古語辞典の語釈を用いたセンター試験古文の内容理解問題解答

横野 光
国立情報学研究所
yokono@nii.ac.jp

1 はじめに

現在，国立情報学研究所では人工頭脳プロジェクト「ロボットは東大に入れるか」を進めている．このプロジェクトでは大学入学試験を解答するシステムの構築を通して，人工知能研究の到達点や未解決の問題を明らかにすることが目的の一つとなっている [1]．試験問題には様々な科目がありそれぞれに応じたアプローチが必要となるため，各科目毎に研究を進めており，我々は大学入試センター試験の国語の古文問題に焦点を当てその解答に取り組んでいる．

センター試験の国語は 4 つの大問で構成されており，第 1 問が現代文の評論問題，第 2 問が現代文の小説問題，第 3 問が古文，第 4 問が漢文となっている．古文問題では，古典作品の一部が問題本文として提示され，古典文法や本文の内容についての問題が出題される．センター試験では受験者に公平であることが重要視されることから，問題として採用される古典作品はあまり有名ではないものが選ばれることが多い．

古文問題で採用される作品は既存のものであり，古典文法で記述されたテキストが今後新しく生成されることはない．そのため，原理的には，既存の古典作品を全て電子化し，現代語訳や品詞情報をアノテーションすることでそれをそのまま問題解答に利用することが可能である．しかし，そのような網羅的なコーパス，アノテーションデータは存在しておらず，また，構築には非常に膨大なコストがかかるため現実的な解決策とは言えない．

利用可能なコーパスが全く存在しないというわけではなく，例えば国立国語研究所が開発を進めている通時コーパス [2] には平安時代の和歌，物語作品，鎌倉時代の軍記物語などが含められており，このような言語資源を活用した解答手法が考えられる．我々はこれまでに現代語訳付きの古典文学作品コーパスを用いて，統計的機械翻訳モデルによって問題本文を現代語に翻訳したテキストを用いて内容理解問題を解答する手法を提案してきた [3, 4]．

センター試験古文問題の一般的な対策としては，まず，助動詞の用法や敬意表現などの文法を覚え，古文単語の中でも重要とされる語句の意味を覚えるということが挙げられることが多い．このことから，受験者にとっては問題本文を完全に理解することが問題解答に必須であるというわけではないと考えられる．

そこで，本論文ではこれまで行ってきた統計的機械翻訳によって本文を全て現代文に翻訳したものを用いる手法ではなく，古語辞典の語釈を利用して本文中の古文単語の一部を現代語に置き換え，それを用いて内容理解問題を解答する手法について検討する．

2 センター試験古文における内容理解問題

センター試験の古文問題は，年度によって細かな違いはあるが，基本的には 1 問目に傍線部表現の現代語訳問題が数題配置され，以降，文法問題が 1 問程度，内容理解問題が 4，5 問程度出題される．また，毎年というわけではないが，文学史に関する問題などが出題されることがある．本節ではこのうち提案手法が対象としている内容理解問題について説明する．

内容理解問題の例を図 1 に示す．センター試験の過去問は大学入試センターの Web サイト[1] や予備校のサイトなどで公開されており，問題本文などは底から参照できる．

古文問題では物語作品が対象とされることが多く，その場合，問題で問われることは登場人物の心情や，行動の理由などである．これは現代文の小説問題と同様のものである．現代文の小説問題では文章として記述されている表層的な情報から例えば常識的な推論などを用いて，その背後に隠れている要素を推定する必要がある．しかし，古文問題では受験者の古文の理解能力を評価することが主な目的とされているため，表層的な意味，つまり現代語での意味が理解できれば解けるような問題と

[1] http://www.dnc.ac.jp/data/kakomondai.html

問３ 傍線部Ａ「また旅の空におぼしめしたちけり」とあるが、中将はなぜそのように思い立ったのか。その説明として最も適当なものを、次の①～⑤のうちから一つ選べ。
解答番号は 23。

① 都を離れている間に亡くなった母が、弔い事をしてもらえないことをうらめしげに訴える夢を見た中将は、悲しさが募り、せめて七回忌の法事だけでも出席したいと考えたから。
② 別れを告げずに都に残してきた母が、うらめしげな様子で自らの死を伝える夢を見た中将は、恋しさが募り、母の言葉どおりであればせめて弔い事だけでもしたいと考えたから。
③ 別れを告げずに都に残してきた母が、出家して西国へ旅立とうとする夢を見た中将は、驚きあわてて、引き留められないまでもせめて後を追わなければならないと考えたから。

(以下省略)

図 1: 内容理解問題の例 (2005 年度センター試験国語 I・II 本試験より引用)

なっている．この点で問題解答に必要な常識的な知識は現代語の問題に比べて少ないと言える．

例えば，図 1 の問題の正解は②であるが，この選択肢中の “弔い事だけでもしたい” という表現が本文中の “せめてものなぐさみに御跡なりとも……” と対応している．選択肢の内容が問題本文と同じである場合，選択肢の表現に対応する箇所が表層的に問題本文中に存在していることが多い．このことから選択肢と本文との類似性を判定することで問題に解答することができると考えられる．だが，当然のことながら問題本文は古語で記述され，選択肢は現代語で記述されているため，一般的には本文を現代語で解釈して問題を解答することになる．

3 提案手法

2 節で述べたように，古文における内容理解問題では解答の手がかりが古語と古典文法による表現として表層に表れているため，それらを現代文に翻訳することができれば，問題に解答できる．また，一般的な勉強法として重要単語の暗記が挙げられており，それらの語句が正しく理解できるかどうかが誤った選択肢の排除に用いられることから，問題解答には網羅的に翻訳できる必要はなく，鍵となる単語が正しく翻訳できれば良いと考えられる．

そこで本論文では，古語辞典を用いて問題本文中の単語を現代語の語釈に置き換え，それと選択肢とを比較し最も類似している選択肢を解として出力する手法を提案する．内容の類似度を測る方法にはコサイン類似度など様々なものがあるが，本手法では単純に本文と選択肢の間で共通している語の個数によって類似度を定義する．処理の流れを図 2 に示す．

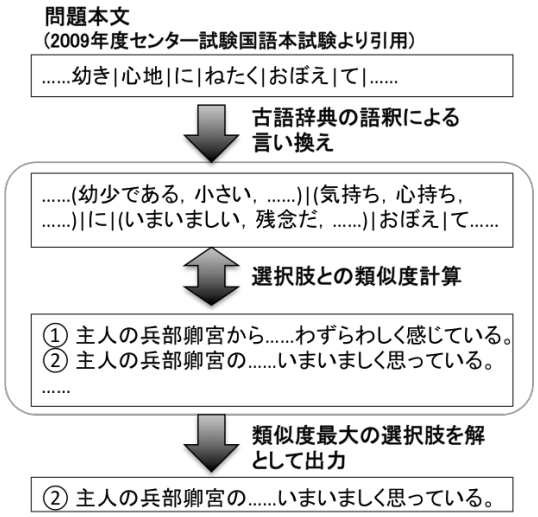

図 2: 提案手法の処理

問題本文を中古和文 UniDic[5] を用いた形態素解析器 MeCab[2]によって形態素に分割し，品詞が名詞，動詞，形容詞，副詞であるもの (以下，これらに該当する語を内容語と呼ぶ) に対して，古語辞典を用いて語釈を検索し，得られた語釈を本文として扱う．古語辞典の語釈は単語である場合や文である場合があるため，実際には語釈に対して UniDic を用いた MeCab で形態素解析を行い，内容語のみを利用する．この処理によって，本文は文章中の古文単語の語釈に含まれる内容語の集合として表現される．選択肢についても同様に UniDic による形態素解析を行い，内容語のみを抽出する．

上記の処理によって得られた語釈の内容語による問題本文の単語集合を S，i 番目の選択肢の内容語集合を c_i とし，S と c_i の間の類似度 $Similarity(S, c_i)$ を以下のように定義する．

$$Similarity(S, c_i) = \frac{w * ol_{VA}(S, c_i) + ol_O(S, c_i)}{|c_i|}$$

$|c_i|$ は集合 c_i の要素数を表す．また，$ol_{VA}(S_1, S_2)$ は単語集合 S_1，S_2 の両方に出現している動詞と形容詞の単語数，$ol_O(S_1, S_2)$ は両方に出現している

[2]https://code.google.com/p/mecab/

動詞，形容詞以外の単語数を表し，w は重みを意味する．

参考書などで重要語とされる語には動詞や形容詞が多いことから，これら用言が正しく解釈できることが問題解答に特に重要だと考えられる．そこで，類似度の計算に関して，本文と選択肢との比較の際に動詞と形容詞の一致に対して重みを与え，その他の語の一致に比べて重要であるとした．重み w の具体的な値は開発データを用いて決定する．

この式によって各選択肢の類似度を計算し，その値が最も高いものを解答とする．問題には “適切でないものを選べ” といった，本文と内容が合致しない選択肢を選ぶものが存在するが，そのような問題に対しては類似度が最小のものを解として出力する．

傍線部の表現に関する問題の場合，解答の手がかりはその周囲に存在することが考えられる．そのため，従来手法 [4] では選択肢と比較する問題文を傍線部を中心に前 l 文，後ろ m 文に限定している．本手法でも同様に，類似度判定に利用する本文をこれら 2 つのパラメータによって決定する．なお，傍線部が対象ではない問題に関しては問題本文全体と比較する．

古語辞典による古語の語釈抽出において，単語によっては複数の語釈が存在するため，本来は文脈にあわせた語義を選択する必要があるが，本論文では語義曖昧性解消は行わず，全ての語義の語釈を利用している．また，辞書の検索は単純に形態素単位での完全一致で行う．

4 評価

提案手法の性能を評価するため，センター試験の過去問 15 回分を対象に問題の正答率で従来手法 [4] との比較を行った．

2 節で述べたように，古文問題には文法問題なども含まれるが，本手法は内容理解問題のみを解答の対象としている．そのため，評価には内容理解問題だけを人手で抽出したものを用いた．傍線文現代語訳問題も内容理解問題と見なし，評価に用いている．

提案手法で利用した古語辞典は旺文社全訳古語辞典第四版であり，今回使用したデータの見出し語の数は約 22000 語である．語釈には反意語や関連する文法事項などが記述されている場合があるが，それらは削除している．

提案手法，従来手法ともにあらかじめ設定すべきパラメータが存在する．このパラメータの値の決定には 2005 年度と 2009 年度のセンター試験の過去問で同様の実験を行い，最も正解率が高かった設定での値 (提案手法に関しては，$w = 3$，本文として問題本文冒頭から傍線部の 2 文後までを利用，従来手法に関しては，単語 3-gram で類似度を計算，傍線部の 1 文前から 1 文後までを本文として利用) を採用した．提案手法の類似度計算の重み w を 1 よりも大きくした方が性能が良かったことから，用言の意味を正しく理解することが正解を選択するために重要であることが分かる．

評価結果を表 1 に示す．問題名の “(本)“は本試験，“(追)“は追試験を表す．

表 1: 評価結果

問題名	問題数	正答数 従来手法	正答数 提案手法
1991 年度国語 (本)	6	2	1
1991 年度国語 (追)	6	0	1
1995 年度国語 (本)	7	1	0
1995 年度国語 (追)	6	1	0
1999 年度国語 I(本)	7	2	3
1999 年度国語 I・II(本)	6	2	3
1999 年度国語 I(追)	7	3	0
1999 年度国語 I・II(追)	7	2	3
2003 年度国語 I(本)	7	2	3
2003 年度国語 I・II(本)	7	3	2
2003 年度国語 I(追)	6	0	3
2003 年度国語 I・II(追)	6	0	0
2007 年度国語 (本)	7	1	1
2007 年度国語 (追)	7	4	3
2011 年度国語 (本)	6	1	2
正答率		0.238	**0.254**

全ての問題において提案手法が従来手法を上回っているというわけではないが，平均的には従来手法よりも良い性能であることが明らかになった．図 2 で例示しているように，ある古語の語釈に対して複数の現代語による表現が対応していることが多く，これによってある程度の類義関係が類似度計算の際に考慮されていると考えられる．

しかし，従来手法では統計的機械翻訳によって本文を現代語訳したものを選択肢との比較に用いているため，翻訳の精度が正答率の精度に影響を与える．そこで開発データに対してではあるが，人手による現代語訳を用意しそれを翻訳結果として利用した従来手法と提案手法との比較を行った．結果を表 2 に示す．

この結果から，従来手法において古文-現代文翻訳モデルの性能を向上させることができれば，提案手法よりも良い結果が得られることが分かる．翻訳モデルの性能を向上させるための基本的な方策としては学習に利用する対訳データの拡充があげられるが，古文のテキストが新しく生成されること

表 2: 人手の現代語訳を用いた従来手法との比較

問題名	問題数	正答数 従来手法	提案手法
2005 年度国語 I(本)	6	2	2
2005 年度国語 I・II(本)	7	4	2
2005 年度国語 I(追)	7	2	3
2005 年度国語 I・II(追)	7	4	4
2009 年度国語 (本)	3	1	1
2009 年度国語 (追)	5	2	1
正答率		**0.429**	0.371

はないため利用できるデータに限りがあるという状況から学習データの規模の拡大には限界があり，翻訳性能の向上は一般的な他言語翻訳に比べて困難であると考えられる．

提案手法は機械翻訳モデルを用いていないためこの問題は回避できるが，一方で，辞書を用いているため，その収録語彙数が性能に影響する．実験で利用した辞書に収録されている見出し語数は約 22000 であるが，これに対して，例えば，従来手法の学習で用いた古典文学作品コーパスに含まれている 40 作品の語彙の異なり数は約 28000 であった．このコーパスに収録されている作品は源氏物語や方丈記など比較的有名なものであり，この他にも古典作品が存在することから，辞書には全ての古語が掲載されているというわけではない．しかし，本実験で利用した辞書は高等学校における古典学習での利用を想定し，大学入試に必要な語彙を中心に収録されている．このことから，本研究で取り組んでいる大学入試問題解答というタスクにおいてはこの辞書は必要量の知識を含む言語資源とみることができる．

5 おわりに

本論文では，センター試験古文の内容理解問題に対して，古語辞典の語釈を利用した手法を提案し，統計的機械翻訳による古語-現代語翻訳を利用した従来手法よりも良い性能であることを実験により示した．

実験では単純な形態素の完全一致によって古語辞典からの検索を行い語釈を抽出した．しかし，前述のように複数の語義を持つ語も存在するため，語義曖昧性解消などによって文脈に応じた適切な語釈を選択する必要がある．また，古語辞典の検索においても，見出し語と中古和文 Unidic の単語単位のずれによって正しい語が検索できないといったことが未解決の問題として挙げられる．今後はこれらに対応する予定である．

謝辞

本研究は国立情報学研究所人工頭脳プロジェクト「ロボットは東大に入れるか」によるものである．また，センター試験過去問データは独立行政法人大学入試センター，株式会社ジェイシー教育研究所から，辞書データは旺文社からそれぞれ提供を受けた．

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? -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, pp. 463–469 (2012).

[2] 近藤泰弘: 日本語通時コーパスの設計について，国語圏プロジェクトレビュー，Vol. 3, No. 2, pp. 84–92 (2012).

[3] 星野　翔，宮尾祐介，大橋駿介，相澤彰子，横野　光: 対照コーパスを用いた古文の現代語機械翻訳，言語処理学会第 20 回年次大会 (2014).

[4] 横野　光，星野　翔: 統計的現代語訳モデルを用いたセンター試験古文問題解答，第 5 回コーパス日本語学ワークショップ (2014).

[5] 小木曽智信，小椋秀樹，田中牧郎，近藤明日子，伝　康晴: 中古和文を対象とした形態素解析辞書の開発，情報処理学会研究報告人文科学とコンピュータ CH-85 (2010).