論文において外部の文献を引用・参照している文の分類

渡邊 晃一朗[1]　矢田 竣太郎[2]　影浦 峡[2,3]

[1] 東京大学 教育学部　[2] 東京大学 大学院教育学研究科　[3] 東京大学 大学院情報学環

[1]kouichirou-watanabe495@g.ecc.u-tokyo.ac.jp　[2]{shuntaroy, kyo}@p.u-tokyo.ac.jp

1 はじめに

本研究の目的は外部の文献を引用・参照することの要件がどのようなものかを示すことである。

学術の世界では，論文という，言語によって構成されるものでフォーマルなコミュニケーションが成り立っており，その中で外部への引用・参照は重要な役割を担っている。外部への引用・参照は論文においては何らかのものを根拠として示す行為であり，不可欠なものである [1]。しかし，やみくもに外部を引用・参照すればいいということではなく，必要のない外部への引用・参照は不適切なものと判断される [2]。

ここで，論文において外部の文献の引用・参照がどのような時に必要なのかが必ずしも明らかでないことが本研究における問題意識である。この問題意識から，本研究では外部の文献を引用・参照することの要件を明らかにすることを目的とする。そのために，既存の論文において外部を引用・参照している文の特徴を明らかにする。外部の文献の引用・参照は外部に存在する知識と内部の言語表現に関わる。外部の文献の引用・参照は論の展開や表現上の特色から決定されることもあるため，内部の言語表現の形式に着目し，本研究では外部に存在する知識を扱わない。

本研究では (1) 外部の文献を引用・参照している文 (以下，出典明示文) と (2) 外部の文献を引用・参照していない文 (以下，非出典明示文) の 2 種類の文を区別する形式的な特徴を発見することを，機械学習において文を 2 種類に分類する素性を発見することとみなす。ここで非出典明示文とは外部の文献の引用・参照を示す書式，例えば「A は B である (田中 [2017])」における「(田中 [2017])」の部分，つまり出典が付された文とし，外部の文献を引用・参照していない文とは，外部の文献の引用・参照を示す書式が付されていない文とする。ただし外部の文献を引用・参照することの要不要を議論する上で，外部の文献を引用・参照していることが明らかな文，例えば「田中 (2017) による と A は B である」,「田中 (2017) は A は B であると述べている」のような文は人名の存在から形式的にも外部の文献の引用・参照が必要であると必然的に判断できてしまうため，本研究では対象から基本的に除外する。これには Juman++[1]の名詞に対する人名判断を利用し，加えて人手によって判断した。

その上で，出典明示文を正例，非出典明示文を負例として，与えられた文をいずれかに分類する 2 値分類問題において，よりよく分類するモデルを構築できるような素性を発見する。これにより，2 種類の文を区別する形式的な特徴を発見する。

2 関連研究

引用をめぐる応用的研究においては，本研究に近いものとして引用部の同定の研究と剽窃検出の研究の 2 つが挙げられる。1 つ目の「引用部の同定の研究」には，Kaplan ら [3] による論文の内容の一貫性を用いた研究や，Athar ら [4] による感情分析を志向した研究，Qazvinian ら [5] による本文内容要約のための研究がある。これらは外部の文献の引用・参照を示す書式が文に付されていることを前提にした上で，外部の文献の引用・参照によって述べられている部分の把握を目的とした研究である。2 つ目の「剽窃検出の研究」には，剽窃を自動検出するための手法の構築の研究がある。Gipp [6] は適切に外部の文献の引用・参照を行なっている論文の表現を用いて剽窃を検出するアルゴリズムを開発している。Soleman ら [7] は 2 つの文書において出典明示文と非出典明示文同士の類似度を tf-idf を用いて計算し，その類似度を基に剽窃の有無を判定している。HaCohen-Kerner らは [8] は剽窃の自動検知における既存の手法を整理した上で，2 つの文書の概要と参考文献のそれぞれの比較や，概要と参考文献の比較を組み合わせて剽窃検出を行なってい

[1]Juman++. available from: http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++ (2017-12-27))

る。Stamatatos [9] は文書のレベルではなく，文，もしくはそれより小さい単位での類似度を計測することによって剽窃の自動検知を行なっている。Eissen ら [10] は外部の文献を考慮せずに，単独の文書の中で抽出可能な特徴のみを用いて剽窃を自動検知する研究を行なっている。

3 実験

3.1 データと手法

本研究で対象としたのは言語処理学会論文誌 LaTeX コーパス[2]のうち，日本語で書かれた論文である。対象となった LaTeX 形式の言語処理学会論文誌 LaTeX コーパス中の論文から注やタイトルを除外し，本文を抽出した。独自にマクロを定義している論文[3]，および外部の文献の引用・参照を示す書式が存在しない論文は除外した。以上の処理の後，各文を抽出し，正例と負例への分類を行った。データの概要は以下の表 1 の通りである。

表 1: データの概要

文書数	文数	出典明示文	非出典明示文
435	100662	4937	100662

2135 文は外部の文献の引用・参照が明示されているとして実験対象から除外した。これは出典明示文には含まれていない。モデルの訓練，テストはデータ，本研究では文をランダムに並び替え，そのうち，70%を訓練データ，残りの 30%をテストデータとして抽出して行っている。

次に手法についてであるが，本研究ではパーセプトロン，ロジスティック回帰，サポートベクトルマシン (以下，SVM，線形分離を行うものと動径基底関数 (RBF) を用いたカーネル法によるもの)，ナイーブベイズ，決定木，ランダムフォレスト，ニューラルネットワーク (MLP)，Adaboost，K 近傍法を使用した[4]うち，本論では実験において安定した性能を示した RBF を用いた SVM を使用した結果のみ記載する。

[2]http://anlp.jp/resource/journal_latex/index.html (参照: 2017-5-8)

[3]pandoc での変換に対応できなかったが，少数なので除外した。pandoc. available from: http://sky-y.github.io/site-pandoc-jp/users-guide/ (2017-12-27)

[4]scikit-learn 0.19. available from: http://scikit-learn.org/stable/ (2017-12-27)

3.2 素性

本研究は論文における言語表現の形式に着目しており，最小単位を文としているため，素性は文レベルのものと文間関係レベルのものを考慮した。

文レベルの素性は，その文の特徴を基に与えた素性である。これには以下の 4 つのクラスがある。それぞれに含まれる素性について説明する。

Expression

- EndSent: 文末表現が「ている」,「ていない」,「た」，現在形，その他の 5 種のうち，どれに当てはまるか
- LenSent: 文の長さ
- LenTerm: 文に含まれる語句の長さの平均
- TypeTokenRatio: 延語数と異なり語数の比率

Grammar

- ProperNoun: 固有名詞の有無
- RatioPoS: 各品詞の割合
- NgramPoS: 品詞についての n=2 の n-gram を作成し，前文との類似度

TermFreq

- TermFreqMean: 出現する名詞の出現頻度 (tf-idf) の平均
- TermFreqDeff: 出現する名詞の出現頻度 (tf-idf) の平均を前文において出現する名詞の出現頻度 (tf-idf) の平均で割った値
- AverWordFreqClass: 最も出現回数が多い語句の出現回数で文に含まれる各語句の出現回数を割り，それを平均した値

Similarity

- TotalSim: 出現する名詞の出現頻度の平均
- SimForePost: 出現する名詞の出現頻度の平均を前文において出現する名詞の出現頻度の平均で割った値

文レベルにおいては EndSent は文末表現が「ている」もしくは「た」の時に外部を引用・参照していた内容を記述している場合が多い [11] ことから，LenSent，LenTerm，TypeTokenRatio，Grammar は Eissen [10] らが剽窃の自動検出の際に文の書き手の統語的特徴を示すとされ，剽窃の自動検出において成果を示していることから使用した。TermFreq は外部を引用・参照している文の内容は筆者の主な主張と同内容であるべきではないとされているので [12]，外部を引用・参照している文は論文の中で中心的な役割を担うことはで

きず，そのため出現する語句は頻度が低いことが予想されることから使用した。Similarity は文の特徴を示し，外部の文献を引用・参照して記述された文は，元は文が属する論文の筆者とは異なる者が記述した文であることから使用した。

文間関係レベルの素性は同一文書中の文同士の関係から抽出される素性のことであり，これには以下の 3 つがある。

Discourse

DisTerm: 文頭の接続詞の有無，あればその種類

DisTermForePost: 前後の文の接続詞の有無，あればその種類

DisTermInhe: 文頭に接続詞を含む時はその接続詞の種類，そうでない場合は前文の値

TopicModel

TopicCos: 前文とのトピックモデルにおける Cosine 類似度

TopicMutNum: 前文と共通するトピックの数

Foregoing

ForegoingTerm: 文に登場する語句が既出か否か

ForegoingSub: 文の内容が内部であるもしくは既出か

Discourse は外部の文献の引用･参照は筆者の論を補助する役割を持ち [12]，理由や具体例を示す論理関係の存在は外部の文献の引用・参照の有無を示す指標となることが予測できることから使用した。TopicModel は Kaplan ら [3] が論文の一貫性を示す指標として使用していたものである。外部を引用・参照している文の内容は補助的な位置付けであるべきとされているので [12]，著者が独自に執筆した文に連続した場合，前文との内容の一貫性が相対的に低くなると予想されることから使用した。Foregoing は一度論文に登場した内容については，再度外部の文献の引用・参照を示す必要はないため文に登場する語句，また文の内容が既出か否かは外部の文献の引用・参照を示すか否かを分ける指標となりうると考えられることから使用した。

3.3 結果

以下，実験の結果を表 2，3，4 に記す。BaseLine として分類に際して正例，負例を等確率で各文に与えた際の精度，再現率，F1 値を使用した。本研究では各クラス，各レベル，全ての素性を使用した実験を行い，その後，文間関係レベルと比べて良い性能を示した文レベルの素性を個別にもしくは組み合わせて実験を行った。

表 2: 各クラスを単独で使用した場合の結果

素性	精度	再現率	F1 値
BaseLine	0.0493	0.4958	0.0896
Expression	0.0775	0.6056	0.1413
Grammar	0.0758	0.6632	0.1388
TermFreq	0.0695	0.6220	0.1251
Similarity	0.0670	0.6399	0.1214
Discourse	0.	0.	0.
TopicModel	0.0552	0.6104	0.1013
Foregoing	0.0538	0.9327	0.1017

表 3: 各レベルの素性を全て使った場合の結果

素性	精度	再現率	F1 値
全てのクラス	0.1070	0.6268	0.1828
文レベル	0.1050	0.6467	0.1806
文間関係レベル	0.0604	0.5836	0.1095

表 4: 文レベルに含まれる各素性の結果

素性	精度	再現率	F1 値
EndSent	0.0508	0.9814	0.0967
LenSent, LenTerm, TypeTokenRatio	0.0804	0.6015	0.1418
ProperNoun	0.0572	0.7572	0.1064
RatioPoS	0.0725	0.6886	0.1311
RatioPoS, NgramPoS	0.0747	0.6755	0.1345
EndSent, ProperNoun	0.0583	0.7427	0.1081

モデルの評価には，対象とするデータが不均衡データのため精度，再現率，および F1 値を用いる。

4 考察

表 2 より，Discourse 以外の素性クラスは単独でベースラインよりも良い性能である。一方，表 3 をみると，各レベルで素性クラスをまとめることにより，単独で用いるよりも高い性能を発揮し，全素性を使うとさらに性能が向上する。したがって，提案した素性は組み合わせて使用することでより高い効果を発揮するといえる。

ついで，表 3 を見ると TermFreq と Similarity を使用した時に文間関係レベルの素性を単独で使用し

た時よりも良い性能であることがわかる。加えて表 4 を見ると RatioPoS，NgramPoS を使用した時の方が ProperNoun を使用した時よりも良い性能であることと，LenSent，LenTerm，TypeTokenRatio を組み合わせて使用した時の方が EndSent を使用した時よりも良い性能であることがわかる。

したがって，文の筆者ごとに異なる統語的な特徴や文に含まれる語句の重要度が有効であるといえる。

5 おわりに

本研究での課題は外部の知識の考慮と文間関係の把握である。外部への引用・参照の必要の有無は外部に存在する知識との関係で決定されうることは容易に想定されるが，本研究では考慮されていない。これについては関連研究の節で述べた剽窃の自動検知で使用されている他の文献との類似度の使用を考えている。次に文間関係の把握については，本研究では文頭の接続詞を考慮する形でしか考慮できなかったが，これまでに提案された日本語における談話構造を把握する手法 [13] の導入を考えている。

参考文献

[1] 高松正毅 (2016) “引用のない論文はない—学術論文における先行研究レビューと引用をめぐって—” 『日本地域政策研究』 vol. 17, p. 74–79.

[2] 中村かおり, 近藤裕子, 向井留実子 (2016) “アカデミックライティングにおける不適切な引用文の分析と課題” 日本語教育国際研究大会, Bali.

[3] Dain Kaplan, Takenobu Tokunaga, Simone Teufel (2016) “Citation block Detarmination using textual Coherence,” Journal of Infomation Processing, vol. 24, no. 3, p. 540–553.

[4] Awais Athar, Simone Teufel (2012) “Detection of Implicit Citations for Sentiment Detection,” Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, p. 18–26.

[5] Vahed Qazvinian, Dragomir R. Radev (2010) “Identifying non-explicit citing sentences for citation-based summarization,” Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, p. 555–564.

[6] Bela Gipp, Norman Meuschke (2011) “Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence,” Proceedings of the 11th ACM Symposium on Document Engineering, p. 249–258.

[7] Sidik Soleman, Atsushi Fujii (2017) “Plagiarism Detection Based on Citation Contexts,” 研究報告情報基礎とアクセス技術, vol. 124, no. 11, p. 1–5.

[8] Yaakov HaCohen-Kerner, Aharon Tayeb, Natan Ben-Dror (2010) “Detection of Simple Plagiarism in Computer Science Papers,” Proceedings of the 23rd International Conference on Computational Linguistics, p. 421–429.

[9] Efstathios Stamatatos (2011) “Plagiarism Detection Based on Structural Information,” Proceedings of the 20th ACM International Conference on Information and Knowledge Management, p. 1221–1230.

[10] Sven Meyer zu Eissen, Benno Stein (2006) “Intrinsic Plagiarism Detection,” Advances in Information Retrieval, vol. 3936, p. 565–569.

[11] 神門典子 “構成要素カテゴリを用いた情報メディアの構造の分析—言語表現に関する考察に基づく分析基準の再検討—” 『Library and Information Science』 vol. 31, 1993, p. 39–49.

[12] 近江幸治 『学術論文の作法—〔付〕小論文・答案の書き方—』 東京, 成文堂, 2011, p. 54.

[13] 梅澤俊之, 原田実 (2011) “センタリング理論と対象知識に基づく談話構造解析システム DIA” 『自然言語処理』 vol. 18, no. 1, p. 31–56.