自然言語処理 115－19
情報学基礎 43－11
（1996. 9. 13）

多様分類情報による検索語拡張

沖電気工業(株) 研究開発本部 関西総合研究所
下畑 光夫, 坂本 仁
{simohata , saka}@kansai.oki.co.jp

概要

検索語を与えて検索を行う際に、検索範囲の拡大のために他の検索語を追加する事が多い. 語の意味体系の表現形式としてシソーラスがよく知られているが、語と語の一般的関係を記述することを目的としているため、様々な観点で用いられることが多い検索語を拡大するには適切な形式であるとは言えない.
本論文では、検索語の拡張を主眼とした多様分類情報について述べる. 多様分類情報は観点によるリンクを特徴とし、観点を選択して検索語拡大を行なうことで適切な検索語拡大が実現できる.
また、多様分類情報を検索者に提示することにより、データベースの語体系の理解を助けるという利点も備えている.

Extension of keywords using varied viewpoint information

Kansai Lab.,Research & Development Group,Oki Electric Industry Co.,Ltd.,
Mitsuo Shimohata, Masashi Sakamoto

Abstruct

When we give keywords to IR system for searching data, IR system usually adds other words to original keywords for extending searching area. Thesaurus is well known for representing system of words. But because their purpose is representation of general relation of words, they are not suitable for extending keywords.
In this paper,we discribe about varied viewpoint information suitable for extending keywords. It is distinguished for viewpoint link between word and word. And proper extension of keywords can be performed by selecting proper viewpoint.
It also has a merit to help searcher to understand word system of database for showing varied viewpoint information.

1 はじめに

検索語を入力して情報検索を行う場合、検索範囲を広げるために検索語に関連する語を論理積で結合することが多い. そのためには、検索語に意味的に関連する語の意味体系をなんらかの形式で保持する必要がある.

語の意味体系を表す形式として、シソーラスがよく知られている. [1][2][3] シソーラスでは類義、対義、上位・下位などの語の関係のリンクを語間に付与している. 例えばWordNet[2]では、「brother」と「sister」の間には対義語のリンクが、「relative」と「brother」の間には上位・下位リンクが、「relative」と「family」の間には全体・部分リンクが付与されている. シソーラスでは普遍的に成り立つ関係を体系として表すことを主眼としており、特殊な状況の基で成立する関係を記述することは少ない.

しかし、検索語の拡大においては、普遍的に成り立つ関係のみならず特定の状況でしか成り立たない関係についても扱えることが重要となる. その意味で、語の関係を主とした語の意味体系は検索語の拡大にはあまり適していないと言える.

本論文では、検索語の拡大を主目的とした多様分類という形式について述べる. 多様分類は観点をリンクとして語の意味体系を表現しており、検索者は観点を選択して適切な検索語拡大を行なうことができる. また、多様分類情報を使用者に提示することで、データベースの語の意味体系の理解に役立つという効果を挙げることについても述べる.

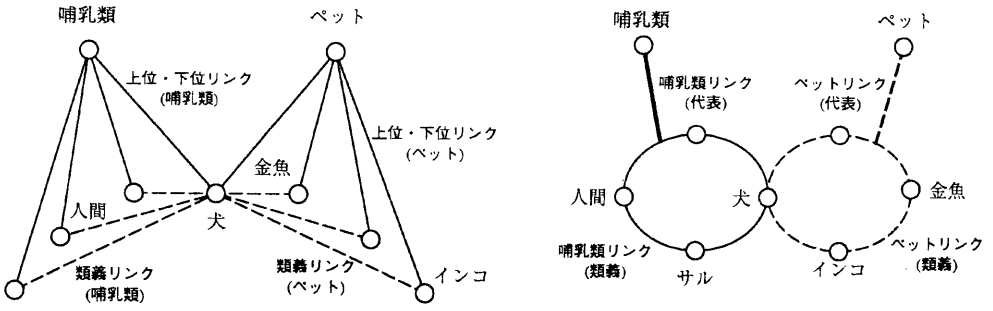
2 多様分類

語には通常、多様な意味(側面、性質)がある. この多様な意味の一つ一つを観点と定義する. 上位・下位や類義や対義などの語と語の間の関係はどの観点で評価するかにより大きく変化する. 例えば、「犬」と「金魚」という2語を考えると、生物学的分類の観点では哺乳類と魚類であり両者の意味的距離は大きいが、ペット用動物と言う観点で見れば両者は意味的に近く類義語としても構わない. しかし、シソーラスは普遍的な語と語の関係をリンクで結合することが基本であり、観点により関係が変化するという面はあまり考慮されていない. (図1a) ただし、必要であればリンクの情報に観点に関する情報を付加することはできる.

検索語を与えて検索を行なう時に、検索範囲を拡大するために元の検索語に他の語を追加して検索することがしばしば行なわれる. このような場合、観点を無視し語の関係を見て追加する語を選ぶより適切な観点に基づいて追加する語を選ぶ方が検索者にとって適切である場合が多い. 「犬」に他の語を追加する時に、観点を無視した(言い換えればすべての観点の)すべての類義語を追加するよりも「ペット用動物」「哺乳類」といった観点を選択する方が適切な拡大が望める.

多様分類は検索語の拡大という目的を考慮し、観点をリンクの主情報とした表現形式である. (図1b) 多様分類情報は類義語集合と代表語から構成される. 類義語集合と代表語は以下のように定義される.

類義語集合	ある観点の下で類似である複数の語の集合.
代表語	類義語集合を代表する語. 適切な語が存在しない場合は省略してもよい.



シソーラスの語間のリンク (a)　　多様分類の語間のリンク (b)

図 1: シソーラスと多様分類の語間のリンク

類義語集合と代表語は、3 章で述べるようにキーワードとして扱われるため、キーワードの語彙から選択しなければならない. また、代表語と類義語集合の関係は上位・下位の関係だけでなく、実体・属性や全体・部分といった関係でもよい. 表 1に「犬」を中心とした多様分類の一例を示す. 表中に分類名の項があるが、これは人間が多様分類情報を見た時に観点を分かりやすく示すためのタイトルであり、任意のタイトルをつけてよい.

この多様分類情報により、任意の観点の下で類義の語の集合を表現する事ができる. 多様分類情報を作成する場合に他の分類情報からの制約はなく、同一の語が複数の分類に類義語または代表語として存在してもよい. また、ある 2 語で代表・類義語の関係が逆転するような分類情報が共存してもよい.

分類名	代表語	類義語集合		
哺乳類	哺乳類	犬	人間	サル
ペット用動物	ペット	犬	インコ	金魚
犬の用途	犬	ペット	牧羊犬	番犬
干支の動物		犬	牛	ネズミ

表 1: 多様分類情報の一例

3　多様分類を利用した検索語拡大

検索語に関連する検索語を追加する時に、多様分類により観点に焦点を当てた拡大を行う事ができる. 多様分類による拡大には、上方、類義、下方の 3 種類がある. 以下に、3 種類

の拡大方法と表1の分類情報と検索語「犬」における拡大例を示す.

上方拡大 検索語を類義語集合に含む分類を収集し、その分類の代表語を追加する.
例)「犬」→「哺乳類」,「ペット」

類義拡大 検索語を類義語集合に含む分類を収集し、その分類の類義語集合を追加する.
例)「犬」→「人間」,「サル」(または「インコ」,「金魚」)

下方拡大 検索語を代表語に含む分類を収集し、その分類の類義語集合を追加する.
例)「犬」→「ペット」,「牧羊犬」,「番犬」

ある語を拡大する場合に上記の3種類の拡大方法のうちどれを採用するかは使用者が選択する.(複数の拡大を選択してもよい.)代表語の存在は必須でないため、実用では類義拡大が中心となる.

類義拡大の例にあるように、複数の分類情報がマッチすることが考えられる.このような場合にどの分類情報を使用するかを決定しなければならない.静的評価方法としては、個々の分類に評価点をあらかじめ設定し評価点の高い分類を採用する方法がある.また、動的評価方法としては、クエリー中の他のキーワードと多様分類中の他の類義語との関連度により選択するという方法が考えられる.

多様分類による検索語の拡大は1レベルしか行わない.例えば「犬」の類義拡大では「人間」という語が追加されるが、さらに「人間」について類義拡大を行うことはしない.シソーラスで用いられる語関係のリンクでは、$W_a = a\_kind\_of(W_b)$, $W_b = a\_kind\_of(W_c)$ であれば $W_a = a\_kind\_of(W_c)$ であるといった推移的性質を持つため多段に渡る展開に有効性があるが、多様分類情報における観点の多段展開は推移性が保証されないため多段の展開は不要な語を混入させる危険が大きい.推移性を持つ語間の関係を表す情報または分類間の関係を表す情報を与えることで、多段展開を有効にすることができると考えられる.

4 多様分類の作成

多様分類情報を作成手段として、人手で作成する方法とテキストデータから自動生成する方法がある.多様分類情報は他の分類情報と整合をとる必要がないため、どちらの方法でも比較的容易に作成することができる.また、後修正も容易である.

多様分類情報がより有効に機能するためには、多様分類情報とデータベースの両者での語の意味体系が一致していなければならない.人手で多様情報を作成する場合には、多様分類情報の作成者はデータベースの語の意味体系を把握し、データベースの語の意味体系に沿って多様分類情報を作成する必要がある.

電子化されたシソーラスから類義語集合を作成すれば労力は大幅に低減されるが、シソーラスは一般的な語について普遍的な意味で体系を記述しており、固有名詞(特に専門用語のような特殊性の高い語)はあまりカバーしていない.このような語を検索語として与えた場合や特定の分野を対象としたデータベースに対しては有用な情報は多くは望めない.また、データベースの語の意味体系と一致しているとは限らないので、後修正はやはり必要であろう.シソーラスを自動作成する方法も種々提案されているが[4][5]、質の高い情報を作成することは困難である.

データベースがテキストデータから構成されている場合は、テキストデータから多様分類情報を自動作成することが考えられる.自動作成により、作成に要する労力の低減と、デー

タベースと多様分類情報の語の意味体系の一致を図ることができる. 2章で述べたように、多様分類はある観点について類似の集合を収集したものである. 自動生成用に言い換えるならば、観点の類似を反映する統計的、言語処理的条件にしたがって語を収集することで多様分類が作成できる. テキストデータからの自動作成は、この「観点の類似を反映した条件」を適切に決定できるかどうかに依存する. 観点の類似を反映した条件としては同じ語を共起語とすることなどが考えられ、コーパスからの共起知識獲得の技術が利用できると考えられる.[6]

5 検索者と検索システムのオントロジーの整合

検索語を与えて検索を行う場合に、検索語が検索システムに認識された(未知語でない)としても検索者と検索システムで語の定義が一致しておらず、検索結果が満足でないことが多い. これは、検索者とデータベースの両者でオントロジーが一致していないためである. 例えば、コンピュータをよく知らない検索者がパソコンに関する情報を収集しようして「コ

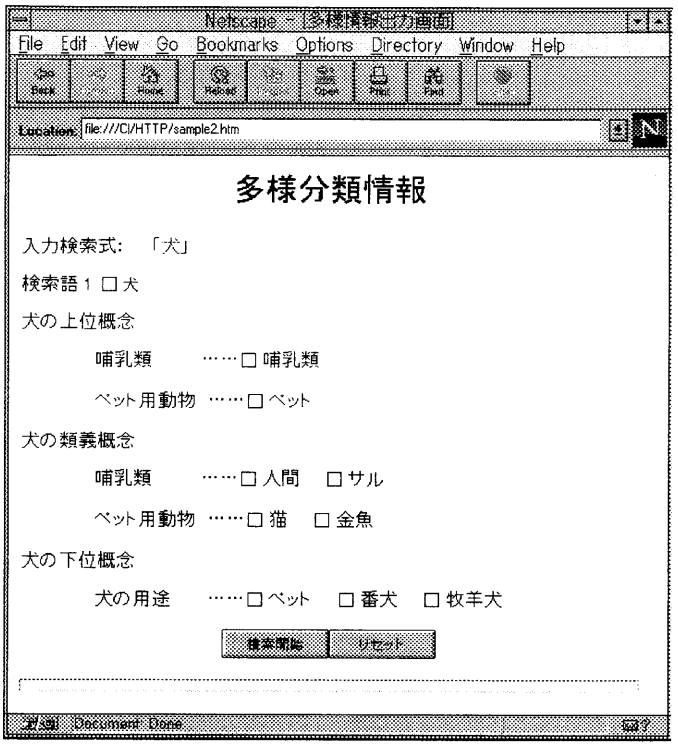


図 2: 多様分類情報の出力画面 †

†ブラウザは、Netscape Communications 社の Netscape Navigator を使用.

ンピュータ」という検索語を入力したとする．検索システムでコンピュータはパソコンやワークステーションやスーパーコンピュータの上位概念として扱われているとすると、検索者へはパソコンのみならず、ワークステーションなどの関連情報も合わせて出力されることになる．検索者は、検索結果を基に自分の要求するデータを得られるように検索式を変更する必要があるが、検索結果からデータベースの語の意味体系を理解して検索語を修正することは容易ではない．このような場合に、検索語の多様分類情報を提示することで、検索語のオントロジーの理解を補助することができ、検索語の修正に大きく役立つと考えられる．

図2にWWW上で作成した試作システムの多様分類情報の出力画面を示す．ここでは、「犬」について上方・類義・下方の3種類の拡大を行った場合を出力している．「犬」の様々な観点による分類が理解できるとともに、任意の語をチェックし、再検索を行うことが可能である．

6 結論

本論文では、多様分類情報により検索語の拡大を観点を基に行なえること、また多様分類情報による展開方法、データベースのオントロジーの理解に役立つことについて述べた．しかし、まだ試作レベルである多様分類情報が実用的になるために解決すべき課題は多い．拡大に使用できる分類が複数ある場合に、使用者の意図に合う分類を選択する方法と、テキストデータから多様分類情報を自動生成する方法の確立が当面の課題であると考えている．新聞記事や技術マニュアルなどのデータベースを対象にし、検索効率の評価、オントロジー教示の効果の評価を行なう予定である．

参考文献

[1] Roget's thesaurus:Longman,1983.

[2] Geoge A.Miller, et al:Introduction to wordnet:an on-line lexical database:

[3] EDR の概念体系辞書,http://www.iijnet.or.jp/edr

[4] Padmini Srinivasan:Thesaurus construction,Information Retrieval,Prentice Hall,1992

[5] 山崎 毅文, マイク パッザーニ: 帰納学習アルゴリズムと階層型クラスタリング手法を用いた概念シソーラスの自動構築及び更新, 情報処理学会情報学基礎研究会,39-7,pp49-56, 1995

[6] 工藤 育男, 井ノ上 直己: コーパスに基づく共起知識の獲得とその応用, 人工知能学会誌, Vol.10,No.2,pp.205-212(1995)