2-4　コーパスからの単語間の意味関係の獲得とその応用

2-4 Acquisition of Taxonomic Relations Among Words from Huge Corpora and its Application

神崎享子　山本英子　井佐原 均
KANZAKI Kyoko, YAMAMOTO Eiko, and ISAHARA Hitoshi

要旨

人手で作成されたシソーラス（語彙の概念体系）の不備や不統一を解消するため、神経回路網モデルを用いた自己組織化マップと、二語間の上位下位関係を求める類似尺度を組み合わせて、大規模なテキストから類義関係と階層関係を自動獲得した。心理実験によって既存のシソーラスと比較し、自動構築した階層構造の妥当性を評価した。この手法を応用して、主題的関係を持つ関連語集合をテキスト集合から抽出し、その関連語集合の検索キーワード群としての有効性を検証した。

Thesaurus is very important lexical knowledge for our inference activity. However, we have only thesaurus compiled by human because we didn't have huge corpora and the algorism to organize concepts using such corpora.

For sake of a verification of an existing thesaurus made by human, we automatically extract lexical knowledge from huge corpora. In our method, we extracted attribute concepts whose instances are adjectives from corpora and calculated similarity relations by Self-Organizing Map and hypernym-hyponym relations by Complimentary Similarity Measures. As a result, we constructed the taxonomic relations of attribute concepts of adjectives on a map. Also we applied our methods to extract related word sets which can be useful for retrieval support. Concretely, in order to extract word sets with thematic relation, we extract related word sets with non-taxonomical relation. Then, we verified the effectiveness of such word sets as key words for information retrieval.

[キーワード]
概念体系，類義・階層関係，自己組織化マップ，主題的関係，検索支援
Thesaurus, Taxonomic relation, Self-organizing map, Thematic relation, Retrieval support

1　まえがき

語彙の体系化の重要性

本研究では、日本語語彙の概念体系をコーパスから自動獲得することを目的とする。また、得られた語彙の関係が情報検索等の応用システムにおいて有効であることを示す。

語彙の意味関係を体系化した辞書は、計算機に推論を効率的に行わせて、必要な情報を取り出すための、非常に重要な基盤の一つである。語彙の意味関係を構造化した辞書というのは、具体的には、単語どうしの類義関係や階層関係、部分全体関係、所有関係などの様々な情報を構造化した辞書ということである。このような情報があることで、一つの単語を手がかりに、関係する情報を見つけ出すことができる。例えば、我々は「自動車」という一つの単語に対して

上位関係：乗り物、…
下位関係：軽自動車、大型自動車、普通自動車
　更に下位関係：トヨタの＊＊＊、日産の＊＊＊、
　　BMW の＊＊＊、など
類義関係：電車、自転車、飛行機、船、…
部分関係：タイヤ、ハンドル、エンジン、ドア、…

などの情報を知識としてもっている。このような

情報があるからこそ、例えば、「車を購入したい」と言えば、「車」の下位関係から何を購入したいのか候補が出てくるし、「車で行くのはどう？」といえば、「他の手段の方がよくないか？」と、類義関係から考えをめぐらすことができ、「車が壊れた」といえば、どこの部分が故障したのか、可能性を自然に考えることができる。

このように、単語の意味関係の構造化は、人間が推論を効率的に行う際に重要であると同様に、計算機にとっても、基盤となる非常に重要な知識なのである。

シソーラスとは何か

単語の意味あるいは概念の関係を体系化した辞書を、「シソーラス」と呼ぶ。

単語の意味関係があるのは名詞だけではない。動詞や形容詞でも同様に意味関係がある。例えば、「赤い」や「白い」は類義関係で、これらは「色」という共通点でまとめられる、「大きい」や「小さい」は類義関係で、これらは「サイズ」という共通点でまとめられる。

一般に内包（所属事例に共通する性質）によって定義する場合に「概念」とよび、外延（所属事例そのものの集合）によって定義する場合に「カテゴリ」と使い分けることが多いとされている[1]。「赤い」や「白い」については、その共通する性質によって定義すると「色」が概念と考えられ、「大きい」や「小さい」については「サイズ」が概念と考えられる。別の見方をすれば、「赤い」や「白い」は「色」というカテゴリの事例であり、「大きい」や「小さい」は「サイズ」というカテゴリの事例である。

「色」や「サイズ」などの概念間の関係を類義関係と階層関係（taxonomy）で体系化したものをシソーラスと呼び、言葉を表層的な文字列ではなく意味を利用して計算機処理する場合に利用される。

これまでのシソーラス

これまで、自然言語処理の分野でも「シソーラス」は大規模に作られてきた。例えば、NICT が配布している『EDR 電子化辞書』や国立国語研究所の『分類語彙表』、NTT の『日本語語彙体系』などの辞書が構築されている。

これらのシソーラスは人数と年数をかけて、日本語語彙を構造化したシソーラスだが、大規模になればなるほど、疑問のある部分もあり、検証し修正する必要があるが、自動的に修正可能な比較的単純な誤り以外に、その体系自体に踏み込んで、変更・修正などができないのが現状である。

我々の研究

近年、利用できる膨大なテキストデータが手に入るようになり、言語処理技術も発展してきた。そこで、我々は現実のテキストから概念体系を自動的にとらえることを試みる。現実の膨大なテキストから概念体系を自動獲得できれば、従来、大規模に人手で作成されてきたシソーラスを検証することができ、修正すべき箇所などを検討することができる。さらに、コーパスからの自動獲得手法がより洗練されれば、未知の大規模データに対しても言語知識を抽出することが可能になる。例えば、年々ニーズが高まっている、医学、生物学、法律、特許などの分野での専門用語の構造化、いわば、専門用語のシソーラスを自動構築できることにもなる。専門用語のシソーラスを構築することで、それぞれの電子化された専門分野の膨大な文書から必要な情報を自動抽出する一助になると考える。我々は医学用語への応用を試みている。

2 方向性ある類似尺度を導入した神経回路網モデルによる自己組織化マップ

我々は、自己組織化マップ（Self-Organized Map；SOM）[2] を用いて、大規模コーパスからシソーラスを自動獲得し、それに基づいて、既存の言語資源である人手構築のシソーラスを検証することを目的に、形容詞の概念をテキストから抽出し概念全体を構造化することを考える。つまり、先に述べた、「赤い・白い」の形容詞の上位概念「色」や、「大きい・小さい」の上位概念「サイズ」をコーパスから抽出し、形容詞の概念を表す「色」や「サイズ」などを自動的に構造化しようと考える。この手法の計算式の説明においては、データから抽出した形容詞の概念「色」や「サイズ」などを、説明の便宜上「語」と呼ぶ。

我々の手法では、自己組織化マップへの入力データを符号化する際に、あらかじめ二語間の意味距離を、上位下位関係のような方向性を求める類似尺度で計算する。これにより、マップ上に、自己組織化による概念の類義関係だけではなく上

位下位関係の分布も得られる。

2.1 入力データ

コーパスから形容詞を範ちゅう化するような抽象的な名詞を取り出すために、形容詞を範ちゅう化する名詞の意味関係をコーパスから探し、データ収集を行った[3][4]。方法は、XがYを範ちゅう化するパターン[5]である「X トイウ Y」という文型を手がかりにXが形容詞、Yが抽象名詞というパターンをコーパスから取り出した。このデータから、該当する形容詞と名詞を、ある程度、人手で取捨選択した。

「形容詞の概念名」として用いる抽象名詞Yは、94、95年の毎日新聞2年分から取り出した。抽象名詞と共起する形容詞、形容動詞は、毎日新聞11年分、日本経済新聞10年分、産業金融流通新聞7年分、読売新聞14年分、新潮文庫100選、新書版100冊の中から用例を調べた。抽出された抽象名詞は365語、形容詞の異なり語が10,525語、延べ語数は35,173語であった。最大共起語数は、「こと」に対する1,594語である。データは、以下のようになる。

［例］

思　い：うれしい　楽しい　悲しい……
気持ち：楽しい　嬉しい　幸せな……
観　点：医学的な　歴史的な　学術的な……

2.2 入力データの符号化

自己組織化マップへの入力データの符号化[6]については以下のようになる。

ここで、一般にω種類の名詞w_i（i=1, …, ω）が存在し、それらのマップを構築すると仮定する。具体的には例えば、思い＝{幸せな、誇らしい、悲しい…}のようなデータを入力データにしてマップを構築する。このような場合、名詞w_iは以下のように共起形容詞のセットで定義される。

$$w_i = \{a_1^{(i)}, a_2^{(i)}, \ldots, a_{\alpha i}^{(i)}\}$$

ただし、$a_j^{(i)}$はw_iと共起するj番目の形容詞で、αiは、w_iと共起する形容詞の数である。これを符号化するために、「相関コーディング法」を用いた。相関コーディング法では、それぞれの名詞間の意味的相関（あるいは意味的距離）を反映するものを求める。

表1　名詞の相関行列

	w_1	w_2	…	w_ω
w_1	d_{11}	d_{12}	…	$d_{1\omega}$
w_2	d_{21}	d_{22}	…	$d_{2\omega}$
⋮		⋱		
w_ω	$d_{\omega 1}$	$d_{\omega 2}$	…	$d_{\omega\omega}$

個々のd_{ij}はある名詞w_i、w_jの二語間を見る場合の関係であり、そのほかの名詞を参照系として考えるときのこの二つの名詞間の関係や、この二つの名詞と他の名詞との関係は、このようなd_{ij}の集合を用いるだけでは反映できない。局所的な意味関係なのである。しかし、このような個々の局所的な意味距離から表1に示すような行列を作成すれば、各行はそれぞれ同一名詞の相関関係を除いた$w-1$個の名詞との局所的意味距離から構成されていることが分かる。すなわち、各行は、ある名詞に対してそれ以外のすべての名詞との意味的な関係を反映していると考えられる。

したがって、ここで提案する相関コーディング法では、名詞w_iをこの行列を用いて以下のような多次元ベクトルに符号化する。

$$V(w_i) = [d_{i1}, d_{i2}, \ldots, d_{i\omega}]^T$$

$V(w_i)$はSOMへの入力であり、この多次元ベクトルを自己組織化によって、それらの間に存在する意味関係を顕在化して二次元空間に表現する。

2.3 二単語間の上位下位関係を求める補完類似度

二単語間の意味距離であるd_{ij}については、我々は、二単語間の上位下位関係を求めるのに有効な補完類似度を利用した[7]。

今、共起形容詞のセットで定義した抽象名詞FとTがあるとする。我々のデータでは、FとTの特徴ベクトルは、双方の共起形容詞の出現状況を0又は1で表現したものに相当する。

それを以下のように表す。

$$\vec{F} = (f_1, f_2, \ldots, f_i, \ldots, f_n)(f_i = 0\text{又は}1)$$
$$\vec{T} = (t_1, t_2, \ldots, t_i, \ldots t_n)(t_i 0\text{又は}1)$$

そして、補完類似度の式は以下のようになる。

$$Sc(\vec{F},\vec{T}) = \frac{ad - bc}{\sqrt{(a+c)(b+d)}}$$

“a”はFとTで共通する共起形容詞の数である。“b”はFとは共起するがTとは共起しない形容詞の数である。“c”はFとは共起しないがTとは共起する形容詞の数である。“d”はFともTとも共起しない形容詞の数である。FがTを完全に包含する場合、c＝0となり、TがFを包含する場合、b＝0となるため、bc＝0となる。補完類似度では、一致情報（ad）と不一致情報（bc）の差分をとるので、包含関係にある二語間の類似度は高くなる。

さらに、補完類似度はFからTの類似度とTからFの類似度が非対称であることも特徴の一つである。Fから Tを見た補完類似度では、bは、Fだけに出現する形容詞の数、cはTだけに出現する形容詞の数である。逆に、TからFを見た補完類似度では、bは、Tだけに出現する形容詞の数となり、cは、Fだけに出現する形容詞の数となる。計算式の分母をみると、FとTがどちらの方向の類似度を計算するかで、bとcに代入される数値の大小が逆転し、それに伴って、類似度も非対称になる。

二単語間の補完類似度値を、**2.2**で述べた意味距離 d_{ij} に代入して相関行列値をとり、自己組織化マップへの入力データとした。

2.4 概念全体の階層関係の構築

補完類似度で得られた結果から、すべての単語の最上位から最下位への階層構築[17]を行い、マップ上にプロットした。手順は以下である。

(1) 包含関係を示す類似度の値の高い順に単語A、Bをつなげる。ここでは、仮に単語Aが上位語、単語Bが下位語という関係とする。

(2) まず、単語Bを上位語として、最高値で下位語となる単語Yを探しBの後ろに連結するというように、A—Bを基点として下位（後ろ）に向かって連結を繰り返す。次に、単語Aを下位語として、最高値で上位語となる単語Xを探してAの前に連結するというように、A—Bを基点にして上位（前）へ向かって連結を繰り返す。一方、上位下位関係は必ず保存する。上位下位関係が壊れる場合は、その関係は連結しない。こうして一本の階層を作る。

(3) 長い階層に完全に含まれる短い階層はマージし、二つの階層が一単語だけ異なる場合は、差異となる二単語の補完類似度が上位下位関係を示せば、それに沿って結合した。

(4) 最後に各階層の最上位に「こと」を結合する。「こと」はすべての形容詞と共起することができ、最も抽象的な概念と考えることができる。計算時間の便宜上、「こと」は最後に各階層の最上位に結合させることとした。こうして、最終的に抽象名詞によって構成される、「こと」を最上位概念とした階層が得られる。

3 形容詞抽象概念の、階層関係を考慮した自己組織化マップ

以上のような手順によって階層構造を反映した自己組織化マップを構築した。概念は上が抽象レベルが最も高く、下へ分布するにしたがって具体名詞となっている。図1は、マップの「感情」に関係する概念階層を示している。右には、感情以外にも性格や状況などを表す形容詞の概念階層例を示している。

4 自動構築による階層とEDR電子化辞書における階層との比較

補完類似度、Overlap coefficient、頻度を考慮した補完類似度それぞれで自動構築された階層と、人手で構築されたEDRの階層とを比較した。被験者は、言語学者や自然言語処理分野、辞書編纂者合計20人であり、シェッフェの一対比較法によって心理実験を行った[8]。その結果、三手法で共通に自動構築された階層は、T検定後有意水準1％でみると、43％が、EDRの階層より妥当かあるいは劣らないと判断された結果になった。また、それぞれの手法独自に作られた階層は、EDRよりの階層より妥当ではないという評価結果になった。今後、さらに、形容詞にかかわる抽象概念の類義・階層関係の構造化手法を洗練し、類義関係、階層関係の妥当性ある体系を自動構築していき、シソーラスの検証に役立てる。

5では、単語間の階層構築手法を専門用語に応

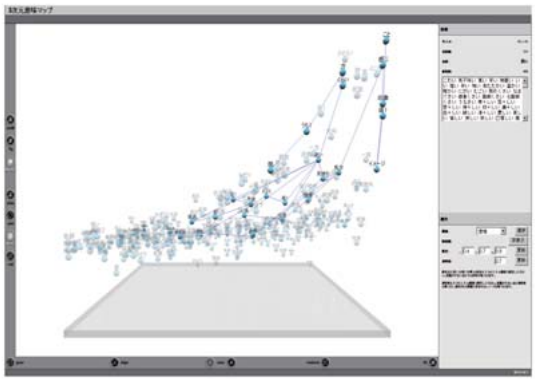

図1　階層関係を反映した形容詞属性概念の自己組織化マップ

［例］　形容詞の概念階層例
階層は、最上位概念→最下位概念の順
優しい
こと-面 1-イメージ-印象-感じ-気分-情感-情愛-心遣い
開放的な,頑固な,先進的な,明るい,誠実な
こと-面 1-性格-気質-気風-気性
難しい,優位な,不穏な,絶望的な
こと-状態-情勢-形勢

例　形容詞からみた概念階層例

用した研究を述べる。

5　自動階層構築方法の関連語集合抽出への応用

次に、**2.4** で示した自動階層構造構築方法[17]の応用を考える。階層関係のほかにも、なんらかの関係を持つ語をコーパスから集めた関連語集合は言語理解や言語生成、情報検索などに有効であると期待される。昨今、コーパスから語彙間の様々な関係を獲得するために、多くの手法が開発されるとともに[9]－[13]、関係を抽出するためのパターンを学習する手法も提案されている[14][15]。関連語集合は情報検索において、有益な情報にユーザを導くための手がかりとなる。Google の検索支援機能のようにユーザが入力したキーワードに関連する語を提示することが考えられるが、入力された語とどのような関係でつながる語を提示すれば、ユーザが適切な情報に到達することを支援できるだろうか。そこで、検索に用いるキーワードとしての有効性の観点で、自動階層構築方法を応用して、文書集合から関連語集合を抽出することを試みた。そして、その関連語集合のキーワード群としての特徴を分析した。

5.1　語彙間にある関係

検索支援において、どんな関係を持っている関連語が追加キーワードとして有効だろうか。語彙の間の関係には、少なくとも「分類的関係（taxonomical relation）」と「主題的関係（thematic relation）」の二つがある。これらの関係は語彙間の関係を認識し、理解するために重要であると報告されている[16]。

「分類的関係」とは、概念の持つ属性の類似性を表す関係のことで、例えば、「馬」、「牛」、「動物」といった単語の間にある関係である。同義関係、反義関係、階層関係などの意味的関係はこの分類的関係に含まれる。一方、「主題的関係」とは、主題的な場面を通して概念を結びつける関係のことで、例えば、「牛」と「ミルク」は「牛の乳を搾る」、「赤ん坊」と「ミルク」は「赤ん坊にミルクをあげる」といった場面を思い出させる、あるいはそのような場面で概念同士を結合する関係である。連想関係、因果関係、含意関係などはこの主題的関係に含まれる。

検索支援として追加される関連語は、よりよいキーワードに言い換えることを目的として、入力されているキーワードと分類的関係にある語彙が使われることが多い。これは既存の辞書やシソーラスにも直接記述されており、比較的容易に獲得し、利用できることも要因である。しかし、検索結果が有益なものに絞り込まれず、かえって結果が意図しないものとなることがある。一方、主題的関係は、文書の内容にかかわる語彙間の関係であるため、このような関連語を追加することで、検索結果を興味深いものに絞り込むことができ、ユーザにとって目新しい情報や知識を与えてくれることが期待できる。この観点での検索支援を目指して、本研究では、主題的関係に焦点を当て、主題的関係を持つと思われる関連語集合を抽出

し、その関連語集合を構成する用語の検索支援における有効性を調査した。

5.2 抽出方法

主題的関係を持つ関連語集合を抽出することを目的として、1) 文書集合から係り受け関係を収集し、実験データを作成、2) 自動階層構築方法を用いて関連語集合を抽出、3) シソーラスを用いて非分類的関係を持つ関連語集合を選別する。

5.2.1 共起関係の収集

文書集合を構文解析し、各文から「A＜の＞B」、「P＜を＞V」、「Q＜が＞V」、「R＜に＞V」、「S＜は＞V」のパターンにあてはまる係り受け関係を収集する。ここで、＜X＞は格助詞、A，B，P，Q，R，S は名詞、V は動詞を表す。収集した関係集合から 3 種類のデータ、具体的には、名詞間の共起関係に基づくデータ（NN データ）、名詞と動詞の係り受けに基づくデータ（格助詞ごと）（NV データ）、主語と目的語の関係に基づくデータ（SO データ）を作成した。

5.2.2 関連語集合の抽出

本論文で提案している自動階層構築方法を拡張し、関連語集合の抽出を行う。この方法は、与えられた二語について、それぞれの共起語との出現パターンの包含関係から語彙間の関係を推定する。前節までに示した単語間の意味関係の獲得においては、階層構造の抽出を目的としているため、用いる共起語をそれぞれの語の下位語に限定していた。本節では、共起関係を階層関係に限定せずに、上記に示した個々のデータに整理した係り受け関係を扱う。これによって、階層構造だけではなく、他の関係を持つ関連語集合も得られる。

5.2.3 主題的関係を持つ関連語集合の選別

最後に、抽出された関連語集合から、分類的関係を持つ関連語集合をシソーラスを使って取り除き、主題的関係を持つ関連語集合を得る。一般にシソーラスに含まれる語彙は分類的関係を表現するように配置されているので、分類的関係を持つ関連語集合は、シソーラス中で同じカテゴリに分類される。つまり、関連語集合がシソーラスに一致するなら、その関連語集合を構成する語彙は分類的に関連していると解釈できる。この考えに沿って、シソーラスに一致する関連語集合を取り除き、残った非分類的関係を持つ関連語集合を、主題的関係を持つ関連語集合として抽出する。

5.3 実験

実験では、医学部ドメインに限定して収集した文書集合（10,144 ページ、225,402 文）を使った。日本語の解析には医学用語辞書や専門用語辞書などは用いなかった。この文書集合から収集した関係集合から作成されたデータの数は、NN データが 225,402、NV データについてはそれぞれ、ヲ格データが 20,234、ガ格データが 15,924、ニ格データが 14,215、未格データが 15,896、SO データが 4,437 であった。シソーラスは Medical Subject Headings（MeSH®）シソーラスを用い、その見出し語とそれらのクロスリファレンスとして付随している類似語を和訳した用語を、抽出する関連語集合を構成する医学用語とした。実験データにはそのうち 2,557 個が現れた。

図 2 に抽出された関連語集合の一部を示す。抽出された関連語集合のうち、三つ以上の用語からなるものを次の選別の対象とした。

卵巣-脾臓-触診	潜伏期間−赤血球−肝細胞
新生児−動脈管開存症−壊死性腸炎	雪−学校−ガス
分泌−胃酸−胃粘膜−十二指腸潰瘍	変化−死−手足
皮膚-アトピー性皮膚炎-ヘルペスウイルス-抗ウイルス薬	病院−角膜混濁−トリアゾラム
皮膚-腹部-頚部-口腔-胸部	反応−アポトーシス−損傷
疲労-子宮筋-妊娠中毒症	研究−調査−味−米
水−酸素−水素−水素イオン	環境−関心−水−肉−下痢
疲労-ストレス-十二指腸潰瘍	権利−資源−心−教育−森林伐採

図2 得られた関連語集合の一部

5.4 分析

抽出した関連語集合が検索に有効であること、すなわち、有益な Web ページに検索結果を限定できることを Google を用いた検索によって調査した。調査の対象は、構成する用語が二つのカテゴリに分布し、そのうちの一つの用語だけが残りの用語と異なるカテゴリに分布する関連語集合とした。そのような関連語集合は、得られた関連語集合 847 個のうち、294 個あった。調査対象とした関連語集合を{X_1, X_2, ……, X_n, Y}と表すとき、X_i は同じカテゴリに分類される用語、Y は X_i と異なるカテゴリ分類される用語とする。このとき、これらの関連語集合それぞれから以下の三種類の検索キーワード群を作成した。

Type 1：異なるカテゴリに分類される Y を除いた{X_1, X_2, ……, X_n}

Type 2：同じカテゴリにある用語のうち一つの用語 X_k と Y を除いた{X_1, X_2, …, X_{k-1}, X_{k+1}, …, X_n}

Type 3：同じカテゴリにある用語のうち一つの用語 X_k を除いた{X_1, X_2, …, X_{k-1}, X_{k+1}, …, X_n, Y}

この三種類は、Type 2 を元となるキーワード、つまり初めに入力されたキーワードとしたとき、Type 1 は Type 2 に同じカテゴリに分類される用語を追加したキーワード群である。追加用語は本研究で使った文書集合において頻度に関する特徴を持ち（高い又は低い頻度を持つ）、Type 2 にある用語と分類的に関連する用語である。一方、Type 3 は Type 2 に異なるカテゴリに分類される用語を追加したキーワード群で、この追加用語は Type 2 にある用語と主題的に関連すると思われる、非分類的に関連する用語である。

まず、Google の検索エンジンが推定し、提示するヒットページ数を使って、量的に検索結果を比較する。具体的には、Type 2 を用いて得たヒットページ数を基準に、Type 2 に一用語追加した Type 1 と Type 3 をそれぞれ用いた場合のヒットページ数を比較する。図 3 と 4 にそれぞれ高頻度と低頻度に関するヒットページ数による比較結果を示す。これらの図において、横軸は元となるキーワード（Type 2）を用いた場合のヒットページ数、縦軸は元となるキーワードに一用語追加した場合（Type 1 又は Type 3）のヒットページ数である。図中の「○」は同じカテゴリにある用語を追加した場合（Type 1）、「×」は異なるカテゴリにある用語を追加した場合（Type 3）のヒットページ数を表す。対角線は Type 2 に用語を一つ追加してもヒットページ数に影響がない場合を示す。

図 3 において、多くの「×」が対角線のかなり下にあることが分かる。これは、異なるカテゴリにある非分類的に関連する用語を追加するほうが、同じカテゴリにある分類的に関連する、高頻度の用語を追加するよりもヒットページ数を減少させる傾向にあることを示している。このことから、有益なページを検索するために、非分類的に関連する用語を追加することは量的に有効であり、その非分類的に関連する用語は分類的に関連する高頻度の用語よりも有益な用語であると考察できる。図 4 においては、図 3 とは対照的に、多くの「○」が対角線のかなり下にあることが分かる。それらの関連語集合を見ると、追加された分類に関

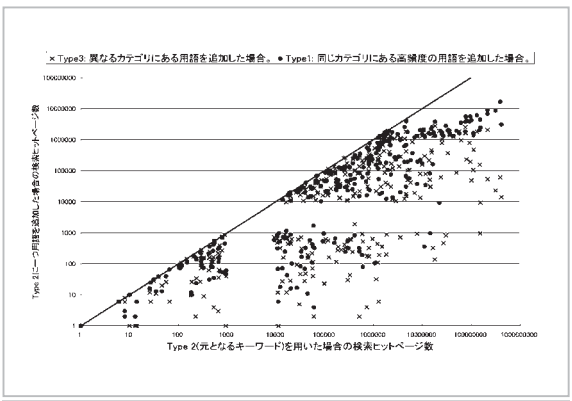

図3 高頻度の用語と異なるカテゴリにある用語をそれぞれ追加した場合のヒットページ数の変動

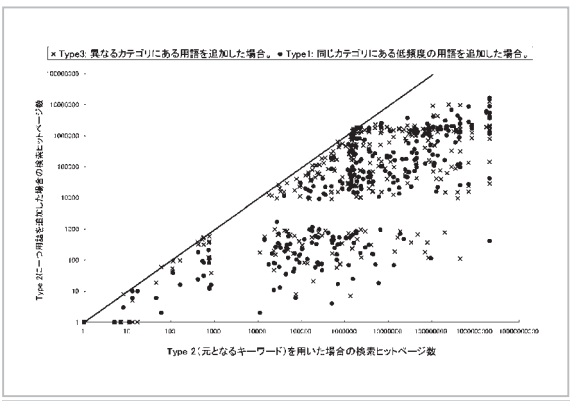

図4 低頻度の用語と異なるカテゴリにある用語をそれぞれ追加した場合のヒットページ数の変動

連する用語がその関連語集合の中で最も低頻度である場合が多かった。これは、低頻度の用語を追加したほうが非分類的に関連する用語を追加するよりもヒットページ数を減少させる傾向にあることを示している。実際に、低頻度の用語はインターネット上でもまれな用語である場合があり、それを含むWebページ自体が少ないと予測できる。したがって、低頻度の用語を追加することはその用語の関係の種類にかかわらず、検索結果に対して量的に有効である。しかし、非分類的に関連する用語を追加した場合と分類的に関連する低頻度の用語を追加した場合の結果の内容を考察すると、そこには大きな違いがある。

例えば、SOデータから得た関連語集合「潜伏期間―赤血球―肝細胞」について考察する。これは、「潜伏期間」がMeSHシソーラスにおいて他の用語と異なるカテゴリに分類される用語で、「肝細胞」が残りの「赤血球」と同じカテゴリに分類される低頻度の用語である。この関連語集合を構成する用語すべてをキーワードとして用いると、検索結果の一位に「マラリアとは?」というタイトルの日本語ページが位置する。「潜伏期間」と「赤血球」を用いた場合(Type 3)も同じページが一位に位置する結果を得る。しかし、「赤血球」と「肝細胞」を用いた場合(Type 1)は、このページは上位10ページ以内には入っていたが、一位ではなかった。他の例として、NNデータから得た関連語集合「卵巣―脾臓―触診」について考察する。これは、「触診」がMeSHシソーラスにおいて他の用語と異なるカテゴリに分類される用語である。この関連語集合を構成する用語すべてをキーワードとして用いると、「卵巣と脾臓の疾患は触診で診断できる。」という情報を含むページが検索される。この結果から、この関連語集合は因果関係を持つと解釈できる。したがって、この関連語集合がユーザの意図を正確に定義し、関連のあるWebページを検索できることを示唆している。

実験において、他の用語と非分類に関連する用語は有益なページに検索結果を限定することに有効であった。これに対して、分類的に関連する用語では、非分類に関連する用語と比べ、高頻度の用語は量的に有効ではなく、低頻度の用語は質的に有意な傾向が見られなかった。今回は最初の試みとして、一つのドメインに限って実験を行い、考察したが、より正確に主題的関係を持つ関連語集合を抽出するために研究を発展させ、より量的かつ質的にその関連語集合の有用性を検証することが今後の課題である。

参考文献

1 河原哲雄,“概念の構造と処理”,人工知能学会誌,Vol.16,No.3,pp.435-440.2001.
2 T. Kohonen, "Self-organizing maps 2nd Edition", Springer, Berlin, 1997.
3 根本今朝男,“「が格」の名詞と形容詞とのくみあわせ”,電子計算機のための国語研究,国立国語研究所,1969.
4 高橋太郎,“文中にあらわれる所属関係の種々相”,国語学103,国語学会,pp.1-16,Dec. 1975.
5 益岡隆志,“名詞修飾節の接続形式―内容節を中心に―”,日本語の名詞修飾表現,田窪行則(編),pp.5-27,くろしお出版,東京,1994.
6 Q. Ma, K. Kanzaki, M. Murata, K. Uchimoto, and H. Isahara, "Self-Organization Semantic Maps of Japanese Noun in Terms of Adnominal Constituents", In Proceedings of IJCNN'2000, Como, Italy, vol.6.:91-96., 2000.
7 山本英子,梅村恭司,“コーパス中の一対多関係を推定する問題における類似尺度”,自然言語処理,Vol.9,No.2,pp.46-75,Apr. 2002.
8 H. Scheffe, "An analysis of variance for paired comparison" Journal of the American Statistical Association, 47, 381-400., 1952.
9 M. Geffet and I. Dagan, "The Distribution Inclusion Hypotheses and Lexical Entailment", Proceedings of ACL 2005, pp.107-114, 2005.

10 R. Girju, "Automatic Detection of Causal Relations for Question Answering", Proceedings of ACL Workshop on Multilingual summarization and question answering, pp.76-83, 2003.
11 R. Girju, A. Badulescu, and D.Moldovan, "Automatic Discovery of Part-Whole Relations", Computational Linguistics, 32(1): pp.83-135, 2006.
12 M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora", Proceedings of Coling 92, pp.539-545, 1992.
13 I. Szpektor, H. Tanev, I. Dagan, and B. Coppola, "Scaling Web-based Acquisition of Entailment Relations", Proceedings of EMNLP 2004, 2004.
14 D. Ravichanfran and E. H. Hovy, "Learning Surface Text Patterns for A Question Answering System", Proceedings of ACL 2002, pp.41-47, 2002.
15 P. Pantel and M. Pennacchiotti, "Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations", Proceedings of ACL 2006, pp.113-120, 2006.
16 E. J. Wisniewski and M. Bassok, "What makes a man similar to a tie?", Cognitive Psychology, 39: pp.208-238, 1999.
17 E. Yamamoto, K. Kanzaki, and H. Isahara, "Extraction of Hierarchies based on Inclusion of Co-occurring Words with Frequency Information", IJCAI 2005, pp.1166-1172, 2005.

神崎享子（かんざききょうこ）
知識創成コミュニケーション研究センター自然言語グループ研究員（旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ研究員）
博士（学術）
自然言語処理

山本英子（やまもとえいこ）
知識創成コミュニケーション研究センター自然言語グループ特別研究員（旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ専攻研究員） 博士（工学）
自然言語処理

井佐原 均（いさはらひとし）
知識創成コミュニケーション研究センター自然言語グループリーダー（旧情報通信部門けいはんな情報通信融合研究センター自然言語グループリーダー） 博士（工学）
自然言語処理