学術論文における関連研究の執筆支援のための被引用論文の推定

小山康平[1]　南泰浩[1]　成松宏美[2]　堂坂浩二[3]
東中竜一郎[2]　田盛大悟[4]　平博順[4]

[1]電気通信大学　[2]NTTコミュニケーション科学基礎研究所
[3]秋田県立大学　[4]大阪工業大学

k1710245@edu.cc.uec.ac.jp, minami.yasuhiro@is.uec.ac.jp
{hiromi.narimatsu.eg, ryuichiro.higashinaka.tp}@hco.ntt.co.jp
dohsaka@akita-pu.ac.jp, e1b17043@st.oit.ac.jp, hirotoshi.taira@oit.ac.jp

1　はじめに

様々な科学技術研究の発展に伴い，分野を超えた研究も多数行われるようになってきた．これに伴い，研究者が学術論文を執筆する際，比較すべき主要アプローチやベース手法を把握するために，関連する数多くの論文を調査することに大きな労力を払うことになっている．我々は，このような研究者の論文執筆を支援するために，論文中の関連研究の章における論文同士の引用関係に着目した．これは，関連研究の章には，その論文に対する，主要アプローチやベース手法の引用がされ，その分野に対する多くの基本情報が含まれているからである．

論文を執筆する際，自分の主張をより強固にするため，引用文献を参照することは重要である．しかし，膨大な論文の中から適切な引用文献を用意することは特に研究者の労力を必要とする．例えば，研究者が自身の専門分野外の事実に関する主張を書くとき，その主張の根拠となる論文を膨大な論文の中から検索しなければならない．加えて，検索した論文を間違った形で引用しないよう注意深く，その論文を読み込まなくてはならない．もし，引用対象として適切ではないと気づいたときには，再度検索からやり直さなくてはならない．

このように考察すると，引用論文を精査する労力を軽減できれば，多くの研究者の支援ができると思われる．本稿では研究者を支援するため，関連研究の文章及びその引用関係の情報を用いて，どの論文を引用するべきかを割り当てるタスクを設定した（引用文献割り当てタスク）．論文に引用を付ける際，筆者が関連文献の候補を用意するだけで，引用文に適切な引用文献を割り当てることができれば，労力を軽減することができる．今回の研究では，これを実現するデータセットとして3.1節に示すデータセットを使用した．このデータセットには論文の引用関係や関連研究の情報が含まれており，サーベイ論文の自動生成タスクなど，複数のタスクで利用することも考慮されている[1]．このタスクの前段階として引用元論文と被引用論文のペアが適切であるかどうかの判定を行うタスクも同時に設定した（引用文・被引用文献ペア適正性判定タスク）．論文執筆者が根拠を示す引用文献を仮に付与した場合，その被引用文献が本当に対応するものであるかを確かめるためには，研究者はその論文を読み込まなくてはならない．その労力を軽減するためにこのタスクを考えた．この引用文・被引用文献ペア適正性判定を最初に行い，その結果を用いて，引用文献割り当てタスクを実現する．

2　関連研究

Michaelらは対象の文章に引用を付けるべきかを推定するモデルの検討をしている[2]．しかし，実際に論文を執筆する研究者を支援するためには，それだけでは不十分である．なぜなら，引用が必要な文であることが分かるだけでは，研究者が引用文献を調べる労力自体は十分に軽減されないからである．今回の実験では対象の文章に，被引用文献のアブストラクトを用いて，適切な引用文献を割り当て

る．これにより，労力の削減を図ることができる．

飯沼らは，最新の研究情報が反映されていないサーベイ論文に，追加するべき新たな引用文献の推定方法を提案している [3]．彼らの研究は，引用文献を推定する点では我々の研究と類似している．しかし，古いサーベイ論文で引用している文献を入力としているために，一から論文を作成する場合の支援には適していない．また，使用しているデータセットの種類が 4 種類の書籍のみであるため，書籍の筆者の記述の傾向を推定に含んでいる可能性がある．我々の研究では，様々な筆者の論文データを使用するとともに，一から論文を作成する研究者に向けたタスクを検討する．

3 タスクの設定

3.1 データ作成

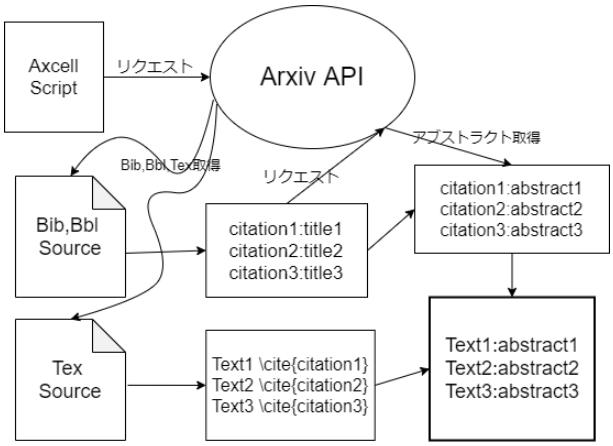

図 1 データ作成の概略

実験に使用するデータセットは，研究者支援の様々なタスクを共通のデータセットで実施できることを目標に作成した．引用元論文のデータとして Axcell[4] のデータを使用した．AxCell は ArXiv から約 10 万件の論文データを取得した．このデータには，論文の Pdf データ，図表データ，Tex ソース，bib，bbl ソースが含まれている．このうち，Tex ソースと bib，bbl ソースを活用することで論文同士の引用関係を抽出する．今回は関連研究の章に着目しているため，Tex ソースから関連研究の章を抽出した．抽出できたデータ数は約 3 万件である．この 3 万件の関連研究の章から，\cite{}を含む文章を取り出し，bib，bbl ファイルからタイトルを抽出した．この際，bib ファイルがない論文が多く存在していたので，主に bbl ファイルに正規表現をかけることによってタイトルを抽出した．抽出し終えたタイトルは，arXiv の API を用いて検索をかけ PDF，Abstract のデータを取り出した．今回はこの工程を通して，(\cite{}を含む文章，引用論文の Abstract) のデータを 2 万件作成した．

3.2 引用文・被引用文献ペア適正性判定タスク

引用元論文と被引用論文のペアが適切であるかどうかの判定をするタスクを行う．入力には引用元論文の引用部分の文章と被引用論文のアブストラクトを使用する．入力の負例として、被引用論文として不適切なアブストラクトを用いたデータを用意する．負例は被引用論文のアブストラクトを同じ論文の関連研究の章で引用されている別の論文のアブストラクトに置き換えて作成した．実際に筆者が関連文献の候補を用意する際，引用元論文と被引用論文のペアは意味合いが近くなっていることが予想される．このタスクの被引用論文を完全にランダムにしてしまうと，意味合いが全く異なるペアのみを負例として学習してしまい，わずかに意味合いが違う負例ペアを区別できない可能性がある．一方，同じ論文で引用されている引用元論文と被引用論文候補のペアは内容が大きく離れていることは少ない．細かい意味合いの違いを抽出するためにも，被引用論文の候補として，同じ論文で引用されている引用論文を使用する．これらの正例・負例データを用いて学習を行い，正例である確率を出力とするモデルの作成をする．

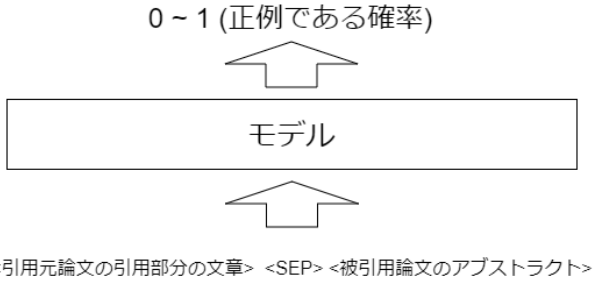

図 2 実験の概略

3.3 引用文献割り当てタスク

論文に引用を付ける際，筆者が関連文献の候補を用意するだけで，引用文に適切な引用を割り当てることができるよう，関連文献の候補の中から正しい被引用論文を推定するタスクに取り組んだ．引用元論文と被引用論文の候補を用意し，被引用論文それぞれに 3.2 節と同様の操作をする．この時，最も適

切である確率の高かった論文候補を適切な被引用論文とする．また，引用文・被引用文献ペア適正性判定タスクで作成したモデルが，細かい意味合いの違いを抽出できるか確認するために，被引用論文の候補は同じ引用元論文の関連研究の章で引用されている被引用論文を使用した．

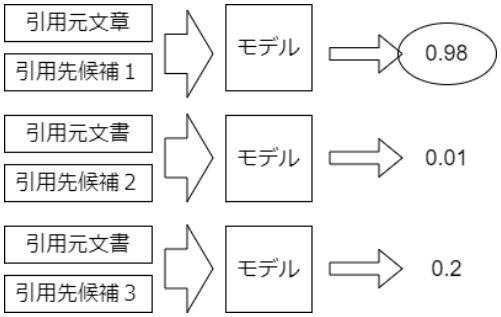

図 3　実験の概略

4　実験

4.1　引用文・被引用文献ペア適正性判定タスク

4.1.1　比較手法

実験に使用したモデルは Random, Word2Vec, BERT, XLNet の 4 種類である．

Random　5 割の確率で正例と判定し，5 割の確率で負例と判定するモデル．この手法を実験のベースライン手法とする．

Word2Vec　ベクトル化した文章同士の類似度を計算することで判定をするモデル．類似度が閾値よりも高ければ正例と判定し，閾値よりも低ければ負例と判定する．閾値は dev データを用いて設定する．入力は，ベース論文の引用部分の文章と被引用論文のアブストラクトである．

BERT, XLNet　huggingface の事前学習モデルをファインチューニングしてモデルを作成した．入力は，(<ベース論文の引用部分の文章> <SEP> <被引用論文の Abstract>) とした．出力は，被引用論文が適切である確率とした．学習率は 1e-4，epoch 数は 75 とした．今回の実験では，モデルが出力した確率が 0.5 以上の時は正例と推定し，0.5 未満と推定した時は負例と推定した．

4.1.2　共通設定

データセットとして，3.1 節で作成したものを正例として使用した．このデータセットを基に (\cite{} を含む文章，同じ論文内の別の個所で言及された引用論文のアブストラクト) の形の負例データを正例と同数作成した．合計データ数は (train/dev/test), (36,000/3,000/3,000) となった．また，評価指標として Accuracy, Precision, Recall, F-measure を計算した．

表 1　引用文・被引用文献ペア適正性判定タスク

	Accuracy	Precision	Recall	F-measure
Random	0.488	0.489	0.492	0.490
Word2Vec	0.558	0.541	0.763	0.633
BERT	0.816	0.822	0.806	0.814
XLNet	0.844	0.846	0.841	0.843

4.1.3　結果

4.2　評価

表 1 で示す通り，引用文・被引用文献ペア適正性判定タスクでは，BERT, XLNet は全ての評価指標でベースラインを大きく上回る結果を出すことができた．これにより，引用文献割り当てタスクでは正しい被引用論文を推定できることが期待できる．一方で，Word2Vec は Recall こそ BERT, XLNet に迫る結果になったものの，Accuracy, Precision はベースラインとなる Random をわずかに上回るだけの結果となった．

4.3　引用文献割り当てタスク

4.3.1　比較手法

実験に使用したモデルは，Random, Word2Vec[5], BERT[6], XLNet[7] の 4 種類である．

Random　被引用論文候補の中からランダムに選ばれた論文を推定された被引用論文とする．この手法を実験のベースライン手法とする．

Word2Vec　4.1 節と同様の処理を全ての被引用論文候補に行う．その結果，類似度が閾値を超えたデータのうち，最も類似度が高い論文を被引用論文と推定する．

BERT, XLNet　4.1 節と同様の処理を全ての被引用論文候補に行う．その結果，適正である確率が 0.5 を超えた被引用論文候補のうち，最も適正である確率が高い論文を被引用論文と推定する．

表 2　引用文献割り当てタスクの結果

	Random	Word2Vec	BERT	XLNet
accuracy	0.280	0.349	0.747	0.795

4.3.2　共通設定

BERT，XLNet モデルは 4.1 で学習させたモデルを用いた．

被引用の候補となる論文の Abstract に対して，4.1 と同様の計算を行い，最も確率の高い被引用候補となる論文を推定された被引用論文とした．被引用の論文候補は，引用元論文のほかの個所で引用されている論文のうち，取得できたデータ全てとした．

データセットには 4.1 で使用した Test データのうち実験に適した 600 件を使用した．評価指標として Accuracy を計算した．

4.3.3　結果

引用部分の文章	Nirkin proposed a system that allows face swapping ...
引用先論文の アブストラクト	We show that even when face images are unconstrained and ...

図 4　XLNet の引用文・被引用文献ペア適正性判定タスク正解例

引用部分の文章	Aziz Elkind and Elkind study the computational complexity of various problems related to weighted voting games
引用先論文の アブストラクト	Coalitional voting games appear in different forms in multi-agent systems

図 5　XLNet の引用文・被引用文献ペア適正性判定タスク不正解例

4.4　評価

表 2 で示す通り，引用文・被引用文献ペア適正性判定タスクと同様に，引用文割り当てタスクでも，BERT, XLNet は Random, Word2Vec を大きく上回り，高精度となった．BERT, XLNet はどちらも 7 割を超える正解率を出すことができ，このデータセットを用いた被引用論文の推定は十分に可能であるといえる．

実際に引用文の適性判定が行われた結果を確認してみる．表 (4) は XLNet が正しく被引用論文を識別できた結果である．引用元論文の文章では，画像中の顔を入れ替える技術について述べられていて，被引用論文のアブストラクトでも画像中の顔に関する言及がされている．どちらも同じ内容であるため，正しく推定できていることが分かる．一方，表 (5) は XLNet が正しく被引用論文を識別できなかった結果である．引用元論文の文章と被引用論文のアブストラクトのどちらの文章でも “voting games"について言及されているため，意味合いの近い文章を識別することはできているようである．

5　おわりに

本研究では，研究者支援の様々なタスクにも使用できることを目的にデータセットを作成し，引用文対引用文献適正判定タスクと引用文献割り当てタスクという二つのタスクを設定し，各種の手法による評価を行った．どちらのタスクでも，BERT, XLNet モデルは高い精度を出すことができた．一方で Word2Vec の精度は，高くなかった．すべての手法での誤認識には，引用部分の文章が被引用論文のアブストラクトに比べるとかなり短いことが影響していると思われる．この改善点として，入力の文章に引用部分の文章だけではなく，範囲をさらに広げて引用部分のパラグラフを使用することがあげられる．今後は，入力やモデルを改善による精度の向上と，共通データセットを用いた更なる論文執筆者支援タスクに取り組みたい．

参考文献

[1] 成松宏美, 小山康平, 堂坂浩二, 田盛大悟, 東中竜一郎, 南泰浩, 平博順. 学術論文における関連研究の執筆支援のためのタスク設計およびデータ構築. 言語処理学会第 26 回年次大会, 2021.

[2] Michael Färber, Alexander Thiemann, and Adam Jatowt. To cite, or not to cite? detecting citation contexts in text. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, pp. 598–603, Cham, 2018. Springer International Publishing.

[3] 飯沼俊平, 難波英嗣, 竹澤寿幸, 広島市立大学大学院情報科学研究科. サーベイ論文作成支援のための引用論文推薦, 2015.

[4] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. Axcell: Automatic extraction of results from machine learning papers, 2020.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5753–5763, 2019.