

レポートや論文に含まれる基礎語彙の分類に関する一考察

A Study of Categorization for in Academic Reports and Papers

内山 清子^{*1}

Kiyoko Uchiyama

^{*1} 湘南工科大学 工学部 コンピュータ応用学科
Shonan Institute of Technology, Applied Computer Sciences

This paper proposes an analysis method for essential words in academic reports and papers. Those documents constitute of various terms and words having individual functions. Three categories for each function are defined as follows; framework, theme and predictive words. The framework words support the logical and rhetorical structure. The theme words have various functions of technical and domain-specific terms. The predictive words express the document's author intentions for emphasis of important points. The words and phrases in academic papers and reports are analyzed and categorized using published academic papers focusing on each function and role. A method is examined for utilizing effective reading and writing in academic activities.

1. はじめに

大学教育において、小論文よりもレポート執筆を要求されることが中心である一方、大学初期段階で専門書を読む、専門文書を書くことに対する教育が十分に行われていない。特に工学系大学において文章執筆に対する教育が軽視されている傾向がある。一方で、工学系の場合は、実験結果をまとめて文章にする機会が多く、論理的に文章を構成し説得力のあるレポートや論文にまとめる必要がある。このような背景において、工学系の教育における専門書を読む、書くことを支援することが重要となっている。

これまで、論文に含まれる専門用語を中心として、抽出や分析に関する研究が行われてきた。専門分野を理解するためには専門用語をまず認識し、理解することが重要である。しかし、実際はまだ分野の基礎的な教育を始めたばかりの学生にとって専門用語の認識は困難である。著者は専門用語の認識について、初心者(学部 3,4 年生と専門分野外の学生および研究者)と専門家(分野の研究者)の間でどのような違いがあるのか、論文読解方法について実験を行った[Uchiyama 2015]。専門分野(情報処理分野と自然言語処理分野)の論文を学部学生や、分野外の学生、研究者に読んでもらい、専門用語について(1) 必須である、(2) 知っておいた方がよい、(3) 後で調べる、(4) Web で調べる、の4つの選択肢を用意し、抽出してもらった実験を行った。この実験では、初心者が専門用語の意味を調べる時に適切に定義などを表示できるようなシステムを開発することを目的として実施した。実験の結果は、初心者や分野外の学生や研究者にとって、専門用語であるかどうかの判断も難しかったという結果が出た。つまり、専門用語の抽出ができないため、理解をする以前の問題であることがわかった。

一方で、研究者は分野外であっても比較的多くの専門用語を抽出することができていた。また、初心者はまんべんなく論文全体を読んでいたので、研究者は最初の 2 ページと最後のページに時間をかけて読む傾向が出ていた。論文の概要を把握するためには、最初や最後を効率的に読む技術が研究者は経験的に培われているのだと考えられる。

この実験の結果を踏まえて、専門用語だと認識するためにはどのような知識が必要であるのかについて考え、実際に論文を文章、論理(章立て)、修辞構造などについて分析を行ってきた。特に修辞構造では、この文章において何を語っているかを、そこに用いられている言語表現とセットで分析をしたところ、同じ役割を持つ文章には似たような表現が使われていることがわかった。そこで、本研究では、文章中に用いられている文法、語彙を中心として、文における役割を 3 つのカテゴリーに分類し、それぞれの役割と、具体的な手がかりとなる表現について抽出し、分析を行った。

以下では、2 章において文の構造と役割について、関連する既存研究について本研究の位置づけとともに整理し、第 3 章で文における語彙や文法表現の役割を分析し、最後に今後の課題をまとめる。

2. 文の役割

論文の構造を分析し、内容を効率的に抽出するための研究がバイオ関連の分野では盛んに行われてきている。テキストにアノテーションをすることによって、薬品や成分の関係を同定し、ドメイン知識を構築することに役立っている。このような方法は工学系の論文にも適用できると考えている。論文の構造と論文で使われている表現の対応について分析し考察した研究[Anita de Waard 2009, 2012] は、本研究の目指すところと類似している。文、文章構造や手がかり表現などの対応付けはこの論文で使用されている分類を参考とする。ただ、詳細に分析を進める上で、必要と思われる項目を追加するなどを行った。文章間の構造(文と文の間の関係)については、今回は行わなかった。

論文に含まれる構造、つまり論理構造(ここでは、章立て)として、「抄録」、「はじめに」、「関連研究」、「実験」、「考察」、「おわりに」に分割して、この章ごとに分析を行う。

3. 文章を構成する役割語

著者はこれまで、専門文書における専門用語について、分野基礎用語、つまりその分野に基礎的な用語を同定し、抽出する方法について研究を進めてきた[内山 2012]。これは、専門文書を読解する上で、分野で最初に必ず知っておかなければならない必須の専門用語であり、この用語の理解なくして、その分野

での研究を理解することができないという最重要用語である。この用語は分野毎にコーパスや教科書などの言語資源がある一定量揃えば同定可能である。その他にも、ある特定論文に特徴的な用語など、専門用語にもいくつかの役割を持っていることがわかった。しかし、専門用語以外にも論文の理解に必要な語彙や表現があると考え、整理を試みた。専門用語といった専門文書の内容を特徴づける語彙の他に、論文の構造を支える骨組み語、内容のキーワードとなりうるテーマ語、論文の主張を述べるために必要な叙述語の三種類に分類した。この分類は、源氏物語の語彙を分類する時に使用されたものであるが[寿岳1967]、今回は工学系学術論文用に解釈しなおして整理した。

3.1 骨組み語

骨組み語は、文章を組み立てる時に使われる語とするものだが、本研究では文が持つ役割を特徴づける表現とした。たとえば、「本研究の目的は」「従来の研究では」「実験の結果」「精度を得た」「今後の課題として」など、これらの表現を含んだ文は、その他の表現に何が来ても、文の役割が変わることはない。また、これらの表現を使って論文を構成することで、読者にもわかりやすく伝えることができる機能を持つという意味で骨組み語という分類とした。

3.2 テーマ語

テーマ語は、専門文書やレポートにおける内容に関連のある用語と定義する。つまり専門用語がこれに当たると考えるが、分野基礎用語よりも、対象の論文のキーワードに対応する、論文の内容を特徴づける重要な役割を果たす用語とする。しかし、実際には、分野基礎用語もテーマ語を導入するために使用されるなど、関連が深いため、今回は分野基礎用語を含んだ文を対象として、テーマ語の機能について分析を行った。

3.3 叙述語

叙述語は、源氏物語のような情景や感情を表現する文章に多く含まれるものであるため、学術論文には不要な語であると考えがちである。しかし、「非常に」「極めて」「最も」などの表現は論文筆者が強調したい部分であり、学術論文のような主観を含めない客観的な文章が多い中における、これらの表現は興味深い。叙述語がどのようなテーマ語と共起して用いられるのかなどを詳細に分析すると、論文全体の位置づけも理解しやすくなると考えられる。

4. 分析

レポート形式の文書として白書が考えられるが、今回は論文形式の文書だけを対象とした。査読による審査があり、著者に馴染みのある分野として、情報処理学会論文誌の中で自然言語処理の分野の論文を100論文収集し、分析対象とした。

まず、論文の全文データを整理し、各章毎に分けた。「抄録」「はじめに」「関連研究」「実験」「考察」「おわりに」の6つに分割した。この際、章のタイトルでは6つの分類項目に分けづらいものもあったが、人手で判断し、該当する章に分けた。また、今回は「抄録」と「はじめに」の章だけを対象とした。その理由として、この二つの章は、研究者が最も長く時間をかけて読んでいる箇所であること、また論文全体の概要や特徴をわかりやすく書いている部分であることから、読者が内容を理解しやすくなるための言語表現を用いているのではないかと考えたからである。

今回の分析は、論文の内容をどう語っているのかを中心に検討を行う目的のため、テーマ語が含まれる文を分析対象とした。自然言語処理における分野基礎用語、あるいは専門用語のリス

トは、以前の研究成果により蓄積されていたため、そのリストを使用した。

100論文のうち、分野基礎用語を含む文が「抄録」に378文、「はじめに」の章に1147文含まれていた。分析手順としてまず文の役割分析として、「目的」「背景」「問題点」「既存研究」などを付与していった。次に各文に含まれる特徴的な表現や語彙を抽出し、骨組み語、テーマ語、叙述語のどれに当たるかを人手で抜き出した。

今回分野基礎用語を含む文だけを抽出し分析対象としたが、論文の内容をある程度把握できる内容であった。その理由として、分野基礎用語が含まれている文は、論文の提案手法や背景などを説明しているものが多かったためである。語彙の役割に着目すると、言語表現として骨組み語として「本論文では」「問題点として」など、文の役割を示す言語的な手がかり語が数多く含まれていた。テーマ語については、分野基礎用語以外の専門用語も多く出現していた。これは、分野基礎用語を使って、より専門的なあるいは論文で提案したい手法の説明を行っていたためである。また、叙述語に関しては、論文の特徴や背景を説明している文に「広く利用されている」「極めて難しい」「非常に効果があった」など、他の箇所には出現しない副詞表現が含まれていた。これらの分析結果から、文や語彙役割を規定する言語表現をある程度特定することができた。

5. おわりに

本研究では、レポートや論文に含まれる語彙について分類し、分析を行った。専門用語といったその文章を特徴づける語彙の他に、論文の構造を支える骨組み語、内容のキーワードとなりうるテーマ語、論文の主張を述べるために必要な叙述語という3つに分類した。論文データを章立てに分けて「抄録」と「はじめに」の章に含まれる文のうち、分野基礎用語を含む文を対象として実際に分析を行った。その結果、骨組み語、テーマ語、叙述語を含むことで、論文における重要表現を構成していることがわかった。今後は、分野基礎用語を手がかりとせず、骨組み語、叙述語から重要表現を抽出できるかどうかを調べていく。また、この結果を初心者が初めて学術論文を読む時に、骨組み語、テーマ語、叙述語を手がかりとして効率的に内容を把握できる仕組みについて検討をしていきたい。

参考文献

- [寿岳章子 1967] 源氏物語基礎語彙の構成、計量国語学, 41, 1967.
- [Uchiyama 2015] Kiyoko Uchiyama, Takeshi Abekawa and Akiko Aizawa, A Study of Analyzing Comprehensive Reading Behavior for Research Activities based on a Web-based Experiment, Pacling2015, 2015.
- [内山清子 2012] 論文の論理構造における分野基礎用語に関する分析, 第二回コーパス日本語学ワークショップ, pp.195-198, 2012.
- [Anita de Waard 2009] Anita de Waard and Henk Pander Maat., Categorizing Epistemic Segment Types in Biology Research Articles. Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009), 2009.
- [Anita de Waard 2012] Anita de Waard and Henk Pander Maat., Epistemic modality and knowledge attribution in scientific discourse: a taxonomy of types and overview of features, ACL'12 Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, pp.47-55, 2012.