対照コーパスを用いた古文の現代語機械翻訳

星野 翔 [1,2]　宮尾 祐介 [1,2]　大橋 駿介 [1,3]　相澤 彰子 [1,3]　横野 光 [1]

[1] 国立情報学研究所　[2] 総合研究大学院大学　[3] 東京大学大学院

{hoshino,yusuke,sohashi,aizawa,yokono}@nii.ac.jp

1　はじめに

統計的機械翻訳には質・量ともに充分な対訳データが不可欠だが、そのような対訳データがいつでも利用可能とは限らない。例えば中古和文（古文）とその現代語訳の対訳データは、書籍として大量に存在するはずであるが、これまでの所、計算機で利用可能な形式に整備された大規模な対訳データとしては存在せず、また対訳辞書も公開されていない。

唯一利用可能な資源として、古代から近世までの文学・歌集作品を収録した小学館新編日本古典文学全集の一部の電子データ（以降、小学館コーパスと呼ぶ）が存在するが、このデータでは段落ごとに古文とその現代語訳が記載されており、統計的機械翻訳に望ましい文対応付けされた対訳データではない。

そこで本研究では、小学館コーパスを古文とその現代語訳の対照コーパスとみなし、対照コーパスから対訳データを作成することで、統計的機械翻訳による古文の現代語訳を可能にした。[*1]

対訳データ作成のための文対応付けには、内山, 井佐原の手法 [10] があり、対訳辞書を利用することで対照コーパスから信頼性の高い対訳部分のみを抽出することができる。しかし、この手法は対訳辞書の利用を前提としているため、今回のようなケースには応用できない。

2　対訳データ作成手法

ここでは前提として、古文とその現代語訳のように、必ずしも文対応付けされていないが、あるセグメント内は対訳になっているデータを対照コーパスとみなして利用する。そのようなデータは、古文と現代語を段落対応付けすることで生成することができる。その場合、セグメントは段落となる。

2.1　句点による分割

古文と現代語の対照コーパスが与えられた時、あるセグメント内で古文と現代語内の句点数が一致していれば、文単位で翻訳がなされているとみなせる。そこでこのヒューリスティックを用いて、句点でセグメントを分割する。句点数が一致しなかったセグメントは使わない。

この手法は、おおむね正しく文対応付けられたデータを生成することができるが、句点数の一致しないセグメントは利用することができない。

2.2　提案手法

そこで提案手法では、句点数が一致する場合に句点による分割を行い、句点数の一致しないセグメントについても、文の数がより多いセグメントをより少ない文の数に分割することで、古文と現代語の文対応付けを行う。

まず、セグメント内の文字列を句点で区切り、古文と現代語のうち、文の数がより多い文集合を N、文の数がより少ない文集合を M と置く。このとき文集合 S に対する文の数を $len(S)$、文 $s \in S$ に対する単語数を $w(s)$ と表し、$len(N) > len(M)$ である。

次に、文集合 N と M の文の数が等しくなるよう、文の数がより多い集合 N を $len(M)$ 個に分割する。その非交差な（各部分集合の要素が隣り合っている）分割を π、部分集合を $p \in \pi$ と表し、また p, M 中の i 番目の要素 (文) をそれぞれ p_i, M_i とする。

部分集合 $p \in \pi$ を評価するためのスコア関数 $score(p)$ を与えた時、これを最大化する部分集合 $p_{max} \in \pi$ を求めるには

$$p_{max} = \underset{p}{\operatorname{argmax}}\, score(p)$$

を計算する。

ここで提案手法では、古文と現代語で単語数と文

[*1] 古文の文章は既出のものであるから、機械翻訳ではなく、計算機に全データを記憶させることも考えられるが、そのような対訳データが存在しないのがそもそもの問題である。

の長さの分布が似ているものが良い分割の部分集合だと仮定し、スコア関数を次のように定義する:

$$f_1 = \sum_{i=1}^{len(M)} (w(p_i) - w(M_i))^2$$

$$\delta(x,y) = \begin{cases} 1 & w(x) - w(y) > 0 \\ 0 & w(x) - w(y) = 0 \\ -1 & w(x) - w(y) < 0 \end{cases}$$

$$f_2 = \sum_{i=2}^{len(M)} (\delta(p_i, p_{i-1}) - \delta(M_i, M_{i-1}))^2 + 1$$

$$score(p) = -f_1 * f_2$$

全セグメントに対してスコア関数 $score(p)$ を計算することにより、それぞれのセグメントでの部分集合 $p_{max} \in \pi$ が求まり、対訳データを生成することができる。

実際には、組み合わせ数が膨大になるのを防ぐため、$30 \geq len(N) \geq len(M)$ の制約を満たすセグメントのみを使用した。

3 実験

提案手法の有効性を実証するため、それぞれの対訳データ作成手法の生成するデータを共通の統計的機械翻訳システムに学習させ、翻訳精度を比較することによって対訳データ作成手法を評価する比較実験を行った。

実験では、対訳データ作成手法に提案手法、句点による分割、さらにベースラインとして対照コーパスをそのまま使用した場合の翻訳結果を比較した。

3.1 実験設定

	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	

表1: 小学館コーパス統計情報

日本霊異記, 古今和歌集, 竹取物語, 伊勢物語, 大和物語, 平中物語, 土佐日記, 蜻蛉日記, 落窪物語, 堤中納言物語, 枕草子, 源氏物語, 和泉式部日記, 紫式部日記, 更級日記, 讃岐典侍日記, 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

表2: 小学館コーパス収録作品

対訳データとしては、前述の小学館コーパスを、セグメント対応のとれた対照コーパスとみなし、そこから各手法を用いて対訳データに変換した。表 1 に小学館コーパスの統計情報、表 2 に小学館コーパスの収録作品一覧を記載する。対訳データの単語分割には MeCab 0.994[7] を使用した。古文の単語分割では辞書として中古和文 UniDic 1.3[1][*2]を用いた。

言語モデルには SRILM 1.7.0[9] を使用し、6-gram 言語モデルを作成した。実験で共通する統計的機械翻訳システムは MGIZA 0.7.3[3] と Moses 1.0[6] を用いて作成した。Moses のパラメータには、予備実験で最良だった distortion limit 0 を指定した。対訳データのフィルタリングやチューニングは行わなかった。

評価尺度には、BLEU[8] と RIBES[4] の 2 つを用いた。

実験では、異なる対訳データ作成手法を均一の評価データで評価するため、以下の手順を用いた:

1. 小学館コーパス 19,102 セグメントのうち、18,602 セグメントを訓練データ、500 セグメントを評価データとする。
2. 訓練データと評価データのそれぞれを、対訳データ作成手法によって分割する。そのとき評価データに限り、均一のデータとするため、句点による分割・提案手法で分割できなかったセグメントもそのまま使用する。
3. 各訓練データで統計的機械翻訳システムを学習させて、対になっている評価データを翻訳する。
4. 各翻訳結果を 1 セグメントずつにまとめる。
5. 1 セグメントを 1 行とみなして評価する。

3.2 実験結果

対訳データ作成手法	行数	BLEU	RIBES
ベースライン	18,602	25.48	76.13
句点による分割	56,436	25.79	75.08
提案手法	84,591	**28.02**	**76.89**

表3: 比較実験結果

行数はそれぞれの対訳データ作成手法での訓練データ量を、太字はブートストラップ・リサンプリング [5] におけるその他全手法との統計的有意性（p<0.01）を表している。

表 3 に比較実験の結果を示す。まず 2 つの評価尺

[*2] http://www2.ninjal.ac.jp/lrc/index.php?UniDic

度に着目すると、BLEU と RIBES の両方で提案手法が他の手法を上回った。特に BLEU スコアには約 2.5 ポイントと大きな差があり、提案手法を用いて作成した対訳データが、翻訳精度の改善に大きく貢献していることが分かる。

RIBES スコアは提案手法で 76.89 ポイントと単言語翻訳にも関わらずあまり高くないが、これは対照コーパスとして使用した小学館コーパスに段落対応付けの誤りが少なくなく、それらが文分割の誤りに影響したためであると考えられる。

次に行数に着目すると、提案手法の行数は最多（1 行あたりの単語数が最小）で、反対にベースラインの行数は最小（1 行あたりの単語数が最多）である。この実験で評価データは均一であるため、適切な行数に分割された対訳データの翻訳精度が最も高いことが分かり、提案手法の有効性が示されている。

3.3 翻訳例の比較と分析

表 4 にそれぞれの対訳データ作成手法で分割した古文、現代語、翻訳結果の例を示す。句点数が一致する場合には句点による分割を、一致しない場合には提案手法を示している。

提案手法とベースラインを比較すると、ベースラインは「宣旨下りたまふ」を「宣旨をなさる」と誤訳しているが、提案手法は「宣旨をお受けになりました」と正しく訳出できており、提案手法とベースラインの翻訳精度の違いが現れている。

一方、ベースラインは「やがて」を「そのまま」と正しく訳出できているが、提案手法では「すぐに」と誤訳されている。これは「やがて」が「そのまま」と「すぐに」の両方の意味を持つ多義語で、語義曖昧性の解消に文脈が必要であるにも関わらず、提案手法での文分割の結果「やがて」が文頭に置かれてしまい、文脈が参照できなくなったためである。

4 おわりに

本研究は、古文を現代語に統計的機械翻訳するための手法を提案し、比較実験によりその有効性を実証した。提案手法は、人工頭脳プロジェクト「ロボットは東大に入れるか」における国語古文問題解答 [2] に利用され、その有効性が示されている。

謝辞

本研究で利用したデータは、国語研究所通時コーパスプロジェクトから提供を受けた。

参考文献

[1] 小木曽智信, 小町守, 松本裕治. 歴史的日本語資料を対象とした形態素解析. 自然言語処理, 20(5):727–748, 2013.

[2] 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.

[3] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, 2008.

[4] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP*, pages 944–952, 2010.

[5] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proc. of EMNLP*, pages 388–395, 2004.

[6] 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. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL Demo and Poster Sessions*, pages 177–180, 2007.

[7] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proc. of EMNLP*, pages 230–237, 2004.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, 2002.

[9] Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. SRILM at sixteen: Update and outlook. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[10] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *Proc. of ACL*, pages 72–79, 2003.

古文（入力）	現代語（参照訳）	翻訳結果（出力）
	ベースライン（1 行）	
御母后、延喜三年癸亥、前坊をうみたてまつらせたまふ。御年十九。同二十年庚辰女御の宣旨下りたまふ。御年三十六。同二十三年癸未、朱雀院生まれさせたまふ。閏四月二十五日、后宣旨かぶらせたまふ。御年三十九。やがて、帝うみたてまつりたまふ同月に、后にもたたせたまひけるにや。四十二にて、村上は生まれさせたまへり。	御母后（穏子）は、延喜三年癸亥（九〇三）に、もとの東宮（保明親王）をお産み申しあげられました、御年十九歳。同二十年庚辰に女御の宣旨をいただかれました、御年三十六歳。同二十三年癸未に朱雀院がお生れになりました。その年の閏四月二十五日、后の宣旨をお受けになられました、御年三十九歳。〔天皇をお産み申しあげなさった同じ月に、そのまま、后にもお立ちになったのでありましょうか。〕このお方が四十二歳の時、村上天皇がお生れになりました。	御母后は、延喜三年癸亥（九〇三）、前東宮をお産み申しあげなさいました、御年十九歳。同二十年庚辰女御の宣旨をなさる。御年三十六歳。同二十三年癸未（九二、朱雀院がお生れになる。閏四月二十五日、大后は宣旨をかぶらせになる。御年三十九歳。そのまま、天皇をお産み申しあげなさる同じ后にお立ちになったのであろうか。四十二に、村上天皇はお生れになった。
	提案手法（6 行）	
御母后、延喜三年癸亥、前坊をうみたてまつらせたまふ。御年十九。 同二十年庚辰女御の宣旨下りたまふ。御年三十六。 同二十三年癸未、朱雀院生まれさせたまふ。 閏四月二十五日、后宣旨かぶらせたまふ。御年三十九。 やがて、帝うみたてまつりたまふ同月に、后にもたたせたまひけるにや。 四十二にて、村上は生まれさせたまへり。	御母后（穏子）は、延喜三年癸亥（九〇三）に、もとの東宮（保明親王）をお産み申しあげられました、御年十九歳。 同二十年庚辰に女御の宣旨をいただかれました、御年三十六歳。 同二十三年癸未に朱雀院がお生れになりました。 その年の閏四月二十五日、后の宣旨をお受けになられました、御年三十九歳。 〔天皇をお産み申しあげなさった同じ月に、そのまま、后にもお立ちになったのでありましょうか。 〕このお方が四十二歳の時、村上天皇がお生れになりました。	御母后は、延喜三年癸亥（九〇三）、前東宮をお産み申されました、御年十九歳。 同月二十年庚辰女御の宣旨をお受けになりました、御年三十六歳。 同月二十三年癸未（九二三）、朱雀院がお生れになりました。 閏四月二十五日、大后の宣旨をお受けになる。御年三十九歳。 すぐに、天皇をお産み申しあげなさる三月に、后にお立ちになったのであろうか。 四十二歳で、村上天皇がお生れになりました。
	ベースライン（1 行）	
中比、某の宰相とかや聞こえし人、才覚も優に、賢人の覚えありけるが、出家して高野山に隠居して、念仏の行をむねとして、真言なんどもうかがひ、道心者の聞こえあるあり。平生の願ひに、「最後の時、念仏すべき用意に、大方の数遍は時によるべし。正しき最後の十念をば、いかに心を澄まして唱へ、第十の念仏一反をば、殊に声を打ち上げて、思ひ入れてのびのびと申して、やがて引き入らばや」と念願して、願ひの如く少しも違はず、念仏して息終りにけり。	中ごろ、某の宰相とかいわれていた人は、才覚も優れ、賢人との評判もあったが、出家して高野山に隠居し、念仏の行を主として真言なども学び、道心者としての評判が高かった。常の願いに、「最後の時に念仏を唱えるための準備として、一通りの数遍はその時々によるべきである。今正に最後を迎えようとする時の十念は、どうにかして心を澄まして唱え、第十の念仏一反を殊更に声を打ち上げ、思いを込めてのびのびと申して、そのまま息を引き取りたい」と念願していたが、その願いの通りに少しも違うところもなく、念仏して息を引き取った。	中ごろ、某の宰相と申しあげたお方は、学識も優美に、賢人の覚えていたが、出家して高野山の山にして、念仏の行をして、真言なども様子をうかがっていたが、道心者がある。平生の願いに、「最後の時、念仏するように、私の遍は時なのであろう。本当にこれが最後の十念を、心を澄まして唱え、第十の念仏の一遍は、特に声を読み上げて考えてのんびりと申して、そのまま引き入らたい」と念願して、願いのとおりに少しも違わず、念仏して息が終ってしまった。
	句点による分割（3 行）	
中比、某の宰相とかや聞こえし人、才覚も優に、賢人の覚えありけるが、出家して高野山に隠居して、念仏の行をむねとして、真言なんどもうかがひ、道心者の聞こえあるあり。 平生の願ひに、「最後の時、念仏すべき用意に、大方の数遍は時によるべし。 正しき最後の十念をば、いかに心を澄まして唱へ、第十の念仏一反をば、殊に声を打ち上げて、思ひ入れてのびのびと申して、やがて引き入らばや」と念願して、願ひの如く少しも違はず、念仏して息終りにけり。	中ごろ、某の宰相とかいわれていた人は、才覚も優れ、賢人との評判もあったが、出家して高野山に隠居し、念仏の行を主として真言なども学び、道心者としての評判が高かった。 常の願いに、「最後の時に念仏を唱えるための準備として、一通りの数遍はその時々によるべきである。 今正に最後を迎えようとする時の十念は、どうにかして心を澄まして唱え、第十の念仏一反を殊更に声を打ち上げ、思いを込めてのびのびと申して、そのまま息を引き取りたい」と念願していたが、その願いの通りに少しも違うところもなく、念仏して息を引き取った。	中ごろ、某の宰相とか申しあげた人は、学才も優美に、賢人の覚えていたが、出家して高野山に離れして、念仏の修行を中心として、真言などもねらい、道心者があるのである。 平生のままに、「最期の時、念仏するはずの用意に、数度は時によるものであろう。 本当に最後の十念を、どのように心を澄ませ、念仏を唱え、第十の念仏一回を、特に声を読み上げていて、のんびりと申して、そのまま奥のほうへ退ったい」と念願して、願いどおりに少しも違わず、念仏を唱えて息が終ってしまったのだった。

表4: 対訳データ作成手法の比較