機械翻訳を用いた中古和文の現代語訳ー分析と課題ー

山田 祐実　大村 舞　岡 照晃　Kevin Duh　松本 裕治

(奈良先端科学技術大学院大学)

Translation of Classical Japanese into Contemporary Japanese Using MT: Analysis and Future Work

Yumi Yamada, Mai Omura, Teruaki Oka, Kevin Duh, Yuji Matsumoto

(Nara Institute of Science and Technology)

要旨

国立情報学研究所による人工頭脳プロジェクト「ロボットは東大に入れるか」において，機械翻訳による古語の現代語訳が行われており，翻訳モデルの学習に平安期から江戸期にわたる古語のコーパスが使われている．しかし，時代によって用法の異なる語がある場合，他の時代の文を翻訳する際に適切な訳語が当てられない可能性がある．また，使用した小学館コーパスには他の作品と比べ敬語表現の多い『源氏物語』が約 55% 含まれるという特徴があった．そこで，学習に使用するコーパスを中古和文に絞り，『源氏物語』の文体が言語モデルへ及ぼす影響を下げるため，BCCWJ や青空文庫によるコーパスを加え翻訳を行った．その結果，翻訳性能の向上が見られた．翻訳結果を分析すると，BLEU による評価方法の見直しや訳語の対応関係の改善が今後の課題となることが分かった．

1　はじめに

現在国立情報学研究所では，現時点での人工知能の達成度と課題を測る試みとして，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 [新井ら 2012]．横野らは，国語の古文問題の解答に取り組んでおり [横野ら 2014]，内容理解に関する問いを解くために統計的機械翻訳を用いて古文から現代文への翻訳を行っている [星野ら 2014]．

統計的機械翻訳は，図 1 のように翻訳モデルと言語モデルを用いて行なわれる．星野らは翻訳モデルと言語モデルをつくるのに，本研究と同様に小学館『新編日本古典文学全集』によるコーパス（小学館コーパス）を用いている．しかしながら，星野らが用いたコーパスには平安期から江戸期にかけての幅広い作品が含まれている．このため，同じ表層形でも時代によって意味の異なる語がある場合，ある時代でよく用いられる意味に高い確率が付与されると，他の時代の文を翻訳する時に適切な訳を当てられない可能性がある．また，言語モデルの学習には小学館コーパスのみを使用し，他のコーパスを使用する試みは行っていない．

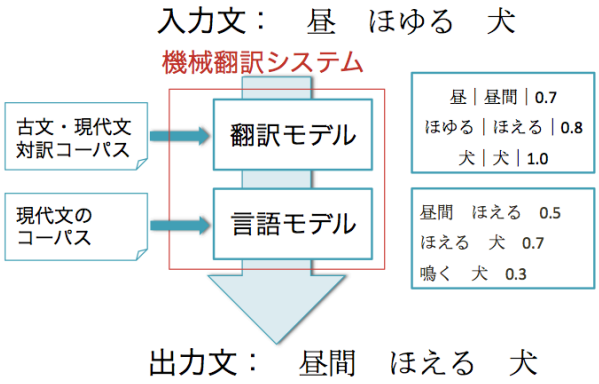

図 1: 統計的機械翻訳の概略

そこで本研究では，中古和文のコーパスを対象に翻訳を行った．さらに，言語モデルの学習用コーパスを複数試し，翻訳結果への影響を調べ，翻訳結果に見られる問題点を分析した．

以下，2 章で統計的機械翻訳について述べる．3 章では今回実験に使用したコーパスについて述べ，4 章では実験設定について説明する．5 章で翻訳の性能と実際の翻訳例を提示し，6 章では 1 文ごとの評価値を調べ，分析を行う．7 章と 8 章で翻訳例を踏まえた今後の課題について述べ，最後に 9 章で本稿のまとめを行う．

2　統計的機械翻訳について

統計的機械翻訳を古文から現代文への翻訳に使う場合，図 1 に示したように，計算機を用いてコーパスから翻訳モデルと言語モデルを生成し，古文の入力に対して適切な現代語の翻訳文を出力するシステムをつくる．翻訳モデルは，単語列間の翻訳関係に確率を付与したものである．この関係と確率が対応づいた表をフレーズテーブルとよぶ．また，言語モデルは，出力文の文としての自然さを確率で評価するものである．出力の際，翻訳候補の文の中から翻訳モデルの確率と言語モデルの確率の積が最も高いものが選ばれる．翻訳モデルは対訳コーパスを用いて作られ，言語モデルは出力言語のコーパスからコーパス内の統計情報をもとに作られる．この過程を一般に翻訳モデルの学習，及び言語モデルの学習とよぶ．

3　使用したコーパス

今回使用したコーパスは，小学館『新編日本古典文学全集』から平安期を中心とした 14 作品，現代日本語書き言葉均衡コーパス（BCCWJ）[Maekawa2008]，青空文庫与謝野晶子訳『源氏物語』[*1] の 3 種類である．

[*1] 青空文庫 与謝野晶子訳『源氏物語』http://www.aozora.gr.jp/index_pages/person52.html

日本霊異記，竹取物語，古今和歌集，土佐日記，伊勢物語，大和物語，落窪物語，平中物語，枕草子，和泉式部日記，源氏物語，紫式部日記，更級日記，讃岐内侍日記，堤中納言日記，蜻蛉日記，大鏡，今昔物語集，将門記，陸奥話記，保元物語，平治物語，方丈記，徒然草，正法眼蔵随聞記，歎異抄，平家物語，宇治拾遺物語，十訓抄，沙石集，曾我物語，近松門左衛門集，洒落本，滑稽本，人情本，俊頼髄脳，古来風躰抄，近代秀歌，詠歌大概，毎月抄，国歌八論，歌意考，新学異見

図 2: 小学館コーパス

表 1: 小学館コーパスの単語数比較

	古文	現代文	計（単語）
星野らの用いた単語数	2,837,101	3,720,257	6,557,358
本研究で用いた単語数	1,071,453	680,464	1,751,917

3.1 新編日本古典文学全集

小学館コーパスに含まれる作品を図 2 に示す．星野らは図中の全ての作品を用いて翻訳モデルと言語モデルを作成したが，平安期から江戸期までの広い時代の言葉が含まれているため，フレーズテーブルから適切な訳語が選ばれにくくなる可能性がある．そこで本研究では，図中の下線で示している中古和文で書かれた 14 作品のみを用いて翻訳モデルと言語モデルを作成した．星野ら及び本実験で使用した小学館コーパスの単語数を表 1 に示す．

本研究で用いた中古和文の 14 作品には「源氏物語」が約 55% を占めているという特徴がある．「源氏物語」の現代語訳は他の 13 作品と比べて敬語表現が多いため，この特徴が言語モデルに影響する可能性が考えられる．「源氏物語」と他の 13 作品の文体の違いを図 3 に示す．「源氏物語」の文には「なさる」や「いらっしゃる」といった敬語表現がよく用いられる．この結果，統計的機械翻訳の評価尺度である BLEU の値が悪くなると予測した．そこで，BCCWJ のコアデータ 58,355 文を言語モデルの学習用コーパスに加え，出力文体への「源氏物語」の影響力を抑えて他の 13 作品の翻訳精度を上げられるかどうか実験を行った．

3.2 現代日本語書き言葉均衡コーパス

言語モデルにおける小学館コーパスの「源氏物語」の影響を押さえるため，BCCWJ からコアデータ 58,355 文を言語モデルの学習に使用した．これは「源氏物語」9,752 文の約 6 倍の規模である．コアデータは，書籍，雑誌，新聞，白書，Yahoo!知恵袋，Yahoo!ブログから構成される．

3.3 青空文庫，与謝野晶子訳『源氏物語』

青空文庫の与謝野晶子訳「源氏物語」17,648 文も言語モデルの学習に使用した．図 3 に示したように，青空文庫の「源氏物語」の方が小学館の「源氏物語」現代語よりも他の 13 作品の文体に近いため，小学館の「源氏物語」を青空文庫の「源氏物語」に差し変えて言語モデルを学習し，翻訳を行った．

13 作品
楊貴妃が、玄宗皇帝の御使者に会って、泣いた顔にたとえて、「梨花一枝、春、雨を帯びたり」などと言ったのは、並一通りではあるまいと思うにつけて、やはりとてもすばらしい点では、他に類があるまいと感じられる。 少しお粥などをさしあげたところ、お召し上がりになりなどしたが、そのうれしさは何にたとえようもない。 耳敏川、これは、またも何をそんなに聞き耳をたてて聞きとったのだろうと、おもしろい。
源氏物語
下草のあれこれ美しく咲いている花々や紅葉などを 手折らせなさって、女二の宮の お目にかける 手土産に なさる。 大殿は廂の御簾の中に いらっしゃる ので、式部卿宮と右大臣だけが おそばにお控えになり、それ以下の上達部は簀子に居並んで、今日は正式の御賀の日ではないので、ご馳走などはそう仰々しくはなくお出ししてある。 源氏の君は、山里の人にも、久しく無沙汰のまま お過しだったことをお思い出しになり、わざわざお使者を お差し向けになった ところ、僧都の返事だけが寄せられる。
青空源氏
林の下草の美しい花や、紅葉を折らせた薫は夫人の宮にそれらを お見せした。 縁側に近い御簾の中に院のお席があって、そこにはただ式部卿の宮が 御同席され、右大臣の陪覧する座があっただけである。以下の高官たちは皆縁側に席をして、そこには形式を省いた饗応の物が出されてあった。 それで源氏の君も多忙であった。北山の寺へも久しく見舞わなかったことを思って、ある日わざわざ使いを立てた。山からは僧都の返事だけが来た。

図 3: 13 作品，源氏物語，青空源氏の文体の違い

表 2: 言語モデルの学習に使用したコーパス

言語モデル	訓練データ	開発データ	評価データ	計（文）
13 作品 + 源氏物語（ベースライン）	17,715	2,211	2,211	22,137
13 作品	7,963	996	996	9,955
源氏物語	9,752	1,215	1,219	12,186
青空源氏	17,648	-	-	17,648
13 作品 + 青空源氏	25,611	-	-	25,611
13 作品 + 源氏物語 +BCCWJ	80,292	-	-	80,292
13 作品 + 源氏物語 + 青空源氏	35,363	-	-	35,363

4　実験設定

本実験では，3 章で述べたコーパスを用いて複数通りのパターンで言語モデルを学習し，古文の現代語訳を行った．翻訳モデルの学習には，対訳になっている小学館コーパスのみを使用した．小学館コーパスは古文とその現代語訳が各作品で段落ごとに対応づいているが，統計的機械翻訳においては一文ごとに対応づいていることが望ましい．そこで，Gale らの方法を用いて一文ごとの対応づけを行った [Gale&Church1993]．実験を行った言語モデルの作成に使ったコーパスの組み合わせを表 2 に示す．

小学館コーパスは，古文・現代文ともに，訓練データ，評価データ，開発データとして 8:1:1 の割合で分割した．訓練データは翻訳モデルと言語モデルを作るのに使用した．言語モデルを学習する際に複数のコーパスを用いる場合，線形補間で複数の言語モデルを組み合わせた．評価データは古文を翻訳システムの入力とし，現代文は出力文の評価で正解データとして使用し

表 3: 実験結果 BLEU 値

	評価用コーパス			
学習用コーパス	13 作品 + 源氏物語	13 作品	源氏物語	小学館（星野ら）
13 作品 + 源氏物語（ベースライン）	22.38	24.81	20.21	-
13 作品	21.09	25.41	17.94	-
源氏物語	20.88	22.71	19.88	-
青空源氏	20.11	22.84	17.61	-
13 作品 + 青空源氏	22.46	24.98	20.24	-
13 作品 + 源氏物語 +BCCWJ	22.41	24.95	20.35	-
13 作品 + 源氏物語 + 青空源氏	21.61	25.55	18.46	-
小学館（星野ら）	-	-	-	28.02

た．開発データは翻訳システムにおける各種パラメータのチューニングに使用した．表中の「13 作品」は「源氏物語」を除いた小学館コーパス中の平安文学 13 作品を，「源氏物語」は小学館コーパスの「源氏物語」を指す．

コーパスの分かち書きには MeCab v0.98 [Kudo et al.2004]，辞書には中古和文 UniDic v1.4 [小木曽ら 2010] 及び UniDic v2.1.2[伝ら 2007]，単語アライメントには GIZA++ v1.0.7[Gao&Vogel2008] を用いた．統計的機械翻訳のツールは Moses v0.91[Koehn et al.2007] を用い, distortion limit は 0 とした．翻訳の際にはエラー最小化学習を用いてパラメータのチューニングを行った．翻訳結果の評価尺度には，翻訳結果と正解語の一致率で翻訳精度を測る BLEU[Papineni et al.2011] を使用した．

5　実験結果

5.1　BLEU スコアの評価

小学館コーパス，BCCWJ，青空文庫の 3 種類のコーパスを 6 通り組み合わせて言語モデルを学習し，古文を現代文へ翻訳した．翻訳結果および星野らの BLEU スコアを表 3 に示す．ただし，星野らは言語モデルと翻訳モデルの学習に図 2 の小学館コーパス全ての作品を用いて翻訳を行っていることに注意してほしい．

出力の評価には，正解文との比較で単語 n-gram の一致度を測る BLEU と呼ばれる評価尺度を用いた．BLEU は出力文に含まれる単語が正解文に含まれる単語と一致しているほど高いスコアを与える．言語モデルによって正解文に近い文体が出力できれば，BLEU も上がると考えられる．

表中の「学習用コーパス」は，言語モデルの学習に用いたコーパスを指す．「評価用コーパス」は，翻訳の入力に用いた評価データの古文のコーパスを指す．言語モデル学習用コーパスに「13 作品 + 源氏物語」を用いた場合をベースラインとして示す．ベースラインでは「13 作品」を翻訳した際に最も評価値が高くなった．

「13 作品」を翻訳したとき，言語モデル学習用コーパス「13 作品 + 源氏物語 + 青空文庫」で最も BLEU が高くなった．いずれの評価データを翻訳した場合も，学習用コーパスに「13 作品 + 青空文庫」や「13 作品 + 源氏物語 +BCCWJ」を用いたときにベースラインより BLEU が

古文	いといみじき心地しけり。
現代文および翻訳結果	ほんとにどうしようもない気持がした。
古文	「などてかくなくぞ」といへど、いらへもせず。
現代文および翻訳結果	「どうしてこのように泣くのか」といっても、返事もしない。
古文	今日いかにまれ、このことを定めてむ。
現代文および翻訳結果	今日どうあってもこのことを決めてしまいましょう。

図 4: 全ての言語モデルで正解データと同じ文に翻訳できた例

古文	その夜は、くろとの浜といふ所にとまる。
現代文	その夜は、黒戸の浜という所に泊った。
翻訳結果	その夜は、また、この美しい 黒戸の浜という所にとまる。
古文	雨降らぬ日、張り筵したる車。
現代文	雨の降らない日に、筵のおおいを掛けた牛車。
翻訳結果	雨は降らない日、張り筵をしている車 をしたのである。

図 5: ①古語と現代語の対応が不適切な例

高くなった．「13 作品」を翻訳した際，ベースラインと比べ「源氏物語」を除いた「13 作品」では 0.6 ポイント上がり，「13 作品 + 青空文庫」，「13 作品 + 源氏物語 +BCCWJ」，「13 作品 + 源氏物語 + 青空文庫」など「源氏物語」の影響を抑えた学習用コーパスを用いたときは 0.14〜0.74 ポイント上がるといったことから，「源氏物語」が「13 作品」の翻訳精度を下げていたと言える．いずれの結果も星野らの BLEU 値と比較して 2.47 ポイント以上低くなっているのは，表 1 で示したように翻訳モデルの学習に用いたコーパスの量が少なかったことが考えられる．ただし，6 章で示すように，BLEU 値では翻訳の性能を測りきれないため，一概に翻訳の性能が劣ったと言い切ることはできない．

5.2　翻訳のうまくいった例，うまくいかなかった例

本章では，評価データ 13 作品を翻訳した結果，正解データと同じように翻訳できた例と正解データとは違う翻訳となった例を示す．まず，どの学習用コーパスでも正解データと同じ訳に翻訳できた例を図 4 に示す．

逆に，評価用データの 13 作品を翻訳して翻訳が正解データと異なる例について①古語と現代語の対応が不適切な例，②主語や目的語など古語で省略されているが現代語では補足されている例，③ある古語に対して正解データの現代語とは表層形が異なるが同義の語が当てられている例，④同じ表層形でも違う意味（語義曖昧性）を持つ例，の 4 種類に分類し，図 5 から図 8 に示す．図 5 の上の例では，入力文である翻訳元の古文にはない「また、この美しい」という句が翻訳結果に出てきている．下の例は，文末に「をしたのである」という句が表出している．これは，フレーズテーブルに不適切な翻訳の対応が多くあることが原因である．図 6 は，古文で主語や目的語などの語が省略されているが，正解データの現代文では補われているために翻訳結果が正解データと完全には一致しない例である．図 7 の 1 つ目と 2 つ目の例は，正解データの現代文と翻訳結果とで意味はほぼ同じだが表層形が異なるものの例である．2 つ目の例では，古文の「あやしき」が正解データの「奇異な」ではなく「不思議な」に訳されている．図 7

古文	御火取に、ひと日の薫物とうでて、こころみさせたまふ。
現代文	中宮さまは、香炉に、先日の薫物を土中から取り出させてお入れになり、出来具合を ためしてごらんになる。
翻訳結果	御香炉には、一日の薫物をとうでになられて、ためしにおさせになる。
古文	それをと思ふなりけり。
現代文	その人をぜひ と思うのだった。
翻訳結果	それをと思うのであった。

図 6: ②正解データに補足語がある例

古文	むかし、二条の后に 仕うまつる 男ありけり。
現代文	昔、二条の后に お仕えする 男がいた。
翻訳結果	昔、二条の后に お仕えしている 男がいた。
古文	その花のなかに、あやしき 藤の花ありけり。
現代文	その花の中に、奇異な 藤の花があった。
翻訳結果	その花の中に、不思議な 藤の花があるのだった。
古文	河 は飛鳥川。
現代文	河 は飛鳥川。
翻訳結果	川 は飛鳥川。

図 7: ③表層形が異なる例

古文	むかし、男、狩の使よりかへり来けるに、大淀の わたり に宿りて、斎の宮の わらはべ にいひかけける。
現代文	昔、男が、狩の使いから帰ってきた時に、大淀の 渡し場 に泊って、斎宮の御殿に奉仕する 童女 に歌を詠みかけた。
翻訳結果	昔、男が、狩の使いから帰ってきたので、大淀の あたり に泊って、そのままかの 子供 に言葉をかけたのであった。

図 8: ④語義曖昧性の問題がある例

の 3 つ目の例は，異なる漢字が対応してしまった例である．図 8 は同じ語でも複数の意味を持つ場合，正解データと異なる意味の語が訳語に当てられた例である．「わたり」には「渡し場」と「あたり」の両方の意味があり、「わらはべ」は文脈により「童女」や「子供」になり得る．

6　1 文ごとの BLEU 評価

実際にどのような翻訳結果の文が BLEU を下げているのか確認するため，ベースラインで「13 作品 + 源氏物語」2211 文を翻訳し，1 文ずつ BLEU で評価した．この結果の分布を図 9 に示す．BLEU は 0 点から 100 点の値で評価を行う．この値は単純に表 3 の全体の BLEU 値と比較することはできない．表 3 に示したような通常用いられる BLEU は 1 文ごとではなく文章全体で算出するためである．BLEU を 1 文ずつ算出する場合，1 文に含まれる単語の数に対して評価データに含まれる単語がマッチする数を計算するため，1 文が短い場合，不当に BLEU が下がることがある．しかしながら，今回はどのような翻訳結果が BLEU を下げているかといった大まかな傾向を考察するためにこの方法を用いる．

図 9 で横軸は 0 点から 100 点まで 10 点ごとに刻んだ BLEU 値を表し，縦軸は各 BLEU 値における文数の分布の割合を表す．

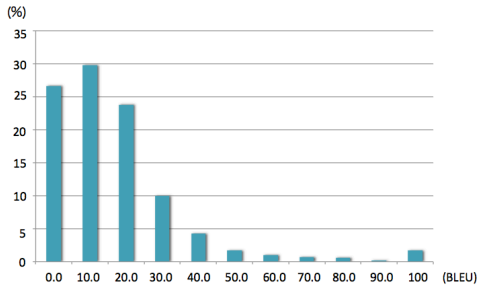

図 9: 1 文ごとの BLEU 値の分布

、 ||| 、 「 それ ごらん なさい 。 ||| 0.25 0.612571 1.231e-05 3.99231e-15 2.718 ||| ||| 4 81235
、 ||| 、 お 答え 申しあげ た 歌 。 ||| 1 0.612571 1.231e-05 2.8906e-16 2.718 ||| ||| 1 81235
、 ||| 、 こう お 答え に なる 。 ||| 0.333333 0.612571 1.231e-05 5.53684e-15 2.718 ||| ||| 3 81235
、 ||| 、 こう ロずさん だ 。 ||| 1 0.612571 1.231e-05 1.64564e-12 2.718 ||| ||| 1 81235
、 ||| 、 こう 書き送っ た 。 ||| 1 0.612571 1.231e-05 9.46122e-13 2.718 ||| ||| 1 81235
、 ||| 、 こう 言い送っ た 。 ||| 0.5 0.612571 1.231e-05 1.42811e-12 2.718 ||| ||| 2 81235
、 ||| 、 こう 言わ れ た 。 ||| 0.333333 0.612571 1.231e-05 7.64404e-14 2.718 ||| ||| 3 81235
、 ||| 、 こう 言わ れ た 。 まして ||| 0.333333 0.612571 1.231e-05 1.19935e-17 2.718 ||| ||| 3 81235
、 ||| 、 こう 詠ん だ 。 ||| 0.473684 0.612571 0.00011079 2.17605e-11 2.718 ||| ||| 19 81235
、 ||| 、 こう 詠ん だ 。 いか ||| 0.25 0.612571 1.231e-05 3.18139e-15 2.718 ||| ||| 4 81235
、 ||| 、 こう 詠ん だ 。 いか で ||| 0.25 0.612571 1.231e-05 7.738e-17 2.718 ||| ||| 4 81235
、 ||| 、 こう 詠ん だ 。 近江 ||| 1 0.612571 1.231e-05 4.04746e-16 2.718 ||| ||| 1 81235
、 ||| 、 こう 詠ん だ 。 近江 なる ||| 1 0.612571 1.231e-05 1.69815e-18 2.718 ||| ||| 1 81235

図 10: フレーズテーブル：対訳の不適切な対応例

図 9 に示した BLEU 値の分布を見ると，50 点台から 100 点台のものが少なく，0 点台から 20 点台に分布する文数が全体の約 80% を占めていることが分かる．BLEU 値ごとに翻訳結果を見ると，60 点台までは元の古文と現代文の間に対応のない語があるために訳せなかったものや，送り仮名や漢字といった表記の違いによるもの，同じ古語に正解データと異なる表層形の現代語が当てられたものが原因で BLEU が下がっている場合が多いことが分かった．対応のある語が訳せているならば翻訳自体はできていると見なせること，また，表記の違いや似た意味の語が翻訳結果に選ばれることは文の大まかな意味を知るためであれば十分な訳といえることから，BLEU による評価方法を見直す必要があると考える．

0 点台から 20 点台を見ると，上記の問題に加え，古文と現代文の評価データが 1 文ずつ正確な対応がとれていないものも多く見受けられた．他には，訳語に不必要な対応が付いているものや，文脈にふさわしくない訳語が選択されていることも BLEU を下げる原因であった．これらは翻訳として不都合であるため，翻訳の過程で改善する必要がある．

次に，学習用コーパスによって翻訳結果に文体や訳語の違いが見られた例を図 11 に示す．この例は，表 3 の BLEU 値と 1 文ごとの BLEU 値に相関の見られたものである．この例でも不適切な語の対応や異なる表層形の語など，上に挙げたような翻訳の問題が見られる．表 3 でも BLEU の低かった「源氏物語」や「青空文庫」で，文頭の「が」をはじめとした不要な対応の他に，「大人ごとに」が「そのうちの年輩ごとに」となっているなどの不適切な対応がある．表 3 で最も BLEU の高かった「13 作品 + 源氏物語 + 青空文庫」では，「這ひ来る」に不適切な訳語が対応していたり，「ほど」や「ごと」など不要な語の表出があるが，ベースラインや他の例と比較すると不適切な対応語の長さが短くなっているなど，全体的な改善が見られる．

7　評価に BLEU を用いる問題点と解決策

6 章で見たように，正解データの現代文にあって古文にない語が翻訳されないために BLEU が下がるという問題点がある．また，現代文の正解データと表層形の異なる似た意味の語が翻訳結果に選ばれた場合，翻訳文としての意味が自然であっても BLEU が下がってしまう．6 章で BLEU の大まかな傾向は妥当であるといえることが分かったが，これらの問題点に対処するためには BLEU による評価を見直す必要がある．たとえば，語の省略に頑強な評価方法として，正解データとの一致率に関する制約を緩めることが考えられる．また，同じような意味の語の評価に関しては，評価における正解を 1 つに絞らないといった対策が考えられる．

正解データ	
古文	二つ三つばかりなるちごの、いそぎて這ひ来る道に、いと小さき塵のありけるを、目ざとに見つけて、いとをかしげなる指にとらへて、大人ごとに見せたる、いとうつくし。
現代文	二歳か三歳ぐらいの幼児が、急いで這って来る道に、とても小さいごみのあったのを、目ざとく見つけて、とても愛らしげな指につかまえて、大人たちに見せているのは、とてもかわいらしい。
翻訳結果	
13+ 源氏	二つ三つぐらいの幼児が、急いでということになりてくる途中、とても小さいの塵ほどのあったのを、目ざとに目をおつけにて、とてもかわいらしげな指にとって、大人ごとにを見せているのは、とてもかわいらしい。
13	二つ三つぐらいの幼児が、急いでこそこそてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。
源氏	が二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしいの指につかまえて、そのうちの年輩ごとにを見せている、それがまことにかわいらしい。
青空	が二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしいの指につかまえて、そのうちの年輩ごとにを見せているの、それがまことにかわいらしい。
13+ 青空	二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとに目をおつけにて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。
13+ 源氏 +BCCWJ	二つ三つぐらいの幼児が、急いでということになりてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。
13+ 源氏 + 青空	二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。

図 11: 文体・訳語の違いと BLEU 値に相関が見られた例

8　フレーズテーブルの問題点と解決策

6 章での分析結果から，古文と現代文とで翻訳の対応が正確に取れていない例も多く見受けられた．これは，フレーズテーブルに不適切な訳語が多く発生したためと考えられる．実際にフレーズテーブルを確認したところ，図 10 に示したように，読点に読点以外の語が対応しているなど多くの不適切な対応があることを確認した．これらの不適切な対訳をフレーズテーブルから取り除く方法は，Johnson らにより提唱されている [Johnson et al.2007]．他にも，一対一の対応を強化するため対訳のコーパスに辞書を追加する方法や，GIZA++ で語の対応を学習する際に不適切な語の対応を適切な語に置き換えることで正確な対訳の確率を上げる方法も考えられる．

9　まとめと今後の課題

本稿では，統計的機械翻訳を用いて古文を現代文に翻訳する際，言語モデルと翻訳モデルの学習に使用するコーパスを中古和文に絞り，言語モデルの学習用コーパスに小学館コーパス以外のコーパスを加えることで翻訳性能の向上を図った．コーパスを加えた結果，星野らよりも評価値は低かったものの，ベースラインよりも翻訳精度は向上した．これは，言語モデルを生

成する際に，小学館作品の中古和文のコーパス内で他と文体の異なる「源氏物語」の影響が少なくなったためと考えられる．また，翻訳結果を 1 文ごとに評価し分析した結果，入力の古語がそのまま訳せている例が見られたこと，古語と現代語で不正確な対応が多くあったことから，BLEU による評価方法の見直しや訳語の対応関係の向上が今後の課題となることが分かった．

謝辞

本研究で使用したコーパス小学館『新編日本古典文学全集』は，国立国語研究所から頂いたものです．関係者各位に感謝致します．

参考文献

[Gale&Church1993] Gale, William A. and Kenneth W. Church (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational linguistics* Vol. 19.1, pp.75-102

[Gao&Vogel2008] Gao, Qin and Stephan Vogel (2008). Parallel Implementations of Word Alignment Tool. In *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing* (ACL2008), pp.49-57

[Johnson et al.2007] Johnson,J. Howard, Joel Martin, George Foster et al. (2007). Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL2007), pp. 967-975

[Maekawa2008] Maekawa, Kikuo (2008). Balanced Corpus of Contemporary Written Japanese. In *Proceeding of the 6th Workshop on Asian Language Resources* (ALR 6), pp.101-102

[Papineni et al.2011] Papineni, Kishore, Salim Roukos,Todd Ward et al. (2011). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (ACL2011), pp. 311-318

[Koehn et al.2007] Koehn, Philipp, Hieu Hoang, Alexandra Birch et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on interactive poster and demonstration sessions* (ACL2007), pp. 177-180

[Kudo et al.2004] Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *EMNLP2004*, pp. 230-237

[Stolcke2002] Stolcke, Andreas (2002). SRILM–an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 901-905

[新井ら 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，27:5，pp.463-469

[小木曽ら 2010] 小木曽智信、小椋秀樹、田中牧郎、他 (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，2010-CH-85:4，pp.1-8

[伝ら 2007] 伝康晴、小木曽智信、小椋秀樹、他 (2007)「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」日本語科学，22 号，pp.101-122

[星野ら 2014] 星野翔、宮尾祐介、大橋駿介、他 (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 回年次大会発表論文集，pp.816-819

[横野ら 2014] 横野光、星野翔 (2014)「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，pp.161-166