テキスト分類用辞書の自動学習

桜井 裕

北陸先端科学技術大学院大学 情報科学研究科

1999 年 2 月 15 日

キーワード： テキスト分類, 情報抽出.

近年、インターネットの普及により、誰もが自由に大量のテキスト情報をインターネットを通じてアクセス可能となったが、その中から必要な情報を素早く見つけ出すことは容易ではない。

これを解決する一つの手法として、テキスト自動編集がある。これは、ユーザの情報探索が容易になるように、あらかじめテキスト情報を編集しておくものである。このようなシステムの一つに、Sun QA-Pack がある。Sun QA-Pack は、Sun ワークステーションを対象としたニュースグループの記事を、幅広いユーザの利用を考慮に入れて、質問応答集としてパッケージ化したものである。Sun QA-Pack のシステムの 1 ステップにテキストの分類がある。これは、分類木構造をとる Sun QA-Pack の分類カテゴリにニュース記事をその内容に応じて分類するものであり、以下のように行なう。まず、記事本文から要約を生成し、分野固有語辞書 (分類に利用される分野固有語とその分類カテゴリを記載した辞書) を用いて、要約中の分野固有語を抽出する。次に、その分野固有語それぞれの分類カテゴリを得る。最後に、それを集計し、最も数の多い分類カテゴリをそのニュース記事の分類カテゴリとする。

このような分類方法を取るため、Sun QA-Pack の分類精度は、分野固有語辞書が分野固有語をどれだけ網羅しているかに強く依存する。しかし、Sun QA-Pack が対象とする分野では、新たな分野固有語が頻繁に出現する。これを手作業でテキストから抽出し、その分類カテゴリを判断し、辞書に追加するのはかなりの労力を要する。

このような背景より、本研究では、テキストから分野固有語を自動的、あるいは半自動的に抽出し、分野固有語辞書に自動的に追加する機能の実現について検討する。これを以下の 2 つを実現することによって達成する。

1. テキスト中から分野固有語を抽出する。

2. 抽出した分野固有語の属する分類カテゴリを推定する。

テキスト中から分野固有語を発見する手法について説明する。分野固有語とは、分類に有効に働く語である。分野固有語に成り得る語として、(a) ある特定の概念 (分類カテゴリ) を表す専門用語、(b) その分類カテゴリに属するプログラム名、システム名、製品名などの固有名詞、の 2 種類が考えられる。(a) は、数が限られているため、あらかじめ用意することが可能である。しかし、(b) は、頻繁に新たに出現する。このことより、テキスト中から製品名をさがし出すことで、分野固有語候補とする。Sun QA-Pack が扱うニュースグループは、質問応答型ニュースグループであり、Sun ワークステーションに関する質問とそれらに対する応答が数多く掲載される。このニュースグループの記事群では、プログラム名、システム名、製品名は、カタカナ、もしくは、英数字の並びで表記されることが多い。このことより、ニュース記事から、カタカナ、英数字列を抽出し、分野固有語の候補とする。

抽出した分野固有語の属する分類カテゴリを推定する手法について説明する。分野固有語候補の分類カテゴリをテキストの文脈を用いて推定する。この方法には、大きく分けて以下の 2 つの方法がある。

- 動詞とその格要素の関係を用いた方法
- 特定語とその前後の分野固有語との関係を用いた方法

動詞とその格要素の関係を用いた方法では、特定の動詞が、特定の分類カテゴリに属する分野固有語を、その格要素に取ることを利用する。

特定語とその前後の分野固有語との関係を用いた方法では、特定の語がその前後に特定の分類カテゴリに属する分野固有語を取ることを利用する。この方法には、大きく分けて以下の 2 つの方法がある。

- カテゴリ明示表現を用いた推定方法
- 並列関係を用いた推定方法

カテゴリ明示表現を用いた推定方法では、日本語には、あるもののカテゴリを明示的に示す表現として、「X という Y」という表現がある。これは、「X」のカテゴリは、「Y」であることを意味する。このことを利用して、「Y」の分類カテゴリが分かっている場合、「X」はその分類カテゴリに属すると推定する。

並列関係を用いた推定方法では、並列関係を構成する語に着目した場合、この語の前後に来る分野固有語は、同じ分類カテゴリに属することが多い。このことを利用して、前後どちらか片方の分野固有語分類カテゴリが既知の場合、もう片方の分野固有語の分類カテゴリを推定する。

これらの分類カテゴリ推定規則は、100%信頼できるものではない。これらの推定規則にそれぞれ信頼度を設定する。信頼度は高い順に a、b、c、d の 4 段階に分ける。この信頼度を分野固有語候補とその分類カテゴリごとに信頼度を集計した後、得点に変換する。

この得られた得点に対して閾値を設定することで、推定結果から、正しい推定結果のみを抽出する。

まず、この分類カテゴリ推定規則を用いて、分野固有語候補を、Sun QA-Pack が対象とする Sun ワークステーションのニュースグループで良く登場する OS、マシン、ハードウェア、ソフトウェア、の 4 つの大分類カテゴリのいずれかに分類することを考える。次に、Sun QA-Pack に使われる分野固有語辞書への学習を考えて、さらに深い階層の分類カテゴリ (詳細分類カテゴリ) の推定を行なう。

この分類カテゴリ推定方法が対象とするテキストは、ニュース記事だけに限定されるわけではない。より多くのテキストに対して適用することで、より多くの推定を行なうことができると期待できる。Web 上には多くのテキストが存在する。Web 上からテキストを取得し、このテキストに対して、分類カテゴリ推定方法を適用することで、より多くの推定を行ない、推定結果をの精度を向上させることを考える。

以上の手法に対して評価実験を行なった結果、本研究で提案した手法を用いることで、ネットーワーク上の色々なテキストを対象にして、そのテキスト中に存在する分野固有語の分類カテゴリをある程度推定することが可能なことが分かった。

今後の課題として、推定精度の向上を目指す。