代替表現として利用することにした．さらに，複数語からなる言語表現の意味をベクトル化する際に各単語のベクトルの単純和ではなく，idf(Inverse Document Frequency) 値によって各単語のベクトルに重みづけする加重和とした．これはイディオムの意味をベクトルで扱う先行研究 [Pershima 15] に基づいている．以上の改良点をまとめるとつぎのようになる．

1. 下線部，および，４つの選択肢の表現に対して英語辞書を引いて代替表現とする（全ての代替表現の組み合わせ照合し類似度最大のものを取る）
2. 複数単語からなる言語表現のベクトル v_p は次式の通りである．ここで，$v_w(w_i)$ は単語 w_i のベクトル．$idf(w_i)$ は単語 w_i の idf 値である．

$$v_p(w_1, \ldots, w_n) = \sum_{i=1}^{n} idf(w_i) v_w(w_i)$$

3.2 下線部の周囲の文脈のみを用いる方法

3.2.1 ngram 尤度に基づく方法

問題文中の下線部を「空欄」に置き換えると，語義推定問題は，文中の空欄を選択肢のいずれかで埋める「空所補充」の問題とみなすことができる．空所補充問題については ngram 言語モデルを用いて，尤度が最大になるような選択肢を選ぶ方法により模擬試験において 70%の正答率を達成している [東中 15]．この方法をそこで，この方法を語義選択問題に適用することにした．なお，先行研究では単語ベースの ngram モデルの他に，単語が動詞であればその直後に動詞の活用形を表すカテゴリ記号を挿入した列に対する ngram である VPOS 法を適用して 10%程度の改善を確認しているが，本研究では単語ベースの ngram モデルのみを試みる．

3.3 下線部を含む文脈の利用

本研究では文脈の範囲を下線部を含む文に限定して検討する．すなわち，下線部を含む文において下線部と各選択肢とを

∗1 https://en.wiktionary.org/

置換した文を作り，元の文と置換後の文の間の意味的同値性を評価し，最も意味の変わらない選択肢を選ぶ．従って，必要な処理は文レベルでの同義性の自動評価である．

二つの文の同義性を評価する処理はパラフレーズ認識と呼ばれる分野で，多くの研究が行われている．これらは 2 つの文（言語表現）が同義か否かを判別する狭義のパラフレーズ認識の他に，翻訳結果が参照訳にどの程度意味的に近いかを数値化することで翻訳品質を評価する**翻訳自動評価**の分野でも検討されている．また，二つの文が同義であるということは，これらの間に双方向の含意関係が成り立つことと考えられるので含意認識手法も利用することができる．

そこで本研究では (狭義の) パラフレーズ認識，翻訳評価，含意認識から一つずつ選んで本問題に適用する．

なお，文を単なる単語列と考えると，3.1 節の word2vec を用いた意味的類似性の評価手法も適用できるが，同義性を比較する 2 つの文は下線部と選択肢の部分以外は全く同じであるため，文脈を無視した場合と同じ結果になる．従って，3.1 節の方法は含まない．

3.3.1 狭義のパラフレーズ認識

パラフレーズ認識には多くの研究が行われている．([Qiu 06],[Socher 11],[Blacoe 12], [Yin 15] など). [Yin 15] によると彼らの方法が最高の精度であり，[Socher 11] より coherence task において 0.7%程度精度が高い．しかしながら，本研究ではソフトウエアが公開されている Socher らの方法を用いることにする．

Socher らの方法は依存構造木を入力として動的プーリング付きの再帰的ニューラルネットを用いて木全体に対するベクトル表現を構成し，これを用いて同義性を判別するものである．彼らの公開しているソフトウエアは与えられた二つの文が同義であれば 1，そうでなければ 0 を返す簡潔なインタフェースとなっている．この方法を，下線部を含む元の文と下線部を選択肢で入れ替えた文に適用して同義か否かを判定し，同義であるものの選択肢の番号を出力する．なお，複数の選択肢に対して同義と判定された場合はその中からランダムに一つを選ぶ[*2]．また，どの選択肢に対しても同義と判定できなかった場合は全ての中から一つランダムに選ぶ．

3.3.2 翻訳結果の自動評価尺度

二つの文の同義性を判定する別の方法として，翻訳結果に対する自動評価手法を用いることができる．翻訳結果の自動評価手法の多く ([Papineni 02], [Denkowski 14] など) は評価対象の翻訳文が参照訳（正解訳）に意味的にどのくらい近いか（類似しているか）を数値化する処理であり，目的言語におけるパラフレーズ認識そのものであると言える．

実際に翻訳自動評価手法をパラフレーズ認識タスクに適用した研究もあり，Yin ら [Yin 15] の実験によると，パラフレーズ認識精度で評価した場合，翻訳評価尺度の中で最も良いのは METEOR であり，最高精度である MultiGranCNN に比べてわずか 2 ポイント低いだけである．METEOR[Denkowski 14] は（翻訳品質の）評価対象の訳文と参照訳とのアライメントを取り，対応の取れた単語を「正解出力」，評価対象訳文の単語を「出力」，参照訳の単語を「正解」とみなして F 値[*3]を計算し評価結果とする．但し，アライメントの際にシソーラス辞書（wordnet) を参照したり，2 言語対応コーパスから学習したパラフレーズ関係を用いたりすることで，柔軟な照合を行っているほか，単語が内容語か付属語かで重みを変えるなどの調整を行っている．

本研究では，下線部を含む元の文を「参照訳」，この文の下線部を選択肢に置き換えた文を「翻訳結果」とみなして METEOR によるスコアを計算し，選択肢のうちで最もこのスコアが高いものを選ぶ．

3.3.3 含意関係認識

二つのテキストが同義であるとは，これらの間で双方向の含意関係が成り立つことであるから，含意関係認識を適用することが考えられる．含意認識を行うソフトウエアはいくつか提案されているが，本研究では含意関係認識として Tian らの TIFMO[Tian 14] を用いることとする．

なお，含意関係の判定対象は文章全体ではなく，下線部を含む文に限定する．これは下線部を含む文以外は全く同じであるからである．さらに，事前検討で双方向の含意関係が成り立つことは条件として厳しすぎることが分かったため，いずれか一方向に固定して実験する．TIFMO の出力は含意関係の有無 (0/1) はなく，含意関係の強さを表す [0,1] の実数値である．そこで，この値の最も大きい文の組に対応する選択肢をシステムの出力とする

4. 実験

4.1 実験の概要

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問題，合計 55 回分 109 問[*4]に対して適用し，正答率を評価した．問題は全て国立情報学研究所で XML 化されたものを使用した．

上述の手法のうち，3.1 節以外の手法を適用するためには，英語の問題文（文章）から下線部を含む文を切り出す必要がある．下線部は XML タグによって同定し，文境界はピリオド等の記号を手掛かりに同定した．

word2vec におけるベクトルデータは大規模なニュース記事から作成されたものを利用した．また idf は New York TImes の 3 年分 (2008-2010 年) の記事[*5]から算出した．METEOR は公開されているツール[*6]をそのまま利用し，パラメータはデフォールト値を使用した．

4.2 実験結果

実験結果を表 1 に示す．この表で正解率は全ての問題に対する正解率，正解率（置換可能）は後述する．また，word2vec の+/-DIC, +/-IDF はそれぞれ選択肢に対する辞書引きの有無，連語に対するベクトルを構成する際に idf 重みを掛けるか否かを表す．また TIFMO の (仮説:下線）とは含意認識の仮説を下線部にした場合，(仮説:選択肢）は選択肢にした場合である．

この表を見ると，下線部と選択肢の類似性を word2vec を使って評価する手法が他の手法を圧倒している．

一方，本問題に対する直接的な処理方法と思われるパラフレーズ認識は今回利用した手法に関する限り，ランダム解答の精度である 0.25 以下となった．パラフレーズ認識の別手法と考えられる METEOR はランダム解答より 7 ポイント，含意認識についてはいずれの場合も 13 ポイント上回っており，一定の効果があることが分かる．また，空所補充問題や語句整除

∗2 実際には最も若い番号を選ぶ

∗3 適合率と再現率の調和平均

∗4 基本的には各試験ごとに 2 問出題されているが，2015 年のセンター追試験のみ 1 問しか出題されていないため 109 問となった

∗5 https://catalog.ldc.upenn.edu/LDC2011T07

∗6 http://www.cs.cmu.edu/ alavie/METEOR/download/meteor-1.5.tar.gz

表 1: 実験結果

手法	正解率 (全問題)	正解率 (置換可能のみ)
word2vec [東中 15](-DIC, -IDF)	0.65	
word2vec +DIC -IDF	0.55	
word2vec +DIC +IDF	0.69	
NGRAM	0.27	0.26
パラフレーズ認識 [Socher 11]	0.18	0.20
METEOR[Denkowski 14]	0.32	0.31
TIFMO[Tian 14](仮説:下線)	0.38	0.28
TIFMO(仮説:選択肢)	0.38	0.41

問題にも有効であった ngram 言語モデルの利用も本問題についてはランダム選択程度の正解率であった.

4.3 考察

まず，文脈を全く考慮せず，下線部と選択肢の意味的類似性を word2vec により評価する手法については，辞書あり，idf ありによって，既存手法を改善することができた．この理由として，英語辞書の語義記述という説明的な情報を導入してバリエーションを増やす一方で，これに伴って混入するノイズを idf の重みで抑えることが有効であったものと考えられる．

逆に，下線部の意味を無視して，文脈における当該表現の妥当性のみを評価する方法は，ngram という局所的な情報だけでは難しいことが分かった．この問題は意味的な妥当性を問うものであるから，語義の曖昧性解消で用いるような文章全体から得られる素性など大局的な情報を用いる必要があると思われる．また，長い文脈がモデル化できる LSTM などの利用も考えられる．

また，パラフレーズ認識手法については．選択肢ごとの文の大きな意味の違いすら適切に把握できているとは言えず，更なる検討が必要である．

文脈を考慮する方法はいずれも下線部を選択肢と置換してその妥当性を評価するため，下線部と選択肢が置換できない問題に対しては精度が低い可能性がある．そこで，置換可能な 91 問のみを対象として正解率を算出したものを表 1 の「正解率 (置換可能)」に示す．表から分かる通り，正解率はほぼ同じであった．理由の一つとして，下線部と選択肢とが置換不可能であっても，選択肢同士で条件は同じであるから相対的なスコアにはさして影響がなかったことが考えられる．

5. まとめ

大学入試センター試験英語において文章中の下線部と意味的に類似した表現を選択肢から選ぶ「未知語語義推定問題」の解法について検討した．本研究では，下線部と選択肢との意味的類似性を用いる方法，下線部の文脈から下線部の位置に出現する語句の尤度を評価する方法，そして，下線部を含む文について下線部を選択肢に置き換えたものが同義かどうか判定するパラフレーズ認識を用いる方法を試した．最初のもの以外は既に提案されている方法である．その結果，下線部と選択肢の意味的類似性を word2vec によって評価する方法が他より突出して優れており，辞書の語義定義文を加えたうえで idf による重みを導入することによりさらに精度が向上して，正解率 69%となることが分かった．

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, 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）．熱心な議論に感謝いたします．

参考文献

[Blacoe 12] Blacoe, W. and Lapata, M.: A comparison of vector-based representations for semantic composition, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 546–556Association for Computational Linguistics (2012)

[Denkowski 14] Denkowski, M. and Lavie, A.: Meteor universal: Language specific translation evaluation for any target language, in *In Proceedings of the Ninth Workshop on Statistical Machine Translation*Citeseer (2014)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318Association for Computational Linguistics (2002)

[Pershima 15] Pershima, M., He, Y., and Grishman, R.: Idiom Paraphrases: Seventh Heaven vs Cloud Nine, in *Proceedings of the EMNLP Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 76–82Association for Computational Linguistics (2015)

[Qiu 06] Qiu, L., Kan, M.-Y., and Chua, T.-S.: Paraphrase recognition via dissimilarity significance classification, in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 18–26Association for Computational Linguistics (2006)

[Socher 11] Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D.: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, in *Advances in Neural Information Processing Systems 24* (2011)

[Tian 14] Tian, R., Miyao, Y., and Matsuzaki, T.: Logical Inference on Dependency-based Compositional Semantics., in *ACL (1)*, pp. 79–89 (2014)

[Yin 15] Yin, W. and Schütze, H.: MultiGranCNN: An Architecture for General Matching of Text Chunks on Multiple Levels of Granularity, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 63–73 (2015)

[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 (¡ 特集¿ 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)

[東中 15] 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 (NLP2015) (2015)