

英和辞典からの知識抽出

下司, 義寛
九州大学理学部

和多, 太樹
九州大学大学院システム情報科学府

廣川, 佐千男
九州大学情報基盤研究開発センター

<https://hdl.handle.net/2324/1526156>

出版情報：情報処理学会第68回全国大会．第68回(平成18年) (3), pp.19-20, 2006-03. 情報処理学会
バージョン：
権利関係：

下司 義寛[†] 和多 太樹^{††} 廣川 佐千男[‡][†]九州大学理学部 ^{††}九州大学大学院システム情報科学府 [‡]九州大学情報基盤センター

1 はじめに

分析対象のテキストに応じてカテゴリ辞書やシソーラスを作成しておけば、分析結果を分かりやすく提示できたり、検索のヒントを与えることができる。従来、このような辞書やシソーラスは多大な人手と費用をかけて作成されている。本発表では、単語の出現頻度を用いて分析対象のテキストに現れる単語について、単語間の概念的な上位/下位の関係を自動的に抽出する方法を提案する。英和辞典の文書を具体的な対象として様々な事柄について知識の抽出を試みる。

2 データとしての英和辞典

筆者らは文書群に現れる単語について、それらの文書頻度を用いて単語の上位下位関係を抽出し、グラフ表示するシステムを開発している。本発表では、英和辞典「英辞郎」に載っている英単語についての説明の文章に現れる単語（日本語、英語の両方）について関連を分析した。英辞郎には1,648,628語の英単語が掲載されている。市販の検索システムだと、「wine」を検索すると図2のような文章が表示される。各単語の説明文章のサイズは平均68.3バイト(34文字)で、そのサイズの分布は図2のようになっている。つまり、非常に短い文章で各単語が説明されている。本発表では、これらの短い文章群に現れる単語の関連をもとめることにより、知識抽出を試みる。長い文章からその要約を求めたり、論理的構造を求める研究はあるが、共通の単語を含むという関連しかない短い文章群から、意味のある事柄を発見する研究は、筆者らの知る範囲では他にない。

3 文書頻度を用いた関連語抽出

本発表で使ったシステムは英和辞典をデータ文書群として[2]のアルゴリズムを実装したものである。利用者がクエリーを与えると、まずシステムは通常の実験を行ない、そのクエリーに関連のある文書のリストを求める。この部分は国立情報学研究所で開発されたGETAを利用した。次のステップでは、クエリーに関



図1: 市販検索システムでの「wine」検索結果

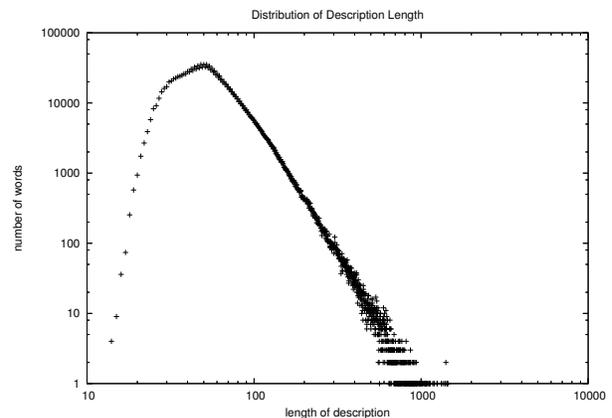


図2: 各単語の説明文章サイズ分布

連の強い単語として、検索結果の文書群に現れていて他の文書群に現れない単語を文書頻度を用いて抽出する。次に、得られた関連語の上位下位関連を、検索結果の文書群における文書頻度を使って求める。例えば「wine OR ワイン」ですると、1496個の文章が求まる。関連単語の個数を出現頻度順に並べると表1のようになる。出現頻度で上位100個まで選ぶと、出現回数数が7回以上の104個の単語が得られる(図3)。

4 文書頻度を用いた上位下位関係抽出

得られた関連語について、「多数決原理」に基づき上位下位関係を求める[2]。すなわち、求まった文書群において、単語Aの方が単語Bより出現数(文書

Limit of Boolean Query for IR

[†] Yoshihiro SHIMOJI(y-shimo@i.kyushu-u.ac.jp)

^{††} Taiki WADA(t-wada@i.kyushu-u.ac.jp)

[‡] Sachio HIROKAWA(hirokawa@cc.kyushu-u.ac.jp)

Department of Physics, Kyushu University ([†])

Graduate School of Information Science and Electrical Engineering, Kyushu University (^{††})

Computing and Communications Center, Kyushu University

([‡])

