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

A Similarity-based Model using Statistical Machine Translation for Solving Questions of Japanese Classics in National Center Test for University Admissions

横野 光 (国立情報学研究所)
星野 翔 (総合研究大学院大学)

1　はじめに

現在，国立情報学研究所で推進している人工頭脳プロジェクト「ロボットは東大に入れるか」は，細分化された人工知能分野を再統合することで分野全体としてどこまでできているのか，また残っている問題は何かということを明らかにすることを目的とし，そのベンチマークとして 2016 年度に大学入試センター試験を，2021 年度に東京大学入学試験を突破することを目標としている [新井，松崎 (2012)]．

センター試験は一般的に，国語，英語，数学，理科，社会の 5 科目からなっており，それぞれに対して解答に必要となる知識，手法は異なる．そのため，問題解答にはそれぞれの科目において検討する必要がある．我々は特に国語の古文問題を対象として，その解答に取り組んでいる．

本論文では，センター試験国語の古文問題ではどのような問題が出題されるのかについて分析を行い，このうち内容理解に関する問題に対して，統計的機械翻訳モデルを用いた解答手法について述べる．

2　センター試験古文問題の分析

センター試験国語古文問題について，本試験と追試験の合わせて過去 16 回分のセンター試験で出題されているか問題の分析を行った．過去に行われたセンター試験の問題は大学入試センター[1]で公開されており，自由に閲覧できる．

問われている内容に基づいた設問の分類の結果を表 1 に示す．出題数は分析に利用した 16 回分の試験問題での小問の数を表す．設問は大きく分けると，文法理解に関する問題，内容理解に関する問題，その他に分類できる．1 回分の試験での古文問題の大まかな構成としては，第 1 問目に本文傍線部の現代語訳，文法問題が 1 問程度，和歌に関する問題や文学史に関する問題は出題されない回もあるがそれぞれ 1 問程度出題される．残りは内容理解に関する問題であり，半分以上の割合を占める．そのため，内容理解の問題が解けるかどうかが高得点を得るためには重要となる．

以降，それぞれの分類について説明する．

2.1　文法問題

文法問題では主に単語の活用や助動詞の用法に関する問題，敬意表現に関する問題が出題される．単語の活用などに関する問題では，傍線部の表現がどのような形態素で構成されているか，それぞれの品詞や活用形，助動詞の場合にはその用法などが問われる．問題の形式としては傍線部の構成要素を問うものの他に，本文中の表層が同じ語のなかで，品詞が違うものを答えるという形式のものもある．

[1] http://www.dnc.ac.jp/modules/center_exam/

表 1: 問題の分類

大分類	小分類	出題数
文法理解	表現の構成要素 (品詞など) の同定	12
	主語の同定	1
	会話文の発話者推定	2
	敬意表現 (敬語の種類など)	4
内容理解	現代語訳	16
	内容理解	29
	心情説明	8
	理由説明	5
	和歌の表現技法	1
	和歌の解釈	5
	文章全体の内容理解	10
その他	古典知識	4

敬意表現に関する問題では，傍線部の表現が尊敬語であるか，謙譲語であるか，丁寧語であるか，またどの登場人物からどの登場人物への敬意であるかなどが問われる．

この他にも文中の述語における主語や発話文の話し手が明示的に現れていない文に対して，それらを同定するという問題もある．

単語の活用などに関する問題は，例えば中古和文 Unidic[小木曽ら (2010)] を用いた形態素解析器 MeCab[2]によって形態素解析を行い，その形態素情報を利用することで解答が可能であると考えられる．しかし，助動詞の用法には多義性があるため，より精度を上げるためには複数の用法を持つ助動詞に対して，その表現中で用いられている用法が何かを推定するモデルが必要となる．

敬意表現に関しては，その表現がどのタイプの敬語であるかを推定する必要がある．解答方法としてはそのような情報を形態素辞書に記述しておき，それを利用して解答するという手法が考えられるが，全ての語が一つの種類の敬意を持つわけではなく，複数の用法を持つ語もあり，そのような語に対しては文脈からタイプを推定しなければならない．また，例えば，天皇は身分関係の上位に必ず来るといったような，人物間の身分に関しての知識少なからず必要となる．

2.2 内容理解

古文の問題の大半は本文の内容理解に関する問題である．実際に問われるものには，単純な現代語訳から，本文中の記述についての理由や説明，そこから推測される登場人物の心情など様々なものがある．このような問題を解くためには，選択肢に書かれてある記述が本文と照らし合わせて正しいかどうかを推定する必要がある．

内容理解に関する問題には，登場人物の心情推定など現代文の小説問題でも出題されるようなものがある．しかし，必要となる知識は古文と現代文では異なると考えられる．例えば，小説の心情推定の問題では，その心情を表す直接的な表現が本文中に出現することはあまりなく，“父は縁側で背中を丸めて足の爪を切っていた。” といった人物の行動に関する表現から，その心情 (この場合は，父親が落ち込んでいる，あるいは悲しんでいる) を推測するというようなことが必要となる．このように本文中に直接的に書かれていないことを使って解かなければならないような問題が小説の問題では多い．

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

一方で，古文の問題における心情推定の問題では，手掛かりとなる表現が小説等に比べて直接現れていることが多い．例えば，図 1 の問題では，①が正解であるが，選択肢中の “兵衛佐の妹なら美

問 5 この文章を通して、兵衛佐の妹に対する兵部卿宮の気持ちはどのように移り変わっているか。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 27。

① 最初はお気に入りの兵衛佐の妹なら美しいだろうから会ってみたいと執心していたのに、何度も贈った和歌をはねつけられて落胆した。しかし、常磐に兵衛佐の邸に忍び込むように進言されて再び希望を見出した。
② 最初はほんの遊び心で臣下の兵衛佐の妹に会ってみようと思ったが、和歌のやりとりをするうちに本気で兵衛佐の妹を想うようになってきた。そしてついに、兵衛佐が不在の夜に勇んで兵衛佐の邸に忍び込むことにした。
(以下省略)

図 1: 内容理解の問題例 (2009 年度国語本試験より引用)

しいだろうから会ってみたいと執心していた” という箇所は，本文中の “女にてかれが妹ならば、いかにいつくしからん。あはれ、見ばや」と、深く御心移りて” に合致し，“落胆した” という箇所は本文中の “宮の思し召し沈み給へる” に合致している．これは論説文や小説のような現代文の問題では，本文で書かれてあることは受験者は大まかなところは理解できるとしたうえで，設問ではより深い理解が必要となる問題が出題されるのに対し，古文の問題では「古文が正しく理解できるか」が焦点となり，現代文として解釈できるかどうかに重点が置かれているからだと考えられる．従って，本文において表層に現れていることが理解できれば，古文の問題の解答は可能であると考えられる．

また，本文中に和歌が出現している場合，その和歌で使用されている表現技法についての説明を求める問題やその和歌の意味を問う問題が出題されることがある．これらの問題では縁語や枕詞といった和歌独自の修辞技法を理解する必要がある．また，古文としての表層的な解釈だけでなく，それによってその歌の詠み手が何を伝えようとしているのかといった，深い解釈が必要となる．

2.3　その他 (文学史)

文法知識，内容理解の問題に比べると数は少ないが，文学史の知識に関する問題が出題されることもある．文学史の問題では本文の内容ではなく，本文に用いられている文献の種類やその成立時期に関係した設問が出されることがある．これらの問題は本文の内容とは直接関係はなく，あらかじめ文学史に関する知識を保有しておき，その知識から選択肢の内容が含意できるかどうかを判定することで解答する．これは世界史や日本史の問題で見られるタイプの問題であり，自然言語処理における含意関係認識のタスク [田，宮尾 (2013)] として解くことができると考えられる．

3　統計的現代語訳モデルを用いた問題解答

2.2 節で述べたように古文の内容理解の問題では古文本文を現代文として解釈できるかどうかが重要であると考えられる．そこで本研究では，古文を機械翻訳モデルによって現代語へと翻訳し，それを用いて内容理解に関する問題の解答を考える[3]．

[3]いわゆる “古典” は新しいものは生成されないため，全ての古典に対してその現代語訳を付与したコーパスを構築するという方法が考えられるが現実的ではない

3.1 古文-現代語翻訳モデル

統計的な機械翻訳では翻訳元の言語のテキストと翻訳先の言語のテキストの対応が取れたコーパスが必要となる．現時点において我々が利用可能な古文-現代文コーパスとしては小学館の新編古典文学全集の電子化データ (以降，小学館コーパスと呼ぶ) がある．このコーパスは古代から近世までの古典文学作品に対して，その本文と現代語訳，注釈が対応づけられたデータである．

このコーパスの古文と現代語訳との対応は文よりも大きなセグメント単位でしかなされておらず，統計的機械翻訳に望ましい文対応付けされたデータではない．そこで対応するセグメント内の文の数に注目し，古文と現代文で文の数が等しければ，前から順に文の対応付けを行い，文の数が等しくなければ，文の数が多いセグメントを少ない方のセグメントの数に合わせて分割することで，文対応の取れた対訳コーパスを作成する [星野ら (2014)]．

学習データ作成に使用した小学館コーパスの統計を表 2 に示す．このコーパスから得られた対訳コーパスの規模は 86684 文対であった．このデータを学習コーパスとして Moses[Koehn et. al. (2007)] を用いて翻訳モデルを構築する．

表 2: 小学館コーパス統計情報

	古文	現代語	合計
単語数	2,837,101	3,720,257	6,557,358
文字数	12,763,402	17,300,081	30,063,483
セグメント数		19,102	

3.2 類似度に基づいた問題解答

内容理解に関する問題では基本的に本文の内容にあった選択肢が正答であると考えられる．そこで 3.1 節で述べた翻訳モデルによって現代語訳した本文と選択肢との類似度を計算し，その値をもとに解を決定する．

本文と選択肢の類似度はコサイン類似度で計算する．コサイン類似度は本文の特徴ベクトル T と選択肢の特徴ベクトル c_i に対して，以下の式で求める．

$$sim(T, c_i) = \frac{T \cdot c_i}{|T||c_i|}$$

本文と選択肢の特徴ベクトルは形態素の n-gram を要素とする．

内容理解に関する問題の多くは，本文中で指示されている傍線部についてのものである．このような問題では，その傍線部の近くに解答の手がかりが存在することが多い．そこで，類似度の計算において問題本文全てを利用するのではなく，傍線部を含む文の前 l 文から後ろ m 文のみを利用する．

提案モデルでは，基本的には類似度が高くなる選択肢を解として出力するが，問題には “適切でないものを選べ” と指示されたものもある．このような問題に対しては類似度が最も低いものを解として出力する．

4 実験

2009 年度，2005 年度のセンター試験国語問題から傍線部の現代語訳問題と内容理解に関する問題を人手で選別し，これらを用いて提案手法の評価を行った．提案手法では本文として傍線部を含む文の前後何文を利用するか，また，n-gram の n の値をどうするかを与える必要がある．これらに対しては別年度の問題を開発データとして決定した．

表 3: 実験結果

問題	正答率
2009 年度国語本試験	0.4(2/5)
2009 年度国語追試験	0.4(2/5)
2005 年度国語 1 本試験	0.33(2/6)
2005 年度国語 1 追試験	0.17(1/6)
2005 年度国語 1・2 本試験	0.2(2/5)
2005 年度国語 1・2 追試験	0.43(3/7)
合計	0.32(11/34)

実験結果を表 3 に示す．括弧の中は正答数と実験の対象となった設問数を表している．実験で使用した問題数が多くないため，この結果から提案手法の有効性を主張することはできないが，1 設問につき選択肢の数は 5 個程度であり，ランダムに答えた場合の正答率が約 0.2 であることを踏まえると，全く見込みがないとは言えないであろう．しかし，試験毎に正答率にばらつきがあり，安定した性能ではないというのは今後の検討課題である．

翻訳モデルの性能に関して，実験で使用した問題の本文とそれに対する翻訳結果を図 2 に示す．

問題文　ある時、中将、昼寝せさせ給ひける御夢に、いづちともなく荻薄生ひ茂りたる野原の、まことに心すごき所に、うす絹のすそ、露にうちしほれたる女房ただひとり立ち給へり。いたはしと思ひて立ち寄り見給へば、わが母にておはせり。中将を見たてまつりて、袖もしぼりあへず、仰せけるは、「都に捨ておき給ひしその嘆きに、月日の行くもおぼえはべらねども、はや六年になりぬ。この思ひゆゑ、われこの世になき身となりにき」とて、さもうらめしげなる気色にて、道もそこはかとなき野中を西へ向きて行き給ふ、とおぼして、夢うちさめぬ。

参照訳　あるとき、中将が昼寝をなさっていたその御夢に、どことはわからない萩やすすきがおい繁っている野原で、本当に人けがなくものさびしい所に、薄衣のすそが露に濡れている女性がたった一人で立っていらっしゃる。中将がお気の毒にと思って近寄ってご覧になると、自分の母上でいらっしゃった。母は中将を見申し上げると、絞りきれないほど袖を涙でぬらして、「あなたが私を都に見捨てたまま消息を絶ってしまわれたその嘆きのために、月日のたつのも忘れておりますけれども、もはや六年になりました。この悲しい思いのために、私はこの世の者ではなくなってしまいました」とおっしゃって、いかにもうらめしい様子で、道もはっきりと分からない野原の中を西方浄土の方へ向かっていらっしゃる、と思われたところで夢からさめた。

翻訳モデルによる翻訳結果　ある時、中将は、ちょうどお昼寝をなさった時の夢に、どこへともなく荻薄の生い茂っている野原のは、本当にうつ所に、うす絹の裾、露に内でうちしおれて、ただ一人で立っていられた。不憫でならないと思って、お立ち寄りご覧になると、ご自分の母でいらっしゃいました。中将のを拝見して、袖も涙でぬらし終らないうちに、おっしゃったことには、「都に捨ておきれたその嘆きに、月日の行方も思われませんが、早く六年になった。この思いから、私はこの世にいない身となってしまった」といって、さも恨めしそうな様子で、道もどこがどうということもない野中を西の方へ向って行った、とお思いになって、夢がさめた。

図 2: 翻訳例 (2005 年度センター試験国語 1・2 より引用)

敬意表現や助動詞に関してはある程度は翻訳できているが，一部の内容語に関しては翻訳できておらず，翻訳元の表現がそのまま用いられている．これは学習に使用したデータが少ないことが原因の一つであると考えられる．このような場合，一般的には対訳データを増やすことによって改善できるが，我々が取り組んでいる古文-現代文翻訳においては，利用可能な電子化された古文データはそれほど多くないため，対訳データを作成するためのコストが大きくなる．このことから，対訳辞書を

構築し，それを用いることによる翻訳精度の向上を図る．

5　おわりに

本論文では，センター試験国語古文問題の分類を行い，そのうち内容理解に関する問題に対して，統計的翻訳モデルを用いて古文を現代文に翻訳し，それと選択肢との類似度に基づいて解答するモデルについて述べた．

2 節で述べたように，古文問題には内容理解に関する問題の他に，文法に関する問題や和歌に関する問題などがある．文法に関する問題に関しては，中古和文 Unidic を用いた形態素解析を行うことで表現の品詞情報を得ることができるため，これを用いて問題に解答するという手法が考えられる．

和歌に関する問題に関しては，本論文で述べた手法と同様に現代語訳をすることで和歌の理解に関する問題に解答することは可能であると考えられるが，和歌には独特の表現技法や訳し方があるため，これをどのようにモデルに組み入れるかが重要であると考えられる．

現状では古文を対象とした言語処理ツールや言語資源は多くなく，また言語資源の規模を拡大していくというのは現代語のテキストと比べると困難である．この，小規模な言語資源しかなく，規模の拡大にコストがかかる，という状況は古文に限らず存在する．そのため，本研究では古文問題の解答モデルの構築という応用を通じて，小規模の言語資源を如何に効率的に活用するかということにも焦点を当て，その方法論の確立に取り組む予定である．

謝辞

本研究では，国語研究所通時コーパスプロジェクトから小学館新編日本文学全集データの提供を受けた．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提供を受けた．

参考文献

[星野ら (2014)] 星野 翔, 宮尾 祐介, 大橋 駿介, 相澤 彰子, 横野 光 (2014),「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 (to appear).

[Koehn et. al. (2007)] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst (2007)「*Moses: Open source toolkit for statistical machine translation*」In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp.177-180.

[新井，松崎 (2012)] 新井 紀子, 松崎拓也 (2012),「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.

[小木曽ら (2010)] 小木曽 智信, 小椋 秀樹, 田中 牧郎, 近藤 明日子, 伝 康晴 (2010),「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.

[田，宮尾 (2013)] 田 然，宮尾 祐介 (2013),「関係代数に基づく推論の含意関係認識への応用」, 2013 年度人工知能学会全国大会 (JSAI2013).