様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　３月２９日現在

機関番号：６２６１５
研究種目：基盤研究(C)
研究期間：2010～2012
課題番号：22500226
研究課題名（和文）分野認知レベルに適した検索のための専門度付専門用語シソーラスの構築

研究課題名（英文）　Constructing a technical term thesaurus for appropriate retrieval depending on the levels of domain knowledge

研究代表者
内山　清子　　（UCHIYAMA KIYOKO）
国立情報学研究所・コンテンツシステム開発室・特任研究員
研究者番号：20458970

研究成果の概要（和文）：本研究は、効率的な検索に利用するシソーラス構築のために、分野における基礎的で必須である専門用語について以下の３点の研究を実施した。
(1)専門用語の専門度（分野基礎性）を示す指標の分析：文書中に出現する専門用語について、分野を理解する上で必須・基礎的なレベルから専門性が高いレベルまでの段階を分野基礎性として客観的な指標について、論文や書籍の文章構造中の出現傾向について分析を行う。
(2) 分野基礎性判定手法の検討：分析結果に基づいて、自動的に分野基礎性が高い用語を抽出する方法を検討する。
(3)システムへの応用の検討:分析に基づいて分野基礎性が高い用語判定を利用してシソーラスを構築し、システムへの応用の可能性について議論した。

研究成果の概要（英文）：This study aims to investigate the occurrence of introductory terms which are defined as basic and essential terms in a target field and study a method for constructing technical terms thesaurus by conducting three topics as below. (1) Analysis of criteria for introductory terms, (2) a method for the degree of introductory terms, and (3) possibility of system implementation. Three basic criteria that are essential for a given target field are proposed as the introductory terms. The proposed criteria are priority, compositionality, and logicality. These criteria can help to clarify the learning strategy and to studying thesaurus based on the given textbooks, the variety of types of and the tokens of constituent in compound words, and the term distribution in the structure of the text.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,000,000	300,000	1,300,000
2011 年度	500,000	150,000	650,000
2012 年度	700,000	210,000	910,000
年度			
年度			
総　計	2,200,000	660,000	2,860,000

研究分野：　総合領域
科研費の分科・細目：情報学、図書館情報学・人文社会情報学
キーワード：分野基礎用語、専門度、分野認知レベル、専門用語

１．研究開始当初の背景

現在、専門用語は一般用語に比較して難解であるのは明らかだが、特定分野において初心者レベルの常識・基礎的な用語か、上級者レベルの専門・応用的内容であるかの段階が曖昧である。専門用語の難しさは文書の読みやすさ（難易度）にも関連する。従来の文書の難易度を判定する研究は、専門的文書の難易度に適応することができない。その理由は従来の研究では、一般の文書（教科書など）を対象に文字種の頻度や一文の長さなどを基準にしているが、専門文書では、専門用語の専門性が難易度に影響するからである。そのため専門用語の専門度を判定する基準が必要となる。本研究における専門度は、分野を知るために必要不可欠である用語は専門度が低く、低専門度の用語知識をより多く必要とする用語は専門度が高い。

本研究の専門度は、先行研究で行われてきた重要語抽出の「重要度」と観点が異なる。従来の重要度とは、分野あるいは書き手（論文の著者）にとっての重要さであるので、重要語抽出では分野に特徴的な語や頻度が高い語、多くの背景知識を必要とする用語（高専門度用語）が抽出される。本研究では、分野の初心者を主に対象とする読み手（検索者）に想定し、特定分野のない背景知識で理解でき、分野必須の用語（低専門度用語）が読み手にとっての重要さであると考える。また、分野によっては最低限必要な基礎的知識の有無で、専門度の低い用語の理解が異なる場合がある。工学的分野における数学の知識があれば、他分野のモデルやアルゴリズムの理解が容易である。この場合は分野認知レベル判定における用語チェックの段階で判別可能である。また、研究における手法がどの分野でいつごろ論文に出現しているかの傾向について可視化した研究があるが、専門度を付与すると更に興味深い傾向が分析できる。

具体的な専門度を判定するために引用情報、文脈情報、語彙情報の３つを重要な要素とする。まず引用情報では、より多くの人・論文から引用される（被引用回数が多い）論文は重要であると考えられる。本研究にもその方法を利用して、被引用回数が多い論文の著者キーワードは重要であり、より広い分野から引用され、長く引用される用語は、分野の基礎的用語（低専門度語）で重要であると考える。

次に文脈情報として、専門用語が文中で表現される語のパターンや共起語に着目する。たとえば、対象とする専門用語 A が「A などの B」の場合、A は B の一種であり、B は基本的な用語で、低専門度語であると判定できる。最後に、「語彙情報」として、専門用語を構成する語の数が長くなると、より詳細な専門用語を表している可能性が高くなり専門度が増す。申請者はこれまで専門用語（複合語）の文脈情報と語彙情報を利用して、専門用語を構成する語の間及び専門用語間の意味関係の解析を行ってきており、その研究成果を本研究に応用し、引用情報という新しい観点を組み込むことで新しい展開が期待できる。

２．研究の目的

本研究は、読み手の分野認知レベルに適した専門用語を提示し、効率的な検索を実現するために、以下の 3 点の構築に取り組む。

(1) 専門用語の専門度（分野基礎性）を示す指標の分析：文書中に出現する専門用語について、分野を理解する上で必須・基礎的なレベルから専門性が高いレベルまでの段階を分野基礎性として客観的な指標について、論文や書籍の文章構造中の出現傾向について分析を行う。

(2) 分野基礎性判定手法の検討：分析結果に基づいて、自動的に分野基礎性が高い用語を抽出する方法を検討する。

(3) システムへの応用の検討：分析に基づいて分野基礎性が高い用語判定を利用して、分野認知レベル判定システム（読み手の分野認知レベルを判定する）や、専門度付き専門用語シソーラス（専門度が確定した専門用語を上位下位、その他の意味的関係を用いて効率的検索に誘導するためのシソーラス）への応用の可能性を議論する。

３．研究の方法

本研究では、専門用語の分野基礎性を定義するために、様々な視点から議論を行った。ここでの分野基礎性とは、その分野における最低限抑えておくべき基礎的かつ必須であることを示す。つまり分野基礎性が高い用語は、その用語を知らなければ、その専門分野のことを理解することができない、他の専門用語も理解することができない用語とする。一方、分野基礎性が低い用語は、基礎性の高い用語の理解を深めた上で、その知識を利用しなければ理解することができない専門性の高い用語であるとする。

まず、分野基礎性の様々な観点の整理、具体的なデータについて分野基礎性を解析を行った。

(1) 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。

(2) 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、4 レベルが 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．研究成果

平成 22 年度は、専門用語の難易度（分野基礎性と言い換えることとした）の指標を決定することを目的として、いくつかの指標を設定した。語彙情報として優先度（初期の段階で学ぶ用語）、経年推移度（年度毎の出現分布が平均している用語）、親密度（頻度が高い語）、網羅度（複数の下位カテゴリで用いられる用語）、語構成度（多くの派生専門用語を生成する用語）、文脈情報として、定義

明確度（手がかり語により導入される用語）の 6 つの指標を設定した。計画には引用情報を組み入れる予定であったが、対象リソースに引用情報が少なかったため、今回は除外した。この 6 つの指標を数値化し、既存の尺度によりどの程度ランキングを行えるかを調べた。対象を自然言語処理分野とし、専門家によりレベル付けをしてもらい正解データとした。

実験に使用したコーパスは、一般のコーパスと専門分野のコーパスとして自然言語処理分野の論文コーパスの 2 種類を用意した。既存の尺度で抽出した特徴語について平均精度を用いて、正解データと比較を行った。その結果、イエーツ補正カイ二乗値が正解データの抽出精度が高かった。この抽出精度をベースラインとし、分野基礎性の独自指標を数値化したものと比較を行った結果、語構成の指標が最も有効であった。この結果から、より多くの派生専門用語を生成する用語は分野基礎性が高いものであることがわかった。今回は専門分野のコーパスとして論文を対象としたが、より基礎的な用語が頻出する教科書となるような書籍を対象とする必要性があると考えられる。

平成 23 年度は、教科書などの基礎性が高い内容の書籍データを利用するために、コーパス拡充を行った。分析対象としている自然言語処理分野の書籍 43 冊分について､本文と、目次、索引、参考文献を電子化し、コーパスやデータとして利用した。自然言語処理分野のコーパスから分野基礎性が高い用語を抽出する手法として、C-Value を用いて抽出、ランキングを行い、既存の指標によって評価を行った。論文などの専門的なコーパスよりも、書籍の目次に含まれる用語を使った尺度の方が効率的に基礎性の高い用語を抽出することができた。また、論文の論理構造を用いた分析として、タイトル、抄録、著者キーワード、本文などの出現箇所による頻度の違いを考慮すると、効率的な抽出が可能になるという知見を得た。論文の論理構造では、本文をひとまとめにして分析を行ったため、もうすこし詳細な区分（はじめに、関連研究、おわりに）に分けて分析をすることも重要であると考え、その区分を分けて、各区分における頻度の出現傾向を分析する準備を行った。

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。

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。

表 1 ：論理構造における分野基礎用語の出現頻度

分野基礎用語	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
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

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。

表 2：論理構造における分野基礎用語を含む文の割合

論理構造	文数	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

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。

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、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。

本研究では、シソーラス構築やシステム応用までには至らなかったが、シソーラス構築に向けた問題点を明らかにし、文章構造を考慮することなど、従来にない分析結果を得ることができた。今後は具体的なシステムへの応用を検討していく予定である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1 件）

① 専門用語の専門性判定に関する一考察, 内山清子, Japio年誌財団法人日本特許情報機構，査読無, pp.152-153，2010.

〔学会発表〕（計 12 件）

① A Study for Introductory Terms in Logical Structure of Scientific Papers, MIC Sorbonne 2012 New Standards for Language Studies, Kiyoko Uchiyama, 査読有, 2012. 11/15, Paris, France.
② A Method for Corresponding Paragraphs with Sentences in Academic Paper's Abstract, Yuan Li, Akihiro Kameda, Kiyoko Uchiyama and Akiko Aizawa, 第 11 回情報科学技術フォーラム(FIT2012)，査読無，2012.9/6, 法政大学小金井キャンパス(東京都)
③ Automatic Translation of Scholarly Terms into Patent Terms Using Synonyms Extraction Techniques, Hidetsugu Nanba, Toshiyuki Takezawa, Kiyoko Uchiyama and Akiko Aizawa, Proceeding of International Conference on Language Resources and Evaluation (LREC2012), 査読有，pp.3447-3451, 2012.5/24, Istanbul, Turkey.
④ 構文パターンを用いた論文の引用文脈からの関係情報抽出，亀田尭宙，内山清子，武田 英明, 相澤彰子，人工知能学会第 26 回全国大会，査読無, pp.1-4，2012. 6/15, 山口県教育会館（山口県）.
⑤ CiNii データベースを用いた研究動向分析システムの構築，福田悟志，難波英嗣，竹澤寿幸，武田英明，相澤彰子，大向一輝，宮尾祐介，内山清子，言語処理学会

第 18 回年次大会，査読無，pp.539-542，2012．3/15，広島市立大学（広島県）

⑥ Analyzing the characteristics of academic paper categories by using an index of representativeness, Takafumi Suzuki, Takafumi, Kiyoko Uchiyama,, Ryota Tomisaka, Akiko Aizawa, Proceedings of PACLIC25: the 25th Pacific Asia Conference on Language, Information and Computation,査読有, pp.587-596, 2011. 12/18, Singapore.

⑦ A Study for Identifying Domain-Specific Introductory Terms in Research Papers, Kiyoko Uchiyama, Proceeding of the 9th International Conference on Terminology and Artificial Intelligence，査読有，pp.147-150, 2011. 11/8, Paris, France.

⑧ 論文中の引用文における構文パターンを用いた論文・概念間の関係抽出，亀田尭宙，内山清子，宮尾祐介，武田 英明，相澤彰子，第 94 回 人工知能学会知識ベースシステム研究会，査読無，pp.25-31，2011. 12/16，慶應義塾大学（神奈川県）．

⑨ オススメ論文検索システム：OSUSUME，内山清子，高須淳宏，相澤彰子，難波英嗣，宮尾祐介，第 25 回人工知能学会全国大会，査読無，pp.1-4，2011．6/1，いわて県民情報交流センター(岩手県)．

⑩ Cross-lingual Recommender System for Research Papers, Kiyoko Uchiyama, Aizawa Akiko, Hidetsugu Nanba and Takeshi Sagara, Proceedings of the 2011 Workshop on context-awareness in Retrieval and Recommendation, 査読有，pp.39-42，2011. 2/13，Palo Alto, USA.

⑪ 専門分野における用語の分野基礎性に関する研究，内山清子，言語処理学会第 17 回全国大会，査読無，pp.1033-1036，2011. 3/10，豊橋技術科学大学(愛知県)．

⑫ 専門用語の分野基礎性に関する一考察，内山清子,情報処理学会自然言語処理研究会 199 回，査読無，pp.1-6，2010. 11/18，広島市立大学（広島県）．

〔図書〕（計 2 件）

① 言語の可能性 3 言語と情報科学，朝倉書店，松本裕治編集，相澤彰子，内山清子，第 4 章「語の共起と類似性」，pp.58-76，2011. 216 ページ．

② からくりインターネット，相澤彰子，内山清子，池谷瑠絵，丸善ライブラリー，pp. 90-112，2010.172 ページ．

６．研究組織

(1)研究代表者

内山　清子（UCHIYAMA KIYOKO）

国立情報学研究所・コンテンツシステム開発室・特任研究員

研究者番号：20458970

(2)研究分担者

なし

(3)連携研究者

相澤　彰子（AIZAWA AKIKO）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：90222447