論文の論理構造における分野基礎用語に関する分析

著者	内山 清子
雑誌名	テキストにおける語彙の分布と文章構造 成果報告書
ページ	3-12
発行年	2013-03-25
シリーズ	国立国語研究所共同研究報告 ; 12-06
URL	http://doi.org/10.15084/00002704

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)

要旨

本論文は、学術論文に含まれる多くの専門用語の中から、分野において必須で重要な用語を分野基礎用語と定義し、その用語の出現傾向について分析を行う。分野基礎用語は特定分野の研究をこれから学ぶような学部の学生は、専門が異なる研究者などに対して、効率的に分野の論文を理解するために、最低限知っておくべき用語を提示することを提案する。この分野基礎用語をどのように選定すべきであるのかについて、様々な観点を想定し、その観点を実際の文章に当てはめて分析を行った。また分野基礎用語が、論文中にどのような出現傾向を示すのか、特に文章の論理構造においてどのような役割を果たしているのかについて分析と考察を行う。

キーワード： 専門用語、分野基礎性、分野基礎用語、論理構造

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

† kiyoko_at_nii.ac.jp

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． 基礎性判定の観点と尺度

3.1 優先度

学問を学ぶ時の学ぶ優先順位がある程度決まってくる。自然言語処理の場合は、形態素解析を学んだ後で、構文解析、構文解析を学ぶといった優先度のことである。こうした学習において共通して初期の段階に優先的に教えられる項目は特に重要で、分野基礎性が最も高い用語として絶対的な尺度であると考えられる。たとえば、複数の教科書に共通する用語や同じ研究分野の大学講義等で複数の先生が共通して初期に教える用語などは優先度が高い語であると考える。

3.2 経年推移度

昔は論文等で頻繁に使われていた用語が年数を経るごとに頻度が減ってきたり、その反対に増えてきたりする用語がある。分野基礎性が高い語は、年数が経っても平均してある一定以上の頻度を保って出現し､分野基礎性が低い語は突然爆発的に使われたとしてもある時期に落ち着いて、以後使われなくなったりする等、出現頻度に安定性がないと考えられる。

3.3 親密度（頻度）

重要な語は、文章で繰り返し使われる語であり、様々な指標に頻度情報が含まれているのはその理由からである。論文や書籍などで頻繁に使われる用語は重要であるから繰り返され、繰り返されることによってその語に親しみを感じ、より使われるようになる。分野基礎性においても同様のことが言え、分野において重要な概念であるため繰り返し出現する用語は、その分野の研究者が親しみ、馴染みのある語として頻繁に用い繰り返される。頻度が高い語は、分野において親密度が高い語であると言える。この傾向から、頻度情報は分野基礎性においても重要な観点であると考えられる。

3.4 下位分野偏り度

同じ分野でもいくつかの下位カテゴリに分類することができ、言語学の場合は形態論、統語論、意味論、語用論などが下位カテゴリとなる。自然言語処理の場合は、人工知能の他、人工知能の一分野としての機械学習、自然言語処理の応用システム（情報検索、機械翻訳、質問応答）、認知科学等複数の分野が関連している。特に言語学は、ある言語現象について自然言語処理を用いて検証することがあるため、関連が深い。また、応用分野では自然言語処理技術を用いていることから、これらの応用にも基礎的な部分が共通している。

複合領域の分野および下位カテゴリの用語をどの程度分野基礎性が高い用語に含めるかも難しい問題である。特定研究分野全体に広く一定した頻度で用いられている用語は分野基礎性が高く、一方で、下位カテゴリの中でしか使われない用語は、分野基礎性はそれほど高くないと考えられる。

教える側から考えると、隣接領域の中でも基礎的な用語について万遍なくカバーして提示するのが必要である。一方、学ぶ側では、背景知識を持った人間であれば、たとえば、言語学を専攻している学生は、言語学の知識があるのでそれは知る必要がなく、その他の領域の知識を探す必要がある。その場合、自分の知識に欠けている下位分野の用語の中から基礎性が高いものから学ぶのと効率が良い。その場合、対象とする分野と下位分野の関連性強さを基準にして考える必要がある。言語学であれば形態論、統語論、生成文法のうち、自然言語処理との関連の強い用語が自然言語処理の分野基礎性の高い用語に含まれることになる。つまり、下位分野単独における基礎専門性ではなく、上位分野との関連性が強い語が重要な語となる。

3.5 語構成度

分野基礎性が高い用語はその用語単独でも多く出現するが、前や後ろに様々な語が接続して多くの新規複合語を構成していることが予測できる。たとえば、「機械翻訳」という用語の場合、後ろに「システム」が結合して「機械翻訳システム」、前に「統計的」が結合して「統計的機械翻訳システム」など、様々な派生の専門用語および新しい複合語（いずれは専門用語として認識され

るものも含む）を生成することができる。この基準は重要度計算の時でも利用されているが、どれだけの語と接続する可能性があるのかで、その基となる用語の重要性が計算できる。ある用語が新規の専門用語を構成している数が多ければその用語の概念は重要であると考えられる。

3.6 定義明確度

定義明確度として、分野基礎性が高い用語はまず定義を明確に述べることが行われる。たとえば、「形態素解析とは、与えられた文を形態素の単位に分割し，その文法機能（一般には品詞および活用情報）を同定する処理を言う.」というように手掛かり語「とは」を用いて定義付けを行う。このように「A とは B」の A に当たる用語は分野基礎性が高いと考えられる。上記の 6 つの観点のうち、定義名義度を除いた 5 つの観点とその尺度について表 1 に示す。個々の尺度は、それぞれが相互に関連しているものと考える。

４．分析

4.1 対象データと抽出単位

3 章における、基礎性判定における観点について、指標の妥当性を検証するために、さまざまなリソースにおける出現頻度、重複度について分析を行った。できるだけ電子的に利用可能なリソースを収集し、分野基礎用語の各リソースにおける出現パターンを調べた。

リソースとして実験的に、教科書や講義資料、事典、論文を題材とした。まず教科書には必ず基礎的な用語が含まれており、単語の頻度ではなく各資料に共通した用語が重要であると仮定する。事典については、特定分野の知識として必須情報を獲得することが可能で、各章の出現頻度やどの章に出現したのかが重要となってくる。一方、論文については、より専門的となり、研究における基礎的な用語や研究動向､流行り廃れなどについて観察することが可能であると考える。ここでは、出現頻度と出現年数が重要であると考える。

このように各リソースでは特徴も重要な観点も異なるため、リソース別の抽出単位と頻度計算の違いを分けた。文書頻度については、各教科書、講義資料で共通する単語の頻度とした。教科書では、索引語の教科書間、講義資料では、講義資料間の重複を調べた。出現頻度では、テキスト中における出現頻度として、論文においては、抄録に出現する単語数、事典では全テキスト中に出現する単語数について調べた。

4.2 分析結果

まず、優先度分析では、教科書や書籍、講義資料で共通して用いられる、あるいは最初に説明される（出現する）用語は基礎性が高いものが多いという仮説のもと、1996 年から 2007 年に出版された自然言語処理の書籍の索引語と、講義資料は自然言語処理関連講義から 3 講義の資料を使ったが、講義資料を入手するのが困難であるのと、著者や講義者によって重点的に説明する箇所が違っているなど、共通事項が少なかった（表 1)。

次に経年推移度では、長期間出現し、頻度が高い語は分野基礎性が高いと想定し、情報処理学

会自然言語処理研究会で 1993 年から 2006 年まで 14 年間の抄録データ 1 MB 分について分析を行った。結果としてはほぼ仮説のとおりであった（表 2)。

親密度、頻度については、言語処理学事典と論文抄録で調査を行ったが、出現年数が長いと累計することで頻度が高くなり、出現年数で頻度を割るとはやりの語が交じることがあった。

表 1：講義資料と教科書の上位頻度語

講義資料の用語	教科書の用語
形態素解析	意味ネットワーク
構文解析	格フレーム
格文法	形態素解析
文脈自由文法	シソーラス
機械翻訳	格文法
LFG	形態素
GPSG	深層格
HPSG	表層格
機械翻訳システム	接辞
情報検索	自然言語処理

表２：経年推移度

用語	頻度	年数
コーパス	477	14
機械翻訳	197	14
形態素解析	188	14
類似度	149	14
文字列	129	14
構文解析	126	14
情報抽出	118	14
翻訳システム	108	14
情報検索	108	14
再現率	98	14

語構成の分析では､多くの専門用語を構成している語は分野基礎性が高いという仮説のもとで、言語処理学事典の索引語を調査した。抽出単位は索引語を構成している 2 グラムの頻度計算を実施した。その結果として索引語は統制されたリストのため、あまり複合語のバリエーションが登録されていなかった（表 3)。本文中では基礎的用語に様々な形式で複合語を合成していたため、語構成調査は、本文のデータが必要であることがわかった。

定義明確度では、定義文で説明される用語は分野基礎性が高いと予測して、言語処理学事典の本文から、手がかり語「～とは」を使った定義文の～に該当する 194 語を抽出した。事典では、重要な用語に対して定義が必ずされていると思ったが､実際はその例が少なく(表 4 に例を示す)、また定義をしていたとしても､「～とは」という定型表現で定義付けをしているわけではなかったので、自動抽出にひっかからなかった例も多かった。また、定義文を論文に適用してみることを少し試みたが、論文ではほとんど定義をしていないことがわかった。定義していたとしても、それは自分の研究で重要な個別の用語であるなど、特殊な例が多く、基礎的な用語に対して、定義付けをしていないことがわかった。この結果から、論文の場合は、分野知識を共有していることを前提としているため、あえて基本的な用語に対して定義することがないことがわかった。

以上のように、各基礎性判定に関わるであろう、観点にしたがって、様々なリソースを使って

分析を行った。この分析は、自動抽出が可能であるかどうかも視野にいれながら行ったが、実際分析をした結果としては、仮説通りの結果であるにしても、観点が多いことや、その観点を自動抽出のスコアリングにどう反映すれば良いのかが非常に難しい。今回は、分野基礎用語をまずは決めてしまい、その用語がこれまで調べてきたリソースの中でどのように出現するのかを詳細に分析することに集中することとした。次の章からは、分野基礎用語の正解をどう決めたのか、また決定した基礎用語が論文中にどのように語られているかなどについて分析と考察をおこなっていく。

表 3　分野基礎用語をベースとした複合語の例

用語	生成頻度	例
コーパス	32	均衡コーパス，話し言葉コーパス
意味論	25	語彙意味論，フレーム意味論
機械翻訳	12	機械翻訳システム，統計的機械翻訳
アルゴリズム	12	EM アルゴリズム，ブースティングアルゴリズム
n グラム	9	n グラムモデル，単語 n グラムモデル
句構造	9	句構造文法，主辞駆動句構造文法
言語学	9	メタ言語学，計算言語学
主要部	9	右側主要部規則，主要部先行型
曖昧性	9	曖昧性解消，構造的曖昧性

表 4：定義文の例

機械学習	1960 年あたりから人工知能の一分野として研究が始まった分野であり，一般に，過去の事例をもとに，それらの中に潜む構造を見出したり，将来の事例についての予測を行ったりするための技術を開発することを目指す.
機械翻訳	コンピュータ・プログラムで，テキストをある言語（原言語という）から別の言語（目的言語という）に翻訳することを指し，自動翻訳や言語翻訳と呼ばれることもある.
固有表現	人名，組織名，地名といった固有の名前を持つ対象を指す表現のことである.

4.3　分野基礎用語の選定

これまで、このように様々な観点から分野基礎用語の出現傾向を分析し、自動抽出を試みてきたが、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、分析と同様に教科書、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。

５．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 5 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。

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。

表 5　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

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 6 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 6　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

６．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