引用・参照の必要性を規定する要素の探索的分析 ―文の要素に焦点を当てて―

渡邊 晃一朗 † 影浦 峡 ‡†
東京大学大学院 † 教育学研究科, ‡ 情報学環
kouichirou-watanabe495@g.ecc.u-tokyo.ac.jp, kyo@p.u-tokyo.ac.jp

1 はじめに

1.1 問題意識及び目的

引用･参照 (以下、「引用」とする) をいつ行うべきかについては、アカデミック・ライティングの教科書や大学のガイドブックなどで説明がなされ、一見すると理解されているように思われる。しかし個々の論文、レポートを対象とした時に、どの部分に「引用」が必要かを明確に判断できるレベルで、「引用」の必要性を規定する要素が何か、またその要素がどのように「引用」の必要性を規定するかを理解できているだろうか。またアカデミック・ライティングの分野では「引用」の必要性の基準について、それが曖昧であると指摘されている。そこでは剽窃の判断が個人によって変わるという問題 [1] や、アカデミック・ライティングの指導内容における課題 [2] が指摘されている。

以上の問題意識から、「引用」の必要性を規定する要素とその作用を、個々の論文において説明することができるレベルで明らかにすることを目的とする。

1.2 タスク

論文において、「引用」は出典を示す記号によって示される。本研究では出典を示す記号の有無は文単位で考え、文単位での分類実験を行う。本研究ではある文が与えられた時、その文の出典を示す記号の有無を判定するという分類問題において素性の有効性を比較することで、どのような要素がある文における「引用」の必要性に関わっているのかを明らかにする。本研究ではタスクにアプローチする最初の段階として、外部情報の存在を考慮せずにタスクに取り組む。

2 「引用」とは何か: タスクの位置付け

ここでは本研究、また本タスクの「引用」の研究における位置付けを示すために、「引用」について既知の事柄である行為として、また記号表現としての「引用」、およびその目的を整理した上で、「引用」に対して何を問いうるのか、またそれに答えるために何を考慮する必要があるのかを整理する。

2.1 行為として、また記号表現としての「引用」とその目的

2.1.1 行為として、また記号表現としての「引用」

引用 (*Citation*) は『最新 図書館用語大辞典』[3] において “自己の著作物中に他人の著作物の一部を挿入して使用すること” と説明されている。しかし、学術論文においては間接引用が存在するため、他者の著作物がその文面を変更されて使用されることがあり、その場合、「引用」している文 (以下、引用文) は引用された文と記号表現として同一性を有さず、また引用符なども付されない。このように文として「引用」の存在が明示されない場合があるが、そのような場合でも学術論文における「引用」は出典を示す記号により明示的に記号表現として示される。つまり、記号表現を扱う限りにおいて学術論文における「引用」は出典を示す記号を付すこととみなされうる。これは出典を示す記号の欠如が剽窃と判断されるということからも明らかである。

2.1.2 「引用」の目的

「引用」の目的の 1 つとして、「引用」している側の文献が、その文献の記述する研究をその研究の該当する研究分野においてその体系の中に位置付けるということが挙げられる [4]。また「引用」する側の論文における “warrant” として、議論の展開において一定の役割を果たすとも言われている [5]。以上のことから、「引用」は科学的な手続きの 1 つとしてある研究分野の体系を形成するため、また議論においても一定の役割を担うために行われてきたことがわかる。加えて、そうした目的で「引用」が行われる時、多くのガイド

ブックで説明されているように、「引用」は著作権法上また研究倫理上、剽窃と判断されることを避けるために適切な形で行う必要がある。

2.2 「引用」についての問

ここでは「引用」に関しての問を考え、本研究の枠組みを整理する基礎とする。

2.1 項を踏まえると、個別の論文を対象とした時、以下の疑問を持ちうるであろう。例えば、どのような文を書いた時に「引用」するのか。常識と考えられる事柄は「引用」する必要はないと言われるが、ある事柄が常識か否かはどのように判断すればいいのか。また、どのような時に、「引用」による根拠が必要なのか。こういった問について、「引用」についての研究は「いつ引用するのか」と「いかなる理由で引用するのか」に答を与えるものであるとされた [5]。しかし、「引用」を始める時、「引用」をしている時、「引用」を終えた時をそれぞれ考えると、以下の問をあげることができる。

1. いつ「引用」が必要なのか、つまり、どのような知識を記述する言語表現に「引用」の必要性があるのか
2. 「引用」する時にどのように「引用」するのか、つまり、「引用」に伴う言語表現はどのように決まるのか
3. 「引用」することでどのような議論が可能になるのか、つまり、「引用」によって「引用」した論文全体の言語表現がどのようになるのか

以上の問について、1 つ目の問は Amsterdamska らの問と共通し、「引用」による議論の展開に焦点を当てている点で 3 つ目の問も実質的に Amsterdamska らの問と共通する [5]。しかし、2 つ目の問については Amsterdamska ら [5] の問と共通しない。

1 つ目の問はある文が記述する知識の位置付けを問うものである。この問が問うものは、言語表現そのものではない。そうではなくて、ある言語表現で示された知識に「引用」の必要性が存在するか否かを問うものである。

次に 2 つ目の問について、間接引用においては引用文は「引用」された文献に含まれる文字列と言語表現としては別のものである。これは、剽窃における一つの形態である “patchwriting” は、「引用」される文献における表現と過度に類似した表現が「引用」する文献の側に存在した場合、出典を示す記号を付していたとしても指摘されることからも明らかである [6]。最近では自然言語処理において引用文を利用した要約の手法が提案されている [7] ように引用文は「引用」された文献の要約であると位置付けることもできるが、その一方で引用文は「引用」された側の文献の中心となる考えを十分に反映していないとする研究も存在する [8]。このように、引用文は単なる「引用」された文献の要約と考えることはできず、それがどのように他の文献を「引用」しているか、つまり、「引用」に伴う言語表現がどのように決まるのかを考察する必要が存在する。

3 つ目の問について、この問は「引用」の機能を問うている点で「引用」に関する問であるように思われるが、「引用」という行為は他の文献からの知識の導入であって、その導入された知識によってどのように議論を進めることができるかは「引用」に限らない一般的な議論の方法において考えられることであり、「引用」の領域に限定した問ではない。

2.3 「引用」に関係する要素

では、「引用」についての問を考える時、どのような要素を考えることが可能なのか、また必要なのであろうか。2.1.2 での「引用」の目的を考慮すると、「引用」を考える際に大きくは 2 つの要素が関係すると考えられる。まず、ある研究分野の体系に研究を位置付けるという目的からは、引用文における外部情報が関係する。ある研究分野の体系に研究を位置付けることを考慮した場合、その研究分野において、他の文献が記述する知識がどのように組織され、どのように引用文が属する文献の記述する研究と関係するかが重要となる。加えて、剽窃を避けることを考えた場合、ある文が記述する事柄が外部に存在するか否かが重要である。

2 つ目は文の要素である。議論においても一定の役割を担うことを考慮した場合、引用文の有する議論の側面における意味内容、またそれが含まれる文献内における位置付けが関係することは予想される。

2.4 本研究のタスクの位置付け

以上のことを踏まえると、本研究で取り組むタスクは 2.2 項で挙げたうちの 1 つ目の問に焦点を当てたものであると位置付けられる。その中でも、2.3 項で挙げた中では文の要素を考慮したものであり、つまり、どのような意味の文が、もしくはどのように文献の中

で位置付けられる文に「引用」の必要性が存在するのかを問うものであると考えられる。

3 実験内容

3.1 対象データと分類器

対象としたデータは言語処理学会論文誌 LaTeX コーパス[1]である。これから論文の本文を抽出し、各文に対して出典を示す記号の有無でラベリングした。実験は文単位で分類を行い、出典を示す記号が付された文についてはその部分を削除したものを分類器に与える。このデータの概要は、文書数 430、総文数 107,898、出典を示す記号が付された文数 7,059、出典を示す記号が付されていない文数 100,839 である。

分類器には線形の Support Vector Machine (SVM)[2] を使用する。これは予備実験において他の分類器よりも高い性能を示したためである。

実験に際しては、10 分割交差検定を行った。

本研究では評価指標として Precision、Recall、F1-score を使用するが、結果の比較においては Precision を重視する。それは、本研究の目的が引用の必要性を明らかにすることであり、Precision が低いこと、つまり引用がない文を多く引用が必要と判断することは分析の観点から避けるべきことだからである。

3.2 素性

本研究では、文の要素を考慮する素性を比較する実験を行う[3]。文の要素としては、対象とした文そのものの意味と、その文の位置付けが考慮できる。対象とした文の位置付けについては、意味的な位置付けと構造的な位置付けを考慮できる。意味的な位置付けとしては、論文における中心的な意味を担うか否かを考慮する。構造的な位置付けとしては、前後の文との関係を考慮できる。そこで、前後の文の情報を考慮する。

本研究で考慮する素性を以下に述べる。本研究では、分散表現として学習済みの分散表現[4]を使用した。

Word Embedding

Word Embedding Sentence (WE-S): 対象とした文に含まれる語句の分散表現の線形和

Word Embedding Noun (WE-N): 対象とした文に含まれる名詞の分散表現の線形和

Word Embedding Verb (WE-V): 対象とした文に含まれる動詞の分散表現の線形和

Word Embedding Conjunction (WE-C): 対象とした文に含まれる接続詞の分散表現の線形和

Similarity

Similarity Abstract (Sim-Abst): 対象とした文が含まれる文献の抄録に含まれる語句の分散表現の線形和と対象とした文に含まれる語句の分散表現の線形和の Cosine 類似度

Similarity Other Sentences (Sim-OS): 対象とした文が含まれる文献の各文に含まれる語句の分散表現の線形和それぞれと対象とした文に含まれる語句の分散表現の線形和の Cosine 類似度の平均

Previous and Next Sentence

Previous Sentence (Prev-S): 対象とした文の前文に含まれる語句の分散表現の線形和

Next Sentence (Next-S): 対象とした文の次の文に含まれる語句の分散表現の線形和

Previous Noun (Prev-N): 対象とした文の前文に含まれる名詞の分散表現の線形和

Next Noun (Next-N): 対象とした文の次の文に含まれる名詞の分散表現の線形和

Previous Verb (Prev-V): 対象とした文の前文に含まれる動詞の分散表現の線形和

Next Verb (Next-V): 対象とした文の次の文に含まれる動詞の分散表現の線形和

Previous Conjunction (Prev-C): 対象とした文の前文に含まれる接続詞の分散表現の線形和

Next Conjunction (Next-C): 対象とした文の次の文に含まれる接続詞の分散表現の線形和

[1] 言語処理学会論文誌 LaTeX コーパス. 入手先 URL: http://anlp.jp/resource/journal_latex/index.html (参照: 2017-5-8)

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

[3] 本研究で扱う分類問題と同様の分類問題を Sugiyama ら [10] が行なっているが十分な成果を示すには至っていない。

[4] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, Armand Joulin (2018) "Advances in Pre-Training Distributed Word Representations," Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), p. 52–55.

表 1: 10 分割交差検定の平均を示した実験結果

素性	Precision	Recall	F1-score
WE-S	0.189	0.603	0.287
WE-S, WE-N	0.240	0.492	0.321
WE-S, WE-V	0.208	0.478	0.284
WE-S, WE-C	0.259	0.367	0.302
WE-S, WE-N, WE-C	0.275	0.376	0.309
WE-S, Sim-Abst	0.182	0.584	0.277
WE-S, Sim-OS	0.190	0.596	0.288
WE-S, Prev-S∗	0.192	0.601	0.291
WE-S, Next-S∗	0.191	0.602	0.290
WE-S, Prev-S, Next-S∗	0.191	0.593	0.289
WE-S, Prev-N, Next-N∗	0.191	0.599	0.289
WE-S, Prev-V, Next-V∗	0.185	0.605	0.282
WE-S, Prev-C, Next-C∗	0.191	0.593	0.288

4 実験結果、考察、結び

表 1 に実験結果を示す。この表 1 が示すように、特定の品詞の語句を考慮した時以外では対象とした文に含まれる語句の分散表現の線形和を単独で使用した時と比較して性能を大きく向上させたものはなかった。

より具体的に見ると、以下のことが指摘できる。まず、WE-S と WE-N、また WE-C 使用時の結果から、対象の文の特定の品詞に着目した時には、対象とした文に含まれる名詞と接続詞を考慮することが有効であることがわかる。特に、対象とした文に含まれる名詞と接続詞を共に考慮した時に Precision は最大の値を示している。この理由としては接続詞によって因果関係が示されること、また「引用」の必要性を決定する文そのものの意味は主に名詞によって表現されるという可能性が考えられうる。次に、WE-S と Sim-Abst、また Sim-OS 使用時の結果から、抄録や他の文との類似度は有効性を示さないことがわかる。ただし、これについては類似関係のみを考えており、他の関係については考慮できていない。最後に、∗ で示した結果から、前後の文の考慮は有効性を示さないことがわかる。ただ、これについても前後の文との論理的な関係を直接考慮できたわけではない。

以上、本論では、「引用」についての問と、それにどのような要素が関係するのかを整理した上で本研究の位置付けを示した。また、探索的に行った実験では対象とした文の名詞と接続詞が重要であることがわかったが、その理由の分析など今後の課題を示した。

参考文献

[1] Charlene Polio, Ling Shi (2012) “Perceptions and beliefs about textual appropriation and source use in second language writing,” Journal of Second Language Writing, vol. 21, p. 95–101.

[2] 吉村富美子 (2013)『英文ライティングと引用の作法—盗用と言われないための英文指導—』 東京, 研究社, p. 8.

[3] 図書館用語辞典編集委員会 (2004)『最新 図書館用語大辞典』 東京, 柏書房株式会社, p. 18.

[4] Charles Bazerman (1988) *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science,* University of Wisconsin Press, Madison, Wisconsin, p. 139.

[5] Olga Amsterdamska, Loet Leydesdorff (1989) “Citations: Indicators of significance?,” Scientometrics, vol. 15, p. 449–471.

[6] Rebecca Moore Howard (1995) “Plagiarisms, Authorships, and the Academic Death Penalty,” College English, vol. 57, no. 7, p. 708–736.

[7] Vahed Qazvinian, Dragomir Radkov Radev (2008) “Scientific Paper Summarization Using Citation Summary Networks,” COLING, p. 689–696.

[8] Marcelo Alves Ramos, Joabe Gomes Melo, Ulysses Paulino Albuquerque (2012) “Citation behavior in popular scientific papers: what is behind obscure citations? The case of ethnobotany,” Scientometrics, vol. 92, no. 3, p. 711–719.

[9] Myriam Herna Ndez-Alvarez, Jose Maria Gomez (2016) “Survey about citation context analysis: Tasks, techniques, and resources,” Natural Language Engineering, vol. 22, no. 3, p. 327–349.

[10] Kazunari Sugiyama, Tarun Kumar, MinYenKan, Ramesh Chandra Tripath (2010) “Identifying citing sentences in research papers using supervised learning,” International Conference on Information Retrieval and Knowledge Management (CAMP), p. 67–72.