修 士 論 文

表題解析による科学技術論文の自動分類

指導教官　佐藤理史 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

今井俊

1999 年 2 月 15 日

要 旨

人間が科学技術論文を分類する場合、論文の表題を見るだけでその論文がおおよそどの分野に属するのかを決定できることが多い。これは、論文表題は論文の最も短い要約となっており、論文の内容に深く関係した専門用語が表題に含まれることが多い、という理由によると考えられる。本研究では、このような考えに基づき論文表題を解析することによって論文に適切な分類カテゴリを付与するシステムを作成した。

作成したシステムは、標準化とコード割当の２つの処理から構成される。標準化では、動詞や機能語を手がかりに論文表題をいくつかの部分要素に分割する。次にコード割当において、それぞれの部分要素に含まれる専門用語を見つけ、その語に対応する分類コードを論文の分類コードとする。本システムは、専門用語集として岩波情報科学辞典の用語の木を用いており、情報科学分野の論文を分類することができる。人工知能学会誌掲載の 369 論文を分類する実験を行った結果、79％の論文を正しく分類することができた。

目 次

第 1 章

序論

近年、安価で高機能な計算機やインターネットの普及により、電子化されたテキストが数多く収集できるようになった。ある特定の事柄に関するテキストが必要となった場合、利用者は収集されたテキストの中からそのようなテキストを特定することになるが、収集されるテキストの数が増すほど求めるテキストを特定することは困難となる。

収集されたテキスト情報から求めるテキスト情報を特定するために、現在、様々なシステムが利用されている。例えば、goo[1]に代表されるような検索エンジンを用いたキーワード検索システムや、Yahoo[2]に代表されるような、あらかじめ分類されたテキスト情報によるディレクトリ検索システムがある。

キーワード検索システムでは、利用者は適切なキーワードの入力することができるということを前提としている。この前提が満たされない場合は、求めるテキスト情報を特定できない欠点がある。それに対し、ディレクトリ検索システムでは、テキスト情報があらかじめ分野別に整理されているが前提とされているが、この前提が満たされる場合、その分野に関するキーワードの知識がなくても、限られた数の分類細目から選択をくり返すことによって目的テキストへたどりつくことができる。しかし、収集された膨大の数を人手で分類することは多大な労力が必要となる。

これを解決する 1 つの方法は、収集されたテキスト情報の分類の自動化を実現することである [1][2]。

従来、テキストの自動分類では、ベクトル空間モデルや TF-IDF(Term Frequency Inverted Term Frequency) 法 [3][4] などを用いてテキストをおおまかなに分類することを実現している。しかしながら、これらの方法は、テキスト全体に出現する単語とその出現頻

[1] http://www.goo.ne.jp/

[2] http:/www.yahoo.co.jp/

度等の統計的特徴のみに依存するため、その正確性には限界があると言わざるをえない。さらに、テキスト全体を処理するため計算量が非常に多いという欠点もある。

一方、実際に図書館などで行われている文献の分類は、「日本十進分類法」[5] に基づいている。これは、人間が表題、目次、要約などを利用し総合的に分類細目（カテゴリ）を決定している。しかし、すべての文献において表題がその内容を適切に表しているとはいえないために[3]、あらゆる分野の文献に対して表題情報から分類細目を決定することは困難であると考えられる。

そこで本研究では、対象を科学技術論文に限定することで、表題情報のみから分類細目を決定することを試みる。科学技術分野の場合、ある分野の専門家は、専門としている分野の論文に関しては、論文の表題を見るだけで、その論文がおおよそ何についての論文なのかを推定でき、適切な分類細目を決定できることが多い。たとえば、人工知能の専門家は、「プロダクション・システムによる線画の解釈」という表題を見ると、『この論文は「線画の解釈」に関する論文で、「プロダクションシステム」を手法として用いているのだろう』という推測できるのが普通である。このような推測が可能なのは、

1. 専門家は専門用語に関する知識を十分に持っている
2. 論文表題は論文のもっとも短い要約となっており、論文の内容と密接に関連した専門用語が表題に含まれることが多い

という理由によると考えられる。もし、この仮説が正しいとすれば、ある分野の専門用語集を用意することによって、論文表題からその論文の分類細目を機械的に決定できる可能性がある。

本研究では、このような考え方にたって、論文表題を解析することにより、その論文の分類細目を決定する方法について検討する。そのために、論文表題を解析しその特性を利用する。論文表題には、(1)「における」「による」など連語として名詞句の関係を表す機能語が多用される、(2) 名詞句には専門用語が含まれることが多い、といった特徴がある。これらの２つの特徴を中心として、表題解析により論文を詳細に分類することを実現する。

本論文の構成は以下の通りである。まず第 2 章では、科学技術論文の表題にはどのような特徴があるかについて調査を行い、その結果に基づき、表題解析による分類手法を検討する。第 3 章では、本研究で作成した自動分類システムの構成と、各機構について述べる。第 4 章では、実際に科学技術論文を分類する実験について述べ、本システムの有効性を検討する。最後に第 5 章で、結論を述べる。

[3] 小説などは表題がその内容を表さないことがよくあるため、表題から内容を推測することは難しい。

第 2 章

表題解析による論文分類

本章では、表題情報からの分類を実現する方法について述べる。まず、論文の自動分類の現状について述べる。つぎに、科学技術論文の表題の特徴について述べる。最後に、これを考慮した分類手法について述べる。

2.1 論文分類の現状

論文の分類には、検索に利用しやすい分類、あるいはそうでない分類が存在するが、検索に利用しやすい分類とは、詳細な分類である。なぜなら、専門分野とは細分化が進んでいるため、抽象的なカテゴリに分類することは、それほど意味をもたない。例えば、「自然言語処理」は、抽象的な概念であるため、「自然言語処理」に分類された論文が求められることは少ない。また、大量の論文を分類するため、1 つのカテゴリに属する論文の数が多い分類は有効に機能しない。一方、自動分類としては、大量の論文を処理したいため、計算量が少ないことが望まれる。

従来、論文の自動分類には、ベクトル空間や TF-IDF(Term Frequency Inverted Term Frequency) 法などを用いてテキストをおおまかに分類することを実現している。このような方法は、論文本文に出現する単語とその統計的特徴のみに依存するため、より具体的な分類細目を付与するに従い、有効には機能しなくなってくる。また、本文全体を処理するため、計算量が多いという欠点がある。

2.2 論文表題の特徴

論文表題を解析しその特性の利用を考えるために、論文表題を調査した。その結果、以下のような特徴がみられた。

特徴 1 専門用語が多く含まれる

論文表題は論文のもっとも短い要約となっているため、論文の内容と密接に関連した専門用語が表題に含まれることが多い。図 2.1に論文表題の例を示す。この図においてボールド体は専門用語を示している。専門用語は名詞句（の一部）として現れることが多い。

- 人工知能向き **オブジェクト指向言語** Monad
- **知識工学**の機械設計 **CAD** への応用
- 建築物安全度査定システム SPERIL-II とその **推論制御**
- 電気ドリル分解・組立て **コンサルタント・システム**
- **類推**の定式化とその実現
- **仮説選定**機構の一実現法
- 構文的予測の分析とその **構文解析**への応用
- **フレーム**型 **データ構造**の一論理的記述について
- **文脈自由言語**パーザへの **Prolog プログラム変換**の応用

図 2.1: 特徴 1 の表題の例（人工知能学会誌より）

特徴 2 機能語により名詞句間の意味的関係を表す

多くの場合、科学技術論文とは、特定の対象について、ある手法を適用した結果の報告であるため、それらを明示するような表現が用いられることが多い。具体的には、図 2.2に見られるように、「による」「を用いた」「における」などの連語が、非常に多用される。これらは、2 つの名詞句（複合名詞句）の意味的関係を表している。例えば、「プロダクション・システムによる線画の解釈」の場合、「による」によって「プロダクション・システム」と「線画の解釈」は、手法と目的の関係にあることを表している。以下では、このような連語を機能語と呼ぶ。

- 故障診断用エキスパートシステムにおける知識獲得
- 拡張ユニフィケーションを用いたパーザ IP の実現手法
- COMEX によるオンライン型エキスパートシステム
- クラス概念によるモデリングを用いた知識型 3 次元ビジョンシステム
- 優先度つきトークンパッシングプロトコルを用いた分散型探索機構
- プロダクション・システムによる線画の解釈
- 動詞の構文-構文意味属性による日本語動詞句内の多義語の同定
- 簡単なパルス回路における不連続変化の定性的解析法
- 関係データベースに基づく演繹データベースの推論実行方式

図 2.2: 特徴 2 の表題の例（人工知能学会誌より）

2.3 分類手法

前節で述べた論文表題の特徴を利用して、表題のみから論文の自動分類を実現することを考える。ここで考える分類とは、論文表題を解析することで、表題が表す内容に対応する専門用語を決定することである。表題の解析は、

- 機能語
- 専門用語集

の利用により行われる。以下では、まず機能語を利用した表題解析について述べ、つぎに専門用語集を利用について述べる。

2.3.1 機能語の利用

機能語を利用した表題解析とは、表題中に含まれる機能語や動詞を手がかりとして、論文表題の構造を決定し、複合名詞句と複合名詞句の意味的関係を明らかにすることである。表題解析の概略を図 2.3に示す。論文表題は、一般の文と比べ構造が単純であるため、機能語や動詞を利用することで、論文表題の構造を把握することができる。具体的には、論文表題の構文的構造を、機能語もしくは「格助詞 + 動詞」を中心とした構造と捉える。また、機能語もしくは「格助詞 + 動詞」は、前後の複合名詞句がどのような役割をもっているのかを表す。論文表題の場合、機能語の直後の複合名詞句は動作の目標としての役割であり、直前の複合名詞句は目標に対する動作の対象範囲であったり、目標の実現の

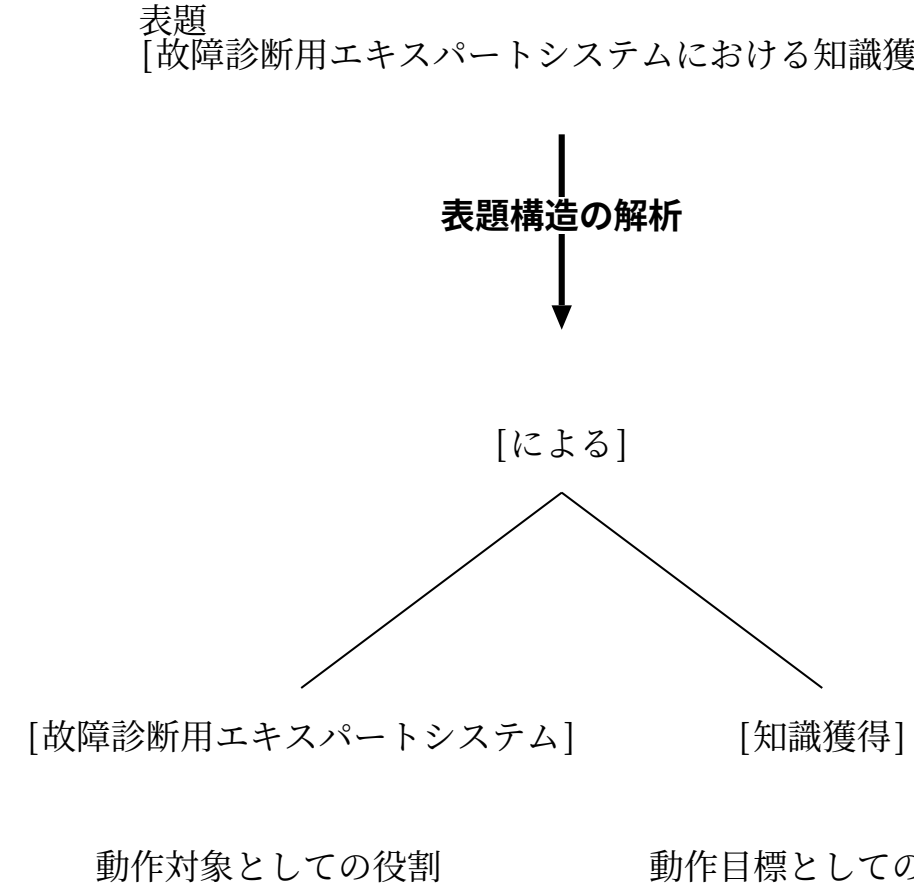

図 2.3: 機能語による表題の解析

ための道具といった役割を持っていることが多い。例えば、「プロダクション・システムによる線画の解釈」において、「による」の直後の「線画の解釈」は、動作（論文）の目標であり、「プロダクション・システム」はその実現のための道具である。

このように機能語（動詞）を利用することで、論文に含まれる複合名詞句間の意味的関係を推定できる。

2.3.2 専門用語集の利用

専門用語とは、ある特定の分野で特定の概念を表す語であり、専門用語集は、その分野で使われる概念のリストに相当する。一方、ある特定の分野の科学技術論文は、その分野における何らかの概念に関連していると考えるのが妥当であり、論文の分類は、その論文

がどのような概念に関連しているかによってなされるべきである。先に述べたように、その分野で使われる概念は専門用語で表現されるため、概念＝分類カテゴリ＝専門用語として論文を分類するという方法が考えられることになる。もし、専門用語が階層的に分類されている専門用語集が入手可能ならば、それに従った分類は階層的分類となる。

以下では、本研究において専門用語集として用いる岩波情報科学辞典 [6] について述べる。

岩波情報科学辞典

本研究では専門用語集として、岩波情報科学辞典の「用語の木」を用いる。「用語の木」とは、それぞれの節点に専門用語を対応づけた木構造であり、専門分野の概念を階層的に整理したものとなっている。図 2.4に示すとおり、用語の木では情報科学を、

- A：基礎
- B：ハードウェア
- C：ソフトウェア
- D：知識システム
- E：情報と社会

の 5 つの大分野に分類し、それぞれの大分野はさらに中分類、小分類と分類されている。用語の木の各節点には 1 つの専門用語と対応する分類コードが置かれていて、分類コードから用語の位置を知ることが出来る。例えば、D22.3 は「知識システム (D)」の部分木の中の「計算言語学 (D2)」の中の「言語理論 (D22)」に属し、その中の「統語論 (D22.3)」というわけである。

節点すなわち用語の総数はおよそ 4,500 語である。

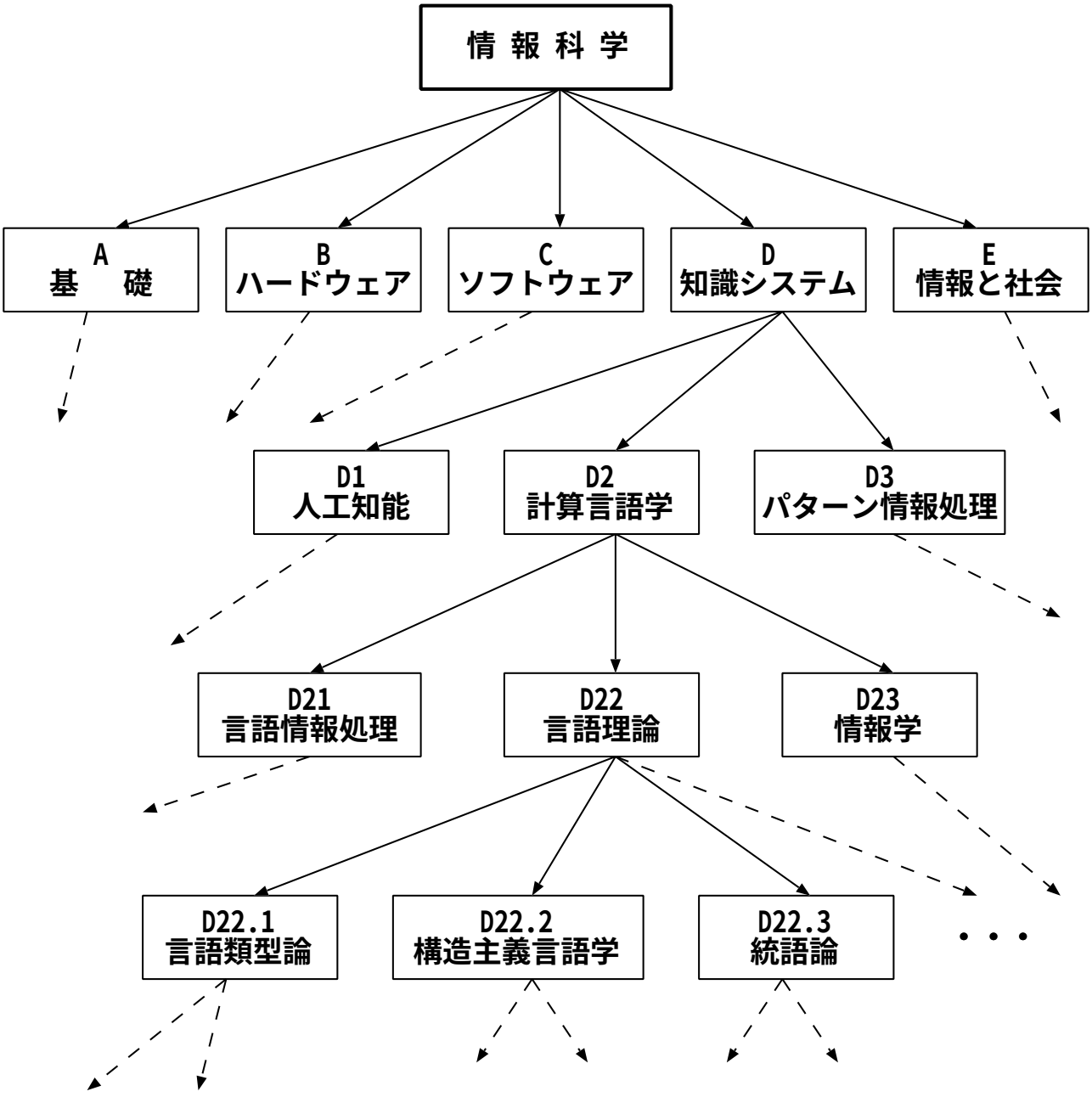
情 報 科 学
A
基　礎
B
ハードウェア
C
ソフトウェア
D
知識システム
E
情報と社会
D1
人工知能
D2
計算言語学
D3
パターン情報処理
D21
言語情報処理
D22
言語理論
D23
情報学
D22.1
言語類型論
D22.2
構造主義言語学
D22.3
統語論
・・・

図 2.4: 用語の木

第 3 章

自動分類システム

本章では、表題解析による自動分類システムがどのような仕組みで実現されているのかについて述べる。まず、自動分類システムの概要について述べる。つぎに、分類システムの出力となる分類コードについて述べる。最後に、分類システムを構成する 2 つのモジュール（標準化、コード割当）について述べる。

3.1 分類システムの構成

作成した自動分類システムの全体構成を図 3.1に示す。本システムは、入力された論文表題に対して、標準化とコード割当の 2 つの処理を行ない、その論文に対する分類コードを出力する。最初の処理である標準化は、動詞や付属語相当句を手がかりに論文表題の構造を解析する処理であり、具体的には、論文表題を名詞句、動詞句、付属語相当句に分割することを行なう。次のコード割当は、標準化によって分割された各要素に該当する分類コードを割り当てる処理であり、具体的には、名詞句や動詞句に含まれる複合名詞と専門用語を照合し、それに対応する分類コードを出力する。

3.2 分類コード

自動分類システムの出力となる分類コードは、

- 主分類コード
- 補助分類コード

の 2 つのコードから構成される。

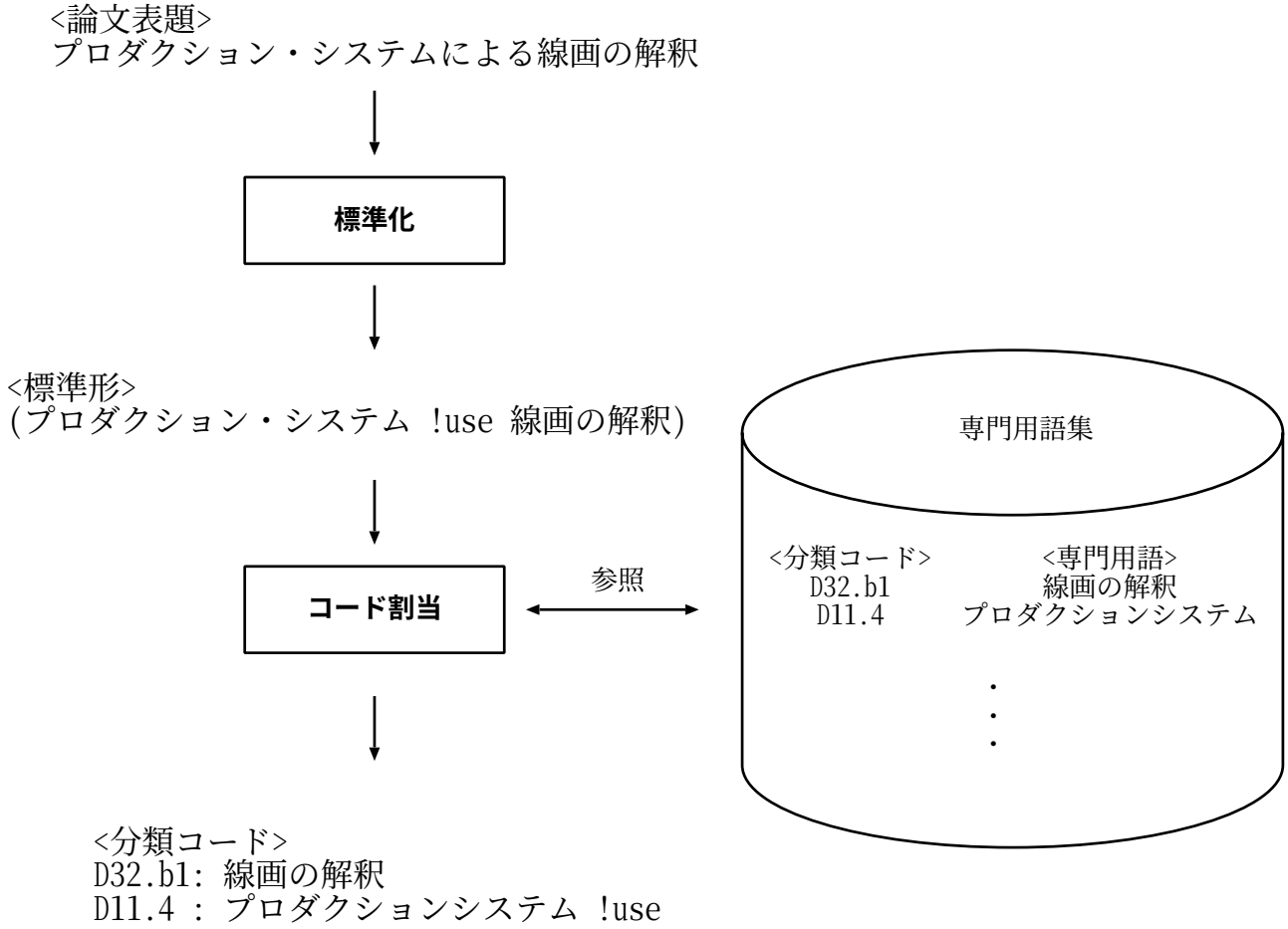

図 3.1: 分類システムの構成

主分類コード

主分類コードは、論文の内容を示す専門的な概念を表し、その概念が情報科学分野に対して、どのような位置にあるのかを表す。具体的には、論文表題に含まれる専門用語に対応するコードが主分類コードとなる。例えば、「類推の定式化とその実現」という表題に対して出力される「D12.35」は、専門用語「類推」に対応しており、その論文は用語の木における「知識システム（D)」の部分木の中の「類推」に関する論文であることを示している。

補助分類コード

補助分類コードとは、論文と専門用語（主分類コード）がどのような関係にあるのかを表す。具体的には、機能語や動詞などの付属語相当語を分類した表3.1を利用する。付属語相当語のうち「を用いた」「を利用した」のようにほぼ同じ意味を持つものはその意味に応じてカテゴリに分類し、そのカテゴリを補助分類コードとする。例えば、ある論文表題の主分類コード「D11.4」に補助分類コード「use」が付加された「D11.4 !use」が出力されたならば、その論文は、D11.4 の専門用語「プロダクション・システム」を手法として利用していることを示している。

表 3.1: 補助分類コードと対応する表現のパターン

補助分類コード	対応する表現のパターン
use	(を (用い |(利用 | 適用 | 応用 | 拡張) し)(た | て)| に (よ (る | り)| 基づ (く | き | いて)))
in	(に (お (ける | いて))| を対象とし (た | て))
about	(に (ついての | 関 (する | しての?)))
for	(の?ため (の | に)| に適した | に対 (する | して)| を扱 (う | った)| を目的と (た | て))
with	を ((持 | も)(つ | った)|(備 | 導入) した | 備えた |(組み込 | 含)(む | んだ))
as	としての

3.3 標準化

標準化処理は、自動分類システムの最初のステップであり、動詞や機能語を手がかりに論文表題の構造を解析する処理である。論文表題に対して、不要部分の削除、置換・変形、分割の 3 つの操作を行うことで、論文表題を名詞句、動詞句、付属語相当句に分割する。具体的に以上の操作は、

1. 文字列処理
2. 単語列処理

の 2 つの処理を繰り返すことによって実現される（図 3.2）。

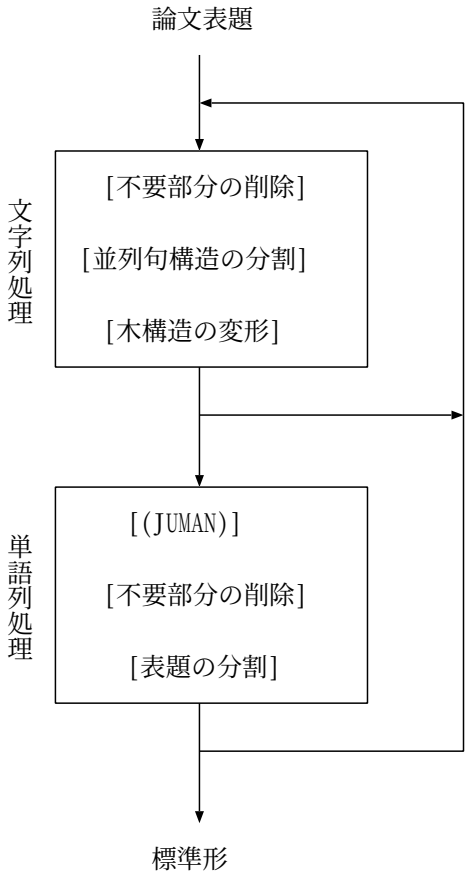

図 3.2: 標準化の概要

3.3.1 文字列処理

文字列処理では、不要な部分の削除、並列句構造の分割、木構造の変形を行う。
以下それぞれの操作について述べる。

不要部分の削除

論文表題では、先頭や末尾でシステムやアルゴリズムなどの固有名詞、末尾で「について」「とその応用」などがしばしば記述される。これらの要素は分類には寄与しないため、これらを前処理の段階で削除する。具体的には、以下の方法で、これらを削除する。

- 論文表題に”：”が含まれる場合、表題を”：”で分割する。得られた文字列のいずれかが英数字列の場合、それを削除する。いずれも英数字列ではない場合は、長さが短い文字列の方を削除する。
- 「(システム | 言語 | アルゴリズム | 手続き | モデル |,)」の直後の英数字列を削除する。
- 「」で囲まれた文字列を削除する。
- 末尾で表 3.2のパターン[1]にマッチする文字列は削除する。

表 3.2: 削除する語のパターン

について
の (実現 | 開発 | 作成)
((について)?の | に関する)(研究 | 提案 | 試み | 検討)
と (その)?(実現 | 応用 | 評価 | 適用 | 利用)(例)?

図 3.3に文字列処理による削除の例を示す。

並列句構造の分割

助詞「と」によって 2 つの名詞句が並列的に接続してできた論文表題がしばしばある。トップレベルにおいて 2 つの名詞句が並列的に接続している場合、それぞれの名詞句が、

[1]”?”は「あってもなくてもよい」、”*”は「任意個」を意味する

- MBT1:実例に基づく訳語選択 -> 実例に基づく訳語選択
- 人工知能向きオブジェクト指向言語Monad -> 人工知能向きオブジェクト指向言語
- 知識獲得のための知識表現「専門家モデル」 -> 知識獲得のための知識表現
- フレーム型データ構造の一論理的記述について
  -> フレーム型データ構造の一論理的記述

図 3.3: 文字列処理による削除の例

もう一方の名詞句の分類に寄与することは少なく、それぞれの名詞句は異なる専門的概念を示すことが多いため、前処理の段階でこれらを分割する。具体的には、論文表題が「XとそのY」(X、Yは名詞句)という形であるならば、「X」「そのY」に分割する。図 3.4 に分割の例を示す。

- 集合束縛変数とその自然言語処理への応用
  -> 集合束縛変数 + その自然言語処理への応用
- 事例ベース推論の対話型モデルとその機械調整支援への適用
  -> 事例ベース推論の対話型モデル + その機械調整支援への適用

図 3.4: 文字列処理による分割の例

木構造の変形

以降の処理を容易にするために、論文表題が、「XのYへの応用」もしくは「YへのXの応用」という構造であれば、「Xを応用したY」に変形する。図 3.5に変形の例を示す。

3.3.2 単語列処理

JUMAN(juman3.5)[7] を用いて形態素解析を行い論文表題を単語列に分解する。その単語列に対して、不要部分の削除、付属語相当句の置換や論文表題の分割を行う。

- 知識工学の機械設計 CAD への応用 -> 知識工学を応用した機械設計 CAD

- 文脈自由言語パーザへの Prolog プログラム変換の応用
  -> Prolog プログラム変換を応用した文脈自由言語パーザ

図 3.5: 文字列処理による変形の例

不要部分の削除

論文表題の末尾には「方式、方法、手法、システム、機構」といった単語が表れることがしばしばある。これらの単語は、直前の単語によって分類に寄与する場合と寄与しない場合がある。例えば、「方式」という単語は、直前の単語が「鍵」の場合、分類細目「公開鍵方式」の一部である可能性があるが、「の（**格助詞**）」+「一（**数詞**）」のあとに続く「の一方式」の場合、分類には寄与しない。以下では、分類には寄与しない単語として削除する場合のパターンを記す[2]。

- 〜**動詞** (方法 | 方式 | 手法 | システム | 機構) → 〜**動詞**
- 〜**動詞** (**数詞**)? の (方法 | 方式 | 手法 | システム | 機構) → 〜**動詞**
- 〜の **数詞** (方法 | 方式 | 手法 | システム | 機構) → 〜

図 3.6に削除の例を示す。

- プラニングにおける連接的目標処理の一手法
  -> プラニングにおける連接的目標処理

- 複数の領域間の関係に基づいて概念を獲得するシステム
  -> 複数の領域間の関係に基づいて概念を獲得する

図 3.6: 単語列処理による削除の例

[2] **太字**は品詞を表す。

表題の分割

多くの論文表題は、いくつかの複合名詞句が機能語や動詞などの付属語相当句によって結び付けられた名詞句である。ある複合名詞句はその他の複合名詞句の分類細目の決定には寄与することが少なく、それぞれ複合名詞句が異なる専門的概念を示すことが多い。そのため、論文表題を複合名詞句、付属語相当句、のいずれかに分割する。具体的には、以下の方法で分割を行う。

- 論文表題が、表3.3のパターンにマッチする付属語相当句を含み、「X 付属語相当句 Y」という形であれば、「X」「付属語相当句」「Y」に分割する。
- 付属語相当句を、表3.1に基づき相当する補助分類コードに置換する（表3.3のパターンには存在するが、表3.1に相当する表現がない付属語相当句は、文の連体修飾と考えて補助分類コードを mod とする）。

表 3.3: 付属語相当句のパターン

の? ため の
を (目的 | 対象) と (した | する)
と して の
に ついて の
格助詞 サ変名詞 する
格助詞 動詞

図3.7に分割の例を示す。

3.3.3 標準化の実行例

図3.8に、標準化の実行例を示す。標準化によって出力されるほとんどの標準形は、

（複合名詞句 補助分類コード）* 複合名詞句

で表される。

• 知識獲得のための知識表現
  -> 知識獲得　のための　知識表現
  -> 知識獲得　!for　知識表現

• 知識工学を応用した機械設計 CAD
  -> 知識工学　を応用した　機械設計 CAD
  -> 知識工学　!use　機械設計 CAD

図 3.7: 単語列処理による分割の例

3.4 コード割当

自動分類の第二のステップであるコード割当処理では、標準化の結果として出力される論文表題の標準形の複合名詞句に含まれる専門用語を認識する。

以下ではまず、コード割当で参照している専門用語集について述べる。つぎに、専門用語との照合について述べ、最後に、分類コードの決定について述べる。

3.4.1 専門用語集

コード割当では専門用語集を参照することで、複合名詞句に含まれる専門用語を認識する。本システムでは、岩波情報科学辞典の用語の木に追加辞書を付加した専門用語集を使用する。専門用語集の一部を図 3.9に示す[3]。追加辞書は、用語の木の項目語である専門用語とそれに対応する類義語から構成される（図 3.10）。類義語の分類コードは、類義語に対応する専門用語のコードとする。

類義語として追加した用語は、

- 辞書で定義されている別名
- 末尾の「法」などを削除した専門用語
- 人手による方法

の 3 種類の方法で決定した。以下では、それぞれの方法について述べる。

[3] 下線部は追加辞書の専門用語

```
[ 1]
人工知能向きオブジェクト指向言語 Monad
-> 人工知能向きオブジェクト指向言語

[ 2]
知識工学の機械設計 CAD への応用
-> 知識工学を応用した機械設計 CAD
->  知識工学 を応用した 機械設計 CAD
->（知識工学 !use 機械設計 CAD）

[ 3]
故障診断用エキスパートシステムにおける知識獲得
-> 故障診断用エキスパートシステム における 知識獲得
->（故障診断用エキスパートシステム !in 知識獲得）

[ 4]
プロダクション・システムによる線画の解釈
-> プロダクション・システム による 線画の解釈
->（プロダクション・システム !use 線画の解釈）

[ 7]
電気ドリル分解・組立てコンサルタント・システム
-> 電気ドリル分解・組立てコンサルタント・システム

[ 8]
拡張ユニフィケーションを用いたパーザ IP の実現手法
->（拡張ユニフィケーション !use パーザ IP の実現手法）
```

図 3.8: 標準化の実行例

```
D o 知識システム knowledge system
D1 o 人工知能 artificial intelligence
D11 o 知識表現 knowledge representation
D12.3 o 推論 inference reasoning
D12.31 o （推論法）
D12.311 o 仮説推論 hypothetical reasoning
D12.311 b ＡＴＭＳ
D12.311 b 仮説選定
D12.312 o 後向き推論 backward reasoning
            :
```

図 3.9: 専門用語集の例

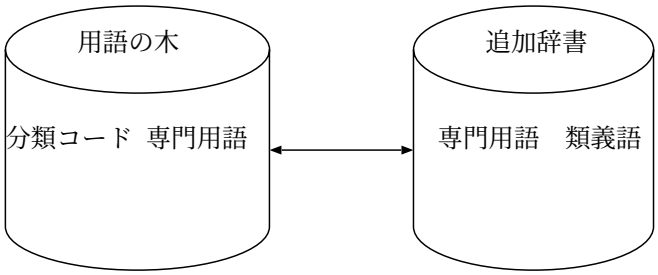

図 3.10: 専門用語集の構成の例

辞書で定義されている別名

岩波情報科学辞典の専門用語の説明文で、別名がしばしば定義される。これらの語は、用語の木の項目語ではないが、一般的に専門用語として論文表題で使用されることが多いため、類義語として辞書に追加する。具体的には、専門用語 A の説明文の中で「B(の略(称)?|(など)?ともいう)」という文が存在するならば、B は A の別名として定義されているため A の類義語として辞書に追加する。専門用語とその説明文の例を図3.11に示す。この場合、「網モデル」の別名として「ネットワークモデル」が定義されているので、類義語として追加する。この方法で、859 の語が追加された。

＜網モデル＞
ネットワークモデルともいう.
データモデルの一つで，データベーススキーマが２層のレコード型の階層構造の組合せで与えられるモデルをいう.
データベーススキーマはバックマン線図で表現される.
データモデルに属する類似の概念として階層モデル，関係モデルがあるが，前者とは１つの階層の下位階層に位置付けられた概念に複数の上位階層概念がありうることで，後者とはそもそも概念間の階層があることで区別される.
網モデルの典型的な例としてはジェネラルエレクトリック社のデータベースシステムＩＤＳ，ＣＯＤＡＳＹＬ仕様，国際規格のＮＤＬで採用されたデータモデルがある.
親子集合を利用して巡航操作が可能なほか，保留属性，挿入属性を指定して実務的に有効な一貫性を自然に表現できる.

図 3.11: 専門用語とその説明文の例

末尾の「法」などを削除した専門用語

専門用語の末尾の単語として「法、方式、システム」がしばしば使用される。これらの単語は、直前の単語が**サ変名詞**の場合、ほとんど意味を持たない場合もあるため、これらを削除したものを削除前の語の類義語として追加する。

- ～ **サ変名詞**（法|方式|システム） -> ～ **サ変名詞**
  例） フィボナッチ探索法 -> フィボナッチ探索

この方法で 93 の語が追加された。

人手による追加

類義語や表記のゆれと思われる 87 の単語を人手で追加した。追加した単語の一例を表 3.4に示す。

表 3.4: 人手で追加した用語の例

専門用語	類義語
（不確実性）	不確実な知識
（概念の学習）	概念形成
パラメター学習	パラメータチューニング
オブジェクト指向プログラミング	オブジェクト指向型プログラミング
トークンパッシング方式	トークンパッシングプロトコル
仮説推論	ＡＴＭＳ, 仮説選定
概念学習	概念獲得
:	:

3.4.2 専門用語との照合

複合名詞句に含まれる専門用語を見つけるために、JUMAN(juman3.5) を用いて形態素解析を行い複合名詞句を単語列に分解する。分割した単語列に対して、以下の方法を適用する。

1. 末尾からの最長一致

複合名詞句には専門用語が含まれることが多い。その専門用語には、

- より重要と思われる語が末尾に記述される。
- 専門用語は、短い文字数に比べより長い文字数の方が専門性が高い。

といった特徴がある。これらの特徴を考慮して、複合名詞句に含まれる専門用語を見つける。具体的には、複合名詞句の単語列に対し、末尾からもっとも長く一致する専門用語を見つける。例えば、複合名詞句「手続き的な知識表現」に対して、末尾の単語が「表現」であるため、「表現」で終わる専門用語にどのような専門用語があるかを調べる。「知識表現」「手続き的な知識表現」「ハイブリッド知識表現」などが候補となるが、「手続き的な知識表現」がより長く一致するので、この複合名詞句に含まれる専門用語として「手続き的な知識表現」が選択される。図 3.12に末尾からの最長一致による照合の例を示す（下線部は専門用語を示す)。

- 人工知能向きオブジェクト指向言語
- 知識工学
- 機械設計CAD

図 3.12: 末尾からの最長一致による照合の例

2. 末尾の削除可能語を考慮

表 3.5に示された単語が、複合名詞句の末尾の単語となることがしばしばある。これらの単語は、一般的にも利用されることが多く、実際に専門用語を決定する際に意味を持たない場合もあるため、これの単語を削除したものに対しても末尾からの最長一致による照合を行う。図 3.13に削除可能語を考慮した照合の例を示す。

- 医療知識ベース **システム**
  -> 医療知識ベース
- 分散型探索機構
  -> 分散型探索

図 3.13: 削除を考慮した照合の例

表 3.5: 削除可能語

機構	方式	方法	法	手法
概念	する	化	環境	階層
過程	定義	構造	上	下
システム	アルゴリズム	アプローチ	モデル	

3. 助詞や区切り記号以降を削除

複合名詞句には、表 3.6に示すような助詞、接尾辞や区切りのための記号がしばしば含まれる。それらの単語以降を削除したものに対しても末尾から最長一致による照合を行う。以下にそれぞれの理由について述べる。

表 3.6: 助詞などの単語

種類	単語
助詞	の、から、へ、に、を、と、で
区切り記号	, ・ / −
接尾辞	用

- **助詞以降削除の理由**
  複合名詞句に含まれる助詞の直前の文節は、複合名詞句を文に言い換えたときの格要素に該当する。例えば、「論理プログラムへの変換」に含まれる「へ」の直前の「論理プログラム」は、「論理プログラムに変換する」のニ格に該当する。また、格要素には専門用語が含まれることがしばしばあるため、複合名詞句の助詞以降を削除したものに対しても末尾からの最長一致による照合を行う。

- **区切り記号以降削除の理由**
  複合名詞句では区切り記号が、名詞を列挙するためにしばしば用いられる。並列に列挙された名詞はそれぞれに専門用語を含むことがあるため、区切り記号以降を削除したものに対しても照合を行う。

- **「用」以降削除の理由**
  複合名詞句に使用される「用」は、「のための」とほぼ同じ意味であり、直前の単語に専門用語が含まれることがあるため、「用」以降を削除したものに対しても照合を行う。

図 3.14に、例を示す。

4.「の、的、型」のスキップ

複合名詞句にはしばしば接尾辞「的、型」や従属接続詞「の」が見られる。これらの語をスキップしたものに対しても照合を行う。以下に理由を示す。

• 論理プログラム **へ**の変換
  -> 論理プログラム

• フレーム問題，非単調論理，イェール射撃問題の間の関係
  -> フレーム問題，非単調論理
  -> フレーム問題

• エキスパートシステム **用**ユーザインタフェイス管理システム
  -> エキスパートシステム

図 3.14: 助詞等以降を削除した例

- 複合名詞句に見られる接尾辞「的」「型」は、直前の名詞を形容詞化して直後の名詞に連体修飾して名詞句を形成しているが、直前直後の単語が直接結びついた複合名詞と意味的には同じである。そのため、これらの語をスキップしたものに対しても照合を行う。
- 複合名詞句に見られる「の」は従属接続詞として名詞と名詞を接続する働きがあるが、これを削除した複合名詞と意味的には同じである。そのため、これをスキップしたものに対しても、照合を行う。

図 3.15に、例を示す。

• 帰納 **的**推論
  -> 帰納推論

• 知識 **の**表現
  -> 知識表現

図 3.15: スキップした例

5. サ変名詞、名詞＋「化」の以降の削除

複合名詞句では**サ変名詞**や**名詞**＋接尾辞「化」といった単語がしばしば使用される。これらは動詞的意味を持ち、複合名詞句を文に書き換えたときの格要素に専門用語が含まれることがある。例えば、「メタ知識定義」は「メタ知識を定義する」の意味であり、そのヲ格「メタ知識」は専門用語である。そのため、**サ変名詞**や**名詞**＋「化」以降を削除したものにたいしても照合を行う。図 3.16に、例を示す。

- データベース論理設計 **支援**エキスパートシステム
  -> データベース論理設計

- データベース **統合化**ツール IKD
  -> データベース

図 3.16: 動詞的概念以降を削除した例

3.4.3 分類コードの決定

前節の方法により複合名詞句に含まれる専門用語を見つけた場合に、複数の専門用語が候補として得られることがある。複数の専門用語が得られた場合、以下の方法で専門用語を決定する。

照合範囲の包含関係

専門用語と複合名詞句の照合の際、その範囲がしばしば重複する。複数の専門用語のうち、ある専門用語の照合範囲がその他の用語の照合範囲を覆う場合、覆われた専門用語は削除する。例えば、「人工知能向きオブジェクト指向言語」に対して末尾からの最長一致により「オブジェクト指向言語」、サ変名詞「指向」以降を削除した最長一致により「オブジェクト」が得られるが、「オブジェクト」は「オブジェクト指向言語」に覆われるため削除される（図 3.17）。

一般的にも使用される語の考慮

「アルゴリズム、ネットワーク」などの用語は専門用語でありながら、一般的な単語としても使用される。複数の専門用語の中で、一般的にも使用される語が専門用語と共に出

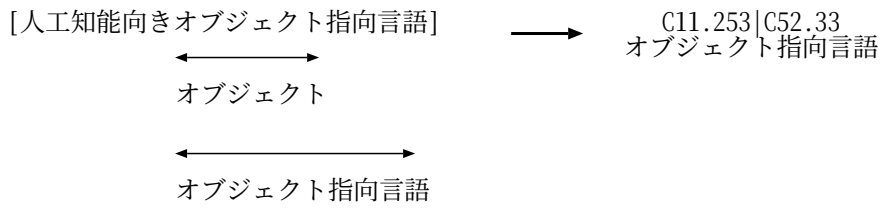

図 3.17: 対象範囲の包含の例

力された場合、一般的にも使用される語を削除する。例えば、「最良優先探索アルゴリズム」に対して「最良優先探索」「アルゴリズム」の 2 語が専門用語として得られた場合、「アルゴリズム」は一般的にも使用される語として登録されているため、「最良優先探索」だけがその複合名詞句の専門用語とする。一般的にも使用される専門用語は、

- 1 単語のサ変名詞
- 人手による選択

の 2 通りで行なっている。一般的にも使用される語として登録した語を図 3.7に示す。

3.4.4 コード割当の実行例

コード割当の実行例を図 3.18に示す。

3.5 分類システムの実行例

分類システムの実行例を図 3.19に示す。

表 3.7: 一般的にも使用される専門用語

サ変名詞	分散、量子化、符号化、信号、変調、 標本化、競合、認識、記憶、学習、 証明、演繹、帰結、否定、含意、 連接、導出、受理、模倣、手続き、 検波、命令、再構成、同期、暗号化、 分割、割付け、配置、配線、 検査、故障、保全、縮退、注釈、 展開、実行、分岐、探索、衝突、 併合、課金、保守、委託、内包、 推論、類推、類比、注意、省略、 分類、微分、立体視、遮蔽、交換
人手	アルゴリズム、ネットワーク、プログラム、 変更操作、再利用、論理、知識、予防保全、 試験、文、語、式、誤り、 雑音、組合せ、内部状態、視覚、集合、 写像、行列、関係

人工|知能|向き|オブジェクト|指向|言語
-> C11.253|C52.33:オブジェクト指向言語

機械|設計|CAD
-> B51.11:ＣＡＤ［１］ D34.622:ＣＡＤ［２］

故障|診断|用|エキスパートシステム
-> D13.1422|D35.23:故障診断エキスパートシステム

プロダクション|・|システム
-> D11.4:プロダクションシステム

電気ドリル|分解|・|組立て|コンサルタント|・|システム
-> D13.141:コンサルテーションシステム

パーザ|IP|の|実現|手法
-> D21.23:構文解析プログラム

類推|の|定式|化
-> D12.35:類推

オンライン|型|エキスパートシステム
-> D13.14:エキスパートシステム

仮説|選定|機構|の|一|実現|法
-> D12.311:仮説推論

フレーム|型|データ|構造|の|一|論理|的|記述
-> D11.6|D21.4221:フレーム

図 3.18: コード割当の実行例

```
[ 1]
人工知能向きオブジェクト指向言語 Monad
-> 人工知能向きオブジェクト指向言語
-> C11.253|C52.33:オブジェクト指向言語

[ 2]
知識工学の機械設計 CAD への応用
-> (知識工学 !use 機械設計 CAD)
-> B51.11:ＣＡＤ［１］, D34.622:ＣＡＤ［２］, D13.1:知識工学 !use

[ 3]
故障診断用エキスパートシステムにおける知識獲得
-> (故障診断用エキスパートシステム !in 知識獲得)
-> D13.13|D14.22:知識獲得,
   D13.1422|D35.23:故障診断エキスパートシステム !in

[ 4]
プロダクション・システムによる線画の解釈
-> (プロダクション・システム !use 線画の解釈)
-> D32.b1:線画の解釈, D11.4:プロダクションシステム !use

[ 5]
電気ドリル分解・組立てコンサルタント・システム
-> 電気ドリル分解・組立てコンサルタント・システム
-> D13.141:コンサルテーションシステム

[ 6]
拡張ユニフィケーションを用いたパーザ IP の実現手法
-> (拡張ユニフィケーション !use パーザ IP の実現手法)
-> D21.23:構文解析プログラム, A41.475|C13.636|C52.517:単一化 !use
```

図 3.19: 分類システムの実行例

第 4 章

実験と検討

本章では、本研究で作成した自動分類システムの有効性を評価するため、システムの出力結果と人間が決定した分類コードを比べ検討する。

4.1 実験

作成した自動分類システムを用いて、実際の論文表題を分類する実験を行った。人工知能学会誌の掲載された 369 編（1986 年〜1995 年）を実験対象とした。

それぞれの論文表題に対して正解（分類コードが複数個の場合もある）を与え、これと分類システムの出力結果の評価を行った。以下では、評価の方法と結果について述べる。

4.1.1 実験の評価

ある論文に対して、分類コードは一つしか存在しないのであれば、正しく分類できたかどうかでシステムを評価できる。しかし、実際には、ある論文に対して、複数の分類コードが存在することも多い。実際にシステムは論文が割り当てたカテゴリの数を表 4.1に示す。また、システムの割り当てたカテゴリが正しい分類カテゴリと誤った分類カテゴリのどちらともいえない場合もある（図 4.1）。

表 4.1: カテゴリ数と論文数

割り当てたカテゴリ数	0	1	2	3	4	5
論文数	42	160	112	40	11	4

仮説時間推論 -> D12.3:推論　（適切な分類コードは「D12.311:仮説推論」）

図 4.1: 正しくも誤りでもない分類の例

本論文では、表題を構成する複合名詞句に対し、

○ 正しい分類

△ 正しい分類とも分類誤りともいえない

× 分類誤り

と 3 段階に評価する。論文が 1 つの複合名詞句からなる表題を持つ場合は、○であればが論文を正しく分類、△や×であれば論文を誤って分類したと評価する。複数の複合名詞句からなる表題に関しては、

- 1 つの複合名詞句に対し○で、かつその他の複合名詞句に対して○もしくは△であるものは、正しく分類。
- それ以外の場合は、分類誤り。

と評価する。2 つの複合名詞句からなる表題の場合、表 4.2のようになる。また、分類すべきでない論文（正しい分類項目が存在しない論文）も存在するため、結果は表 4.3のように整理される。 ここで、分類すべきでないものを分類した場合、それが正しい分類になることはないので、この表において値が入るのは A から E の 5 箇所となる。

表 4.2: 2 つの複合名詞句からなる論文の評価

1. 複合名詞句	2. 複合名詞句	論文
○	○	○
○	△	○
○	×	×
△	△	×
△	×	×
×	×	×

表 4.3: 分類の評価

	分類した		分類しない
	正しい	誤り	
分類すべきもの	A	B	C
分類すべきでない	-	D	E

A 分類すべき項目が存在し、システムは正しく分類した。

B 分類すべき項目が存在するが、システムは誤って分類した。

C 分類すべき項目が存在するが、システムは分類しなかった。

D 分類すべき項目は存在せず、システムは誤って分類をした。

E 分類すべき項目は存在せず、システムは分類しなかった。

評価の際には、3 つの精度について考える。

- カバレジ：このシステムはどれだけの論文に対して分類することができるか

$$\text{カバレジ} = \frac{A+B+D}{A+B+C+D+E} \tag{4.1}$$

- 分類精度：分類された論文のうち、正しく分類された論文はどれくらいか

$$\text{分類精度} = \frac{A}{A+B+D} \tag{4.2}$$

- 実用精度：このシステムはどれだけの論文に対して正しく分類することができるか

$$\text{実用精度} = \frac{A}{A+B+C+D+E} \tag{4.3}$$

補助分類コードの種類が少ないため、付加されるべきすべての主分類コードに対して補助分類コードは正しく付加できた。そのため、補助分類コードの有無について、分類の評価には考慮しない。

表 4.4: 分類実験の結果

	分類した		分類しない
	正しい	誤り	
分類すべきもの	292	37	31
分類すべきでない	-	0	9

4.1.2 実験の結果

実験結果を、図 4.4に示す。精度を以下に示す。

$$\text{カバレジ} = \frac{329}{369} = 89\% \tag{4.4}$$

$$\text{分類精度} = \frac{292}{329} = 89\% \tag{4.5}$$

$$\text{実用精度} = \frac{292}{369} = 79\% \tag{4.6}$$

4.2 検討

正しく分類出来なかった（B, C, D, E）論文について調査を行った。その結果を、表 4.5 に示す。ここでは、分類誤りを (a)〜(f) の 5 つに分類した。

表 4.5: 誤り原因の分析

誤りの原因	B	C	D	E	計
a. 専門用語がない	6	8	0	0	14
b. 抽象的な専門用語	7	4	0	0	11
c. 形態素解析の誤り	5	0	0	0	5
d. 標準化の誤り	2	0	0	0	2
e. 人間でも分類不能	0	0	0	9	9
f. その他	17	19	0	0	36
計	37	31	0	9	77

a. 専門用語がない

複合名詞句に含まれる専門用語が、しばしば辞書に登録されていない。科学技術論文は新しい技術の報告であるため、岩波情報科学辞典の出版以降に発生した新しい専門用語が、表題中に使用されることがある。例えば図 4.2の「ユーザモデル」「CBR[1]」は、現在は専門用語として一般的に認知され表題でもしばしば使用されているが、辞書には登録されていない。

この問題は新しい語を辞書に追加することにより解決できると考えられる。。

1. ユーザモデルを利用した説明文プランニング
（ユーザモデル !use 説明文プランニング）
説明文プランニング -> D12.25:（計画作成）
ユーザモデル ->

2.CBR システムの構築環境 ->

図 4.2: 専門用語が存在しないの例

b. 抽象的な専門用語

専門用語では、抽象的な単語がしばしば使用される。具体的には、「例からの学習」という専門用語の「例」「学習」は抽象的な表現で具体的に示すものは無い。しかし、実際、論文表題では具体的な事例が記されることがあるため、分類誤りの原因となる。図 4.3における「組み立て文」「組立手順の生成」を抽象的な単語で表すと「例」「学習」となる。

1. 組み立て文からの組立手順の生成

図 4.3: 抽象的な専門用語の例

c. 形態素解析の誤り

複合名詞句に専門用語が含まれているのに、形態素解析の誤りのため認識できないことがある。本システムで形態素解析に用いた JUMAN(juman3.5) は連続するカタカナや

[1]Case Based Reasoning（事例ベース推論）

1. 交換ソフトの領域モデルに基づくデバックエキスパートシステムの開発
（交換ソフトの領域モデル !use デバックエキスパートシステム）

2. 設計と試験を統合的に支援する知的 CAD/CAT システム
（設計と試験を統合的に支援する !mod 知的 CAD/CAT システム）

図 4.4: 形態素解析の誤りの例

英数字を分割しない。そのため、図 4.4に示すように、「デバックエキスパートシステム」「CAD/CAT」はそれぞれ 1 単語となり、「エキスパートシステム」「知的 CAD」を得ることができない。

d. 標準化の誤り

標準化で、分割しない方がよいと思われる表題を分割してしまうため、分類誤りとなった。例えば図 4.5の表題の複合名詞句「単一例による学習」において、「による」が付属語相当句と認識され分割される。この場合の「による」は、「からの」とほぼ意味で用いられているため、標準化の前処理の段階で置換されることが望ましい。

1. 単一例による学習とパターン認識
（単一例 !use 学習とパターン認識）
パターン認識 -> D31:パターン認識

図 4.5: 標準化の誤りの例

e. 人間でも分類不能

人間でも表題情報のみからでは、適切な分類カテゴリを決定することが困難だと思われる論文（図 4.6）。

1. 属性値の差異に基づくカテゴリー形式モデルの実験的検討
（属性値の差異 !use カテゴリー形式モデルの実験的検討）

図 4.6: 人間でも分類不能の例

f. その他

論文表題とそれに相当する専門用語の表層上の一致が少ないため、専門用語を決定するには高度な推論が必要であると考えられる。例えば、図 4.7の例 1. は、「実例に基づく翻訳」から「機械翻訳」が分類カテゴリとして正しいと考えられるが、これは「情報科学分野の論文であるから、『実例に基づく翻訳』を機械的に実現する方法に関する論文であろう」という推論が必要と考えられる。また、例 2. は、「概念項目の自動抽出」から「概念学習」が適切と考えられるが、これは「『項目』は意味をもたないので削除し、『自動抽出』は『学習』に含まれるだろう」という推論が必要である。このように表層情報以外から推測することは困難である。

1.MBT2：実例に基づく翻訳における複数翻訳例の組合せ利用
実例に基づく翻訳における複数翻訳例の組合せ利用

2. 対訳辞書からの概念項目の自動抽出
対訳辞書 -> 辞書 [1]

図 4.7: その他の例

第 5 章

結論

従来、論文の分類には、論文本文に出現するキーワードの集合とその頻度に基づいて、その論文がどの分野に属するかを決定する方法が用いられている。しかし、実際に人間が論文を分類する際、その表題やキーワード情報を利用するだけで、適切な分類細目（カテゴリ）を決定できる。それは、科学技術論文の表題は、論文の最も短い要約となっており、論文の内容に密接に関連した専門用語が表題に含まれることが多い、という理由によると考えられる。そこで、専門用語集を用意することによって、論文表題からその論文の分類細目を機械的に決定できる可能性がある。

本研究では、このような考え方にたって、専門用語集を用いて表題を解析することにより、科学技術論文を自動的に分類する手法を提案した。

作成した自動分類システムは二段構成をとる。

- **標準化：**動詞や機能語を手がかりに論文表題をいくつかの部分要素に分割し、整形する。具体的には、論文表題を複合名詞句、補助分類コードのいずれかに分割する。
- **コード割当：**標準化の結果として得られる論文表題の部分要素に含まれる専門用語をみつけ、分類コードを決定する。具体的には、専門用語集を用いて複合名詞句に含まれる専門用語を見つけ、それを割り当てるべき分類コードとする。

表題解析による自動分類手法の有効性を調べるため、人工知能学会誌に掲載された 369 論文に対して、分類実験を行った。専門用語集と分類体系として岩波情報科学辞典の用語の木を用いた。実験の結果、369 論文中の 292 論文に対して、システムは正しい分類コードを割り当てることが出来た。その精度は、79%となる。分類誤りの 77 論文のうちの 36 論文は、正しい分類コードの決定のために、何らかの推論が必要なものであった。25 論

文は辞書の問題であり、5 論文は形態素解析の誤り、2 論文は標準化の誤りであった。9 論文は人間でも表題から分類カテゴリを決定することはできないものであった。

今後の課題としては、実験規模の拡大による本手法の有効性の検討や、キーワードの利用、より詳細な表題の解析 [8]、言い換え [9][10] の導入により精度の向上を図ることで、分野別の論文集の自動生成などに応用できると考える。

謝辞

本研究を進めるにあたり、多くの御教示を賜りました佐藤理史助教授に深く感謝致します。そして、日ごろから技術的にも精神的にも支援してくださいました知識工学講座の皆様に心から感謝の意を表します。

参考文献

[1] Philip J. Hayes and Steven P. Weinstein, Construe-TIS: A System for Content-Based Indexing of a Database of News Stories, In Alain Rappaport and Reid Smith (eds.), *Innovative Applications of Artificial Intelligence 2*, pp.49–64, AAAI Press/MIT Press, 1991.

[2] 佐藤円, 佐藤理史, ネットニュース記事群の自動パッケージ化 情報処理学会論文誌, Vol. 38, No. 6, pp.1225–1234, 1996

[3] 長尾眞, 黒橋禎夫, 佐藤理史, 池原悟, 中野洋, 言語情報処理, 岩波書店, 1998

[4] 長尾眞編, 自然言語処理, 岩波書店, 1996

[5] もり・きよし原編, 日本十進分類法 新訂９版, 日本図書館協会, 1995

[6] 長尾眞 他編, 岩波情報科学辞典, 岩波書店, 1990.

[7] 松本裕治, 黒橋禎夫, 妙木裕, 新保仁, 長尾眞, 利用者定義可能な日本語形態素解析システム JUMAN 使用説明書, 京都大学工学部長尾研究室, 1991.

[8] 松村敦, 池田和幸, 高須淳宏, 安達淳, 構造化インデクスを用いた情報検索システム, アドバンスト・データベース・シンポジウム '97 論文集, pp151-158, 1997

[9] 佐藤理史, 論文表題を言い換える, 情報処理学会研究報告, 98–NL–127, pp187–194, 1998.

[10] 近藤恵子, 佐藤理史, 奥村学, 「サ変名詞＋する」の動詞への言い換え, 情報処理学会研究報告, 98–NL–127, pp179–186, 1998.