論文の論理構造における分野基礎用語に関する分析

内山 清子（国立情報学研究所）†

An Analysis of Domain-Specific Introductory Terms in Logical Structure of Scholarly Papers

Kiyoko Uchiyama (National Institute of Informatics)

１．はじめに

学術論文には分野で使われる専門用語や、著者が自分の研究を特徴づけるために作り出す独自の専門的な複合語などが数多く含まれる。これらの語は、分野の初心者にとって初めて遭遇する用語であり、その用語の意味を理解した上で論文を読み進めることが必要となる。しかし、分野初心者にとって、専門用語はすべて未知の語であり、どの語が重要な語であり最初に学ぶべき用語であるのか、また対象論文の研究内容の手がかり語となる用語であるのかなどの区別ができない。こうした専門用語に対して優先度を示すことにより、分野初心者が論文を読んで理解するための支援になるのではないかと考えた。そこで、本研究では、対象分野において最初に必ず学ばなければならない語、その分野における基礎的・必須である専門用語を分野基礎用語と呼び、分野基礎用語の選定方法を検討し、論文の論理構造における出現分布について分析を行う。

２．関連研究と分野基礎用語の位置づけ

従来、分野の用語（専門用語）については、専門性や重要性といった指標や関連用語収集などのテーマで研究がおこなわれてきた。まず、専門度を推定する研究として、専門外の人に対して専門用語を使わずに平易な用語に置き換えるために、専門外の人から見て比較的専門的な用語か、かなり専門的な用語かの 2 段階に分けたものがある。次に用語の重要性については、複合語を構成している単語の種類や隣接する単語の数をベースにして用語らしさとしての重要性を計算する手法が提案されてきた。また、関連用語収集として、複数の書籍に共通する用語をシードワードに設定して、その用語から関連する用語を自動的に収集する研究が行われた。この研究におけるシードワードは、本研究における分野基礎用語と一部一致している。

本研究において、論文を理解するために効率的な用語として分野基礎用語を位置づけるために、分野基礎用語から始まり専門性・難易度が高い用語に至る学習段階を想定し、自分の知識と目標レベルに応じた以下の 4 段階の知識・学習レベルを設定した。

(1) 一般、大学学部生、他の研究分野の研究者
(2) 大学学部生（その分野を専門に学びたい学生）
(3) 大学院修士（修士論文テーマ探し）
(4) 大学院博士、研究者（博士論文、研究論文テーマ探し）

まず、第一段階の一般、大学学部生、他の研究分野の研究者に対しては、分野知識を持っていないことを前提として、分野の全体的な概略を説明した解説文や理解しやすい教科書などに掲載されている用語を提示することが有効であると考える。次は学部 3 年生を想定して、卒業論文をまとめるために必要な分野の成り立ちも含めた詳細な概要を把握する必要がある。この段階では分野でよく利用される用語の理解を深めることが重要となる。第 3 段階は、大学院修士の学生が自分の修士論文のテーマを探すために、その分野の最新動向も踏まえて、興味のあるトピックに関する論文を読む必要性が出てくる。この段階では、論文を読むために、よく使われる用語に関連した専門性の高い用語を学ぶ。最後の段階では、大学院博士課程の学生や研究者として、過去の詳細な研究成果も含めた狭く深い

† kiyoko_at_nii.ac.jp

情報が重要となってくる。この段階では、分野の中の特定のトピックに対する専門家が使っている専門性と難易度の高い知識を持っていることが前提となる。本論文では、このような4つの知識・学習段階を考えた中で、分野初心者に重要な最初のレベル（1 と 2）に必要な用語を分野基礎用語と位置付ける。

３．分野基礎用語の選定

分野基礎用語を抽出する対象分野を自然言語処理とした。これまで実験的に自然言語処理の研究者一名に、分野基礎性の定義を説明した上で、重要な自然言語処理用語を 308 語選定し第 2 章で説明した 4 段階に分類してもらった。内訳は 1 レベルが 20 用語、2 レベルが 186 用語、3 レベルが 89 用語、レベル 4 が 13 用語である。この正解セットを用いた自動抽出手法として、一般コーパス（毎日新聞）と専門コーパス（情報処理学会自然言語処理研究会で発表された論文）を比較して、対数尤度比、カイ二乗値、イエーツ補正カイ二乗値等の各尺度の平均精度や、C-Value による用語らしさの検定をして実験を行ってきた。

結果的に出現頻度に基づいて分野基礎用語を自動的に抽出することは難しく、精度が低かった。また分析結果から、抽出時のスコアランキングで基礎性の度合いをつけることが現実的ではないことがわかり、正解セット自体を再検討することにした。理想的な選定方法としては、専門家に分野基礎用語を選定してもらい、多くの専門家が共通して選定した用語は分野基礎用語であると決定することが考えられる。しかし、専門家の意見を数多く集めることが難しいため、専門家の判断と同等であると見なせる客観的な基準を検討した。

そこで分野基礎用語を抽出する対象として、教科書、事典、論文の 3 種類を用意した。用語は、形態素解析を行い品詞が名詞あるいは名詞の連続であるものを抽出した。この 3 種類とも専門家が執筆したものであるため、これらのリソースから抽出した用語は複数の専門家の判断と同等であると考えられる。詳細は以下の通りである。

(1) 教科書 :「自然言語処理」分野の日本語の教科書 39 冊の目次に出現する用語（異なり語数 694 語）
(2) 事典 :「言語処理学事典」の目次に出現する用語（異なり語数 463 語）
(3) 論文 : 情報処理学会自然言語処理研究会で発表された論文のタイトル、抄録、キーワードに含まれる用語（異なり語数 13493 語）

教科書と事典の目次に出現する用語に着目した理由として、目次は初心者にもわかりやすい表題および学んでほしい用語を必ず著者が選定する、つまり著者が考える分野基礎用語は目次に含まれると考えたためである。この 3 種類のリソースに共通して出現する用語は 90 語であり、この 90 語を分野基礎用語と選定した。

４．論文の論理構造における分野基礎用語

論文の論理構造において、分野基礎用語がどのような出現パターンを示すのかを調べた。本論文における論理構造とは、「抄録」、「はじめに」、「関連研究」といった論文を構成している章に関連している意味のあるまとまりのことを指している。分析対象の論文コーパスは、分野基礎用語の選定時に利用した論文とは異なり、情報処理学会の論文誌に掲載された自然言語処理分野の論文の中から抄録で「実験」、「評価」、「精度」、精度の数値「%」などを含んでいる 100 論文を選んで論文コーパスとした。実験を扱った論文に絞ったのは、論理構造が比較的わかりやすく、論文の流れもある程度パターン化できるのではないかと仮定したためである。本論文では、論理構造の要素を「抄録」「はじめに」「実験」「関連研究」「おわりに」「その他」の 6 種類に分けた。「その他」は多くの場合、「関連研究」の記述の後から、「実験」記述の前までのまとまりを指している。

分析対象の論文コーパスを論理構造の要素に分割し、それぞれの要素の中における分野基礎用語の出現傾向を分析した。表 1 に出現頻度 100 以上の用語について、論理構造別の出現頻度を示す。なお、ある用語が別の用語の部分文字列となっている場合は（「文字」「文字列」など）、重複している数を差し引いて数えている。

表 1　論文の論理構造における分野基礎用語の出現頻度

	抄録	はじめに	実験	関連研究	おわりに	その他	合計
意味	54	231	360	93	49	561	1348
コーパス	64	160	448	79	59	330	810
品詞	33	116	339	28	34	361	550
辞書	30	103	310	43	36	239	522
日本語	40	136	182	45	38	225	441
生成	19	80	186	46	36	324	367
未知語	15	50	167	22	20	160	274
知識	28	101	88	16	37	105	270
言い換え	17	93	99	25	29	185	263
形態素解析	25	60	131	14	20	122	250
文字	7	26	89	24	9	65	220
シソーラス	9	39	89	34	16	39	187
アルゴリズム	16	43	73	14	17	252	163
照応	7	36	65	40	6	68	154
固有表現	5	14	108	4	7	91	138
形態素	6	27	87	2	10	124	132
文字列	8	22	82	13	4	186	129
クラスタリング	7	30	66	14	7	23	124
語義	7	35	57	7	16	45	122
機械学習	16	43	25	15	13	34	112
構文解析	7	45	35	13	9	46	109
機械翻訳	14	51	22	13	8	27	108
言語処理	16	59	9	12	10	31	106
決定木	11	34	40	11	9	74	105
言語モデル	10	19	53	12	10	70	104

最も出現頻度が高い「意味」は、一般的な文章にも使われる単語であるため、用語と見なすことが難しいが、実際に出現している文を読むと、「意味」が他の分野基礎用語と共起して出現するなど、重要な役割を果たしていることがわかった。自然言語処理において「意味」を理解することが目的でもあるため、本論文では用語と扱うことに意義があると考える。このように表 1 のリストを見ると、分野初心者でも意味がわかるような「品詞」「辞書」「文字」などの単語が並んでいる。これらは分野基礎用語の定義である、「必ず学ばなければならない語、その分野における基礎的・必須である専門用語」という基準からはずれることになる。しかし、これらの単語は、研究の背景など導入部分を記述するためには必須の語、および重要な手がかり語の役割をはたしていることがわかった。

次に、分野基礎用語が出現する文が全体のどのくらいの割合を占めているのかを調べ、表 2 に示す。分野基礎用語が一つの文に複数出現することもあるため、文単位での傾向を分析した。その結果、「抄録」、「はじめに」の論理構造の要素では、全体の半分以上を占めていることがわかった。次いで「おわりに」「関連研究」の要素で 4 割以上に分野基礎用語が含まれている。これは分野基礎用語の 90 語のうち頻度 0 を除いた 74 語が、「抄録」や「はじめに」などの論文の重要な部分を説明する文章に半分以上含まれるということになる。この結果を見ると、「抄録」や「はじめに」に多く出現する用語が分野基礎用語なのではないかと予測されるが、これまで行ってきた実験では「抄録」の中で高頻度な用語が、分野基礎用語にはなっていなかった。今回はこれまでと正解セットや分析対象コーパスが異な

っているため、単純に比較することはできない。しかし、今回の対象コーパスが論文誌に採択された実験論文であるため、論理構造がはっきりしていることや、用語の使い方や表現も推敲を重ねるなど、質の高い文章であることから、分野基礎用語の出現傾向が特徴的になったのだと考えられる。

これまで、分野基礎用語は分野特有の専門用語で、分野初心者がその分野を理解する上で必ず学ばなければならない用語と考えていた。しかし、客観的な指標による分野基礎用語の選定および実際の論文中に出現する傾向を分析すると、必ずしもその用語自体を学ぶ必要はなく、むしろその用語が手がかり語となって周辺の用語との関連により、その分野の理解を深める役割を果たしていた。つまり、分野基礎用語をベースとして、周辺用語との関連を示してあげることにより、分野初心者への論文理解を手助けすることができるのではないかと考えられる。

表２　論文の論理構造における分野基礎性用語を含む文の割合

	文数	分野基礎性用語を含む文数	割合
抄録	656	362	0.552
はじめに	2448	1284	0.525
実験	8931	2701	0.302
関連研究	1222	542	0.444
おわりに	805	394	0.489
その他	11965	3439	0.287
合計	26027	8722	0.376

５．まとめ

本論文では、その分野で必ず学ぶべき用語や手がかり語となる分野基礎用語の選定基準と、実際の論文における出現パターンの分析を行った。選定の基準は、多くの専門家が執筆した本や事典の目次、論文のタイトル、抄録、キーワードの中から共通して出現するものとした。この客観的な基準に従って抽出した分野基礎用語が論文の論理構造の要素別に出現する頻度に基づいて分析を行った。

分析の結果から、今後は分野基礎用語が出現する文が研究のどのような内容を表現しているのか（研究の背景、動機、既存研究の比較など）をさらに詳しく分析し、分野基礎用語と共起する用語との文法的関係（主語、目的語、補語、修飾語など）と意味的関係（目的、手法、対象など）を付与するなど、論文の内容理解の支援をする表現方法を検討していく。

文　献

中川裕志、森辰則、湯本紘彰(2003)「出現頻度と連接頻度に基づく専門用語抽出」、自然言語処理、Vol.10 No.1、pp.27-4

佐々木靖弘、佐藤理史、宇津呂武仁(2006)「関連用語収集問題とその解法自然言語処理」、Vol.13 No.3、pp.151-175

千田恭子、篠原靖志、奥村学(2005)「技術成果を効果的に伝える表題作成支援手法：開発と評価」、情報処理学会論文誌、Vol.46 No.11、pp.2728-2743

内山清子(2010)「専門用語の分野基礎性に関する一考察」、情報処理学会自然言語処理研究会報告、2010-NL-199(15), pp.1-6

Kiyoko Uchiyama(2011)、「A Study for Identifying Domain-Specific Introductory Terms in Research Papers」、Proceeding of the 9th Terminology and Artificial Intelligence、pp.147-150

自然言語処理学会、『言語処理学事典』(2010)、共立出版株式会社