専門用語管理支援システムの研究
概要

小山 照夫
(国立情報学研究所)
KOYAMA Teruo
(National Institute of Informatics)

キーワード： 専門用語、用語管理、データベース
Keywords： terms, term management, database system

1 はじめに

様々な研究分野において、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として、用語管理には専門知識を持つ人間の多くの作業が必要とされることがある。多くの専門家を擁して本格的な用語管理を行うためには相当の資金を必要となり、このことが大きな問題となっている。

またもう一つの問題として、用語管理を支援するための枠組みや環境が整備されていないことも挙げられる。

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。

本研究では、第一に用語管理のための環境の問題に着目し、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を提供するシステムの要件を明かにするとともに、その実現を目指してシステムを構築していく中で、要件の妥当性を評価することを目的とする。特に用語管理支援の目的で、専門分野テキストからの用語候補抽出機能が、有効に利用できるかどうかについて検討を行う。

本研究では第二に、専門分野テキストの構造を統計的に調べることを通して、テキストからの用語抽出性能を向上させることも目的の一つとしている。

専門文書からの用語抽出は、文書から用語となりうる文字列を選び出す際の参考になり得ると期待できるが、この時、用語抽出性能はできるだけ高いことが望ましい。用語抽出の性能を向上させることができるならば、より有効性の高い支援手段となることが期待できる。

現在までに代表研究者等は日本語文書からの用語抽出アルゴリズムを開発してきているが [1],[2]、この研究ではさらに性能を高めた用語抽出アルゴリズムの実現を目指す。

以下ではまず、用語管理システムを構築する上で配慮すべき事項を検討し、管理システムの要件を明かにする。その後に、用語抽出性能を向上させる方法に関して検討を行う。

2 用語管理の背景

2.1 用語の分野性

用語データベースを検討するに当たって、研究分野と用語の関係をまず明かにする必要がある。

用語は、基本的には研究分野ごとに定義されると考えられている。ただしここで言われる分野がどのようなものであるかはそれほど簡単ではない。それは例えば自然科学全体というような膨大な領域を想定するものから、単一の疾患に関する医学的問題のみを扱うなどの、より狭い領域に特化したものまで、さまざまなカバー範囲と議論の対象を擁する領域を想定することが出来る。

領域の問題は、どのような立場からどのような問題に関わるかに依存しており、ユーザの視点に応じて領域の範囲は様々に異なると考えられる。

見方を変えるなら、領域の範囲はユーザごとに適合的に選択できる必要がある。このことをシステムの視

点から見るなら、独立した複数の領域について、それぞれの用語を並列して管理することが可能な枠組みが要請されていると言える。システムはユーザの目的に合わせて、適切な領域における用語集合を、複数並列して管理できなければならない。

2.2 協同作業としての用語管理

一つの分野が選択されたとして、そこで管理すべき用語は、通常は膨大な数にのぼる。多数の用語を管理するに当たって、一人でその全てを管理することには無理がある。

したがってそれぞれの分野の用語は、複数人で協力して管理できる事が必要となる。一般に多数の用語を複数人で管理することには個々の管理者の大きな労力を必要とすることになるが、システムはそれぞれの管理者の労力を最小とするものでなければならない。

このことを可能にする一つの方法は、ネットワークに接続された環境を、データ共有可能な形で整備することにより、協力関係にある複数の用語管理者がどこでからでも作業ができるシステムを準備することである。

3 システムの要件

用語管理の擁する以上の特質から、用語管理システム一般に対して要請されるいくつかの要件が導かれる。

3.1 分野別データベース環境の整備

構築するシステムはまず、複数の様々な分野について、分野ごとに独立に用語や用語間関係を管理できるものでなければならない、このためにはそれぞれの分野の用語を蓄積･検索･修正できる複数の独立したデータベースが定義できる形で、システムが構築されている必要がある。

3.2 Web 環境によるユーザインタフェース

広範なデータの管理には複数作業者の協力が不可欠であるから、一つのデータベースに対して、それぞれ複数の作業者が同時にアクセスして利用できることも重要である。

実際の利用者の作業を考えるなら、システムは利用者の側で特別な環境を整備する必要なく、例えばネットワークに接続されたパソコンを用いて、どこからでも利用可能なものであることが望まれる。

この目標を達成する代表的な方法は、システムを Web アプリケーションとして実装することである。今回開発するシステムでは、複数のデータベースを独立して管理できる基本的なデータベース管理の枠組みを構築するとともに、このデータベースにアクセスするためのユーザインタフェースを Web アプリケーションとして実装し、この UI を通してデータベースを操作する枠組みの構築が必要である。

3.3 ユーザごとのアクセス制限

複数データベースが管理されている状況であっても、個々のシステム利用者が操作の対象とするデータベースは、ほとんどの場合一つか、あるいはせいぜい少数のものであると考えられる。

したがって個々のユーザが、それぞれその対象とする適切なデータベースにアクセスでき、かつ一方では、権限のないデータベースにアクセスする事が許されない枠組みが必要である。

このことを可能にするためにはシステムにユーザアカウントを設け、それぞれのユーザがユーザごとの特定のアカウントを用いてシステムにアクセスできるようにするとともに、アカウントごとにアクセスできるデータベースが制限できなければならない。

3.4 用語収集資料と用語定義支援

用語集編纂に当たっての問題点の一つは、用語候補をどこからどの程度収集するかであり、この問題が本格的な用語集を編纂する上での大きな問題となっている。システムは用語収集を支援する機能を備えることが望ましい。

各研究分野の成果は、基本的には文書の形で報告されるため、用語候補はそれらの文書から収集されるべきである。一方で人手だけで多くの文書から適切に用語となりうる文字列を切り出すことは用意な作業ではない。

このためシステムは、まず用語収集の大本となる文書群を管理する必要がある。次に、管理された文書からどのような用語候補が考えられるかを提案する機能を備えることが望ましい。

以上の点から今回のシステムでは、ユーザがそれぞれの分野における研究抄録を登録し、全文検索可能な形でデータベースに蓄積して、用語候補の用例の検索を可能にすると同時に、蓄積された抄録から用語と考えられる文字列を抽出する機能を組み込み、抽出され

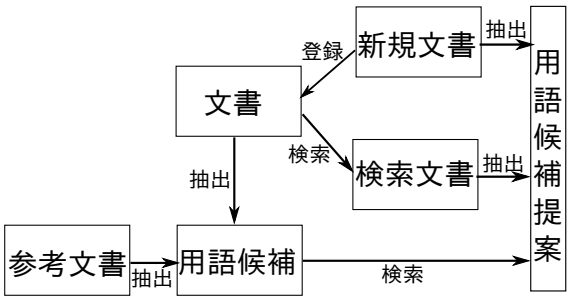

図 1: 用語候補提案の概要

た用語候補を表示することによって、用語登録を支援する機能を用意することとした。ここで用語候補の抽出に当たっては、研究代表者らが開発してきた、用語抽出アルゴリズムを使用する [1],[2]。

なお、これとは別個にあらかじめ参考文献を用意しておき、この文献から初期の用語候補を抽出する機能も用意している。これによって、とりあえず用語となる可能性のある候補を事前に一定程度用意する事が出来る。

用語候補提案に関する以上の概要をまとめると、図.1に示す形で用語候補の提案を行うことができる。

3.5 オフラインでのデータ操作

用語集編纂に当たっては、用語をオンラインでデータベースに登録する機能に加えて、オフラインで編集したデータをデータベースにアップロードする機能も必要である。

いくつかの分野では既に、部分的には用語の整理が行われていることもあるし、基本的な用語については相当程度分かっている部分もある。これらの既に整備されているデータについては、その全てをオンラインで入力することは効率的とは言えない。むしろ、オフラインでまとめられたデータを、入力を外部委託するなどして、機械可読なデータを作成することが効率的な場合も想定される。

システムがオフラインで整備されたデータのアップロードを行う機能を備えることにより、出発点となる基本的データを効率的にデータベースに蓄積することができる。

3.6 用語辞書データの参照

定義された用語データは、辞書の形で参照できる必要がある。また、このデータに基づいて各分野での用語辞書を作成する目的で、編集された辞書データをダウンロード出来る必要がある。

今回のシステムでは、現在の辞書情報の参照と、辞書データのダウンロードを行う機能を提供する必要がある。

3.7 用語抽出と形態素解析

用語候補抽出のためには形態素解析と形態素辞書を必要とする。これらのうちで形態素辞書は、本来は対象分野の特異性に合わせて調整すべきものであり、ユーザがその内容を変更できる事が望ましい。

今回のシステムでは、初期の形態素辞書として既存のものを使用するが、必要に応じて形態素辞書に形態素を追加／削除する編集機能を備えることにより、用語候補抽出に用いる辞書を変更する機能を用意している。

以上のシステム要件を考慮した上で、実際に Web ユーザインタフェースを通して用語および用語関連データを、協調的に編集する機能を持つプロトタイプシステムを開発した [3]。このシステムの詳細は項を改めて述べる。

4 システム開発におけるその他の試み

システム開発に関して、システム実装に加えて、システム評価や保守性の向上、性能の向上に関していくつかの検討を行った。

4.1 システムの評価

システム機能、特に用語候補提案機能に関して、用語関連の検討を行っている情報科学技術協会 (INFOSTA)SIG ターミノロジ部会 [4] において、用語管理の専門家にシステムの紹介を行い、意見を聴取した結果、システムの基本的機能および用語候補提案機能については高い評価が得られた。一方でユーザインタフェースに関してはまだ改善の余地があると指摘された。

4.2 システム保守性改善の試み

システムに関して、メンテナンス性向上を目的として、Web フレームワークの利用を検討するとともに、フレームワークの提供するデータベース一貫性管理の枠組みを利用するための、データ構造の変更を試みた。

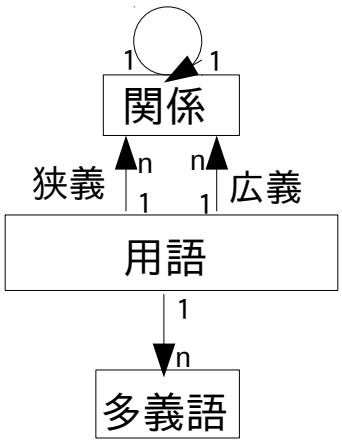

図 2: 用語関連レコード間の依存関係

一つの用語レコードは、可能性として広義および狭義という二つの用語間関係レコードを子レコードとして持つ場合が考えられる図.2。

このことは、一つの用語レコードに依存する 2 種類の用語間関係レコードが存在することを意味するが、このままの形では多くのフレームワークの提供するレコード間一貫性管理の枠組みを使用することは困難である。

また、もう一つの問題として、用語間関係で「広義」および「狭義」の関係は、常に逆関係を持つことになり、一方の定義に依存して他方も同時に変化する。この関係は本来冗長なものであることを意味しており、一貫性管理を行う上で問題となりうる関係である。

そこで、用語間関係レコードについては逆関係までまとめて一つのレコードとすることにより、冗長なレコードを排除すると同時に、各用語レコードに対してダミーのレコードを追加することにより、例えば Ruby on Rails などのフレームワーク内で一貫性管理の枠組みが適用可能となる形でデータベースの実装を変更し、動作確認を行った [5]。

4.3 その他

システム変更の試みとしては、この他にディスクアクセスを高速化する効果を評価する目的で、SSD の導入を試みた。

結果としてローカルな処理および形態素辞書コンパイルなどディスクアクセスの大きい処理については、ある程度の効果が見られたが、一般的なネットワーク越しの、入力や検索処理については、効果はそれほど顕著ではなかった。

また、今後のユーザ環境の変化を考慮して、タブレット端末によるシステム利用の検討も行った。データ参照にはある程度利用ができるが、データ操作には、キーボードがないこともあり、あまり適しているとは言えない。また、タブレット端末の特徴であるタッチスクリーンインタフェースを活用するためには、ユーザインタフェースの大幅な変更が必要となる。

これらの問題は今後システムを改良していく上での課題である。

5 用語抽出問題

研究代表者らはこれまでに日本語専門文書からの用語抽出に関して複合語に注目した用語抽出方式を提案し、実際の文書に適用してその有効性を確認してきた [1],[2]。

この手法を適用するにあたっては、形態素解析の精度が問題となる。

一般に利用可能な形で公開されている日本語形態素解析システムと形態素辞書は、一般的な日本人が日常的に触れる可能性の高い文書に対して最適化されている。一方で専門文書は、そこで利用される形態素集合にも、また構文一般にも、一般的な日本語文書とはやや異なる特徴を持っており、既存の形態素辞書や解析アルゴリズムでは解析精度に問題を生じる可能性がある。

この問題に対処する本来の方法は、分野ごとに語彙や構文を調査し、結果に基づいて解析アルゴリズムや形態素辞書を最適化することである。しかしながら、分野の特殊性に合わせた改変は、相当程度大規模なものとなることが予想され、大きな労力を必要とする可能性がある。

このことを考えるなら、このような基本的な改変を行うことは必ずしも実際的とは言いきれない。今回は限られた労力の中で、実際的に用語抽出性能を向上させるいくつかの方法を試みた。

5.1 代表的用語の形態素登録

今回はまず、分野によって頻出する形態素を調べることにより、分野に固有の形態素がある程度明らかになることを示すとともに、比較的頻度の高い用語を仮に形態素として登録することにより、用語抽出性能が一定程度向上することを明かにした。ただし、性能向上は限定的なものに留まっている [6]。

抽出性能が改善された要因を調べてみると、形態素として登録した用語の出現する周辺で形態素解析結果

が変化していることが分かる。この変化は、多くの場合、用語抽出性能を改善する方向で変化しているため、性能向上が見られたと考えることができる。この結果に基づき、より広範に形態素誤りを修正できる枠組を検討した。

5.2 形態素解析誤りの修正

用語の一部を形態素として登録することにより、形態素解析結果がどのように変化するかを精査した結果、一般的な日本語文書にはあまり使われないが、専門文書には相当程度の頻度で出現する、いくつかの形態素の直後で、系統的な解析誤りが生じていることが明かとなった。

具体的には、今回用いた形態素解析器 chasen[7] の形態素分類で「記号-アルファベット」と分類されたもの、および接尾辞「化」の直後で系統的な解析誤りが生じやすいことが判明した。

これらの系統的誤りのうち、ある程度頻度が高く、かつ、正しい結果が容易に推定できるものとして、「記号－アルファベット」の直後 13 種類、および接尾辞「化」の直後 9 種類の誤りについて、解析結果を事後的に書き換えて用語抽出を行った。結果として相当程度の抽出精度改善が達成できることが明らかとなった [8]。

5.3 外来語の取扱い

日本語専門文書、特に近年の理工学系統の文書では、多くの外来語が使用されている。用語抽出問題についても、外来語まで含めた用語抽出を行うことが望ましいが、これまでの方法では外来語については十分な配慮を行っていなかった。

外来語は、日本語とは異なる形態素から構成され、日本語とは異なる複合語構造を持つため、日本語の形態素辞書と複合語構造に基づく用語抽出はうまく働かない。

正式な形で外来語用語を扱うためには、外来語として出現する形態素の登録と、元言語の複合語構成規則を考慮した複合語判定機能が必要となるが、これを十分実用になるレベルで整備するには多大な時間と労力を必要とする。これ自体は重要な課題であるが、一方でとりあえず利用可能な、より簡便な方法があるなら、当面はそのような方法を利用することが考えられる。

日本語文書に外来語が出現する場合の特殊性として、基本的には議論の対象となるものごとを指示するために使われているのであり、文章の論述構造自体は日本語のままとなっていることが大部分である。このことは、日本語文書に外来語が出現する場合、その多くは議論の対象となる用語となっていると考えられることを意味している。

このため、実際には外来語の範囲を明かにし、その全体を用語と考えると同時に、それらと日本語形態素との複合型を考慮すればほとんどの場合に用語判定は可能である。外来語の範囲は、アルファベットまたはカタカナのみからなる文字列か、あるはそのような文字列同士が特定の区切り記号で仕切られた形であると考えて、およそ間違いはない。以上の考察の元に外来語と外来語の関連する用語を抽出するアルゴリズムを作成した。結果として、日本語用語の判定よりは適合度が落ちるが、実用上十分と考えられる外来語関連用語の抽出が可能となった [9]。

6 まとめ

本研究において研究代表者らは、用語管理システムの要件を整理し、要件を満足する形で、実際に操作可能なプロトタイプシステムを作成し、実際に動作することを確認した。

作成したシステムについて、識者の評価を求めた結果、今回のシステムの特徴である用語候補提案機能については、高い評価を得たが、ユーザインタフェースに関しては改善の余地があると指摘を受けた。

システム実装に関する今後の課題としては、Web システムフレームワークの導入による保守性の改善と、データ一貫性管理をより徹底したものとすること、システム事態のパフォーマンス向上、タブレット端末装置に適したユーザインタフェースの検討などが挙げられる。

本研究の目的の一つである、用語候補提案機能の中核となる用語抽出機能の改善に関しては、

- 用語のいくつかを形態素として登録する方法
- 形態素解析の系統的誤りを修正する方法
- 外来語の取扱い

に関して検討を行った結果、抽出性能を相当程度向上させることが可能となった。

形態素解析誤りや外来語の混在に基づく用語抽出性能の制約については、本来は形態素辞書の整備や解析アルゴリズムのパラメータ調整、外来語用語の取扱いアルゴリズムの追加などによって対応すべき問題では

あるが、実際問題として大きな労力を必要とすることから、時間的／コスト的制約のある中では本格的な対処は困難である。

今回試みた方法はいずれも簡便ではあるが、相当程度の効果が見られることから、有効な手法であると言えよう。ただし、本来の形での性能向上に関しても、今後はさらに検討を進める必要があると考えている。

参考文献

[1] 小山照夫他: “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”, 情報処理学会自然言語処理研究報告, 2006-NL-176, pp.55-60, 2006.

[2] 小山照夫他: “候補の接続関係を考慮した複合語用語抽出”, 情報処理学会自然言語処理研究報告, NL-193, 2009.

[3] 小山照夫他: “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.

[4] http://www.infosta.or.jp/

[5] 濱田宏平他: “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.

[6] 小山照夫他: “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.

[7] 小山照夫他: “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-220, 2015.

[8] http://chasen-legacy.sourceforge.jp

[9] 小山照夫他: “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.