Vol.2012-NL-209 No.2
2012/11/22

論文間参照情報のデータベース化に基づく参照タイプの同定

小出　寛史†1　　高橋　真也†2　　松本　惇†2
横堀　圭太†2　　韓　東力†1

学術研究において文献サーベイを行うことは必要であるが、研究に必要な文献を探索することは非常に時間がかかる。そのため論文関の関連性を明確にすることにより文献サーベイの効率化を図ることができると考えられる。既存研究ではルールベースの手法や機械学習により論文間の参照タイプを同定する手法が提案されているが、参照タイプの分類数か訓練データ数が不十分といったような問題が残っている。本研究では論文中の文献参照情報をもとにデータベースを構築した後、論文間の参照・被参照関係を見やすくするためのツールを構築する。さらに、論文間の参照タイプを機械学習により事前に定めた９つのタイプに分類する手法を提案する。

Citation Classification Using Information around the Citing Spot

HIROSHI KOIDE†1　SINYA TAKAKHASHI†2　JUN MATSUMOTO†2
KEITA YOKOBORI†2　DONGLI HAN†1

Searching appropriate papers is an important but difficult task in the process of scientific research. Knowing the specific relationship between a certain paper and another paper it cited is convenient and avoids unnecessary time consumption. Previous studies proposed methods based on rules or machine learning to classify the citing reasons, whereas negative factors like inefficient learning data or lack of reasonable citation classifications still remain. In this paper, we first build a database based on the information extracted from papers around the citing spots. Then we try to make the citing-cited relationships between papers visualized through an interface. Finally, we propose a machine learning based method to classify the citations into 9 citing reasons that we have defined in advance, and make some experiments to verify the effectiveness of our method.

1. はじめに

既存文献への参照が学術論文の執筆において極めて重要な一環となっている。参照理由にはさまざまなものがある。執筆中の論文（以下では起点論文と呼ぶ）と深く関わっているものもあれば、研究背景の一つとして紹介しているだけのものや起点論文の正当性を謳うために複数の文献を同時に列挙しているものもある。前者の参照理由においては、さらに「アルゴリズムの参照」や「実験データの利用」、「研究手法の対比」など多岐に分かれている。

起点論文における他の論文への参照理由を手動で判定しようとする試みが社会科学において以前からたくさん存在していた[1][2][3][4][5][6]。文献情報学においても参照理由の同定に様々な利用価値がある。たとえば、大きな学術論文データベースでは学術論文の価値として「影響因子」（impact factor）を計算している場合が多い。ある論文の「影響因子」の計算は他の論文がその論文をどれだけ引用（参照）しているかにより機械的に行われている。すなわち、学術的な理由以外の原因で参照されたケースも全体の被引用回数にカウントされてしまうため、本当の影響度が求まらない。そこで、起点論文からある論文を引用（参照）した理由を機械的に把握することが可能になれば、総引用数にカウントするかどうかの判断をその理由により行うという仕組みを導入することでより正確な「影響因子」を求めることができる。さらに、論文間参照理由の自動判定の身近な応用の一つとして、論文サーベイにおける利用も挙げられる。研究者があるテーマについて研究を始める際に、そのテーマに関連して文献サーベイを行うのが一般的である。文献サーベイに使用される手法の 1 つに、起点論文を 1 つ選定し、その論文が参照している論文（以下「被参照論文」と呼ぶ）や、さらに被参照論文が参照している論文という順にサーベイの対象を広げていくという手法がある。しかしながらこの方法では、研究に必要な文献を探索するのに非常に時間がかかる。もし起点論文と被参照論文の関連性をある程度把握することができれば、文献サーベイに費やされる時間や労力を大幅に減らすことが出来ると考えられる。

既存研究には起点論文と被参照論文の関連性を論文間の「参照タイプ」としてとらえ、それを「論説根拠型」、「問題点指摘型」と「その他型」の 3 種類に分類しているもの[7][8]や、機械学習により論文間の参照タイプを分類するもの[9][10][11]がある。前者では論文間の参照タイプ数が少なく具体的な参照理由を把握することが難しいため、本研究の目標である文献サーベイの効率化を実現するのに不十分であると思われる。後者[9][10][11]のいずれも著者が独自

†1 日本大学大学院総合基礎科学研究科
Graduate School of Integrated Basic Sciences, Nihon University
†2 日本大学文理学部情報システム解析学科
Department of Comouter Science and System Analysis,
College of Humanities and Sciences, Nihon University

の観点でトレーニングデータとテストデータを選定しており、データ規模が小さい上、具体的な論文情報や学習データの加工方法などもほとんど非公開となっているため、再現性が極めて低いという大きな問題がある。また、既存研究のすべてが英語で書かれた論文を対象にしており、既存の手法が日本語において有効かどうかという実験分析が見当たらない。

このような背景に踏まえ、我々は論文間の参照タイプを従来の 3 種類から、「歴史、類似研究、理論、研究手法、実験・データ、結果」の 6 種類に細分化し、日本語の論文を研究対象として研究を進めてきた[12]。具体的には、参照個所、論文タイトル、文献種類と位置情報の 4 つの観点から論文間の参照タイプを判定している。しかし、参照タイプを細分化した際の基準が明確ではなく、参照タイプの判定に使用したアルゴリズムにおける加点方式も主観性が強かったため、十分に信頼できる結果を得ることができなかった[12]。

本研究では今までの経験成果を踏まえ、より実用性と信頼性の高いシステムを構築することを目標とする。具体的には、論文間の参照タイプを Teufel [9]が提唱した 12 種類のうち、日本語に対応できると判断した「weak」、「coco」、「cocoGM」、「cocoRo」、「PBas」、「PUse」、「PModi」、「PMot」と「Neut」の 9 タイプへと再構成する。各タイプの詳細は表 1 にて示す。また、既存研究におけるデータ量不足の問題に対してはデータベースを作成することで対応し、さらにデータベースから論文間参照関係の自動抽出機能を実現することでシステムの応用性を高める。最後にデータベースから取得した情報を機械学習に用いることにより参照タイプの同定を試みる。

表 1　各参照タイプの説明

weak	既存研究を取り上げ、その既存研究を他の研究と比較することなく問題点を挙げている論文
coco	既存研究を取り上げ、その既存研究と起点研究を比べたうえで、問題点を指摘している論文
cocoGM	取り上げた研究に対して、既存研究・起点研究との目的や方法を比較するために参照している
cocoRo	既存研究を取り上げ、他の既存研究・起点研究との結果を比較するために参照している
PBas	起点研究を行う上で、取り上げた既存研究をもとに行われている研究を参照している
PUse	既存研究で作られたツール・アルゴリズム・データを起点研究で使用した論文
PModi	既存研究で作られたツール・アルゴリズム・データを起点研究でアレンジしたのちに使用した論文
PMot	起点研究を行う際に、取り扱う手法やデータが既存研究にて有効性が示されていると明記されている論文
Neut	取り上げた既存研究に対して紹介するだけにとどまり、中立的な記述がされている論文

2. 論文間参照情報のデータベース化

ほとんどの既存研究ではルール生成や訓練データ作成のために独自のデータコレクションを構築しているが、データ量が少ないか特殊なドメインに限定されており再現性の低いものとなっているケースが多い。本研究では、一定の規模を保ち、かつ広く公開されている論文データソースを選定し、それをデータベース化していくことにより研究成果の汎用性と再現性を大幅に向上させることを目標とした。

2.1 データソースの選定

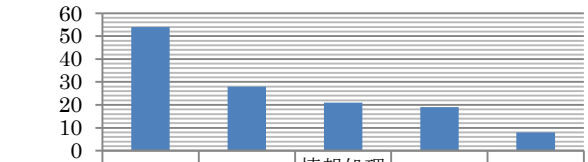
本研究では、「言語処理学会年次大会発表論文集」(以下「NLP」と記述)をデータベースのソースとして選定している。「NLP」に収録されている論文を起点論文として、後述の条件を満たす刊行物に収録されている論文を参照した箇所を抽出し、そこに含まれている参照情報を項目ごとにデータベースに格納していく。

探索の起点とした論文（起点論文）は、以下の条件を満たすものとする。

- 「NLP」の最近 9 年分（2004 年~2012 年）に収録されているもの
- 論文の本文が日本語で書かれているもの

「NLP」の 2003 年までの論文は紙媒体のものしかなく、電子作業を行うことが困難なため、今回のデータベースから除外している。

被参照論文の範囲を決めるにあたり、事前に起点論文の集合から 1 年ごとに 10 篇ずつの論文をランダムに抽出し、参考文献の収録されている刊行物を調査したうえ、論文集や論文誌に対してのみ集計を行った。図 1 は事前調査の結果のうち、出現回数で上位 5 つの刊行物をグラフにしたものである。



	言語処理学会年次大会	自然言語処理	情報処理学会・自然言語処理研究会報告	情報処理学会論文誌	電子情報通信学会論文誌
出現回数	54	28	21	19	8

図 1　事前調査の結果

この調査結果に基づき以下の条件を満たすものを実際に「被参照論文」として抽出した。同じ被参照論文でも起点論文の著者によって書式が完全に異なっていたり、著者名や論文タイトルなどの書誌情報が間違っていたりするので、同一の論文がシステム上で複数の ID を持たないように人間による確認を徹底した。

- 集計結果のうち上位 5 位に含まれる以下の刊行物に収録されているもの
  - ➢「言語処理学会年次大会発表論文集」
  - ➢「自然言語処理」
  - ➢「情報処理学会・自然言語処理研究会報告」
  - ➢「情報処理学会論文誌」
  - ➢「電子情報通信学会論文誌」
- 発行年が 2000 年以降のもの
- 論文の本文が日本語で書かれているもの

2.2　データベースの構成項目

今回作成したデータベースが 2 種類の情報から構成されている。一つは論文中にある被参照論文を引用している箇所およびその前後にあるテキスト情報であり、もうひとつの情報は論文のタイトルやセクションの総数など論文全体に関する情報である。主に以下の項目から構成されている。

- 論文番号：各起点論文と被参照論文を唯一に識別できる通し番号である。
- 論文のタイトル：各起点論文と被参照論文のタイトルである。ここでは正確な情報を得るために、J-GLOBAL[13]やCiNii[14]を参考にした。
- 論文の著者：各起点論文と被参照論文の著者である。ここでも正確な情報を得るために、J-GLOBAL や CiNii を参考にした。
- 参照文：文献参照の印が含まれる文である。
- 前文：一定の条件を満たす参照文の前の文である。
- 後文：一定の条件を満たす参照文の前の後の文である。
- セクション総数：起点論文に含まれるセクションの総数である。
- セクション名：参照個所（前文＋参照文＋後文）が含まれるセクション（章・節）の見出しである。
- 被参照論文発表年：被参照論文が収録されている刊行物が発行された西暦の年号である。
- 参照範囲の種類：参照個所が「本文」や「脚注」、「セクション名」のいずれに含まれているかを表す。
- 非本文参照個所：参照範囲が本文以外であった場合の該当個所である。
- 箇条書き：参照個所が箇条書きに含まれているか、また参照個所の前後に箇条書きが存在するかを表す。

起点論文から参照箇所を抽出して論文自体の情報とともに項目ごとにデータベースに格納していく。全項目のうち、論文タイトルやセクション総数など現時点で利用されていないものもあるが、今後利用していく予定である。

2.3　データベースの利用

データベースは「言語処理学会全国大会論文集」に収録されている論文を中心に参照情報をまとめたものなので、論文間の参照・被参照関係に関連する各種のツールやアプリケーションを作成することが可能である。

図 2　論文サーベイ補助ツールの検索画面

図 3　論文サーベイ補助ツールの参照関係表示画面

図 2 と図 3 は我々がデータベース検索を基に試作した論文サーベイ補助ツールの検索画面と論文間参照関係の表示画

面である。このツールは指定された起点論文から関連性のある論文を探索する際に有用であるが、論文間の参照タイプが不明のままではツールとしての有用性も限定されてしまう。今後は論文間参照タイプの分類結果を加えることにより論文サーベイの効率向上を目指したい。以下第３章と第４章ではこのデータベースに格納されている情報を用いて、機械学習により論文間参照タイプを分類する手法を提案する。

3. 機械学習に用いる学習データ

本研究ではデータベースの各レコードから、参照文、参照文前文と参照文後文を学習範囲として抽出し、参照タイプごとのuni-gramデータとbi-gramデータを作成している。そして、被参照論文ごとの参照タイプは人手で付与している。参照タイプの付与は著者のうちの４人による共同作業で行った。参照タイプ決定の際に意見が分かれた場合には、見解が違った理由をすべて挙げたうえ、その１つ１つについて討論を重ねて各自の意見を尊重しながら全員が納得するまで続けた。それでも決まらない場合には、多数決により決定する。仮に２対２で意見が対立してしまった場合は第三者の意見を聞き、再度討論し５人で判定を行う。

この方法を採用するのは参照タイプの分類数が多く、どちらが正解なのかを９つの選択肢から判定することは人間にとってもとても難しいことが理由である。安易に正解を付与してしまうと、機械が正確に学習できなくなることを最大限に防いでおきたいと考えていた。このような方式を導入した結果、大変多くの作業時間が参照タイプの正解付与に費やされ、第５章でも述べられるように現時点ではたった４００個の学習データしか完成せず、データ数に大きく依存するという機械学習の実験結果に大きな影響を与えてしまっている。

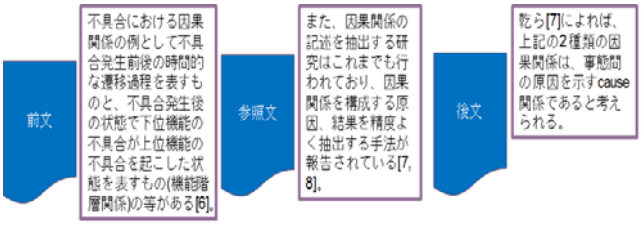


図４　参照箇所の例

参照箇所は前文、参照文と後文に分かれ、図４に示されるような形となっている。図４に示した例は文献[15]より抽出したものを使用している。前文には参考文献番号が付与されているが、参照文でも参考文献番号が含まれているので、前文の参考文献番号を無視する。そして後文は前文と同様の扱いで参考文献番号は入っていても無視する。各文章を形態素解析器Mecab[16]を使用して解析させ、その結果から「名詞」、「動詞」と「形容詞」の基本形を抽出する。抽出した単語列からuni-gramとbi-gramデータを作成し、学習データの集合に保存する。参照文からbi-gramデータを作成する流れを図５の通りになる。ここで使用している例は図４にある参照文である。

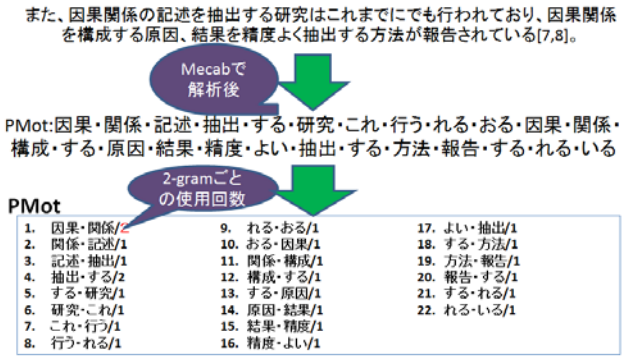


図５　参照文から bi-gram 作成の流れ

4. 機械学習による参照タイプ判定

論文間の参照タイプを判定するには、ナイーブベイズ法という機械学習の手法を用いてみることにした。式(1)においては、「cat」が９タイプのカテゴリである「weak」・「coco」・「cocoGM」・「cocoRo」・「PBas」・「PUse」・「PModi」・「PMot」・「Neut」のいずれかを表し、「doc」は通常文章のことを表すが、ここでは参照文およびその前文と後文のことを表している。

$$P(cat|doc) = \frac{P(cat)P(doc|cat)}{P(doc)} \quad (1)$$

$P(doc)$が各カテゴリに共通なため、$P(cat)P(doc|cat)$が最大となるカテゴリを求めればよい。P(cat)は学習データのうち各参照タイプが占める割合で求めることができる。P(doc|cat)はある一つの参照タイプが与えられた際に、「doc」によって表された文章が参照箇所に現れる確率を表し、式(2)と式(3)のように近似される。ここでは、式(2)ではuni-gram モデルを、式(3)では bi-gram モデルを使用している。

$$\begin{aligned} &P'(doc|cat) \\ &= P(word_1 \cap word_2 \cap \cdots \cap word_{k-1} \cap word_k|cat) \\ &\approx \prod_i P(word_i|cat) \end{aligned} \quad (2)$$

$$\begin{aligned} &P''(doc|cat) \\ &= P\big((word_1word_2) \cap (word_2word_3) \cap \cdots \\ &\qquad \cap (word_{k-1}word_k)|cat\big) \\ &\approx \prod_i P(word_{i-1}word_i|cat) \end{aligned} \quad (3)$$

さらに、$P(word_i|cat)$と$P(word_{i-1}word_i|cat)$はそれぞれ式

(4)と(5)のように求められる。

$$P(word_i|cat) = \frac{T(cat,word_i)}{\sum_{word\in V,} T(cat,word)} \quad (4)$$

$$P(word_{i-1}word_i|cat) = \frac{T(cat,word_{i-1}word_i)}{\sum_{word'word''\in V} T(cat,word'word'')} \quad (5)$$

$P(word_i|cat)$は uni-gram の条件付き確率であり、あるカテゴリ cat が与えられた際に$word_i$がどれくらい出てきやすいかを表す。V が訓練データ中の全 uni-gram の集合を表し、$T(cat,word)$はカテゴリ cat に単語$word$が出てきた回数である。同様に、$P(word_{i-1}word_i|cat)$は bi- gram の条件付き確率である。V が訓練データ中の全 bi- gram の集合を表し、$T(cat,word'word'')$はカテゴリ cat に単語$word'word''$が出てきた回数である。参照タイプごとに上記の計算を繰り返して行い、すべての確率値から最も点数の高かった参照タイプを結果として出力する。

5. 実験と評価

本節では、提案された論文間参照タイプの判定手法の有効性を検証するために行われた評価実験の概要と結果を述べる。

5.1 基本実験

実験用のテストデータもデータベースから作成していく。学習データと同じように形態素解析器で解析させた上、「名詞」、「動詞」と「形容詞」の基本形から n-gram データを作成する。この実験では bi-gram モデルを採用しているため、bi-gram データのみを作成する。uni-gram モデルと bi-gram モデルの有効性の比較については 5.2 節で詳しく述べる。

実験はクローズドテストとオープンテストの 2 種類を行った。クローズドテストでは、論文間参照タイプの正解が付与されている 300 個のデータから算出されたパラメーターを用いて、同じ 300 個のデータを分類してみたところ、予想通り比較的に高い精度が得られた（表 2 の All:Close を参照）。

オープンテストでは、学習データ数と分類精度の関係を把握するために、正解付きのデータ集合からそれぞれ 30 個、100 個、300 個と 3 回に分けて学習データを取得し、それぞれのパラメーターを算出した。テスト用のデータとして上記学習データと異なる 30 件のデータを使用した。正解付きの 400 個のデータ集合に、参照タイプの分布が均一ではなく、各参照タイプ間における標本データ数の格差が大きい現状を踏まえ、ここではデータ集合における参照タイプごとの割合を計算し、その割合に基づき学習データとテストデータをランダムに取得している。表 2 はクローズドテストとオープンテストの結果を示している。表中の 30-Le:Open、100-Le:Open と 300-Le:Open はそれぞれ 30 個、100 個と 300 個のデータを学習に用いるときの実験を表している。

表 2　オープンテスト・クローズドテストの結果

	テストデータ数	学習データ数	システムの正解数	正解率
All:Close	30	300	24	80.0%
30-Le:Open	30	30	8	26.3%
100-Le:Open	30	100	13	43.3%
300-Le:Open	30	300	16	53.0%

クローズドテストで得られた結果(80.0%)に対し、オープンテストでは 26.3%、43.3%と 53.0%となり、学習データ数の増大につれて分類精度が徐々に増えていく様子が伺える。分類数の多さ（9 タイプ）と学習データの少なさを加味して総合的に考えると、機械学習による日本語論文間参照タイプの自動判定が十分に可能な範囲内にあることが言える。しかし、現時点での最大 50%台という正解率が決して高いとは言えない。この問題を解決するカギとなるのは学習データの規模であることから、今後はいかに正解付与の効率を向上させるかを模索する必要がある。

提案手法が各参照タイプに対する個別の効果を分析するために、上記各実験における参照タイプごとの適合率と再現率を表 3 にて示す。

表 3　参照タイプごとの適合率・再現率

All:Close	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	100%	0%	83%	50%	67%	88%	0%	100%	100%
再現率	43%	0%	100%	100%	67%	100%	0%	100%	100%

30-Le:Open	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	33%	0%	83%	0%	67%	0%	0%	0%	0%
再現率	33%	0%	28%	0%	0%	0%	0%	0%	0%

100-Le:Open	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	33%	0%	50%	0%	67%	63%	0%	0%	50%
再現率	20%	0%	50%	0%	33%	63%	0%	0%	40%

300-Le:Open	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	100%	0%	50%	50%	33%	63%	0%	50%	50%
再現率	38%	0%	50%	50%	50%	63%	0%	100%	67%

一定数以上の学習データを確保することで比較的に安定していると思われる All:Colse と 300-Le:Open の共通点として、適合率はともに「weak」の精度が非常に高いというところである。しかしこの 2 つの実験の再現率を見てみると、「weak」の精度のみが 50%を下回っていることが分かる。これらの事実から、本手法により「weak」に分類されたも

のについては信憑性が高いことが分かる。それに対し、「PBas」に分類されたものの多くはおそらく「weak」に分類されるべきであるということもいえよう。このような知見を得ることは大変有意義である。たとえば、機械で分類した結果を論文サーベイに利用する際に信頼できる参照タイプとそうでないものを区別することにより、サーベイの効率を上げるとともに正確な関連文献に到達することも容易になる。

5.2 拡張実験

5.1 節では本手法の bi-gram モデルによる有効性について検証したが、本節では uni-gram モデルとの比較や参照箇所の範囲変動が分類結果に及ぼす影響について検証する。

表 4 は bi-gram モデルの「参照文のみ」、bi-gram モデルの「参照文＋前後文」、uni-gram モデルの「参照文のみ」と uni-gram モデルの「参照文+前後文」の 4 つの組み合わせで行った分類実験の詳細を示している。

表 4　n-gram モデルと参照範囲の変更に関わる検証結果

	テストデータ数	学習データ数	システムの正解数	正解率
bi-gram 参照文＋前後文	30	300	16	53.0%
uni-gram 参照文＋前後文	30	300	12	40.0%
bi-gram 参照文	30	300	12	40.0%
uni-gram 参照文	30	300	9	30.0%

実験結果からでは、基本実験で採用した bi-gram モデルと「参照文＋前後文」の組み合わせは分類精度が最も高かったことが分かった。さらにそれぞれの参照タイプごとの適合率・再現率を表 5 にて示す。

表 5　参照タイプごとの適合率・再現率

bi-gram 参照文+前後文	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	100%	0%	50%	50%	33%	63%	0%	50%	50%
再現率	38%	0%	50%	50%	50%	63%	0%	100%	67%

uni-gram 参照文+前後文	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	67%	0%	50%	50%	33%	38%	0%	0%	50%
再現率	33%	0%	27%	100%	25%	60%	0%	0%	67%

bi-gram 参照文	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	67%	0%	33%	0%	0%	100%	0%	0%	0%
再現率	50%	0%	100%	0%	0%	53%	0%	0%	0%

uni-gram 参照文	weak	coco	cocoGM	cocoRo	PBas	PUse	PModi	PMot	Neut
適合率	33%	100%	33%	0%	0%	63%	0%	0%	0%
再現率	50%	14%	100%	0%	0%	50%	0%	0%	0%

表 5 の結果から確認できることは、uni-gram モデルはほぼすべての参照タイプで適合率と再現率との両方において bi-gram モデルより低下していることである。これには、bi-gram モデルにおける特徴的な接続パターンが uni-gram モデルでは単語単位に分解され、うまく機能しなくなってしまっているのではないかと推測できる。

また、参照箇所の範囲変動については、やはり前後文の併用が参照タイプの分類精度の向上につながっていることが判明した。たとえば、「weak」や「PMot」は後文に重要な単語が存在することが多いので、後文を削除することで、精度が低下してしまうケースが起きている。たとえば、文献[17]から抽出した以下の参照箇所を考えよう。

- 前文：「そこで，ルールを用いて略語を生成することで， 前述のような複合語にも柔軟に対応できるようにする手法が考えられる.」
- 参照文：「このような研究としては村山らの研究[2]が挙げられる.」
- 後文：「[1][2]どちらも，略語の読みや漢字の音訓の情報が用いられてない.」

後文に「[1][2]どちらも,略語の読みや漢字の音訓の情報が用いられてない.」と書かれており、ここで初めて「weak」であると判断できる。参照文だけでは「列挙」もしくは「紹介」の印象しか取れず、「weak」として判定することは無理であろう。

6. おわりに

本研究では、論文サーベイの効率向上を目標に次のことを試みた。まずは既存の研究成果をもとに、論文間の参照タイプを日本語にも対応できるように「weak」、「coco」、「cocoGM」、「cocoRo」、「PBas」、「PUse」、「PModi」、「PMot」と「Neut」の 9 タイプへと再構成した。次に、広く公開されている論文データソースを選定し、それをデータベース化することにより、関連論文の検索や参照・被参照関係の抽出など論文サーベイの際に頻繁に手動で行う必要のある作業を自動で行えるようにした。最後に、この論文サーベイの効率をさらに向上させるべく、論文間の参照タイプを機械学習を用いて上述の 9 つのものに分類する手法を提案した。

実験結果での正解率を単純に数字として評価すると、決して満足できるものとは言えない。ただし、今回提案した手法における参照タイプ数の多さと学習データの少なさを考慮すると、日本語論文間参照タイプの判定を機械学習により行う可能性が十分にあるのではないということがいえよう。

今回の実験では、学習データとテストデータを合わせても 400 文しかない。これは 3 節でも述べたようにデータ数

に大きく依存するという機械学習の実験結果に大きな影響を与えてしまっている。そのため今後はデータ量を増やすことが急務であると考えている。また bi-gram モデルの使用方法にも改善の余地が残っていると思われる。bi-gram データを作成する際に「動詞＋形容詞」の組も登録されてしまっているなど、文法上ありえないような組み合わせが存在し、なおかつ現在使用している品詞が「名詞」、「動詞」と「形容詞」のたった 3 つしかないというのも問題となっているのではと考えている。今後は使用する品詞数を増やしながら、bi-gram データの作成方法を検討していく。また、現在未使用となっている「論文タイトル」や「セクション名」などの情報がすでにデータベースに登録されているので、こちらも参照箇所と同様に素性として機械学習に導入し、その効果を実験で評価していきたい。

参考文献

1) Garfield, E. Citation Index: Its Theory and Application in Science, Technology and Humanities. New York, NY:J. Wiley. (1979)
2) Weinstock, M. Citation Indexs. Encyclopedia of Ligrary and Information Science, 5, pp.16-40. New York, NY:Dekker. (1971)
3) Moravcsik, M. and Poovanalingan M. Some Results on the Function and Quality of Citations. Social Studies of Science, 5, pp.88-91. (1975)
4) Chubin, D. and Moitra, S. Content Analysis of References:Adjunct or Alternative to Citation Counting? Social Studies of Science, 5(4), pp.423-441. (1975)
5) Spiegel-Rosing, I. Science Studies: Bibliometric and Content Analysis. Social Studies of Science, 7, pp.97-113. (1977)
6) Oppenheim, C and Susan P.Renn. Highly Cited Old Papers and the Reasons Why They Continue to Be Cited. Journal of the American Society for Information Science, 29, pp.226-230. (1978)
7) 難波英嗣，神門典子，奥村学，「論文間の参照情報を考慮した関連論文の組織化」，情報通信学会論文誌，42(11)，pp.2640-2649. (2001)
8) 難波英嗣，奥村学，「論文間の参照情報を考慮したサーベイ論文作成支援システムの開発」自然言語処理，6(5)，pp.43-62. (1999)
9) Teufel, S, Advaith S, and Dan T. Automatic Classification of Citation Function. In Proceedings of EMNLP-06. (2006)
10) Teufel, S. The Structure of Scientific Articles –Applications to Citation Indexing and Summarization. CSLI Publications. (2010)
11) Radoulov, R. Exploring Automatic Citation Classification. Master thesis in University of Waterloo. (2008)
12) 小出寛史，橋本陽平，秦野福己，韓東力，「論文間参照タイプ判定の細分化に基づくサーベイ補助システムの構築」，言語処理学会第 18 回全国大会論文集，D4-6，pp.999-1002. (2012)
13) http://jglobal.jst.go.jp/
14) http://ci.nii.ac.jp/
15) 大森信行，森辰則，「不具合事例文からの製品・部品を示す語の抽出―語の実体性による分類―」，電子情報通信学会論文誌，J95-D(3)，pp.697-706. (2012)
16) http://mecab.sourceforge.net/
17) 岡田真，高橋幹浩，「漢字を中心とした複合語の略語の自動生成–音訓を考慮したルールを用いて–」，言語処理学会第 14 回全国大会論文集，C4-4，pp.787-789. (2008)