ことばの単位に基づく翻訳レベルの定義と古文－現代文間の文節アライメント

臼井 久生 [1] 尾崎 太亮 [1] 古宮 嘉那子 [1] 小木曽 智信 [2]
東京農工大学大学院 生物システム応用科学府 [1] 国立国語研究所・総合研究大学院大学 [2]
{h-usui,hiroaki-ozaki}@st.go.tuat.ac.jp,
kkomiya@go.tuat.ac.jp, togiso@ninjal.ac.jp

概要

本研究ではことばの単位に基づく翻訳レベルの定義を提案する。例えば、単語のように粒度が小さい単位で翻訳を行うと、正確性が上がるが流暢性や読みやすさのレベルが下がり、段落のように粒度が大きい単位ごとに翻訳を行うと、流暢性や読みやすさのレベルは上がるが正確性が下がる。そのため、本研究では、翻訳レベルを定義することでこれらを整理し、流暢性や正確性を含めて翻訳を総合的に評価・検討することができるようにする枠組みの構築を目指す。また、本研究では日本語文法において、言葉を細かく区切った際に不自然にならない最小単位である文節レベルの翻訳に注目し、日本語の古文と現代文の翻訳を対象に、文レベルのアライメントの取れた対訳から、LLM を用いて文節レベルのアライメントの試行を行った。

1 はじめに

古文の翻訳は様々な種類があるが、多くの現代語訳と呼ばれるものは訳註や補充を含む意訳になっている。源氏物語の冒頭を例に挙げると、

> 『いづれの御時にか、女御、更衣あまたさぶらひたまひける中に、いとやむごとなき際にはあらぬが、すぐれて時めきたまふありけり。』

という原文に対して、小学館の新編日本古典文学全集における現代語訳は次のようになる。

> 『帝はどなたの御代であったか、女御や更衣が大勢お仕えしておられた中に、最高の身分とはいえぬお方で、格別に帝のご寵愛をこうむっていらっしゃるお方があった。』

この現代語訳において、文脈や語からの類推で容易にわかるため帝という語が何度か補記されている。しかしこれにより原文に含まれない箇所と含まれる箇所の峻別が容易でなくなっている。古文とその現代語訳の語や句レベルのアライメントをとることが容易な訳として大学受験での古典分野などでの逐語訳が知られている。これは単語一つ一つの意味をとり現代日本語として不自然にならないように訳したものである。この方法で先ほどの文を翻訳すると、

> 『どなたの御代であったか、女御や更衣が大勢お仕えしておられた中に、非常に高貴な身分とはいえないが、格別にご寵愛をこうむっていらっしゃるお方があった。』

となり、訳語のアライメントが取りやすくなっている。しかし、逐語訳は流暢な現代語訳と比べるとやはり読みづらく、不自然であると言える。また、直訳という方法もある。この翻訳手法は辞書の現代語との対応する訳をそのまま並べたものであるが、原文に忠実な訳である反面、文として不自然になってしまう。したがって、どのような翻訳が良いかを考える上ではこのような翻訳スタイルを含めて考える必要があると言える。

さらに、機械翻訳を念頭に置くとできる限り原文とアライメントが取れる訳文がある方が、BLEU などの機械的な指標で翻訳性能を評価することができるため好都合であるが、流暢性の観点も含めて総合的に翻訳を評価・検討する枠組みが必要となると考えられる。

本研究では、様々な翻訳を自然言語処理の観点から評価・検証するために訳文に対する翻訳の粒度や定義を明確にした 5 つのレベルを提案する。特に本研究では、日本語文法において、言葉を細かく区切った際に不自然にならない最小単位である文節レベルに注目する。現状、文節レベルでアライメントのとれたコーパスは存在せず、また、人手で大規模なコーパスを作成することは困難である。そのた

表 1　翻訳レベル別の翻訳例: 上部の原文は源氏物語『桐壺』の冒頭部分を Web 茶まめ [1] の中古和文 UniDic[2] を用いて分かち書きしたものである。下部はそれぞれの翻訳レベルに基づいて翻訳した例である。

原文	いづれ／の／御／時／に／か／、／女御／、／更衣／あまた／さぶらひ／たまひ／ける／中／に／、／いと／やむごとなき／際／に／は／あら／ぬ／が／、／すぐれ／て／時／めき／たまふ／あり／けり／。
単語レベル	いつ／の／御／代／に／だったか／、／女御／、／更衣／数多く／お仕え／なさって／た／なか／に／、／とても／高貴な／身分／に／は／あら／ない／が／、／際立っ／て／盛んな時期／にみえ／られる／いる／そうだ／。
文節レベル	いつの／**時代だったか、**／女御、／更衣／数多く／お仕えなさっていた／中に、／たいそう／高貴な／身分では／**居られないが、**／際立って／**寵愛を受けられて**／いる／そうだ／。
節レベル	いつの時代だった**ろうか**、女御、更衣**が**数多くお仕えなさっていた中に、／たいそう高貴な身分ではないが、**特に**寵愛を受けていらっしゃった。
文レベル	**どの帝の**時代だったろうか、女御、更衣が数多くお仕えになっていた中で、たいそう高貴な身分ではないが、特に寵愛を受けていらっしゃった。
段落レベル	**帝は**どなたの御代であったか、女御や更衣が大勢お仕えしておられた中に、**最高の身分とはいえぬお方で、格別に帝の**ご寵愛をこうむっていらっしゃる**お方があった。**

め、本研究ではまず古文の文レベルでのアノテーションが取れた対訳コーパスから LLM を利用して文節アライメントを行う。また、現状文節アライメントされている対訳コーパスがなく正解例がないため、LLM を用いてその評価を試みた。

本研究の貢献は、次の 2 点である。

- 流暢性および正確性を意識した翻訳の違いを区分し、評価・検証することが可能な翻訳レベルをことばの単位に基づく形で新たに提案したこと
- 文節レベルの対訳を作成するために、LLM を用いて古文－現代文間の文節アライメントを行い、その評価と修正の試行を行ったこと

2　関連研究

BERT のモデルを応用して単語アライメントを取る研究は複数存在する（Lai ら [3] 他）。しかし、昨今の Decoder のみの LLM をそのまま用いてアライメントを行う研究は少ない。陳ら [4] は単語アライメントの誤った対応から歌ことばのコノテーション検出を行っている。彼らは、現代語と古文を比較し、和歌の原文には出現せず現代語訳にのみ出現する意味の単位のことをノンリテラル要素として定義している。これらの要素は翻訳レベルの定義においても重要な要素であると考えられる。

現在、古文－現代文対訳コーパスは星野ら [5] によって作成された文アライメント済みのデータセットが存在している。このデータセットは小学館コーパスに含まれる古典作品から構成されている。この古文－現代文対訳コーパスを利用して実際に古文から現代文への文対文翻訳を行なった例として Usui ら [6] の研究がある。また、Ozaki ら [7] は国語研長単位及び文節区切りを Transformer ベースのモデルで行い、文節区切りにおいて 97.5%の精度を達成した。これにより、古文における文節区切りの自動化が可能となった。

また、多言語間で共通の構文構造をアノテーションする枠組みに Universal Dependencies (UD)[8] がある。本研究の構想にはこの、多言語において共通の構文構造というレイヤー構造が一助となった。

3　ことばの単位に基づく翻訳レベルの定義

1 節で述べた通り、粒度を明確にした 5 つのレベルでの定義を行い、それぞれの粒度についてその他語順整序などの要素を含めたものを翻訳レベルとする。翻訳レベルは以下の 5 つとする。 これらとは別に、指定のレベルを超えたことばの語順整序、主語の補充、一対多対応などをそれぞれのレベルに付け加えて表記し、レベル 1, 語順整序ありなどとする。

レベル１：単語レベル　日本語においては国語研短単位を基準にしてアライメントをとった翻訳を行う。このレベルでは基本的に語順整序は認めず、短単位ごとに対応する語を割り当てる。実装手法としては語義曖昧性解消が近く、all-words WSD のように全ての短単位に対して対応する意味を候補の中から選択し、文とするような翻訳を行う。

レベル２：文節レベル　国語研の文節認定を元にアライメントをとった翻訳を行う。現状公開されて

いる日本語 UD コーパスには文節境界が付与されているため、それを利用して文節ごとにアライメントをとった翻訳を行うことが可能である。語順整序は、文節内のみ認める。この文節について、文節は日本語のみに存在する概念である。

レベル３：節レベル　単文、あるいは復文における１つの節を基準にアライメントをとった翻訳を行う。この時の一つの節は、根ではない動詞を含む節を指す。語順整序は認めるが、主語の補充などは認めない。

レベル４：文レベル　単文復文問わず、一つの句点までを文としてアライメントをとった翻訳を行う。文レベルにおいては、語順整序は認めるが、主語の補充は文法上判別可能な場合のみ認める。

レベル５：段落レベル　段落ごとのアライメントをとった翻訳を行う。基本的に語順整序に加え、主語の補充や背景知識の付与なども認める。もっとも制限のない、流暢な翻訳である。

表 1 の下部は冒頭の源氏物語の一節をレベルごとに訳した例である。低いレベルでは流暢性が低いが単語など小さい単位での原文とのアライメントが取れており、高いレベルでは流暢ではあるが原文にないものが付与されていることがわかる。この例のように、翻訳レベルを定義することにより翻訳の粒度を明確にすることができるうえ、そのレベルにするような翻訳や、訳文のそれぞれのレベルの付与を行うことができるようになる。

4　LLM を用いた文節アライメント

本研究においては、レベル 2 の文節レベルでそれぞれの文節の順序を入れ替えることを認め、一対多対応を認める翻訳レベルでのデータ作成を行うことを目指す。現状、近代以前の日本語と現代日本語訳のアライメントの取れているデータセットに星野らの作成した文アライメント済みのデータセット [5] がある。このデータセットを用いて、文レベルから文節レベルへのアライメントを行う。このデータセットは源氏物語、枕草子、古今和歌集などの古典作品から構成されており、合計で 86,684 文のペアが存在している。データセットを文節に区切るため、尾崎ら [7] の提案した Monaka[1] を用いて、付録 B の設定で文節区切りを行った。このデータを用いて GPT-4o で文節アライメントを行った。プロンプトは付録 C に示す。古文と現代文の文節をそれぞれ入力し、対応する文節ペアを zero-shot で出力させた。出力形式は古文と現代文の文節の indices を利用し、それぞれの古文文節に対応する現代文文節を辞書形式で出力させた。この実験では付録 A に示す、源氏物語『澪標』の一節を原文として分析した。この文の原文の文節数は 16 であり、現代語訳の文節数は 27 であった。アライメントの結果を評価するために、古文の翻訳のために訓練された T5 モデル [6] のエンコーダー部分を用いて類似度を計算した。具体的には、アライメントされた各文節ペアに対して、それぞれの文節をエンコードし、コサイン類似度を計算した。この類似度スコアを用いて、アライメントの品質を評価した。T5 として retrieva-jp/t5-base-long[2] を古文から現代文に翻訳するためにファインチューニングしたモデルのエンコーダー部分を用いた。モデルには 1 文をすべて入力し、文節にあたる部分のトークンのベクトルをそのまま、あるいは複数トークンであれば平均化して文節ベクトルとし、古文と現代文それぞれの文節ベクトル同士のコサイン類似度を計算した。

5　実験結果

表 2　GPT を用いた際の古文と現代文の文節アライメント例。多くが一対一対応であるが、一部一対多対応が見られる。

古文	現代文
入道の	入道が
思ひかしづき思ふらむ	たいせつに かわいがっているだろうと、
ありさま	様子を
思ひやるも	思いやるに
ほほ笑まれたまふ	ほほえまれずには いらっしゃれない
こと	ことも
多く、	多く、
また	また、
あはれに	しみじみと
心苦しうも	おいたわしく
ただ	この
この	姫君の
ことの	ことばかりが
御心に	お心から
かかるも、	離れないのも、
浅からぬにこそは。	浅からぬ

表 2 に GPT でのアライメント結果を示す。また、付録 E に T5 によるアライメント例を示す。この結果から、GPT では 9 つの現代文文節がアライメントをとられていないことがわかった。これらのうち、翻訳をされた際に追加された『姫君』という主語に

1）https://github.com/komiya-lab/monaka

2）https://huggingface.co/retrieva-jp/t5-base-long

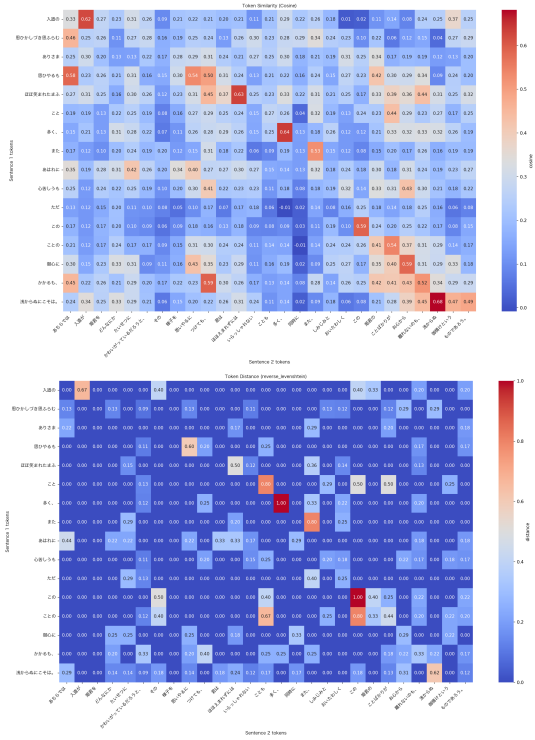

図 1　（上）T5 による類似度行列（下）編集距離スコア：この図においては比較する文節それぞれの文字が近ければ近いほど高く、遠ければ低いスコアとなる。

注目すると、対応する古文の文節が存在しないにも拘わらず誤ってアライメントが取られていた。また、係り結びである『こそは。』で文末が終わっているが、対応する現代語訳である『あろう。』を適切にアライメントさせられていないこともわかった。

次に、類似度スコアの結果について図 1 の上部に T5 を用いた類似度行列を示す。また、付録 D に BERT を用いた類似度行列を示す。T5 では高い類似度が連なって見られるが、これらの類似度の高いものに「あはれに，しみじみと」などの意味的に近いと思われるペアが含まれておらず、意味より表記が近いため値が高くなっていると予想した。この表記の近さを測るために、編集距離を正規化し、1 から引いた値を図 1 の下部に示す。図 1 のふたつの図の比較から、編集距離スコアと類似度スコアの高いマスが複数一致していること、つまり、類似度スコアが高いマスの中には表記の近いものが多く含まれていることがわかった。

次に、GPT のアライメントと T5 による類似度スコアを比較する。まず、主語の補充について見る。GPT によるアライメントの結果では『姫君の』を『この』に誤ってアラインメントしている。T5 による類似度スコアでもこの語を適切にアライメントさせることはできなかった。また、GPT によるアライメントの結果では「あはれに、しみじみと」という意味的に近いペアがアラインメントされているが、T5 による類似度スコアは高くない。このことから、T5 に比べ、GPT は意味的な類似性を考慮したアラインメントを行っている可能性が示唆される。

6　展望

本研究の文節アラインメントは萌芽的な試作である。以下に将来の展望を示す。まず、今回の実験では GPT を用いて文節アライメントを行ったが、他のファインチューニングを行い古文の学習をした LLM を用いた場合との比較や、プロンプトの工夫なども今後検討したい。また、LLM によるアライメントの評価として類似度行列による方法を試行したが、この方法でのアライメントの評価は難しいことがわかった。このアライメント評価を行うためには、人手によって文節が付与された対訳コーパスが必要である。人手によるアライメントの補助として、LLM による結果を示しそこからの修正を人手によって行う方法を考えている。

次に、ことばの単位についての展望を述べる。今回の翻訳レベルはことばの単位としたが、日本語 UD コーパスには、文節係り受け情報や主辞情報なども付与されており、この手法を応用することで複数レベルの翻訳を UD 対訳コーパスから作成することも可能である。さらに、UD を用いることによりソース・ターゲット間の言葉の単位を揃えることができるため、この手法は UD 対訳コーパスや UD 解析機のある諸言語にも応用可能である。そのほかに最も低いレベルでの翻訳と人手によるアライメントの補助の方針を活用することで、インターリニアグロスの作成補助など人文学への応用も可能である。今後、これらに向け翻訳レベルそれぞれにおける翻訳の作成と評価の方法を検討したい。

7　おわりに

本研究では、様々な翻訳を評価・検証するために訳文に対する翻訳の粒度、定義を明確にした翻訳レベルを定義した。また、これらのレベルのうち文節レベルにアラインメントの取れたコーパス作成を目指し、LLM による古文と現代文における文節アライメントを行った。今後、多言語への同じ手法の応用や、他分野での活用など様々な発展に向け評価手法などの検討を重ねる。

謝辞

本研究は、JST 次世代研究者挑戦的研究プログラム JPMJSP2116 の支援を受けたものです。また、本研究は、公益財団法人 三菱財団 人文科学研究助成「自然言語処理を利用した古文解析」、国立国語共同研究プロジェクト 「開かれた共同構築環境による通時コーパスの拡張」、国文学研究資料館 令和 7 年度国文研プロジェクト型共同研究 「大規模言語モデルを用いた OCR 読み取り結果のエラー訂正と現代語への翻訳」の助成を受けたものです。

参考文献

[1] 智昭堤, 智信小木曽, Tsutsumi Tomoaki, Ogiso Toshinobu. 複数の unidic 辞書による形態素解析支援ツール『web 茶まめ』の実装と運用. 情報処理学会論文誌, Vol. 64, pp. 749–757, Mar 2023.

[2] 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, Vol. 20, No. 5, pp. 727–748, 2013.

[3] Siyu Lai, Zhen Yang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Cross-align: Modeling deep cross-lingual interactions for word alignment. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 3715–3725, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[4] 旭東陳, ボルホドシチェク, 啓史山元, Chen Hilofumi Yamamoto Bor Hodoscek Xudong. 単語アライメントの誤り対応を用いた歌ことばのコノテーション検出. じんもんこん 2022 論文集, Vol. 2022, pp. 111–118, 12 2022.

[5] 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 回年次大会発表論文集, pp. 816–819, 2014.

[6] Hisao Usui and Kanako Komiya. Translation from historical to contemporary Japanese using Japanese t5. In Mika Hämäläinen, Emily Öhman, Flammie Pirinen, Khalid Alnajjar, So Miyagawa, Yuri Bizzoni, Niko Partanen, and Jack Rueter, editors, **Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages**, pp. 27–35, Tokyo, Japan, December 2023. Association for Computational Linguistics.

[7] Hiroaki Ozaki, Kanako Komiya, Masayuki Asahara, and Toshinobu Ogiso. Long unit word tokenization and bunsetsu segmentation of historical Japanese. In John Pavlopoulos, Thea Sommerschield, Yannis Assael, Shai Gordin, Kyunghyun Cho, Marco Passarotti, Rachele Sprugnoli, Yudong Liu, Bin Li, and Adam Anderson, editors, **Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)**, pp. 48–55, Hybrid in Bangkok, Thailand and online, August 2024. Association for Computational Linguistics.

[8] Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. Universal Dependencies. In Alexandre Klementiev and Lucia Specia, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts**, Valencia, Spain, April 2017. Association for Computational Linguistics.

A　実験に用いた全文

表 3　実験に用いた全文

原文	入道の思ひかしづき思ふらむありさま思ひやるもほほ笑まれたまふこと多く、またあはれに心苦しうもただこのことの御心にかかるも、浅からぬにこそは。
翻訳文	あちらでは入道が姫君をどんなにかたいせつにかわいがっているだろうと、その様子を思いやるにつけても、君はほほえまれずにはいらっしゃれないことも多く、同時にまた、しみじみとおいたわしくこの姫君のことばかりがお心から離れないのも、浅からぬ御情けというものであろう。

B　Monaka の各種設定

Monaka について、model は all_in_one、古文用辞書は wabun、現代文用辞書は gendai、入力形式は tsv ファイル、出力形式は bunsetsu-split として実行した。

C　GPT4o を用いたアライメントの試行

プロンプトは次のとおりである。src_block には文節区切りを行った現代文、tgt_block には文節区切りを行った古文がそれぞれ改行区切りで入力される。

```
"You are an expert in establishing correspondences
     between bunsetsu (phrases) in classical and modern
     Japanese.\n"
"The first sequence of tokens represents the classical
     Japanese translation of the second sequence, which
     is written in modern Japanese.\n"
"Each element in both sequences corresponds to a bunsetsu
     .\n"
"Establish correspondences between bunsetsu in the
     classical Japanese text (Target) and the modern
     Japanese translation (Source).\n"
"A single bunsetsu may correspond to multiple bunsetsu in
      the other sequence.\n"
"Return ONLY a JSON object with the following format:\n"
'{ \"alignments\": [ { \"src\": int, \"tgt\": int }, ...
     ] }\n'
"- Use 0-based indices for both sequences.\n"
"- Align every Source bunsetsu to at least one Target
     bunsetsu.\n"
"- Prefer monotonic, semantically faithful alignments.\n"
"- Do not include explanations, comments, or code fences.
      Output pure JSON only."
f"{src_block}\n\n{tgt_block}\n\nReturn JSON now."
```

D　類似度行列の BERT との比較

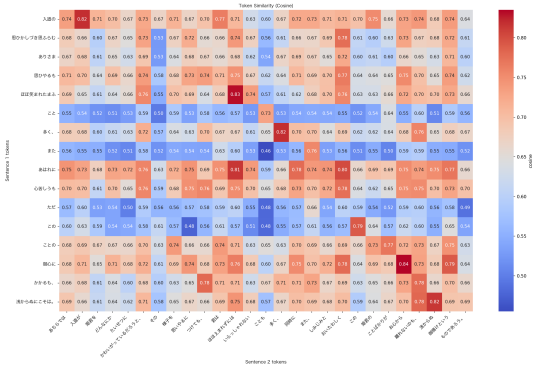

図 2　BERT を用いて類似度を計算した結果の類似度行列
BERT には、tohoku_nlp/bert-base-japanese-whole-word-masking を用いた。この類似度行列について、T5 の類似度が高い部分が BERT でも同様に高くなっている。しかし、BERT ではより多くのマスが高くなっている。また、値の幅が BERT は T5 に比べて狭いことがわかる。この BERT は T5 と比べて古文の学習量が少ないため、T5 に比べて類似度の評価が難しいと考えられる。

E　T5 によるアライメント例

表 4　T5 による類似度の高い文節ペアの例。ほぼ全ての現代文文節に対応する古文節が存在しているが、『ただ』という古文文節に対応する現代文文節が存在していない。これは類似度を計算した際に本来対応する文節より他の文節が高くなってしまったためである。

古文	対応する現代語（類似度降順）
入道の	入道が
思ひかしづき思ふらむ	かわいがっているだろうと、/いらっしゃれない/同時に
ありさま	
思ひやるも	あちらでは/思いやるに/姫君の
ほほ笑まれたまふ	君は/ほほえまれずには
こと	ことも
多く、	多く、/しみじみと
また	その/また、
あはれに	姫君を/たいせつに/様子を
心苦しうも	おいたわしく
ただ	
この	この
ことの	ことばかりが
御心に	お心から
かかるも、	つけても、/離れないのも、
浅からぬにこそは。	どんなにか/浅からぬ/御情けという/ものであろう。