用語管理システムの開発

小山 照夫[1,a)]　竹内 孔一[2,b)]　濱田 宏平[2]

概要：多くの研究分野において用語の整理と管理の重要性が指摘されながら、実際には十分な対応がなされているとは言い難い。この理由として、そもそも用語を管理する枠組み自体が十分に整備されておらず、また実際の用語管理を継続的に遂行していくにあたっては多大な人的資源を必要とすることなどが挙げられる。本研究では、適切に選定された研究領域ごとに、用語管理を総合的に行うデータベースを構築した上で、用語管理に関わる労力を低減する目的で、用語候補自動抽出機能を利用した、用語登録支援機能を備えたシステムを開発する試みについて発表する。

1. はじめに

現在では、多くの研究分野において精力的な研究が進められており、様々な成果が発表されている。これらの成果は一般に研究文献の形で公開されるが、文献中で取り扱われる主要な内容の多くは専門用語を利用して記述されている。このことから文献情報を適切に選択して高度利用するためには、当該分野でどのような専門用語が用いられているかを把握し、重要な用語については適切に整理し、管理する必要がある。

しかしながら、この用語の整理と管理をすべて人手によって行うことには大きな労力が伴う。どの研究分野においても、これまでに蓄積されてきた膨大な知識に対応する用語が用いられているし、さらに新しい研究成果によって新しい概念が提示されるにしたがって、その概念を記述するための新しい用語を追加していく必要も生じる。これらの作業に多大な労力を必要とする結果として、多くの研究分野では、その重要性が認識されながら、包括的な用語の整理と管理が十分に行われているとは言い難い状況が存在する。

この状況を改善するには、用語管理を行うためのフレームワークを備えたデータベースを構築するとともに、用語の登録に関する負担を緩和するための枠組みを確立することが重要である。本研究では、用語管理のためのデータベースシステムを構築し、システムに用語候補自動抽出機能を用いた、用語登録支援機能を統合することにより、この用語管理の問題を緩和する試みについて発表する。

[1] 国立情報学研究所
NII, Chiyoda-ku, Tokyo 101–8430, Japan
[2] 岡山大学
Okayama Uniersity, Okayama 700–8530, Japan
[a)] t_koyama@nii.ac.jp
[b)] koichi@cl.cs.okayama-u.ac.jp

2. システムの要件

本研究では、実用的な用語管理システムを実現するにあたっての要件として、次のものを想定している。

- 用語は本来分野特異性が高いものであり、広範に過ぎる研究領域について、そこで用いられるすべての用語を矛盾なく管理することは容易ではない。用語管理にあたっては、適切な適用範囲を選択できる必要がある。一般にはこの適用範囲としては、学会等を領域として選択することが考えられるであろう。システムはこれらの、ユーザが想定する領域ごとに用語管理データベースを構築することを可能とする必要がある。
- 一方、研究分野全般の幅広さを考えるなら、そのできるだけ多くの部分について用語の体系的整理が行われることが望まれる。ここでは、個別の領域をそれぞれ適切に選択した上で、できるだけ多くの領域に関するデータベースを平行して作成・管理できる必要がある。このことからシステムは、任意の数の領域について、独立したデータベースを個別に管理できる枠組みを提供する必要がある。
- 用語管理にかかる労力を考えるなら、それぞれのデータベースに対して唯一人しか保守を行うことができないのは現実的とは言えず、個々の領域データベースに対してそれぞれ複数人のメンテナが協同して同時に作業を進めることが可能でなければならない。このためには各領域のデータを管理する枠組みとして、データベース管理システムを用いることにより、同一のデータベースに対して複数人が同時にアクセスできる環境を提供する必要がある。

用語	関係	関係2	読み／コメント
インタプリタ言語			インタプリタゲンゴ
衛星画像			エイセイガゾウ
ＡＴＭ			エイティーエム poly
	Ｐ		Asynchronous Transfer Mode
	Ｐ		Automated Tellers Machine
	Ｒ	ＢＴ	情報端末装置
	Ｒ	ＢＴ	通信プロトコル
エキスパートシステム			エキスパートシステム
オブジェクト指向言語			オブジェクトシコウゲンゴ

図 1　データ内容例

- 実用的観点からは、実際の用語管理作業を行うにあたって特別なハードウェアやソフトウェアを必要としないことも重要である。このためには現状では Web 環境を利用することにより、各メンテナが、一般の PC 上のブラウザを用いて、ネットワークを介して用語保守作業ができる環境を構築することが妥当であろう。
- 用語や用語間関係（シソーラス）登録を個別に人手によって実行することはメンテナの大きな負担となることが予想される。システムはこの負担を低減させるための支援機能を備える必要がある。

今回紹介するシステムは、これらの要件を満足することを目標とするものである。以下ではシステムの中心となる用語関連情報のデータ管理と、用語候補自動抽出機能を用いた、用語登録支援機能についてその概要を述べる。

3. システムの基本データ

用語管理に当たっては、管理すべき基本データとして、用語その物に加え、一般にはシソーラスと呼ばれる用語間の関係、および用語の多義性に関するデータを管理することが要求される。前節で述べた要件から、これらのデータはそれぞれデータベース中の独立したテーブルの形で管理する必要がある。今回のシステムでは、各研究領域ごとに、用語テーブル、関係テーブル、多義語定義テーブルを用意している。システムの基本機能は、これらのテーブル定義に基づいて、各データを登録、検索、編集することとなるが、これらのデータの管理に当たってはいくつか考慮すべき問題がある。

3.1 多義語と概念

用語を取り扱う上で多義語の問題は避けるわけにはいかない。多義語が存在するということは、その語義ごとに区別される、相互に異なる概念があるということである。そこで、特に用語間関係などを厳密に扱うためには、表記と語義ごとの概念を区別して管理する必要がある。言い換えれば、同一の表記に対して考えられる語義ごとに複数の項目を設定し、それぞれについて保守を行わなければならない。しかしこのことは、保守に当たってメンテナが、対象とする表記に対してどのような語義が存在するかを常に意識しなければならないことを意味する。これはメンテナにとって大きな負担となることが予想される。

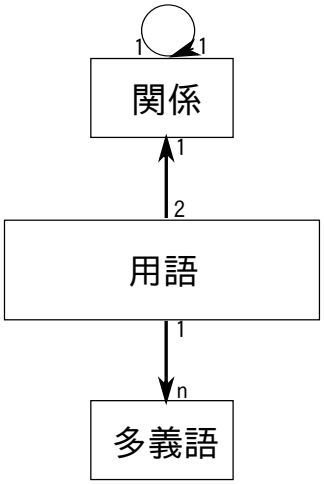

図 2　基本データ間の依存関係

一方で、管理されている用語を利用するのはもっぱら人間であり、機械による利用は別途方策を考えるとするなら、結果を利用する上で人間の柔軟な判断が働くと考えてよい。この場合、用語間関係を単に表記に関連付けて定義しても、大きな問題は生じないと考えられるであろう。そこで今回のシステムでは、単一の用語レコードに対して任意数の多義語レコードを別テーブルで定義する形をとることとし、用語間関係データは用語表記に対して定義することとした。図 1. は実際のデータベースから、多義語や用語間関係定義まで含む一部の項目を抜き出して表示した例である。ここで、関係に P とあるのは多義語定義、R とあるのは用語間関係を表す。また、BT は広義 (Broader Term) を表している。

ここに示すデータが人間の解釈の下に利用されることを想定するなら、このような形式で表示しても問題は生じないと考えられる。

3.2 データ間の依存関係

用語表記と多義語の関係も含め、データベースの各テーブルに格納されるデータの間には一定の依存関係が存在する。データの一貫性や健全性を考慮するなら、これらの依存関係を意識したデータ操作が必要となるが、ここで関係レコードと用語レコードの間の関係はやや特殊なものである。関係レコード自体は、関係の種類として同義、広義、狭義、その他関係、という四種類の関係を想定しているが、ここで一つの関係レコードは常に二つの用語レコードに依存することになる。複数のレコードが別テーブルの単一のレコードに依存する関係はごく一般的なものであるのに対して、単一のレコードが複数のレコードに依存する関係はあまり一般的なものとは言えない。用語間関係テーブルのレコードはまた、「その他関係」以外の関係は、逆関係も持

つことになる結果、一つのテーブルの中で相互依存関係にあるレコードが存在することになる。

依存関係は多義語レコードと用語レコードの間にも存在するが、この関係は一般に多く見られる、単純な一対多の関係となり、特に問題になる関係ではない。以上のことから、主要テーブルの間には図 2. に示す関係が存在することになる。実際のシステム実装においては、これらの依存関係を意識した上で正しく保守することが必要となる。

テーブル間の依存関係を正しく保つためには、本来は依存関係を記述するデータモデルを定義し、モデル上の制約関係として一貫性管理を行うことが望ましいが、実際にはやや特殊な性格を持つ依存関係が含まれているところから、現状では厳密な依存性のモデル化には至っていない。結果として現状の依存関係管理は、SQL レベルでの一連の手続きの形で実装されている。しかしながらこの方法では、様々な場所に記述された個別の処理を管理する必要を生じるところから、システム保守性という観点からは、将来的に問題を生じる可能性が否定できない。現在、適切なデータモデルに関する検討を行っており、将来的には明確に定義された制約関係を用いる形で一貫性を保証する枠組みの確立を目指す必要がある。

4. 用語登録支援機能

用語やシソーラスの登録はそれ自体大きな労力を要する作業であり、ここで何らかの支援手段を提供することが望まれる。本システムでは第一に、既に定義済みの用語データや用語間関係データが機械可読な形で存在する場合には、フォーマットを整えてシステムにアップロードする機能を用意している。これに加えて、特に新規用語登録を支援する手段として、著者等が以前に開発した日本語用語候補抽出システム [1] を利用することにより、新規用語定義の支援機能を実現することを試みている。本節ではその概要について述べる。

4.1 用語抽出のためのテキストデータ

用語候補抽出機能を利用するためには、用語抽出の対象となる文献（抄録）情報が必要となる。今回のシステムでは、参照文献テーブルと文献テーブルという二つのテーブルを用意することとした。ここで文献テーブルとは、新規用語登録の元となる文献データを随時登録するためのものであり、参照文献テーブルはそれ以外の、変更を必要としない関連文献データを蓄積しておくためのものである。これらの文献データから抽出された用語候補データはすべて用語候補テーブルに蓄積される。

4.2 ユーザに対する用語候補の提示

文献データから抽出される用語候補は、様々な局面で利用することができる。システムの先頭画面では、これまでに登録された文書および参考文書から抽出済みの用語候補を検索するフォームを用意しており、候補の部分文字列と抽出頻度を指定して、該当するものを検索することができる。

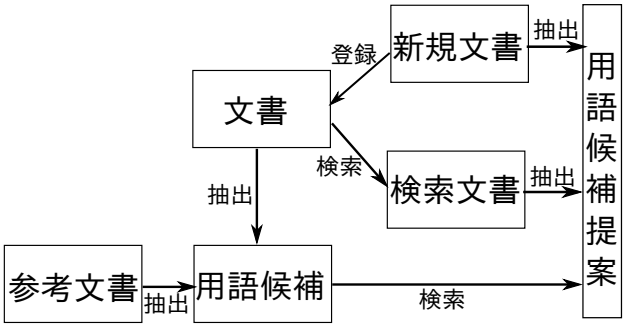

図 3 用語候補提示機能

ユーザは新しい文書を登録することができるが、これは文書を一件づつ登録することも可能であるし、また、あらかじめ編集したファイルに含まれる複数文書を一括してアップロードすることも可能である。いずれにせよ新規に文書が登録されると、新規文書ないしは文書集合から抽出される用語候補がその場で抽出される。ここで抽出された候補は、ユーザからのリクエストに応じて表示が可能である。

ユーザは登録された文書を検索することができる。検索された文書は、編集または削除が可能であるが、これ以外に、検索された集合に含まれる用語候補や、個々の文書に含まれる用語候補を表示させることができる。以上の用語候補提示の枠組みを図 3. に示す。

4.3 用語候補の登録

前節で述べたように、いくつかのタイミングで用語候補が表示されるが、表示された用語候補のそれぞれに対して、その取扱いを指定することができる。提示された各候補にはラジオボタンで、用語、非用語、保留を指定できるようになっており、用語ボタンを選択して登録することにより、そのまま用語として登録される。用語と判定された候補は、それ以降は候補として表示されることはない。同様に非用語と指定された候補も、それ以降は表示されなくなる。該当する用語候補は最大で 20 件まで表示されるようになっており、そのすべてを一括して一画面で処理できるようになっている。

用語登録にあたり、候補提示から直接用語登録が可能となることの効果は、メンテナがキーボード入力の大部分を省略できるところにある。検索や文書登録に当たって、一般的なものか、あるいはある程度分野を絞ったものかのコントロールも可能であり、候補の中に適切なものが存在するなら、個別にキーボード入力するよりははるかに少ない労力で用語登録が可能となる。一方で用語抽出が不成功に終わった用語については、個別に登録する必要がある。

5. システムの現状

現在、上記の機能を備えたシステムの実現を目指して開発を進めている。具体的には情報処理分野を例として、試験的データベースを構築し、部分的な機能を Apache Web サーバと、サーバ上の CGI によって実装し、機能毎に動作確認と機能評価に関する評価実験を進めている。

実装に当たって、データベース管理システムとしては、Mysql に全文検索機能をプラグインとして付加する mroonga[2] を使用している。また、参照文書として NTCIR-I[3] 学会発表データから抽出した、情報処理学会の抄録を利用している。この試験環境を用いて基本動作の確認とユーザインタフェースの評価を進めているが、今回のシステムの最大の特徴である、用語候補の提示については、十分な参照文献があらかじめ登録してあれば、用語登録支援の目的で有効であると考えている。

6. 今後の課題

システムの今後の課題としては、まず第一に個別の機能試験レベルにとどまっているシステム実装を、統合的なものとして、実際に運用可能なものとする必要がある。このシステム構築に当たっては、現在の CGI レベルのステートレスな実装では、実行効率やシステム保守の面で問題が多いと考えられる。

基本データの間の依存関係が明確なモデルとして確立されていないことは、もう一つの大きな問題である。SQL レベルでのアドホックな管理では、データの整合性を保証する上で明らかに不十分である。また、システム保守の面から考えても、保守性を低下させる要因となっている。

データモデルの問題を離れて、一般的にシステムの保守性を向上させることはもう一つの重要な課題である。単純な CGI による実装は保守性が高いとは言えず、保守性を向上させるためには、Web システム開発のためのフレームワークの導入等も検討しなければならない。

今回のシステムでは、とりあえず用語登録に関する支援機能だけを取り上げているが、実際には用語間関係（シソーラス）登録についても支援機能を実現することが望まれる。用語候補間の入れ子関係などを用いた支援機能の有効性についても検討を進める必要がある。

用語候補抽出にあたっては、形態素辞書に起因する抽出漏れが問題となる可能性がある。実際に文書で使用される形態素は、領域ごとに特徴を持っているはずであり、領域文書に適合した形態素辞書を使用することが望ましい。このためには、形態素辞書の内容を管理し、必要に応じて辞書を変更する機能を備えることが望ましい。現在、本稿で述べるシステムとは独立した形で、Chasen の形態素について編集と編集結果の確認を行うシステムを作成し、評価を進めているが、この枠組みについても、システムへの統合を検討したい。

システム評価についてはこれまでのところ、部分的実装を進めながら開発者の視点から機能的評価を中心としたものに限られているが、今後は実際にシステムを使用すると想定されるユーザの立場から、機能的側面に加えてユーザインタフェースの使い勝手についても評価を進める必要がある。

以上をまとめるならば、第一に、データモデルを確定した上で、保守性の高いフレームワークの下に、MVC 等の明解なモデルに基づく、効率のよいステートフルな実装に移行することが挙げられる。次に、実際のユーザにシステムを操作してもらい、機能面及びユーザインタフェースの側面からの評価を受けてシステムの改良を行う必要がある。さらに、用語間関係定義支援や、対象領域に適合させた形態素情報保守機能などについても検討を進め、最終的には有効な支援機能を備えた、総合的な用語管理システムの実現を目指したい。

謝辞：本研究は科学研究助成事業、基盤 (C)24500303 の援助の下に行われている。

参考文献

[1] 小山照夫、影浦峡、竹内孔一: “日本語専門分野テキストコーパスからの複合語用語の抽出”, 情処研報, 2006-NL-176, pp.55-60, 2006.
[2] http://mroonga.org/ja/
[3] KANDO, N., and NOZUE, T. eds.: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Proc. NTCIR Workshop I, 1999.