7C-4

英和辞典からの知識抽出

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検索を行ない、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

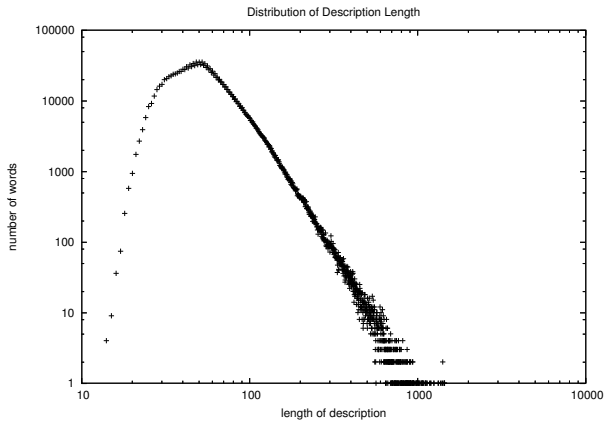

図 2: 各単語の説明文章サイズ分布

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 (‡)

ワイン (1304),wine(583), 白ワイン (233), 醸造 (150), シャトー (135),chateau(128), 名産地 (68), 名産 (68), 赤ワイン (64), 原産地 (60), 品質管理法 (57), ブドウ酒 (54), ワイン醸造元 (54),origen(54), 原産地呼称 (53),dominacion(52),dominacion de origen(52), chardonnay(48), ボルドー (48),blanc(47), シャルドネ (47), ブラン (46), ブルゴーニュ(40), （中略） sancerre(7),spatlese(7), サンセール (7), ベーレンアウスレーゼ (7), ポイヤック (7), ワイン醸造品質管理協定 (7), エブロ河流域 (7),classico(7), ソアーヴェ(7), ボルドーワイン (7)

図 3: 文書頻度上位 100 までの関連語

頻度	～100	～90	～80	～70	～60	～50	～40
個数	6	6	6	6	10	17	23
頻度	～30	～20	～10	9	8	7	6
個数	25	30	71	79	89	104	129
頻度	5	4	3	2	1		
個数	175	217	317	753	2525		

表 1: 頻度ごとの関連語数

頻度）が大きくて、さらに、単語 B が出現する文書の過半数に単語 A も出現しているとき、A の方が上位とする。グラフとして表示するときには、単語 B の上位にある単語の中で、最小なものだけに線を引く。図 4 は図 3 の関連語で文書頻度が上位の 20 個について上位下位の関連の表示したものである。左側の単語の方が文書頻度が高く上位になっている。文書頻度を下げるとより多くの単語についても関連が得られる。図 5 はこうして得られた大きなグラフにおいて「白ワイン」の下位の一部を表示したものである。白ワインの種類やそれぞれの産地名などが現れていることが分かる。

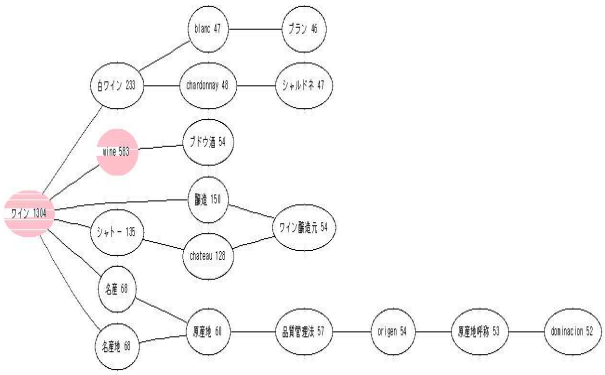

図 4: ワイン関連語概念図

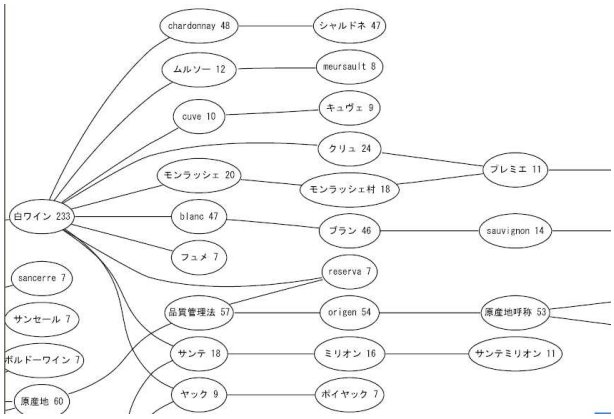

図 5: 白ワインより下位部分 (詳細)

5 まとめと今後の課題

160 万個以上の短い文章群からなる英和辞典を使って、知識を構築するシステムを提案した。Web 文書や専門的な書籍の索引情報からキーワードや知識を抽出する研究 [1, 5] はあるが、本発表のような非常に短い文章群から単語の関連抽出を試みるものはない。文書頻度を利用して単語間の階層構造を求める [4, 6] との比較は今後の課題である。

参考文献

[1] 藤井, 石川, World Wide Web を用いた事典知識情報の抽出と組織化, 電子情報通信学会論文誌, Vol.J85-D-II, No.2, pp.300-307, 2002

[2] 廣川, 下司, 和多, 文書群からの概念グラフの構成, 情報処理学会第 169 回自然言語処理研究会,pp.79–84, 2005

[3] 中西 他, 特定分野を対象とした連想検索のための書籍の索引部を用いたメタデータ空間生成方式, 電子情報通信学会論文誌, Vol.J88-D1, No.4, pp.840–851, 2005

[4] Y. Niwa et al., Topic Graph Generation for Query Navigation, NLPRS'97, pp.95–100, 1997

[5] 桜井, 佐藤, ワールドワイドウェブを利用した用語説明の自動生成, 情報処理学会論文誌, Vol.43, No.5, pp1470–1480, 2002

[6] P. Srinivasan, Thesaurus Construction, in W.B. Frakes and R. Baeza-Yates eds, Information Retrieval: Data Structures and Algorithms, Prentice-Hall, 1992.