Japio

Web と特許情報を事典的に活用するシステム

Wikipedia の分析による用語説明モデルの導入

東京工業大学大学院情報理工学研究科准教授
藤井　敦

PROFILE
1998 年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009 年より現職。2003 年 IPA から「天才プログラマー／スーパークリエータ」を受賞。自然言語処理、情報検索、音声言語処理の研究に従事。2009 年度特許版産業日本語委員会委員。
✉　TEL

1 はじめに

知的財産権の一つである特許権は、高度な発明の保護を目的としている。日本では年間約 40 万件の特許が出願され、多様な専門分野に関する発明が蓄積されている。特許に内在する人間の英知を体系化し、活用することができれば、今日の高度情報化社会において産業上の価値が高い。

特許には発明に関する新語や専門用語が多く含まれている。筆者は、World Wide Web と特許情報から種々の用語に関する説明情報を抽出し、さらに複数の説明情報を組織化することで、百科事典的なコンテンツを自動構築する研究を行ってきた [1]。また、構築したコンテンツに対して、見出し語、関連語、同義語、質問文、関連語マップといった多様な手段によって検索する機能を開発した。当該研究成果は、事典検索システム Cyclone として一般に公開している [2]。

科学技術や文化の急速な発展によって、様々な用語について調べる機会が公私を問わず増えている。そのため、World Wide Web 上の様々なツールを使うことが多い。代表的なツールには、Google や Yahoo! などの「検索エンジン」と Wikipedia などの人手で編集された「事典」がある。両者には、情報の量と質という点において、それぞれ長所と短所がある。

検索エンジンは、億単位のページ集合が検索の対象であり、提供される情報の量が多いという利点がある。しかし、検索される情報が体系化されておらず、必要のない情報も含まれるため、情報の質が低い。

事典は、説明を目的とした情報に限定され、項目等によって情報が統制されているため、情報の質が高いという利点がある。しかし、人手による編集に依存しているため、調べたい言葉が必ず登録されているとは限らない。また、説明の内容が著者の視点に偏るという問題もある。すなわち、情報の量において問題がある。

本研究の目的は、検索エンジンと事典の長所を統合して、有用性が高い調べ物のツールを実現することである。今回新たに、既存の事典である Wikipedia を分析することによって「用語説明が編集される仕組み」を解明し、用語説明に関するモデル（用語説明モデル）を構築した。さらに、そのモデルに基づいて検索エンジンの結果を組織化し、「事典的な検索」を実現した。

用語説明モデルの構築において、「動物名」や「病名」といった対象によって説明に必要な観点が異なる点に着目した。例えば、「動物名」は「生態」や「形態」、「病名」は「症状」や「治療」といった観点に基づいて説明される。そこで、Wikipedia から用語の種類に応じて異なる観点の構造を学習し、さらに観点ごとに固有の単語分布を学習する。その結果、例えば、動物の「ハクビシン」に関する Web 検索の結果に含まれる複数のページやスニペットから、「生態」、「形態」、「分布」などの観点に対応するテキストを抽出し、「ハクビシン」に関する事典的な情報を組織化することを可能とする。

ここで、「Wikipedia の存在が前提であれば、Wikipedia の記事だけを読めばよいのではないか？」という疑問が生じるかもしれない。この問いに対する答

えは「No」であり、本研究には２つの意義がある。まず、Wikipedia の未登録語について、Wikipedia と同じような観点の構造で説明を得ることができる。さらに、Wikipedia の登録語に対しても、別の用語で使われている観点を補ったり、一般の Web ページや特許情報から幅広く説明を収集することができる。

2 本研究の位置付け

筆者が開発した Cyclone は、Web や特許情報から説明情報を抽出し、事典的な調べ物を支援する検索エンジンである。Cyclone には要約機能が実装されている。具体的には、説明の「観点」に着目し、ある用語について説明している複数の段落から、観点ごとに代表文を抽出し、多観点の要約を生成する。

観点の設定と観点ごとに代表文を抽出する規則の作成は人手で行っている。現在は、情報処理用語を対象とし、岩波情報科学辞典などを参考にして、「定義」、「例示」、「目的」など 12 の観点を用いている。

図 1 に、用語「XML」について Cyclone で要約された説明情報を示す。図 1 に示された観点ごとに代表文を読むことで、必要な情報をなるべく落とさずに、複数の段落に含まれる冗長性を排除することができる。

しかし、上記の要約手法には２つの問題がある。1 つ目の問題は、観点の定義を人手で行う点にある。さらに、定義すべき観点は用語の種類によって異なる。例えば、病名について説明する場合は、「原因」や「症状」といった観点が必要であり、定義すべき観点の種類が情報処理用語と根本的に異なる。そこで、情報処理用語以外の用語に幅広く対応するためには、観点を定義する負荷が大きい。

２つ目の問題は、観点ごとに代表文を抽出する規則を人手で作成する点にある。抽出規則は、特定の単語や表現を手掛かりとする。例えば、用語の後ろに「とは」という表現があれば、「定義」に関する文として抽出する。しかし、用語の種類によって必要な観点が異なるため、新しい観点に対して常に規則を人手で作成しなければな

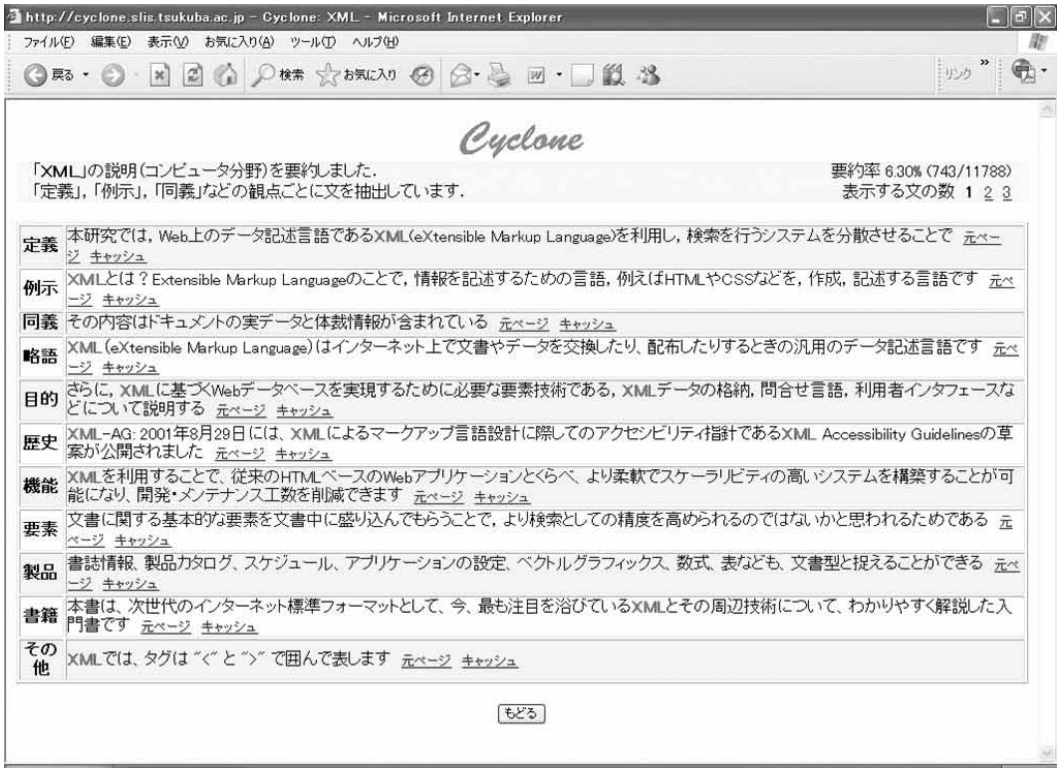

図 1　「XML」について Cyclone で要約された説明情報

らず、負荷が大きい。

以上の理由から、Cyclone の要約は情報処理用語に限定されている。本研究は、この問題を解決するために、Wikipedia から様々な用語の種類について観点の構造を抽出し、「用語ごとに観点を定義する負荷」と「観点ごとに抽出規則を作成する負荷」を削減する。

3 事典的検索の手法

3.1 概要

本研究で提案する事典的検索の概要を図 2 に示す。図 2 は、事前に行う「用語説明モデルの構築（上）」と検索結果のテキスト集合が与えられた段階で実行する「検索結果の組織化（下）」に大別される。用語説明モデルは Wikipedia の記事集合を用いて構築する。検索結果の組織化では、「りんご病」のような用語を検索質問として Web を検索した結果から、複数のテキスト（ページまたはスニペット）を収集し、用語説明モデルに基づいて観点ごとに個々のテキストを分類する。さらに、観点ごとに代表的なテキストをユーザに提示する。

3.2 用語説明モデルの構築

本研究で構築する用語説明モデルとは、「病名」や「動物名」といった用語の種類に応じて、説明の観点を列挙したプロトタイプである。さらに、ある用語について検索されたテキストが与えられると、用語の種類を特定し、その見出し語に対応する観点の候補から適切な観点を特定するための分類器である。

まず、「人名」、「動物名」、「病名」といった用語の種類ごとに、Wikipedia の記事から観点の構造を抽出する。ここでは、Wikipedia の記事にある「目次」に着目する。図 3 は「破傷風」に関する Wikipedia の記事である［3］。図 3 の「目次」には、「1. 原因」、「2. 症状」、「3. 治療」などの項目（セクション）が並んでいる。本研究では、1 つのセクションを 1 つの

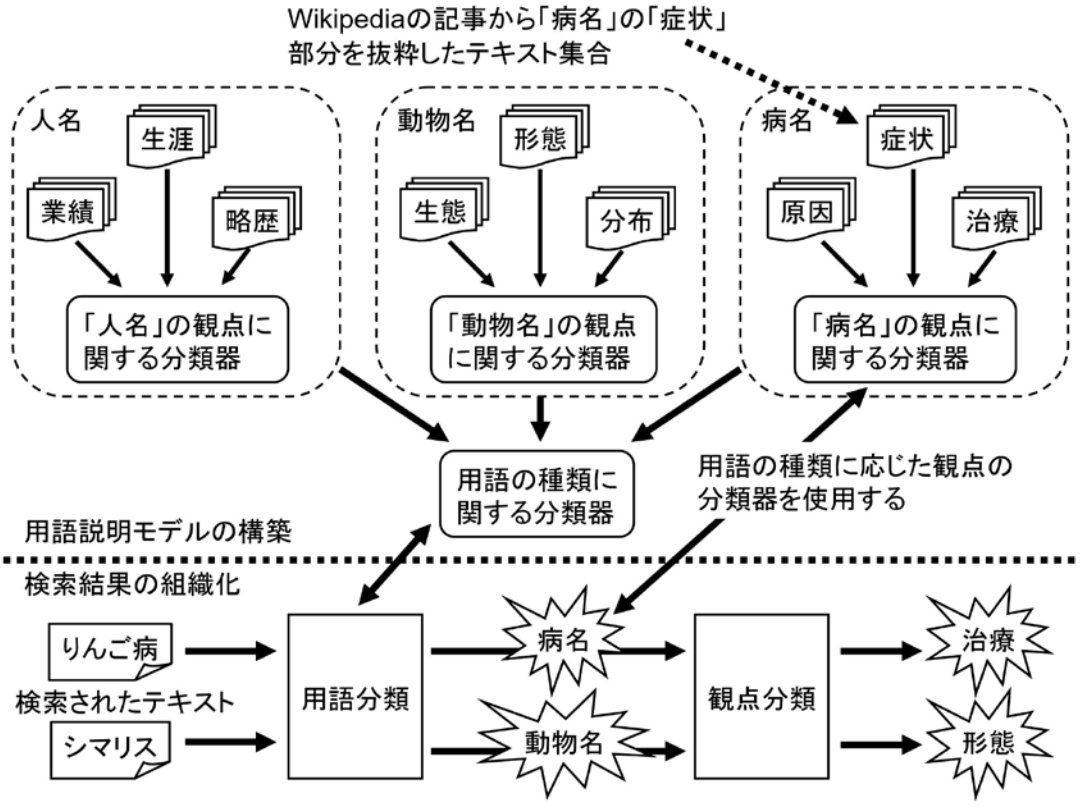

図 2　用語説明のモデル化による事典的検索の概要

図 3　「破傷風」に関する Wikipedia の記事

観点として使用する。

しかし、あらゆる病名の記事でセクションが完全に一致するわけではない。そこで、ある用語の種類（例えば「病名」）ごとに複数の記事を収集し、使用頻度が高いセクションを観点として選択する。観点の選択基準として、使用頻度に関する閾値を設定する方法や、使用する観点数の上限を決めておく方法があり、目的等に応じて適宜使い分ける必要がある。

さらに、該当する Wikipedia の記事集合から複数の分類器を学習する。ただし、分類の目的によって使用する学習データが異なる。「用語の種類」を分類するためには、用語の種類ごとに記事集合をまとめて 1 つのカテゴリに対応する学習データを作る。他方で、「観点」を分類するためには、観点ごと記事集合をまとめて 1 つのカテゴリに対応する学習データを作る。

分類器の学習にはサポートベクターマシン（SVM）を使用し、さらに多値分類に拡張した。そこで、用語の種類や観点は任意の数でよい。多値分類への拡張方法、素性、素性の値にはそれぞれ選択肢がある。予備実験の結果、One-Vs-Rest 法を使って多値分類に拡張し、単語ユニグラムを素性として、素性の値は全て定数とした場合に分類精度が最も良かったため、これらを採用している。

3.3　検索結果の組織化

3.3.1　概要

検索結果の組織化では、「りんご病」などの用語について検索されたテキストを入力し、「用語分類」と「観点分類」を順番に実行する。用語分類では、用語に関する分類器を用いて、「りんご病」の説明テキストを「人名」、「動物名」、「病名」のいずれかに分類する。観点分類では、分類された用語の種類に応じて観点を分類する。図 1 の例では、「病名」に分類されたため、病名に対応する「原因」、「症状」、「治療」のいずれか

に説明テキストを分類する。SVM に基づく One-Vs-Rest 法では分類結果ごとにスコアが計算されるため、観点ごとにスコアが高いテキストを抽出し、ユーザに提示する。

本研究における「検索結果の組織化」は、与えられた文字列を何らかの観点に分類するという抽象度の高い処理である。そこで、事典的検索を達成するための応用方法は 1 通りではなく、運用状況等に応じて使い分ける必要がある。3.3.2 ～ 3.3.4 では、3 通りの応用方法を提案する。後になるほど Cyclone での運用方法から逸脱する度合いが大きくなる。

3.3.2　Cyclone の要約手法を自動化する

Cyclone の要約手法において、人手に依存している部分を自動手法に置き換える方法がある。具体的には、ある用語に関する説明の段落を文に分割し、文単位で観点に分類して、最後に観点ごとに代表文を選択する。Cyclone の枠組みでそのまま利用できるという利点があるものの、文単位では素性が少ないため分類精度が高くない可能性がある。また、1 つの観点が複数の文にわたって説明されている場合には、文単位での分類は本質的に困難である。

3.3.3　Cyclone の段落を観点に分類する

Cyclone で検索される段落の単位で観点に分類する方法がある。Cyclone では、段落を「機械」や「化学」といった技術分野や「スポーツ」や「芸能」といったジャンルに分類し、ユーザは検索された情報を絞り込むことができる。これと同じ発想で、段落を観点に分類することで、ユーザは自分が知りたい観点に基づく説明だけを選択的に取得することが可能になる。ただし、1 つの段落に複数の観点が混在する場合があるので、観点に順位を付けて、上位から一定数の観点を段落に付与する必要性が生じる可能性がある。

3.3.4　一般の検索エンジンと組み合わせる

一般的な Web 検索エンジンや特許検索システムで検索された文書集合を観点に基づいて分類する方法がある。Cyclone とは独立しているものの、検索された情報を説明の観点に基づいて体系化するという目的は達成される。

具体例を用いて説明する。Yahoo! で「ハクビシン」を入力して検索した上位 100 件のスニペットを分類した。本研究で作成した用語説明モデルには、「動物名」に対して、「生態」や「分布」など合計 7 種類の観点が対応していた。本稿執筆当時、Wikipedia における「ハクビシン」の記事には、「形態」、「歴史」、「分布」、「分類」に関するセクションはない。しかし、本手法は動物名に関する記事の集合から代表的なセクションを観点として抽出するため、Wikipedia において個別の記事では欠落している観点を補うことができた。

分類したスニペットには、上記 7 種類以外の観点として、「ハクビシン」をテーマにした「著作」や「名称の由来」があった。これらの用語説明モデルに含まれない観点は、検索されたスニペットの内容を分析して抽出するしかない。今後は、カテゴリ分類とクラスタリングの併用について検討する必要がある。

多義語の例として、「キーウィ」で検索されたスニペットを分類した。用語分類において、鳥のキーウィについて記述されたスニペットは「動物名」に分類され、果物のキーウィについて記述されたスニペットは「植物名」に分類された。さらに、これらのスニペットは観点分類によって動物名や植物名に固有の観点に分類された。本手法によって多義や多観点を考慮した検索を実現できることが分かった。

4　おわりに

本研究は、Wikipedia という既存の事典を用いて、用語説明が編集される仕組みをモデル化した。さらに、構築した用語説明モデルを用いて、検索エンジンで得られたテキスト集合を説明の観点に基づいて分類し、事典的な検索を可能にした。本手法の特長は、与えられたテ

キストがどのような用語について書かれているかを特定し、その結果に基づいて分類すべき観点の候補を変更する点にある。Wikipedia のように協調的な人手編集による事典は今後も発展するだろう。しかし、爆発的に増える情報を自動的に統制する技術も必要である。情報の統制において自動化が困難な事象を特定し、人手編集との棲み分けや共存について検討する必要がある。

謝辞

本研究の一部は、文部科学省科研費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（課題番号：21013003）、NEDO「産業技術研究助成事業」によって実施された。

参考文献

[1] 藤井敦．特許情報を専門用語辞典として活用するシステム，Japio 2008 YEAR BOOK, pp.196-199, 2008.
[2] http://cyclone.cl.cs.titech.ac.jp/
[3] http://ja.wikipedia.org/wiki/ メインページ