1Q-7

学術論文からの研究資源情報の自動抽出

井上 慧 † 松原 茂樹 †† 長尾 確 ††
† 名古屋大学 工学部電気電子・情報工学科 †† 名古屋大学 大学院情報科学研究科

1 はじめに

研究活動においては研究資源の利用が不可欠である。ここで研究資源とは、研究で用いられるデータや資料、ツール、さらには既存手法や評価法などを意味する。これまでに膨大な量の研究資源が構築されており、その多くは様々な用途に利用することができる。しかし、現状では既存の研究資源が有効に活用されているとは言い難く、その理由として、研究資源に関する情報が十分に共有されておらず、利用者が自らの目的に合致した研究資源を容易に見つけることができないことが考えられる。

そこで本稿では、研究資源を検索・選択する際に有用となる情報を機械的に取得し提供することを目的として、学術論文から研究資源の名称を自動抽出する手法を提案する。研究資源名の抽出の手がかりとして、研究資源名の形態的特徴の他に、研究資源名以外の研究資源情報を用いる。

2 研究資源情報とその抽出

2.1 研究資源情報とは

本稿では、研究資源の検索・選択において特に有用となる情報を研究資源情報と呼ぶ。学術論文の本文には、研究で用いられた研究資源の名称（以下、研究資源名と呼ぶ）や、研究資源がどのようなものなのかを示す表現（以下、定義表現と呼ぶ）、どのような目的・用途で使用されたのかを示す表現（以下、用途表現と呼ぶ）などの研究資源情報が含まれている。例えば、「Radio Button は，拍打に基づくバトン型演奏プログラムである.」という文では、研究資源「Radio Button」に対して、「拍打に基づくバトン型演奏プログラム」の部分が定義表現に相当する。また、「単語への分割と品詞の推定には ChaSen を利用している.」という文では、研究資源「ChaSen」に対して、「単語への分割と品詞の推定」の部分が用途表現に相当する。

本稿で提案する抽出手法では、ツール、ウェブサービス、コンピュータ言語、データ・資料、電子機器・電子部品、汎用的な手法・指標、定理、法則などの研究資源の名称を取得対象とする。これまでに、特定の専門分野コーパスから専門用語を自動抽出する手法 [1] が提案されているものの、研究資源名を対象とした自動抽出については未だ実現されていない。

2.2 定義表現および用途表現の利用

研究資源名の抽出には、研究資源名の形態的特徴が手がかりとなる。研究資源には作成者が固有の名前を付けるケースが多く、研究資源名に未知語が含まれる可能性が高い。また、「法」や「アルゴリズム」など、研究資源名の末尾によく登場する表現も抽出の手がかりとなる。しかし、「自己組織化マップ」のように未知語や末尾表現を含まない研究資源名も存在するため、未知語や末尾表現のみを手がかりにするのでは不十分である。

一方、定義表現や用途表現を含む文には、研究資源名が記述されている場合が多い。2008 年度人工知能学会全国大会の発表論文集に収録されている 50 論文から研究資源名、定義表現および用途表現を人手で抽出した結果、定義表現を含む 127 文の全ての文に、また、用途表現を含む 152 文のうち 143 文に研究資源名が含まれていた。このことから、研究資源名の抽出には、定義表現および用途表現の持つ特徴が有用な手がかりとなると考えられる。

3 研究資源名の自動抽出手法

提案手法では、まず、研究資源名の可能性がある形態素列を、研究資源名の形態的特徴や既知の定義表現および用途表現などを手がかりにして、論文の本文テキスト中から取得する。取得した研究資源名候補の集合から、SVM を用いて研究資源名を抽出する。

3.1 論文 PDF からの本文テキストの抽出

抽出の前処理として、まず、pdftohtml[2] を用いて論文の PDF ファイルを HTML 形式に変換する。次に、文字列のフォントサイズ・位置情報を手がかりにして HTML から本文テキストを抽出し、文に分割する。

3.2 研究資源名候補の取得

以下の 3 種類の手法によってそれぞれ取得した形態素列を研究資源名の候補とする。

1. 未知語に基づく抽出

   形態素解析された本文テキストから未知語を取得する。その未知語の前後に、名詞・接頭辞・接尾辞など研究資源名の一部になりうる形態素が存在すればそれも取得し、得られた形態素列を研究資源名候補とする。図 1 の場合、未知語「Ｖｉｃｏｎ」を含む形態素列「Ｖｉｃｏｎモーションキャプチャシステム」を研究資源名候補とする。

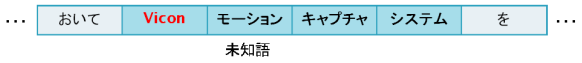

図 1: 未知語に基づく研究資源名候補の抽出

2. 既知の末尾表現に基づく抽出

   本文テキストから、「法」や「アルゴリズム」など、研究資源名の末尾によく登場する既知の表現を取り出す。末尾表現の前に、名詞・接頭辞・接尾辞など研究資源名の一部になりうる形態素が存

Automatic Extraction of Information on Research Resources from Technical Documents

† INOUE, Kei (kinoue@nagao.nuie.nagoya-u.ac.jp)
†† MATSUBARA, Shigeki (matubara@nagoya-u.jp)
†† NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)
Dept. of Information Engineering, School of Engineering, Nagoya University (†)
Graduate School of Information Science, Nagoya University (††)

在すればそれも取得し、得られた形態素列を研究資源名候補とする。図2の場合、既知の末尾表現「アルゴリズム」が末尾となる形態素列「遺伝的アルゴリズム」を研究資源名候補とする。

図 2: 末尾表現に基づく研究資源名候補の抽出

3. 既知の定義表現および用途表現に基づく抽出

同一の定義表現および用途表現が学術論文の集合中に複数回出現することは極めて稀であるため、定義表現および用途表現そのものを研究資源名の抽出に利用することはできない。そこで、定義表現は末尾の名詞（以下、定義名詞と呼ぶ）を、用途表現は末尾の動詞・サ変名詞（以下、用途動詞と呼ぶ）を研究資源名抽出の手がかりとする。例えば、定義表現「2値分類問題に対する有効な分類モデルの導出手法」の場合は「手法」が、用途表現「記事を主観的な意見を含むレビュー記事と非レビュー記事に分類する」の場合は「分類」が手がかりとなる。

また、研究資源名と定義表現が一文中に含まれている場合、研究資源名と定義表現の間には「とは」や「である」などの表現がよく出現する。同様に、一文中の研究資源名と用途表現の間には「を用いて」や「のために」などの表現がよく出現する。これらの表現（以下、手がかり表現と呼ぶ）は、研究資源名、定義表現および用途表現と、連接関係あるいは係り受け関係にあるため、研究資源名の抽出において有用な手がかりとなる。

抽出手法としては、まず、構文解析された本文テキストから既知の手がかり表現を取得する。手がかり表現の直前の形態素が既知の定義名詞あるいは用途動詞であった場合は、手がかり表現を含む文節の係り先の最後の形態素から、名詞・接頭辞・接尾辞など研究資源名の一部となりうる形態素を取得していき、最終的に得られた形態素列を研究資源名候補とする。図3の場合は、手がかり表現「である」の直前の形態素が定義名詞「手法」であるので、「である」を含む文節の係り先から研究資源名候補「自己組織化マップ」を取得する。

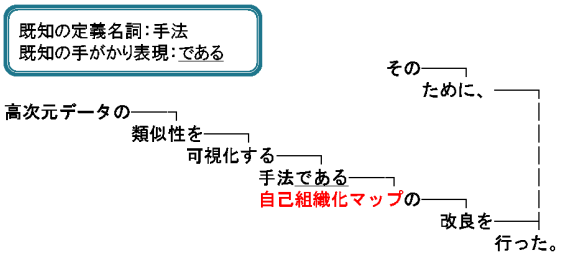

図 3: 定義表現に基づく研究資源名候補の抽出

手がかり表現を含む文節の係り先に、既知の定義名詞あるいは用途動詞が含まれている場合は、手がかり表現の直前の形態素から研究資源名の一部となりうる形態素を取得していき、最終的に得られた形態素列を研究資源名候補とする。図4の場合は、手がかり表現「を用いて」の一部「用いて」の係り先に用途動詞「構築」があるので、「を用いて」の直前の形態素から研究資源名候補「不完備情報ゲーム」を取得する。

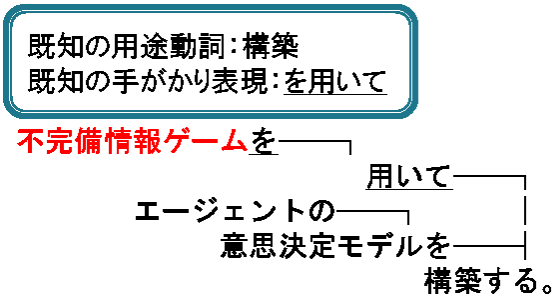

図 4: 用途表現に基づく研究資源名候補の抽出

3.3 SVM による研究資源名の抽出

本研究では、SVM.NET[3] を利用して学習を行う。学習用データの作成には、学術論文から人手で抽出した研究資源名、定義表現、用途表現、手がかり表現を用いる。前節で述べた手法によって研究資源名候補を取得し、人手で抽出した研究資源名に該当するものを正例、該当しないものを負例とする。SVM の素性としては、研究資源名候補の文字数、形態素数、使われている文字の種類、品詞の種類、既知の末尾表現の有無、一文中で既知の手がかり表現と共に現れるか、既知の定義名詞と共に現れるか、既知の用途動詞と共に現れるか、などを用いる。

4 おわりに

本稿では、学術論文からの研究資源名の抽出手法を提案した。今後の課題としては、人手で抽出した研究資源名を正解データとした提案手法の評価実験、研究資源名以外の研究資源情報の抽出手法の考案などがある。

参考文献

[1] 中川 裕志, 湯本 紘彰, 森 辰則, 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol.10, No.1, pp.27-45, 2003.

[2] pdftohtml: http://pdftohtml.sourceforge.net/

[3] SVM.NET: http://www.matthewajohnson.org/software/svm.html