Sentence BERT を用いた源氏物語内の古今和歌集引用の検出

叶内琉聖　古宮嘉那子
東京農工大学工学部　東京農工大学大学院
{s212110v@st.go,kkomiya@.go}.tuat.ac.jp

概要

本研究では Sentence BERT を用いて、源氏物語における古今和歌集の序文や和歌からの引用の検出を行った。コーパスから、源氏物語の一文と古今和歌集の一文あるいは一句のペアを作り、この二文をモデルへの入力とした。このとき、源氏物語の一文に古今和歌集の一文あるいは一句が引用されていれば正例、引用されていなければ負例として、二値分類を行った。負例の数は、正例の数との比で設定した。実験の結果、正例と負例の比を 1:58 にしたシステムにおいて、69.6%の F 値で古今和歌集からの引用を検出した。また、実験によってシステムが、表記の異なる言葉や作中に何度も登場する語句による引用の検出を苦手としていることが分かった。

1　はじめに

古文作品は、当時の教養・文化をもとに書かれている。そのため、当時すでに読まれていた他の古文・漢文作品からの引用が作品に含まれていることがあり、ときには作品の理解に重要な役割を果たしている。ゆえに、古文作品を理解するには、古文特有の文法や単語に関する知識に加え、他の古文・漢文作品についての知識が必要となる。しかし、対象が多く、すべての引用に気が付けるほど十分な作品知識を得ることは難しい。そこで本研究では、ある古文作品における他の作品からの引用を検出し、古文作品への理解に役立てたいと考えた。

本研究では、源氏物語における古今和歌集の序文や和歌からの引用を対象として、Sentence BERT[1] を用いた検出を行う。与えられた源氏物語内の一文に対し、ある古今和歌集の一文・一句からの引用の有無を分類するシステムを作成し、これを古今和歌集のすべての序文・和歌に用いることで、最終的に、与えられた源氏物語内の一文から古今和歌集からの引用を検出する。

2　関連研究

関連研究として、源氏物語における古今和歌集からの引用をベクトル検索によって検出する、近藤 [2] の研究がある。この研究では、源氏物語の全文と古今和歌集内の和歌にそれぞれベクトル埋め込みを行い、近傍検索と N-gram による文字列の一致を組み合わせることで、引用の検出を行っている。

また、Riemenschneider ら [3] は、SPhilBERTa と呼ばれる、トリリンガル SentenceRoBERTa モデルを提案している。このモデルは、古代ギリシア語・ラテン語・英語を横断した、別言語間の意味理解と似た文の識別に優れており、古代ギリシア語文章とラテン語文章の間の引用検出に役立つことが期待されている。

Konstantinidou ら [4] は、HoLM リソースと機械学習の 2 つの手法で、ホメロスの叙事詩「イリアス」と「オデュッセイア」の間の関係性を調べた。HoLM リソースは、ホメロス詩を統計的にモデル化したもので、作品内におけるある詩行の特殊度を評価できる。この研究では、詩行に対し、ほかの作品との類似度の調査や、機械学習による分類を行うことで、叙事詩間の関連性を調べている。

西洋の古典の引用検出についてはさらに Liebl ら [5] の研究がある。

3　Sentence BERT を用いた引用検出システム

本稿では、Sentence BERT を用いた二値分類によって、引用の検出を行った。Sentence-BERT は、BERT を文章単位でベクトル変換できるようにファインチューニングしたモデルで、文章同士の比較に優れている。本稿で作成したシステムは、源氏物語の一文と古今和歌集の一文あるいは一句のペアを作り、モデルに対し二文入力を行うこのとき、入力した二文のうち、源氏物語の一文に古今和歌集の一文あるいは一句が引用されているかどうかを分類する。

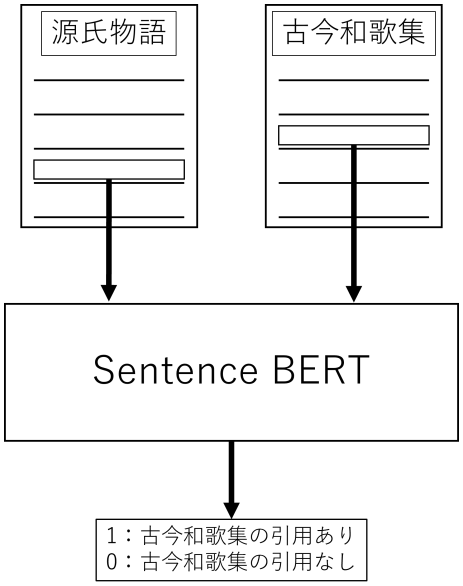

図 1　システムの模式図

システムの模式図を図 1 に示す。

4　データセット

データセットは、小学館コーパス（小学館新編日本古典文学全集の一部の電子データ）から、源氏物語と古今和歌集を用いて作成した。小学館コーパスは、古文作品それぞれにつき、古典本文・現代語訳・注釈が段落ごとにまとめられている。源氏物語のコーパス内の注釈には古今和歌集からの引用についての言及が複数あり、引用された仮名序や和歌なども具体的に示されている。本研究ではこの注釈を用いて、正例のペアを作成した。源氏物語コーパスの注釈と古今和歌集コーパスの古典本文では、同じ和歌でも一部に表記の違いがみられたため、古今和歌集の部分をすべて古今和歌集コーパスの古典本文における表記に変更した。例えば、源氏物語コーパスの注釈における「天の川紅葉を橋にわたせばや七夕つめの秋をしも待つ」という和歌は、古今和歌集コーパスの古典本文内の「漢河紅葉を橋にわたせばやたなばたつめの秋をしも待つ」という表記に変更した。これは、負例との差異を無くすためである。また、源氏物語コーパスの古典本文と、古今和歌集コーパスの古典本文のうち仮名序部分と和歌部分から、負例のペアを作成した。

5　実験

本研究では、次の 2 つの実験を行った。全ての実験において、テストデータに対する、出力の正解率と正例についての F 値によって、評価を行った。

5.1　正例/負例の比率を変えた分類実験

実験 1 として、正例/負例の比率を変えた 5 種類のデータセットを作成し、それぞれのデータセットを用いて、二分割交検証を行った。これにより、データセットにおける正例の事前確率の影響を調べる。本研究では、正例と負例との比率が、1:1、1:2、1:3、1:4、1:58 となるように、負例のペアを作成した。ここで、1:58 の比率設定は、実験における、源氏物語の総文数 21,888 文と、その中で、古今和歌集の引用が含まれていた 371 文との比率、約 59:1 に基づいている。

Sentence BERT には、日本語用 Sentence BERT モデルである sentence-bert-base-ja-mean-tokens-v2[1)] をファインチューニングして用いた。

システムの入力の単位は、文または和歌単位とした。ただし、源氏物語は鍵括弧などが多く含まれるため、読点のみで文を区切ると、文中に鍵括弧などが残る可能性がある。そこで実験 1 では、「源氏物語の一文」を、次のルールで区切った。

1. 読点「。」で区切る
2. 鍵括弧記号、二重鍵括弧記号で区切り、それぞれの括弧を消す
3. 後ろの鍵括弧（」や』）の直前には、読点が無いので、新たに加える

また、作中に登場する和歌は、前後を〔〕で囲んだ。

負例の作成は、次のようにして行った。正例と負例との比率が 1:1〜1:4 の場合は、源氏物語全体と古今和歌集の仮名序・和歌から、ランダムに選び、負例のペアを作成した。また、1:58 の場合は、データセット内の古今和歌集の仮名序・和歌ごとの偏りを減らすため、古今和歌集の全ての仮名序・和歌それぞれに、源氏物語の一文をランダムに組み合わせてペアにすることを繰り返した。

本実験では、データセットを 2 つに分け、二分割交差検証を行った。この時データセットは正例ごと、負例ごとに分割し、学習時に正例と負例との比率が変化しないようにした。また、学習データのうち、1/5 を、バリデーションに用いた。

実験は、エポック数 30、バッチサイズ 3、学習率 [0.01、0.001、0.0001] の設定で行った。各エポックごとにバリデーションを行い、F 値が最も高かった

1）https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

表 1　実験 1 結果

	適合率	再現率	F1	正解率
1:1	0.927	0.924	0.926	0.926
1:2	0.91	0.904	0.907	0.938
1:3	0.915	0.919	0.917	0.958
1:4	0.876	0.889	0.882	0.953
1:58	0.826	0.612	0.696	0.991

表 2　実験 2 結果

	適合率	再現率	F1	正解率
1:4	0.00453	0.9	0.00902	0.916
1:58	0.0934	0.592	0.161	0.997

エポック数・学習率の組み合わせを最終的なモデルとした。最適化関数には SGD、損失関数にはクロスエントロピーを用いた。

5.2　引用検出のための分類実験

実験 1 において作成したシステムのうち、正例と負例との比率を 1:4、1:58 と設定したシステムに対し、次の実験を行う。各システムのテストデータから、正例と負例のペアを 10 用例づつ取り出し、各ペアの源氏物語部分に対し、古今和歌集のすべての文、つまり仮名序 73 文および和歌 1111 句との全通りのペアを作成し、正例でないものはすべて負例として引用かどうかを分類する。つまり、源氏物語の文 20 例（10 例は古今和歌集の引用を含み、10 例は引用を一切含まない）に対し、古今和歌集のすべての文とのペアの中から、引用を含むペアをどの程度検出できるかを調べた。なお、この際、訓練データは 1:4 または 1:58 の正例/負例の比率であるが、テストデータ中の正例は 23680 用例中の約 10 用例にすぎないことに注意されたい。

この実験は、もし仮に源氏物語の新章が見つかった場合、どの程度の性能で古今和歌集の引用が検出できるかを調べるタスクである。源氏物語の一文に対し、古今和歌集の全文・全句をそれぞれペアとしてシステムに入力することで、古今和歌集からの引用の有無を調べる。

6　結果

表 1 に実験 1 の結果を示す。また表 2 に実験 2 の結果を示す。

表 1 より、データセットに含まれる負例の数が多くなればなるほど、F 値は下がり、正解率は上がることがわかる。正例と負例との比率が 1:4 のときと、1:58 のときを比べると、再現率が大きく下がっているのに対し、適合率に大きな低下は見られない。テストデータに含まれる負例の数が大きく増加したことを踏まえると、負例の数が増えたことで、正例に対する分類には誤りが増えたが、負例に対する分類の精度は上がったといえる。

また、表 2 より、学習データに含まれる負例の数が増えたことで、F 値と正解率が上がっていることがわかる。実験 2 におけるテストデータの数は、正例と負例との比率で変わらないので、適合率と再現率の変化から、正例に対する分類は間違いが増えたが、負例を正例と誤って分類することが減ったといえる。

以上のことから、データセットに含まれる正例の事前確率を低くすると、正例に対する分類性能は下がるが、負例に対する分類性能が上がり、結果的に、全体的な正解率は上がることが分かった。負例の数が増えたことで、正解率が上がったのは、負例を負例として当てているだけだからだと考えられる。

また、正例と負例の比率を 1:58 にした実験では、古今和歌集の全文を訓練データとして利用している点で他の実験とは異なる。他の条件をそろえて実験していないので正確なことは分からないが、この点が性能向上に寄与した可能性がある。

7　考察

実験 2 における、正例と負例との比率が 1:58 のシステムの出力のうち、正例として正しく分類できなかった例を、表 3 に示す。

1、2、3、4 に関しては、表記の違いが原因で引用として分類できなかったと考えられる。例えば 2 では、「とこの山なる」が引用部分だが、古今和歌集では「鳥籠の山なる」となっており、鳥籠（とこ）の部分の表記が異なっている。本研究のシステムは、日本語用 Sentence BERT を用いたものなので、ChatGPT のように広範な知識を持たないため、鳥籠（とこ）のような特殊な読み方に対応していなかったと考えられる。

5、6 に関しては、引用された語句が、源氏物語内で何度も登場したためだと考えられる。源氏物語コーパスを調べたところ、源氏物語内で「行く方」は 42 回、「まじり」は 77 回登場しており、語句が珍しくないので、引用とみなされなかったと思われる。しかし、人間の読者には、珍しくない語句が出てきても引用だと分かるということは、周囲に古今和歌集だと理解させる何らかの手掛かりが他にある

表 3　実験 2 1:58 用例

	源氏物語	古今和歌集
1	とこの山なる。	犬上の鳥籠の山なる名取河いさと答へよわが名洩らすな
2	御車よりはじめて、御前など、大将殿よりぞ奉れたまへるを、なかなかまことの昔の近きゆかりの君たちは、事わざしげきおのがじしの世の営みに紛れつつ、えしも思ひ出できこえたまはず。	世の中にある人、ことわざ繁きものなれば、心に思ふことを、見るもの聞くものにつけて、言ひ出せるなり。
3	〔たましひをつれなき袖にとどめおきてわが心からまどはるるかな〕外なるものはとか、昔もたぐひありけりと思たまへなすにも、さらに行く方知らずのみなむ。	身を捨ててゆきやしにけむ思ふよりほかなるものは心なりけり
4	世人のなびきかしづきたてまつるさま、かく忍びたまへる道にも、いとことにいつくしきを見たまふにも、げに七夕ばかりにても、かかる彦星の光をこそ待ち出でめとおぼえたり。	漢河紅葉を橋にわたせばやたなばたつめの秋をしも待つ
5	〔たましひをつれなき袖にとどめおきてわが心からまどはるるかな〕外なるものはとか、昔もたぐひありけりと思たまへなすにも、さらに行く方知らずのみなむ。	わが恋はむなしき空にみちぬらし思ひやれども行く方もなし
6	〔かきくらし晴れせぬ峰の雨雲に浮きて世をふる身をもなさばや〕まじりなば。	郭公峰の雲にやまじりにしありとは聞けど見るよしもなき
7	らうたげなるありさまを見棄てて出づべき心地もせず、いとほしければ、よろづに契り慰めて、もろともに月をながめておはするほどなりけり。	わが心慰めかねつ更級や姨捨山に照る月を見て
8	恨めしと言ふ人もありける里の名の、なべて睦ましう思さるる、ゆゑもはかなしや。	わが庵は都の辰巳しかぞ住む世をうぢ山と人はいふなり
9	いみじき武士、仇敵なりとも、見てはうち笑まれぬべきさまのしたまへれば、えさし放ちたまはず。	力をも入れずして天地を動かし、目に見えぬ鬼神をもあはれと思はせ、男女の中をも和らげ、猛き武士の心をも慰むるは歌なり。

可能性が高い。今後はそれらの手掛かりが何であるかを専門家の手助けを得て特定し、利用したいと考えている。

7 や 8 については、源氏物語と古今和歌集の間で、語句の変化が大きく、引用が分かりにくかったためだと考えられる。しかし、源氏物語コーパスを確認したところ、本実験で用いた文の前後で、「姨捨山の月」や「宇治」など、引用元の和歌と関連する語句がみられた。本実験は源氏物語を文単位で区切っており、これよりも長く源氏物語の入力をとれば、システムはこれらも引用として分類できるかもしれない。

9 に関しては、「武士」という言葉自体は一般的で、現代でも使われており、引用とみなされなかったと考えられる。

8　結論

本研究では、源氏物語における古今和歌集からの引用を、Sentence BERT を用いたシステムによって行った。正例の事前確率を低くすることで、負例に対するシステムの分類精度を上げ、全体の正解率を向上できるが、正例に対するシステムの分類精度が下がってしまうことが分かった。また、分類できなかった正例から、システムが検出を苦手とする引用の種類として、表記の違いにより同じ言葉と認識できなかった場合や、対象の語句が作品内で何度も登場した場合、源氏物語の区切り方が短かった場合などが分かった。今後は、これらの場合でも引用を検出できるように、システムを改良していきたい。

今回は Sentence BERT を用いてシステムを作ったが、今後は、RAG を使った手法でも、引用検出システムを実装し、比較したいと考えている。RAG は、既存の大規模言語モデルに別の検索システムを組み合わせる手法である。大規模言語モデルを用いているので、例えば表記が異なる語句でも、同じ語句として認識することが期待できる。

また、今後は、引用と剽窃との特徴の違いを見つけ、検出に活かしたい。引用と剽窃は、ほかの文章の内容を自分の文章に取り込んでいることは同じだが、目的が異なっており、引用は読者に気づいてもらうことを、剽窃は読者に気づかれないことを目的としている。この違いを、検出に活かせれば、引用に特化した検出を行えると考えている。

謝辞

本研究は JSPS 科研費 JP22K12145、公益財団法人三菱財団 自然言語処理を利用した古文解析の助成を受けたものです。また、エラー分析にあたり、青山学院大学 名誉教授　近藤 泰弘先生にご意見をいただきました。御礼申し上げます。

参考文献

[1] Aoife Cahill, James H Fife, Brian Riordan, Avijit Vajpayee, and Dmytro Galochkin. Context-based automated scoring of complex mathematical responses. In **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 186–192, 2020.

[2] 近藤泰弘. 『源氏物語』の引き歌をベクトル検索によって検出する方法. 人工知能学会全国大会論文集 第 38 回 (2024), pp. 1N4OS1803–1N4OS1803, 2024.

[3] Frederick Riemenschneider and Anette Frank. Graecia capta ferum victorem cepit. detecting latin allusions to ancient greek literature. In **Proceedings of the Ancient Language Processing Workshop**, 2023.

[4] Maria Konstantinidou, John Pavlopoulos, and Elton Barker. Exploring intertextuality across the homeric poems through language models. In **Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)**, 2024.

[5] Bernhard Liebl and Manuel Burghardt. “shakespeare in the vectorian age” – an evaluation of different word embeddings and nlp parameters for the detection of shakespeare quotes. In **Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature**, 2020.