様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ３ 日現在

機関番号：６２６１５
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００３０３

研究課題名（和文）専門用語管理支援システムの研究

研究課題名（英文）Research on a Term management Support System

研究代表者
小山　照夫（KOYAMA, Teruo）

国立情報学研究所・情報社会相関研究系・教授

研究者番号：８０１２４４１０

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：どのような研究分野においても、そこで使用される用語を把握し、管理することは重要な課題である。しかしながら特にわが国では、用語管理を十分な形で行っている研究分野は多くはない。この理由の一つが用語管理のための環境が整備されていないことにあると考えられる。本研究では用語候補提案機能を備えた用語管理データベースを、Web環境で利用可能な形で実装し、識者からの聞き取りによって、システムが有望な環境となり得ることを確認した。また、用語候補提案の基礎となる用語抽出機能の改善を併せて行った。

研究成果の概要（英文）：In research fields, administration of terms is an essential task for the advance of researches. In this study we implement a term management support system with term recommendation utility on a term database management system. The system is evaluated to be useful for term management tasks.
We also modify our term extraction algorithm so that the performance of extraction is highly improved.

研究分野：データベース

キーワード：用語抽出　専門用語　用語管理　データベース　外来語

１．研究開始当初の背景

様々な専門研究分野において、用語の整理と管理は重要な課題であると認識されているにもかかわらず、実際には多くの分野で用語の管理が十分に行われているとはいえないのが現状である。

用語管理が十分に行えない背景の一つとして、用語管理を行うための環境や、管理を支援するための枠組みが十分に整備されていないことが挙げられる。

用語管理の問題は、多くの作業者と、作業者の間の協調作業が要求される複雑な工程を伴うため、特定の研究分野において、用語管理の目的で実際にある程度の人手が確保できたとしても、複数の作業者が協調して用語辞書の編纂が出来る環境が整備されていなければ、作業者それぞれの負担はさらに大きなものとならざるを得ない。

様々な分野の用語専門家がそれぞれの分野で、複数の協同作業者とともに共通の用語集を管理できる、専門用語管理支援のための環境としてどのようなものが適切であるかを明らかにし、環境構築の方向性を明らかにすることが望まれていた。

一方で専門用語管理支援を行う上で、用語登録の負荷を軽減することは重要な課題である。この問題の解決に、専門文書からの用語抽出機能を活用することが有効であると期待される。

日本語専門文書からの用語抽出については、代表研究者等が複合語としての用語抽出方法を提案し、一定の成果を挙げていたが、一方でより抽出性能の高い方式を開発する必要性も存在していた。

２．研究の目的

研究開始当初の問題点把握に基づいて、実際に稼働する用語管理支援システムの枠組みを評価するためのプロトタイプシステムの構築と、その有効性を評価することが、研究の第一の目的である。

構築すべきシステムは、多くの研究分野の用語について分野ごとに独立した形で、関連情報を網羅的に管理できるとともに、各分野ごとに複数作業者が協力して並行的にデータ管理が行えるものでなければならず、また、利用にあたって利用者側で新規機器購入などの特別な環境整備の必要がないものであることが要求される。

このような目的で利用できるシステムのプロトタイプを実際に実装し、システムの動きを確認するとともにその有効性と課題とを明らかにする必要がある。

つぎに、既存の用語抽出アルゴリズムを改良し、より抽出精度を高めることにより、用語管理支援を高度化することがもう一つの目的となる。登録すべき用語のすべてを作業者が当該分野のテキストを参照しながら手作業で入力することは作業者の大きな負担となることが予想されるものであり、用語登録の元となるテキストに含まれる用語候補を信頼できる精度で示唆してくれる手段があるなら、用語登録にあたって価値が高いと考えられる。

研究代表者らはこれまでに複合語を中心とする日本語用語抽出の研究を進めてきているが、より一層の抽出性能の向上が望まれている。

このためには、用語抽出の元となるテキストや、その形態素解析結果を統計的に解析し、問題点を把握した上で、アルゴリズムの改善を図ることが必要であり、これが研究の第二の目的となる。

３．研究の方法

用語管理を支援する枠組みとして、第一に用語と用語間の関係、用語の元となる文献を統合的に管理・活用する枠組みを実装する必要がある。このためには、複数者が同時アクセス可能で、全文検索可能なデータベース管理システム上に必要なデータを登録して管理する環境を整備する必要がある。

このデータベース上のデータを、ユーザ側で特別な環境を整理することなく利用するために、インターネットを介して、通常のパソコンからブラウザ経由でアクセスが可能なWebアプリケーションとしてのユーザインタフェースを構築することが有効であると考えられる。

一方で、用語情報管理にあたって用語定義を支援する目的で、用語候補抽出機能を組み込むこととする。用語候補抽出は、あらかじめ登録された文献データから、最初の候補を取り出すだけではなく、必要に応じてユーザが追加する文献データに適用することにより、新規登録文献に含まれる用語候補を随時抽出し、蓄積・表示する機能を用意する必要がある。

用語データおよび文献データは、利用者が手持ちのパソコンから、ブラウザを介してオンラインで登録するほかに、ローカルに編集したデータをアップロードする機能も備えることが望ましい。

以上の要件を満足するシステムのプロトタイプを実際に実装し、それぞれの機能に関して識者の意見を聴取しながら評価を行うこととする。

用語抽出アルゴリズムの改善については、システムに統計解析システムへのインタフェースを組み込むことにより、文献情報や文献情報の形態素解析結果に対して統計解析を実施することにより、抽出精度を低下させている要因としてどのような物があるかを解明するとともに、精度をより高める方法を明らかにし、実際のシステムに組み込むこととする。

４．研究成果

用語管理支援システムとして、全文検索可能なデータベース管理システム上に、用語情

報、用語間関係情報、文献情報、用語候補情報等を管理する環境を実装した。ここではデータベース管理システムを中心とすることにより、複数分野の情報を、個別に管理できる環境を構築できる。各研究分野のユーザは、それぞれ指定されたデータベースにアクセスすることにより、独立した形で目的のデータのみを操作することができるようになる。

各分野のデータベースは、それぞれ複数のオペレータを登録することが可能であり、同時に複数の作業者が同一データベースにアクセスして共同作業を行うことが可能である。

このシステムを利用するためのユーザインタフェースとして、一群の Web アプリケーションを実装している。システムのユーザは、手持ちのパソコンからブラウザを用いることにより、インターネットを介してシステムにアクセスし、データベース内容を登録・変更することができる。

データベース管理システムは、複数データベースを同時に管理することが可能となっている。ここで一つの分野に一つのデータベースを対応させることにより、単独のシステム上で複数分野の用語をそれぞれ独立した形で管理することが可能となる。

システムを利用する際には、あらかじめ割り当てられたアカウントとパスワードが要求される。所定のアカウントにはそれぞれ対応するデータベースが定められており、一人のユーザはアカウントの対応する特定分野のデータベースに限ってデータの追加・編集が可能となっている。

用語データ、用語間関係データ、文献データは、それぞれブラウザを介してオンラインで登録・編集することも可能であるが、必要に応じてローカルに編集したファイルをアップロードする機能も備えている。この機能により、初期の情報登録をスムーズに行うことが可能になる。

用語に関する検討では、対象となる用語が出現する文献の調査が必要となる。システムは、ユーザが必要と判断する研究抄録文献データを登録し、必要に応じて活用することが可能である。

システムは用語候補を提案するための用語候補データを管理している。このデータはユーザが登録した文献から自動的に抽出されるほかに、オプションとしてあらかじめ用意された文献を登録しておくことにより、これらのデータからも用語候補抽出を行った結果を蓄積しておくこともできる。

用語候補テーブルに登録されたデータは随時検索・参照可能であり、検索結果表示画面から、直接用語として登録することも指定できる。

実際に実装したプロトタイプについて、情報科学技術協会で発表の機会を与えていただき、識者の意見を求めた結果、システムの有用性については高い評価が得られた。一方で Web ブラウザを介したユーザインタフェースについては、必要な機能は一通り備わっているが、より直感的に理解しやすいユーザフレンドリな形に改善することが望ましいという意見も得られた。

これらの要望を満足させるためには、Web アプリケーション実装のためのフレームワークの採用等が必要になると考えられる。

データベースの実装にあたっては、データモデルの選択が必要となるが、用語データと用語間関係データの間には、１対２の関係が存在しており、また用語間関係データ同士では逆関係を持つものもある。これらの関係はデータベーステーブル間のデータ依存関係としてはやや特殊なものであり、例えば Ruby on Rails を始めとする多くの Web アプリケーションフレームワークの標準的なデータ依存性管理の枠組みでは、必ずしも適切な一貫性管理の方法が提供されていない。

このことを考慮して、中間的なデータモデルを考案し、フレームワークで用意された依存性管理の枠組み内で一貫性の管理ができる枠組みを提案した。

用語管理では、単に用語と用語間関係を管理するだけではなく、文献における用語出現の傾向を把握する統計処理の枠組みが用意されることが望ましい。今回のプロトタイプシステムでは、統計解析手法を提供する R システムとのインタフェースを組み込むことにより、様々な統計処理の枠組みを提供するための環境も整備した。

これらの成果に基づき、実際に用語管理支援システムのプロトタイプを構築し、その有効性を評価した結果、基本的なシステムの有効性が確認された。

用語候補抽出アルゴリズムの改善により、用語抽出の精度を向上させることは、今回の研究のもう一つの目的である。

文献情報や文献の形態素解析結果情報を統計的な視点から検討することにより、いくつかの要因が抽出精度を低下させていることが明らかとなった。

抽出精度を低下させる一つの要因は、専門分野で特徴的に出現する形態素が、一般的な形態素辞書に登録されていないことである。これらの形態素を辞書登録してやることにより、抽出精度の向上が期待できる。

反対に、一般的な日本語文書の解析を目的とする場合には相対的に有用な形態素情報であっても、専門文書の解析にあたって誤りを生じる原因となる場合もある。このような形態素は逆に辞書から削除することにより、抽出精度が向上する可能性がある。

そこで第一の試みとして、形態素辞書の一部を追加し、問題となる可能性のある形態素情報を削除することを試みた。結果として、一定の抽出性能向上が達成できることが明らかとなった。

次に、形態素辞書項目を入れ替えた形態素解析結果と、従来の解析結果を比較すること

により、形態素解析器 Chasen と形態素辞書 IPADIC2.7.0 を使用した形態素解析では、専門文書に特有の文書パタンのいくつかについて、系統的な誤りを生じる可能性のあることが明らかとなった。

一例として、接尾辞「化」は一般的な日本語文書ではそれほど出現頻度が高くないのに対して、多くの専門分野文書では、複合語語尾として高頻度で出現する。この接尾辞の直後に接続助詞「および」や複合的格助詞「について」などが接続する場合、現在のシステムではこれらをしばしば動詞「およぶ」や動詞「につく」と誤って解析する傾向が見られる。

そこで、このような系統的な誤りパタンとして比較的頻度の高い 22 通りを取り上げ、解析結果中にこれらのパタンが出現した場合、強制的に正解と推定されるパタンに置き換えたうえで用語抽出アルゴリズムを適用する方法を採用した結果、抽出性能の大幅な改善を達成することができた。

専門分野のうちで特に理工系の文書では、外来語が数多く使用されているが、従来の手法では外来語については特別な取り扱いをせず、形態素解析器の解析結果をそのまま利用していた。

しかしながら現在一般に利用可能な日本語形態素解析器（Chasen, Mecab, Juman 等）では、外来語の扱いについて必ずしも適切な形にはなっていない。

外来語を含む文書からの用語抽出に関しては、本来は外来語部分に対する用語性判定が必要と考えられるが、日本語文書中に外来語が出現する場合、その多くが外来語部分全体で一つの用語ないしは複合語用語を構成する単位となっていると考えてよい。

このことから、アルファベット文字列およびカタカナ文字列が、空白やハイフンなどの一定の区切り記号を挟んで連続する場合、その全体を一つの外来語単位として置き換えた後、用語抽出を行う方法を試みた。結果は外来語、また外来語を複合語要素として含む用語の多くを、妥当な形で抽出することが可能となった。

以上をまとめたところ、本研究では用語とその関連情報を管理するデータベースシステムと、システム利用のための Web アプリケーションとしてのユーザインタフェースを実装することにより、用語管理支援システムのプロトタイプを作成し、その有効性について有望であるという評価を受けた。また、用語抽出アルゴリズムの見直しを行った結果、外来語を含む用語抽出をはじめとして、従来のシステムの性能を大幅に改善することが可能となり、結果として用語管理支援の目的で、より有効な手段を提供することが可能となった。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0 件）

〔学会発表〕（計 5 件）

① 小山照夫、竹内孔一: 外来語の扱いを考慮した日本語専門文書からの用語抽出、言語処理学会第 21 回年次大会(NLP2015)、 2015 年 3 月 18 日、京都大学大学院情報学研究科（京都府京都市）.

② 小山照夫、竹内孔一: 形態素解析の系統的誤りと用語抽出、情報処理学会自然言語処理研究報告、NL-220、2015 年 1 月 19 日、九州大学医学部百年講堂会議室（福岡県福岡市）.

③ 小山照夫、竹内孔一: 専門用語抽出における形態素辞書変更の効果、情報処理学会自然言語処理研究報告、NL-218、2014 年 9 月 1 日、首都大学東京（東京都八王子市）.

④ 濱田宏平、竹内孔一、小山照夫: 用語間関係を一貫して登録できる用語管理システム、 言語処理学会第 20 回年次大会(NLP2014)、2014 年 3 月 18 日、北海道大学工学部（北海道札幌市）.

⑤ 小山照夫、竹内孔一、濱田宏平: 用語管理システムの開発、情報処理学会自然言語処理研究報告、 NL-212, 2013 年 7 月 18 日、はこだて未来大学（北海道函館市）.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：

番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://reserach.nii.ac.jp/~koyama/official/tmdb/

６．研究組織
(1)研究代表者
小山　照夫（KOYAMA Teruo）
国立情報学研究所・情報社会相関研究系・教授
研究者番号：80124410

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
竹内　孔一（TAKEUCHI Koichi）
岡山大学大学院自然科学研究科・講師
研究者番号：80311174