分野情報を利用した特許文の分類*

4B−6

磯山秀幸†

NTTデータ通信株式会社‡

1 はじめに

テキストの分野分類の精度を高めるには、分野特有の情報をいかに利用するかが重要である。我々は、分野情報を二段階に適用して分類を行なう方法を用いることにした。すなわち、分野ごとのキーワード等を用いたパターンマッチングによる得点づけと、他分野との関連を考慮したルールに従う分野決定である。本報告では、日本語の特許文を分野分類の対象として、分類実験を行なった結果と、この適用性について報告する。

IPC(国際特許文類)とは

今回の分類実験に用いた特許文(明細書)は公開特許公報のテキストデータである。公開特許公報には、国際特許分類記号(IPCコード)が付与されて発行されている。

国際特許分類の構造は、全技術分野を8大区分(セクション)に分け、その中をクラス、サブクラス、メイングループ、サブグループと、さらに細かく展開するという方法を採用しており、各階層ごとに記号が割り当てられている。例えば、今回分類に用いた「G06F 15/20」というIPCコードは仮名漢字変換やワードプロセッサなどの特許が属する分野に対応するが、これは、表1のような構成になっている。

G(セクション)	物理
06(クラス)	計算;計数
F(サブクラス)	計算の少くとも一部は電気的に行なわれるデジタル計算機;計算機デジタルデータを取り扱う装置
15(メイングループ)	グループ7/00と少くとも他の1つのメイングループとに含まれる機能の組合せを特徴とするデータ処理装置
20(サブグループ)	特定の用途に適合される計算部分の設計または構成

表1: IPCの階層的構造の例

今回の特許文の分類実験では、処理対象とした特許文にIPCコードを付与することで、その特許文がどの分野に属しているかを表すことにした。

2 分類のアルゴリズム

分野情報を二段階に適用して分類を行なう方法について、説明する。はじめに、パターンマッチング用のルールと、分野決定用の制御ルールを作成しておく。

パターンマッチングルール

各々の分野に各分野ごとに、その分野に特徴的な表現・用語をパターンマッチングルールとして登録する。

制御ルール

制御ルールには、分野仮定の基準となる各分野ごとの得点のthresholdと、分野決定用の「if ～ then」ルールを記述する。

特許文ではある分野と別の分野が、一方的にまたは相互に関連している場合がある。この場合、一方の分野が仮定されていれば(thresholdに達していれば)、もう一方の分野の得点がthresholdに達していなくても、該当分野として採用するようなルールを記述する。表1の「15(メイングループ)」の記述によると、'G06F 15/'と'G06F 7/'が関連した分野と考えられる。

第一段階: 入力されたテキストをスキャンし、定義したパターンにマッチする度に対応する分野に得点を加算していく。その分野の得点がthresholdに達すると、現在処理している特許文がその分野に属すると「仮定」する。

第二段階: 仮定された分野と他の分野との関連を調べ、妥当なら、その分野に属するとして「決定」する。

3 分類実験および考察

今回の実験では、'G06F 15/20(仮名漢字変換やワードプロセッサ等に関する分野)'の分野を分類するルールを64件の特許文から作成し、ルール作成に用いたものとは別の特許文115件[1]を用いて、分類実験を行なった。

自動抽出キーワードを用いた分類

制御ルールについてはプログラマが作成しなければならないが、パターンマッチングルールについては、できるだけ人手に頼るのを避け、自動作成することを目指した。ルール作成には、以下のようなBottom Up方式を用いた。

1. 目的の分野および関連分野に属する特許文を選ぶ。
2. 既存のキーワード抽出ツールを用いて、各々の特許文からキーワードを抽出する。

*Classification of Patent Documents using Domain Specific Information

†Hideyuki ISOYAMA

‡NTT DATA COMMUNICATIONS SYSTEMS CORP.

[1]1988年以前に公開された特許から89件。1992年に公開された特許から27件。

3. 最下位のサブグループレベルにおいて、同一のキーワードが異なる分野 (IPC) に出現していれば、そのキーワードは上位のメイングループレベルのキーワードとする。
4. メイングループレベルにおいても、3. の処理を行ない、サブクラスレベルのキーワードを求める。同様にして、最上位のセクションレベルまでのキーワードを求め、パターンマッチングルールに登録する。

分類実験の結果を表2に示す。

IPC	Recall(%)	Precision(%)
G	98	99
G06	98	99
G06F	76	94
G06F 15	22	38
G06F 15/20	27	29

$$Recall = \frac{Correct}{Expected}, Precision = \frac{Correct}{Assigned}$$

表 2: 自動抽出キーワードを用いた分類の結果

表2によると、セクション〜サブクラスまでのレベルの分類には、Recall, Precision ともに良好な精度が得られている。これは、任意の特許文に対して、コンピュータに関連する特許を分類することができる程度である。しかし、メイングループ以下のレベルについては、十分な分類精度が得られなかった。これには、以下の理由が考えられる。

1. パターンマッチングルールの元になったキーワードは、プログラムによって抽出したものであるため、必ずしもすべてのキーワードが、その特許文の特徴を正しく表しているわけではない。
2. キーワードを抽出した特許文の量が十分ではない。

ルールをチューンアップした場合

上記で作成したパターンマッチングルールについて、不適当なルールを削除し、以下のようなルールを加えることで、パターンマッチングルールのチューンアップを行なった。チューンアップ後の分類結果を表3に示す。

パターン1 パターンマッチングルールに登録されている語に対して、同義語・類義語を補う。

例: 機械翻訳 → 言語翻訳 or 翻訳機械 or 自動翻訳

パターン2 単純な単語だけでなく、複数の語からなる表現を加える。

例: 文書作成 → 文書 … 作成

(「文書を効率良く作成できる」のように“文書”と“作成”の間に他の語が含まれてもマッチする。)

これらのパターンマッチングルールを追加することで、メイングループ以下のレベルに対しても分類精度を向上させることができた。今回は人手によってパターンマッチングルールのチューンアップを行なったが、パターン1の方法については、同義語・類義語辞書を用いることで、またパターン2の方法についても形態素解析技術を応用することで、パターンマッチングルールを自動的に展開することは可能である。

IPC	Recall(%)	Precision(%)
G	100	99
G06	100	99
G06F	100	93
G06F 15	34	54
G06F 15/20	50	62

表 3: チューンアップ後の分類結果

4 さらなる改善案

さらに、特許文の分類精度を向上させるのに有効と思われる方法を述べる。

キャッチワードを用いた分類　通称「キャッチワード」と呼ばれる『技術用語による特許分類索引』が、実際のIPCコード付与作業において利用されている。よってキャッチワードは、特許文の分野別情報を広範囲に網羅していると考えられる。キャッチワードをパターンマッチングルールに利用することで、分類精度の向上が期待できる。

項分け記載を利用した分類　従来の特許明細書の本文は、【特許請求の範囲】【発明の詳細な説明】の項にしか分けられていなかったが、近年では、「項分け記載」として、【産業上の利用分野】【従来の技術】【発明が解決しようとする課題】【発明が解決しようとする課題】【課題を解決するための手段】などの項分けが追加されている。今回の実験では、特に利用しなかったが、パターンマッチを行なう際に、これら項分け情報を利用することで、分類精度・処理速度の向上が期待できる。

5 おわりに

日本語の特許文の分類を行なう処理において、分野別情報を利用した分類ルールを用いることで、国際特許分類の中分類までは、実用的な精度を得ることができた。またルールのチューンアップを行なうことで、更に下位分野の分類についても、精度向上が可能なことが分かった。今後は上記の改善案を検討し、大量データによる検証を行っていく。

参考文献

[1] 特許庁 編集, 社団法人 発明協会 発行: IPC(第五版) 特許 実用新案国際特許分類表
[2] 特許庁 編集, 日本特許情報機構 発行: 技術用語による特許分類索引