E-046

用語の説明を収集抽出・分類するシステム

The system that collects and classifies explanation of term

矢島健司†，佐川 雄二，田中 敏光
Kenji Yajima, Yuji Sagawa, Toshimitsu Tanaka

1. はじめに

従来，わからない用語について知りたいとき，図書館などでそれについての本を探したり，国語辞典や百科事典などの紙を媒体とした辞書を使用して調べてきた．しかし，新しい用語，特に専門用語などは収録されていないことが多い．紙媒体の辞書では用語の説明は正確であるが，収録・編集・出版に時間がかかるためである．

近年では，インターネットの普及によって Web 検索を利用し，用語を調べるという方法も可能になった．Web 上に掲げられる情報の更新速度は速く，また様々な分野にわたる情報が存在している．したがって，求める情報がそこに存在している可能性は高い．

しかし，Web 検索で得られる情報は膨大なもので，必ずしも求める検索結果に関係のある情報を含むものであるとは限らず，参考にならない情報も多い．検索で得られた多くの情報源から必要な情報を見つけ出すのは，ユーザにとってとても困難で手間のかかる作業である．

本研究では，Web 検索で得られた検索結果から用語の説明文を抽出し，それらを自動的に分類するシステムの構築を目指している．これまでは，一つの用語が分野によって複数の意味を持つ場合，それらを分類するシステムを作成した[1]．本報告では，説明文の抽出能力向上のための改善法とその結果，説明文の分類法の改善案について述べる．

2. 説明文抽出

2.1 説明を含むページの収集

本研究では，Web 検索エンジン Google を用いて用語検索を行い，情報を収集する．このとき，用語に対する情報をより得やすくするため，検索は要求された用語に対し，「は」「とは」や「について」などの付属語を付加したクエリを作成し，拡張検索を行う．各検索結果の上位 50 件を用語の説明を含んだ Web ページであるとし収集する．

2.2 説明文抽出

収集した Web ページが必要とする用語の説明以外を情報も含む場合，閲覧に手間がかかると考えられる．2.1 で収集された情報には，そうした可能性があるため，用語の説明であると考えられる部分のみ抽出を行う．従来[1]は，クエリがヒットした部分を含む一段落の文章を説明文であるとして抽出してきた．

† 名城大学 大学院理工学研究科

2.3 説明文抽出の改善手法

説明文が複数段落にまたがる場合や，Web ページのレイアウトのため改行が必ずしも文の切れ目で行なわれていない場合がある．これらの場合，2.2 の説明文抽出手法では，説明文の取りこぼしが発生する．そこで取りこぼしを減らすために以下のように手法を改善する．

2.3.1 手法 1 文脈の切れ目に注目する手法

段落は文章のまとまりの一つではあるが，強いまとまりではない．そこで，より強い文脈の切れ目に注目することで，文章が複数行にまたがっている場合でも，どこまでが説明部分であるかが判定できると考えられる．文脈の切れ目を特定する判断材料としては“また”や“ところで”といった接続表現を使用する．これらを用いて，Web ページのクエリを含む文章から，話題が変わったと判断出来るところまでを説明部分であるとして抽出する．“また”などの並列を示す表現であれば，その後に続く文も説明文であるとして収集する．“ところで”などの転換を示す表現であれば，そこから話題が変わるので，その直前までを説明文として抽出する．

2.3.2 手法 2 既存の辞書を手本とした手法

用語の説明には，ある程度典型的な記述の仕方がある．そこで，既存の辞書の説明文を手本として Web ページから手本と似た文章部分を判定することで，説明部分を特定し抽出することができる．手本とする辞書として Wikipedia[4] による用語の説明を用いることとする．類似度判定のため，Web ページを文単位に分割し，文ごとにベクトル化を行なう．このベクトルと指標となるベクトルとの類似度をベクトル空間法を用いて求め，文間の距離が近いものを説明文であるして収集する．ベクトルの要素には，名詞・動詞を用い，指標には手本の説明文をベクトル化したものを利用する．また，言い換え表現の辞書を作成することで手本と多少違った言い回しの文章も収集できるようにする．

2.4 改善手法の評価

改善手法により，システムの抽出した説明文が元のページ全体のどの程度の割合かを調査した結果を表 1 に示す．

元のページの文章量によらず，ほぼ同程度の分量の文章を抽出しており，内容を主観的に評価した結果からも，説明文として妥当なものであった．この結果から改善手法は，説明以外の情報も含むページから，説明文だけを正しく抽出できていることがわかる． 3.2 の手法では，用語を含む 1 文 2 文は抽出することができるが，例などの文章は例が手本と似ていないと抽出できない．現在，名詞・動詞の重み付けを行わず，１文単位で抽出を繰り返しているが，重み付けや文章単位で見る方法を試してみる必要がある．

表 1　抽出される説明文

ページ全体	抽出された説明文	説明文の割合
10,000 字程度	200～600 字	2～6%
4,000 字程度	300～500 字	7～12%
300 字程度	200～300 字	66～100%

3.　説明文の分類

同じ用語でも分野によって異なる意味を持つ場合がある．また，同じ分野の同じ用語の説明でも，想定して要る読み手などにより説明のしかたが異なる場合がある．これらを分類するために関連用語を使用する．説明文の内容が異なれば，説明文内で使用される関連用語が異なるからである．

3.1 関連用語抽出

2.2 で抽出した説明文それぞれに対して形態素解析を行い，与えられた用語の関連用語を抽出する．形態素解析には，MeCab[2]を用いる．解析結果から，名詞を関連用語として抽出する．また，分野に固有の専門用語の多くは，既存の語構成要素に基づいた複合語として定義される[3]．本システムは，名詞が連続して出現した場合，それらを複合語としてひとまとめにして扱い，関連用語として抽出する．

3.2 マッチングによる分離

3.1 で作成した関連用語集同士の関連用語の完全マッチングを行う．関連用語の一致する割合がカント解析により得た閾値より高ければ用語集同士の関係が強いとし，グループの作成を行う．図 1 は関連用語集 a から b，c，d，･･･へのマッチング一致の割合であり，作成されるものはグループ A＝{a，b，d，･･･}のようになる．

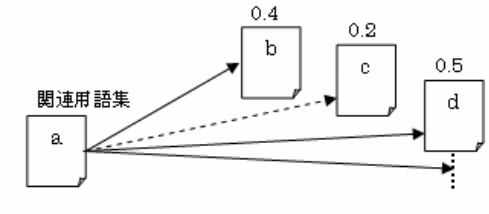

図 1．マッチングの例

3.3 包含関係による絞込み

3.2 の方法のみによる分類では説明文の総数を n とすると n 個の分類になってしまうため，現実的ではない．そこで 3.2 で作成したグループの包含関係を考え，あるグループに完全に属するグループは消去する方法をとる．消去の対象がなくなるまで包含関係を調べ，残ったグループを分類されたグループとして抽出する．

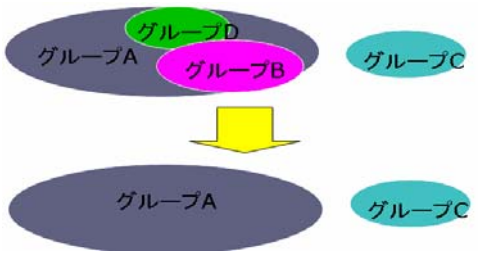

図 2．グループ化の例

3.4 分類実験と改善案

以上の手法により説明文分類を行なった結果，分野による分類は，分野を混同することなく分類することが出来，言い回しによる分類でもおおむね良好な結果が得られた．しかし，複数の分野にわたり説明が存在していた場合，自分の知りたい分野の用語の説明にたどり着くまでに時間がかかることが予想されるので，分類を階層化することを検討中である．

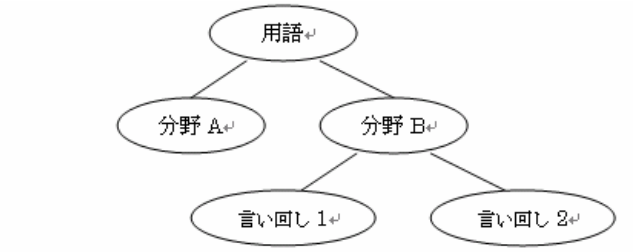

図 3．分類の階層化

4.　今後の課題

- 分類法の考察
- 言い換え表現の辞書の作成

参考文献

[1]　矢島健司，“用語の説明を収集・分類するシステム”，電子情報通信学会，2007 年総合大会 D-4-1，2007

[2]　MeCab，京都大学情報学研究科，http://mecab.sourceforge.net/

[3]　小山照夫，“動詞の挿入による日本語複合語の構造解析”，NII Journal, No.2, p.39-44, 2001

[4] Wikipedia，http://ja.wikipedia.org/wiki/