用語間関係を一貫して登録できる用語管理システム

濱田 宏平　　竹内 孔一　　小山 照夫
岡山大学工学部　岡山大学大学院　国立情報学研究所
koichi@cl.cs.okayama-u.ac.jp, t_koyama@nii.ac.jp

1　はじめに

テキストから専門用語を抽出する手法は多く提案されているが (例えば文献 [1])，専門文書から常に用語を収集しつつ人手で用語を管理するには適切な管理ツールが必要である [11]．そこで日常的な用語管理作業という観点に立ち，日本語の専門文書に対する用語収集機能を持ち，さらに用語間の関係を一貫して登録できる用語管理システムを提案する．

2　背景

専門用語辞書を構築する研究は大きく分けて (A) 単言語の用語管理に重視した管理システム [2]，(B) 多言語における専門用語翻訳対を重視した研究 [8])，(C) 用語を ontology として概念を固定させて派生関係などの集約に重視した研究 [4] [7] (D) 用語内の語構成の文法的・意味的性質に着目した研究 [5][6][3] が行われている．このように用語を単に抽出するだけでなく用語を管理して，文書構成や他言語訳語辞書の基本データ，ならび専門分野を理解する上での基本知識として利用することが期待されている．

上記で取り上げたシステムは英語を含む欧米の言語が中心であり，形態的な手掛かりを元にした用語抽出手法が基礎として利用されている．一方で，日本語に関する用語管理システムでは人手で用語を登録することが前提になっており，用語抽出システムを取り込んでいる物が見受けられない[1]．また目的として専門用語の翻訳対の構築が重視されており [9]，欧米のシステムのように概念や語構成といった学術的な側面よりも語の登録や入出力の利便性に重点が置かれている．よって現状では日本語の用語管理システムは (1) 用語抽出システムを取り込んだ用語管理システムの設計や，(2) 用語間の関係を登録する機能などが提案されていないように見受けられる．

[1]例えば google で「用語管理ツール」で検索した結果を参考にした．

そこで本研究では用語抽出機能を持ち，用語間の関係を簡単にかつ一貫して記録できる用語管理システムを提案し，ブラウザベースのインターフェースを利用できる Ruby on Rails 上で構築したので報告する．

3　用語管理システム

まず始めに本研究で仮定する用語管理を行う枠組みを設定した後，必要要件を整理し，具現化するための手法とモジュールを明らかにする．

3.1　仮定する用語管理の枠組み

用語かどうかの最終判断は人手で行う．よって判断以外の作業，例えば用語の一貫した記録や確認のための表示などはすべてシステム側の処理として構築する．用語は単独では存在しない場合が多いので [1]，用語間の関係を人手で一貫して定義して表示できる機構を持つ．作業者が用語かどうかを判定するための基礎データとして対象とする専門分野のテキストコーパスが存在すると仮定する．用語候補をテキストから取り出す用語抽出システムを内部に持ち，候補を作業者に提示して選択する機構を持つ．この時，日本語は分かち書きのために形態素解析システムを利用するが，専門分野では特定のパターンで区切り誤りを起こすことが想定されるので形態素の辞書を修正する機構を持つとする．

3.2　要求に基づく用語管理システムの機能の具現化

前節で設定した用語管理の枠組みを具現化するために用語管理システムとして下記の手法や機構を利用する．

m1 Web ブラウザと SQL によるインターフェース
人手による作業を効率的に行い，かつコストを掛

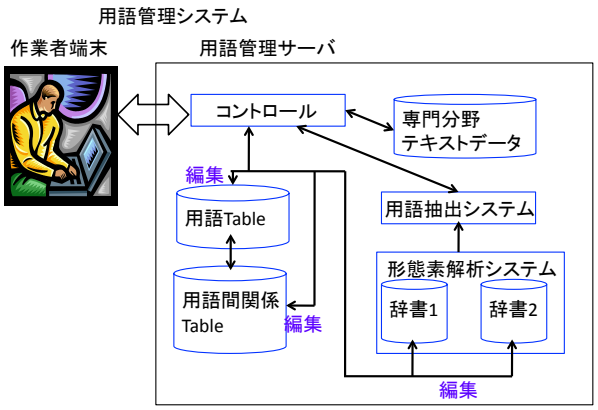

図 1: 用語管理システムの全体像

けず構築するため Web サイト構築で利用されているフレームワーク (具体的には Ruby on Rails) を利用する．専門分野テキストや用語，用語間関係は MySQL によるデータベースに登録して利用する．

m2 用語抽出機能
登録する専門分野テキストから形態素パターンを利用して候補となる用語をランク付けして表示する．このため形態素解析システムを内部で利用する．

m3 形態素解析システム修正機構
分野に応じて形態素解析システムの出力を変更するため，辞書拡張が容易な ChaSen を利用する．具体的には形態素辞書の登録を作業者が変更することで ChaSen の出力を変える．辞書の登録前・後で形態素解析システムの出力がどう変わったか差分を表示できる機構を持つ．

m4 用語間の関係の定義
用語間の関係として基本的な上位，下位関係，同義，関連の 4 つを仮定する．用語の関係は作業者が気がついた時に自由に登録できるようにする (4 節参照)．その時，重複した登録や矛盾が無いようにする．また検索の際にも上位・下位関係が全て提示できるようにする．手法として Rails の ActiveRecord を利用した登録・参照を利用する．

4 用語管理システム

上記で設定した機能を含む用語管理システムの構成を図 1 に示す．

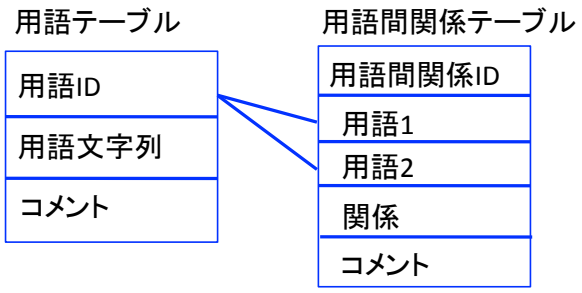

図 2: 用語テーブルと用語間関係テーブルの構造

用語管理システムは専門分野テキストデータ，用語，用語関係，形態素解析システム辞書 (2 つ) を DB 化して持っており，作業者から編集することが可能である．専門分野テキストデータは NTCIR-1 および 2 で作成された論文著者抄録データを利用する．現段階では情報処理分野の抄録を入れてシステムの挙動を確認している．形態素解析システムは辞書を切り替えて出力させることが可能で，辞書を編集した場合の形態素解析結果の差分確認が行えるようになっており先行研究ですでに実装している [10]．今回は用語登録モジュールと用語間関係登録モジュールについて新たに実装したので次節以降で具体的に記述する．

4.1 用語テーブルと用語間関係テーブルの構造

用語間の関係は基本的なものとして，上位 (Broader Term (BT))，下位 (Narrower Term (NT))，同義 (Synonym (SS))，関連語 (Related Term (RT)) の 4 つについて人手で一貫して登録する．用語間の関係を定義づける用語と用語テーブルは一致する必要があるので，用語間関係の内容は用語リストの ID 間の関係を記述することになる．よって図 1 に示したように，用語テーブルと用語間関係テーブルの間には 1 対多関係で関連づける必要がある．用語テーブルならびに用語間関係テーブルの構造を図 3 に示す．用語間関係テーブルでは 2 つの用語を用語 1，2 としてそれぞれ対象する ID を記録する．また関係は用語 1 から見た用語 2 に対する関係を登録する．例えば「プログラミング言語」から見て「アセンブリ言語」に対する関係は NT とする．

4.2 用語間関係の一貫した登録と検索

用語間の関係を登録する際の問題として，関係は用語間に対して定義する物であるのに対して MySQL におけるテーブルで実装する場合，用語 1 と用語 2 という方向性のある記述で登録する必要があるという点である．これにより，用語間関係テーブルとして 1 つのテーブルで管理したとしても，登録の際に用語 1 と用語 2 を入れ替えた関係を重複して登録されてしまうと矛盾が生じたり，異なる意味関係を複数登録されることになる．そこで，作業者が用語間関係を指定する際，入力された 2 つの用語を文字列順でソートし，順位が低いものを用語 1，順位が高いものを用語 2 として一貫して登録する (図 3)．ソートの際に作業者が入力した順序と異なった場合は，用語間関係が NT や BT の場合には入れ替えて登録する．

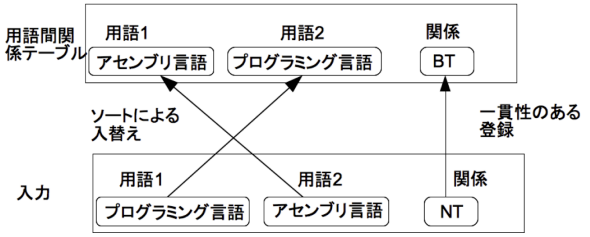

図 3: ソートを利用した用語間関係の一貫した登録

次に，用語間関係の検索を考えた場合，検索対象とする用語が用語 1 に登録されている場合と用語 2 に登録されている場合のそれぞれを検索して集約する．集約するには用語 1 で検索した場合と用語 2 で検索した場合では，用語間関係 NT と BT は入れ替えて関係性を統一する．表示も同様で，検索と指定に対して統一して表示を行う．例えば,「プログラミング言語」について関係ある用語を全て取り出す場合で説明する．用語間関係テーブル内には図 4 に示すように用語 1 と用語 2 にそれぞれ登録されているとする．検索アルゴリズムは用語 1 側の場合と用語 2 側の婆の場合のそれぞれを取り出し，関係を統一して表示を行う (図 5)．図 5 では検索用語を中心に上位語を左側に，それ以外を右側に配置し，視認性を高めるため関係を記号で記述している．さらに，用語関係の記号を検索用語からみてどいう関係になるかで統一して記述している．

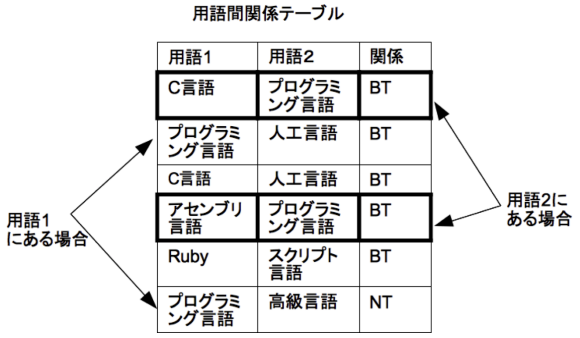

図 4: 用語間関係テーブル内の様子 (「プログラミング言語」について検索した結果)

5 まとめ

本稿では一貫性を考慮した用語登録，および用語間関係登録機能を持つ専門用語管理システムの実装について提案した．本提案システムは作業者が対象とする専門テキストデータをシステムに登録すると用語抽出器による候補の提示や，作業者が登録した用語に対して関係性を一貫して定義することが可能になる．本稿では特に，用語間関係を取り扱う際には上位，下位関係といった非対象の関係では常に関係を一貫して登録，検索する必要があることを明らかにした．また，これらをシステムに実装して動作が正しいことを確認した．

今後，用語抽出に関する形態素解析システムのチューニングや収集した用語に対する評価などの操作について検討していく予定である．

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た．

参考文献

[1] Béatrice Daille. Conceptual sctructuring through term variations. In *Proceedings of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 9–16, 2003.

[2] Jody Foo and Magnus Merkel. Cumputer aided term bank creation and standardization. In Marcel Thelen and Frieda Steurs, editor, *Terminology in Everyday Life*, pp. 163–179. John Benjamins, 2010.

図 5: 用語間関係の表示例 (「プログラミング言語」について検索した結果)

[3] Jose Manuel Urena Gomez-Moreno, Pamela Faber, and Miriam Buendia Castro. Frame blending in specialized language. *Teminology*, Vol. 19, No. 2, pp. 175–201, 2013.

[4] Koen Kerremans, Peter De Baer, and Rita Temmerman. Competency-based job descriptions and termontography. In Marcel Thelen and Frieda Steurs, editor, *Terminology in Everyday Life*, pp. 181–193. John Benjamins, 2010.

[5] Marie-ClaudeL ' Homme. A lexico-semantic approach to the structuring of terminology. In *Proceedings of the 3rd International Workshop on Computational Terminology*, pp. 7–14, 2004.

[6] Paul Sambre and Cornelia Wermuth. Instrumentality in cognitive concept modeling. In Marcel Thelen and Frieda Steurs, editor, *Terminology in Everyday Life*, pp. 233–254. John Benjamins, 2010.

[7] Mari Carmen Suárez-Figueroa, Guadalupe Aguado de Cea, and Asunción Gómez-Pérez. Lights and shadows in creating a glossary about onotology engineering. *Teminology*, Vol. 19, No. 2, pp. 202–236, 2013.

[8] Rita Temmerman and Sancho Geentijens. Ontologial support for multilingual domain-specific translation dictionaries. In Marcel Thelen and Frieda Steurs, editor, *Terminology in Everyday Life*, pp. 137–146. John Benjamins, 2010.

[9] 山本ゆうじ. シンプルな用語集形式 utx とその活用. 電子情報通信学会技術報告書「言語理解とコミュニケーション研究会」NCL2013, pp. 17–21, 2013.

[10] 豊嶋弘樹. 専門用語抽出のための形態素解析辞書管理システムの構築. 岡山大学工学部情報工学科特別研究報告書, 2013.

[11] 小山照夫, 竹内孔一, 濱田宏平. 用語管理システムの開発. 自然言語処理研究会, *2013-NL-212(2)*, pp. 1–4, 2013.