

意味属性に基づくテキストベース検索方式†

松尾 比呂志†† 内 野 一††

検索指示文と各テキストの見出し文(検索対象文)との文間類似度を用いてテキストを検索する方式について述べる。テキストの検索方式として、文の中からキーワードを抽出して比較する方式が知られているが、単語間の比較であるため、検索指示文「頭が痛い」でキーワード「頭痛」が付与されたテキストを検索できない。また、検索指示文および各検索対象文を意味構造に変換して比較する文照合方式も試みられているが、「頭痛がする」と「頭が痛い」の関係のような、複数の単語の組合せが一方の文の単語と同義の関係にある場合の検索は行えない。このような類義関係も扱える方式として、文の意味を意味属性の集合で近似的に表すことにより、検索指示文と検索対象文との間の類似度を算出して検索する方式を提案する。具体的には、①文中に含まれる単語の意味を、複数の意味属性の集合で表現する、②文の持つ意味を、文中に含まれる単語の意味属性の集合として扱う、③意味属性の集合間の類似度を、文間の類似性の尺度とする。本方式を適用したテキストベース検索システム(TBRIS)を実現し、「けがや病気の応急手当」に関するテキストを対象に評価実験を行った。その結果、約90%の検索成功率を得ることができ、本方式の有効性を検証した。

1. はじめに

新聞記事検索や文献検索等のように複数のテキストを対象に検索するテキストベース検索方式として、各テキストに人手であるいは自動的に^{1),2)}付与したキーワードを基に検索する方式(キーワード検索方式)が知られている。検索条件を指定する方法には、検索論理式を直接入力する方法と自然言語による検索指示文を入力する方法があるが、一般のユーザでも簡単に検索できるという点で、日常使う言葉でそのまま入力できる後者の方法が有利である。検索指示文から検索条件を生成する方法としては、抽出したキーワードの集合を直接使用する方法やそれらのキーワードから検索論理式を生成する方法が考えられている。前者は、テキスト側もキーワードの集合で表して、キーワードの集合間の類似度により検索を行う³⁾のに対し、後者は、検索指示文を意味構造に変換した後、世界知識等を用いて検索論理式を生成する⁴⁾⁻⁵⁾。

これらの検索方式を多様な表現での検索という観点からみると、キーワードの表面上の違いを吸収するために、類義語辞書等(シソーラス)を用いている。しかし、広範囲の語を類義語として扱うと検索結果にゴミが多くなり、一方、狭すぎると検索漏れが生じるなど、検索精度を上げることが困難である。

検索精度を上げるため、秋山らは、テキスト内容が見出し文に集約されていることを利用して、見出し文

と検索指示文との間の照合(内容マッチング)により、キーワードで検索した結果を絞り込んでいる^{4),5)}。この方法は、検索結果のゴミを除くことは期待できるが、キーワード検索の段階で、検索漏れが生じる。例えば、検索指示文「頭が痛い」からキーワード「頭」と「痛い」が抽出されても、「頭痛」がキーワードとして付与されたテキストが検索されない。このように、検索指示文中の複数のキーワードの組合せと同義であるキーワードを持つテキストが検索されないなど、検索漏れを防ぐという観点から考えると、従来のキーワード検索方式には限界がある。

そこで、本論文では、多様な表現の類似関係を扱うために、単語の意味属性に基づいて、検索指示文と各テキストの見出し文との意味的類似性により検索する方式を提案する。

文間の類似性の判定を行う方法としては既に、述語論理式で表現して照合を行う秋山らの方法⁵⁾や、複合名詞句表現に対応する表現に変換して照合を行う高松らの方法⁶⁾が提案されているが、いずれも、「頭が痛い」と「頭痛がする」のように、1つの単語が表す概念が比較対象の文に含まれる複数の単語を複合した概念である場合には、意味的には一致するにもかかわらず、類似した文として扱うことができない。これに対し、本論文の方法は、単語の意味属性の集合で文を表すことで、複数の単語の組合せが一方の文の単語と同義の関係にある場合でも、意味的な類似性の判定が可能となる点に特徴がある。具体的には、①単語の意味を、意味属性の集合で表現する、②文の持つ意味を、文中に含まれる単語の意味属性の集合として扱う、

† A Method for Text-base Retrieval Based on Semantic Categories by HIROSHI MATSUO and HAJIME UCHINO (NTT Communication and Information Processing Laboratories).

†† NTT 情報通信処理研究所

③意味属性の集合間の類似度を、文間の類似性の尺度とすることにより、多様な表現の類義関係を考慮した検索が実現されることを示す。類似度計算では、意味属性と検索対象文との対応を示すインデックステーブルを用いた高速な方法を示す。さらに、本論文では、本方式を適用したテキストベース検索システム(TBRS)の例を示し、その評価実験結果を基に本方式の有効性を示す。

2. テキスト検索における意味属性の有効性

2.1 意味属性

検索指示文と見出し文(検索対象文)との間の意味的類似性を考慮した検索を実現するため、文の意味を単語の意味の集合として近似する。また、単語列も文の特殊な場合として考え、同一の処理で扱う。

単語の意味に対しては、以下のように考える。単語は、ある概念に付けられたラベルであり、その概念が単語の意味であると考え、意味的類似性を考慮した検索を可能とするには、概念と概念との類似関係を表現する必要がある。本論文では、この類似関係を以下の2つの考え方に基づいて表す。

(1) 概念の階層化

複数の概念において、多少の異なりを捨象して、これらに共通する性質や構成要素を表す概念(すなわち、上位概念)を考える。さらに上位概念に対する上位概念を考えることによって、概念のツリー状の関係が記述される。このような考えに基づいたものとして、意味素性やシソーラスがこれに相当する⁷⁾。

本論文では、一般に意味素性などで分類されている概念(これを意味分類と呼ぶ)よりも、より詳細に分類した概念を想定しており、ある2つの概念に対し、親概念(すぐ上位の概念)が共通であるならば、意味的類似性が高い概念だと考える。例えば、同義語は、単語自身が表す概念の親概念が同一な語として考える。ここでは、単語自身が表す概念より上位の概念に付けられたラベルを意味属性と呼ぶ。意味属性を単なる番号で表現してもよいが、本稿では、意味属性がどのような概念に対応するかが判るように、意味属性に対応する概念を【 】で囲って表現する。

単語自身が表す概念の親概念に相当する意味属性を最下位の意味属性と呼ぶ。原則として、各単語には最下位の意味属性を付与するが、その上位概念が検索対象分野において具体的な対象や具体事象を指す重要な概念である場合には、その上位概念に属する意味属性

も付与する。例えば、単語「膝」は概念「ひざ」に属する単語であり、概念「足」は概念「ひざ」の上位概念である(図1(a))。例えば、対象テキストが「けがや病気の応急手当」に関するテキストの場合は、概念「足」は、「けが」の具体的な場所を表す重要な概念である。このとき、単語「膝」に対して、その最下位の意味属性【ひざ】とその上位概念に対応する意味属性【足】を付与する。

(2) 複合概念

複数の概念の組合せで表される概念がある。例えば、複合語は複数の単語が連結されたものであり、各単語が示す概念の組合せで複合語が示す概念を表すことができる。複合語が表す概念を複合概念と呼ぶ。一般には、単語がそのまま連結された語を複合語と呼んでいるが、複合概念という観点で考えれば、一般には複合語と呼ばない単語でも、複合概念を表していると考えてよい単語もある。例えば、図1の(b)において、単語「頭痛」は「頭の痛み」という意味を持ち、単語「頭」が表す概念と単語「痛み」が表す概念の組合せからなる複合概念と考える。これにより、単語「頭痛」の意味を、単語「頭」と単語「痛み」に付与された意味属性の集合【頭】、【痛み】で表現する。

2.2 単語分割と意味属性の抽出

単語の意味は、単語辞書中に、複数の意味属性を登録することで記述する。意味属性は自立語(名詞・形容詞・動詞)に対して与える。ただし、「する」「ある」のように単独では具体的な意味を表さない単語には意味属性を与えない。また、活用語に対しては、語幹に対して付与する。

図2は、単語の表記と登録した意味属性の例(「検索のための意味属性」の部分)を示している。例えば、単語「頭痛」には、【頭】と【痛み】という2つの意味属性が登録され、単語「頭」には、意味属性【頭】、単語「痛い」の語幹「痛」には、意味属性【痛み】が登録されている。

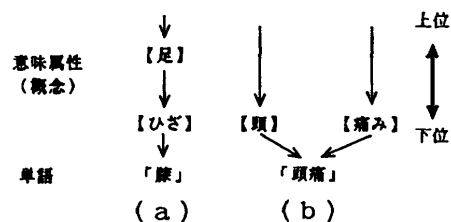


図1 意味属性と単語の関係

Fig. 1 Relation between semantic categories and a word.

単語表記	検索のための意味属性	システム辞書の意味属性(意味分類)
頭	【頭】	【頭】【先】【毛髪】
痛(い)	【痛み】	【痛み】【苦しみ】他
頭痛	【頭】【痛み】	【病気類】
嘔吐	【吐く】	【病気類】
しもやけ	【凍傷】	【病気類】【けが】
凍傷	【凍傷】	【病気類】【けが】
薬	【医薬品】	【医薬品】【薬品】

入力装置	【in】	} 専門用語
出力装置	【out】	
入出力装置	【in】【out】	
故障	【障害】	【不調】【衰微】

図2 単語辞書による意味属性の登録
Fig. 2 Semantic categories in the word dictionary.

検索対象文および検索指示文は、形態素解析処理により単語に分割され、各単語に対して意味属性が抽出される。形態素解析処理は、文献 8) で示されている方法に基づいて行う。なお、テキスト登録時に検索対象文の形態素解析処理が、テキスト検索時に検索指示文の形態素解析処理が行われる。

● 複合語の扱い

複合語も、形態素解析処理で単位語(辞書にはこの単位で登録されている)に分割される⁹⁾。例えば、「頭痛薬」を単位語として登録しない場合、この複合語「頭痛薬」は単位語「頭痛」と単位語「薬」に分割され、単語「頭痛」から意味属性【頭】【痛み】が抽出され、単語「薬」から意味属性【医薬品】が抽出される。一方、「頭痛薬」を単位語として登録する場合には、意味属性として、【頭】【痛み】【医薬品】を登録する。これにより、いずれの場合でも、単語「頭痛」と「頭痛薬」は意味属性【頭】と意味属性【痛み】とで関連づけられる。

● 活用語の扱い

活用語に対しては、その語幹と活用形が単語辞書に登録され、形態素解析処理によって語幹と語尾に分割される。意味属性は、語幹に対して付与される。例えば、「痛い」は、語幹「痛」と活用語尾「い」に分割され、語幹「痛」に意味属性【痛み】が付与される。

2.3 文の意味の表現とその有効性

(1) 意味属性の集合による文の意味の表現

検索対象文および検索指示文は、形態素解析処理により単語に分割され、各単語の意味属性が抽出される。抽出された意味属性の集合として、文の意味を表す。図3に、図2で示した単語辞書を使用した場合の文の意味の表現例を示す。例えば、文1「頭が痛い」の場合には、単語「頭」から意味属性【頭】、単語「痛

	文	文の意味
文1	頭が痛い	{【頭】、【痛み】}
文2	頭痛がする	{【頭】、【痛み】}
文3	入力装置が故障した	{【in】、【障害】}
文4	出力装置が故障した	{【out】、【障害】}
文5	入出力装置が故障した	{【in】、【out】、【障害】}

図3 意味属性による文の意味の表現
Fig. 3 Expressions of sentence meaning by semantic categories.

(い)」から意味属性【痛み】が抽出され、文1の意味は以下のような集合で表現される。

{【頭】、【痛み】}

(2) 意味属性による表現の有効性

文の意味を図3のように、意味属性の集合で表すことにより、文間の意味的類似性を考慮した検索を、意味属性の集合間の類似性を判定することにより行うことができる。

例1: 図3において、文1と文2は、同じ集合で表現され、検索指示文「頭が痛い」で検索対象文「頭痛がする」を検索でき、また、その逆も可能である。

例2: 文3に対しては、例えば一致する意味属性の個数で判定すれば、文5が最も類似し、文5に対しては、文3と文4が同程度に類似していると判定できる。

3. 意味属性の付与方法

本方式では、意味属性をいかに的確に付与するかが高い検索精度を得る上で重要なキーとなる。

意味属性は、単語辞書に登録することによって付与するが、1つの辞書ですべての分野をカバーすることは、困難なので、単語辞書として、システム辞書と分野辞書の2つの辞書を用いる。システム辞書には、一般に使用される語とその語が表す概念の上位概念に対する意味属性(すなわち、意味分類)を登録しておく(図2を参照)。分野辞書では、対象分野における重要な語に、細分化した概念や、複合概念を構成する概念、を表す意味属性を登録する。すなわち、システム辞書で付与されている意味属性を詳細化するという考え方で辞書の構築を行う。このためには、できるだけ概念が詳細に分類されている辞書をシステム辞書として用いたほうがよい。概念が詳細に分類された辞書として、文献 8)、9) で用いている辞書がある。この辞書では、日英翻訳のための訳し分けのために、階層的に分類された約2,800の意味属性が用意され、30万語が登録されている。本論文では、この辞書を基に作成

した 15 万語の辞書をシステム辞書として用いる。以下に、意味属性の具体的な付与方法について述べる。

(1) 分野辞書に登録すべき単語の識別方法

①未知語（辞書に登録されていない文字列）:

テキスト登録時に見出し文（検索対象文）が形態素解析される。この形態素解析処理では、未知語がある場合には、その文字列を指摘できる。未知語として指摘された文字列は、辞書に登録する。

②意味属性を詳細化する語:

形態素解析された後、検索対象文中の単語とその意味属性をテーブルに保存する。このテーブルを検索することにより、表記が異なり意味属性の集合が一致する単語を指摘できる。意味属性の集合が一致する単語は、ほぼ同義語として扱われるため、適用分野によっては、これらの単語の意味属性の集合が一致しないようにする必要がある。例えば、システム辞書では、「頭痛」にも「腹痛」にも意味分類として意味属性【病気類】が付与されている。しかしながら、病気に関する文を検索対象として扱う場合には、「頭痛」と「腹痛」を区別しなければならない。

(2) 意味属性の各単語への付与方法

以下の 4 つの方法の組合せにより決める。

①登録すべき単語から連想される語を思い浮かべ

(例: 単語「救急車」に対して、「緊急」「病院」「車」などの語を思い浮かべる)、連想した語に付与されている意味属性の中から、登録すべき単語に適した意味属性を選択することによって行う。選択の候補となる意味属性は、連想した語に対し形態素解析を行うことによって取り出せる。また、意味属性に対応する概念や、その意味属性が付与された単語例を表示することによって、ユーザは容易に選択すべき意味属性を決定できる。

②登録すべき単語の意味を文で表し (例: 単語「頭痛」に対し、「頭が痛いこと」と文で表す)、①と同様に、その文中に含まれる単語の意味属性の中から選択する。

③上位概念と下位概念との関係を記述したツリー状のソーラスを辿ることにより、付与すべき意味属性を見つける。例えば、登録すべき単語「フロッピーディスク」に対し、意味属性【装置】の下位概念を辿ることにより、【周辺装置】や【ディスク】の意味属性を見つけて付与する。

④①から③の方法で適切な意味属性が見あたらなかった場合には、③のソーラスに登録された概念

を細分化する新たな概念を設け、新規の意味属性を付与する。

以上により、一般ユーザでも意味属性を付与でき、適用分野に応じた分野辞書を容易に作成できる。

4. 文間の類似度

以下の観点により、検索指示文と検索対象文との間の類似度を算出する。

(1) 検索指示文の意味属性が多く含まれる検索対象文ほどその類似度を高くする。

(2) 検索指示文と同じ表記（活用語の場合にはその語幹）を持つ単語を含む検索対象文は、その類似度をより高くする。

(1) は、検索指示文と検索対象文との意味属性の集合間の一致度として計算される。意味属性は、その種類が膨大になることを防ぐため、多少の意味の違いは無視して付与する。この多少の意味の違いは、(2) により対処する。例えば、「凍傷」と「しもやけ」に対して同じ意味属性を付与し、表記の違いにより、多少の意味の違いを吸収するため(2)を考慮する。ただし、意味属性で与える得点よりも低い得点を与える。

検索指示文と検索対象文 m との類似度 SIM_m を以下のように定義する。

$$SIM_m = S_m / S$$

ただし、

$$S_m = \sum_j (\alpha * (\sum_i US_{ijm}) + \beta * UW_{im})$$

$$S = \sum_j (\alpha * (\sum_i D_{ij}) + \beta * E_i)$$

α, β : 各々意味属性と表記に対して各単語ごとに与える基本点で、 $\alpha > \beta$ である。

$$US_{ijm} = \begin{cases} D_{ij}: IS_{ij} = OS_{km} \text{ となるペアが存在する場合 (意味属性の一致)} \\ 0 : \text{その他} \end{cases}$$

$$UW_{im} = \begin{cases} E_i: IW_i = OW_t \text{ となるペアが存在する場合 (表記の一致)} \\ 0 : \text{その他} \end{cases}$$

$IW_i (i=1, 2, \dots, P)$: 検索指示文中の i 番目の単語 (自立語のみ)

$IS_{ij} (j=1, 2, \dots, Q_i)$: 単語 IW_i に付与された j 番目の意味属性

$OW_{tm} (t=1, 2, \dots, T)$: 検索対象文 m 中の t 番目の単語

$OS_{km} (k=1, 2, \dots, R)$: 検索対象文 m 中の単語に付与された k 番目の意味属性

$$D_{ij} = W_{ij}/Q_i$$

W_{ij} : 意味属性 IS_{ij} の重み。(後で述べる評価実験では、すべての重みを1としているが、抽象度が高い意味属性やテキスト中での出現頻度が高い意味属性には低い値を与えることを想定している.)

E_i : 単語 IW_i の重み (品詞によって異なり、固有名詞は一般名詞より高い値を取る.)

Q_i : 単語 IW_i に付与された意味属性の個数

この類似度は、検索指示文と検索対象文の意味属性および表記が完全に一致すれば1となり、一致する意味属性や表記がなければ0となる。一般には、0と1の間の値を取り、この値が高い文ほど類似性が高い。

分母 S は、検索対象文には依存しないため、相対的な類似度 S_m のみを計算すればよい。

検索指示文の各単語に対して基本点 α を割り当て、さらに、表記が一致した場合には基本点 β を与える。検索指示文の1つの単語に対して複数の意味属性が付与されている場合には、その意味属性の個数で割った値を各意味属性に対して割り当てる。これは次の理由のためである。例えば、検索指示文が「頭痛や目眩」であり、検索対象文として文1「頭痛がする」と文2「目眩がする」があり、「頭痛」には2つの意味属性が、「目眩」には1つの意味属性が付与されていた場合に、単純に意味属性が一致した個数で類似度を算出すると文1のほうが文2より高い類似度となる。これを防ぐために、基本点を配分して、「頭痛」に付与されている意味属性【頭】と【痛み】の各々に得点 $\alpha/2$ を割り当てている。

図4にその例を示している。文1は【頭】【痛み】の意味属性でそれぞれ得点 $\alpha/2$ 、【吐く】という意味属性で得点 α が与えられ文1の類似度は 2α となる。文2は、意味属性による得点 $2 \times \alpha/2$ と表記「頭痛」の一致による得点 $\beta (< \alpha)$ で $\alpha + \beta$ となり、文1のほうが

が高い類似度が与えられる。

5. 類似度算出法

検索対象文の1つ1つを比較しながら各検索対象文の類似度を算出する方法では、汎用計算機上で、大量の文を検索しようとする、膨大な時間がかかり実用的な時間内での検索が困難になる。そこで、意味属性と検索対象文、表記と検索対象文との対応を記述したインデックステーブルを用いて高速な類似度計算を行う。具体的には、テキスト登録時に、あらかじめ検索対象文からインデックステーブルを自動的に作成しておき、このインデックステーブルを用いて計算を行う。

相対的類似度 S_m は以下のように変形される。

$$S_m = \sum_i \sum_j (A_{ij} * \delta_{ijm}) + \sum_i (B_i * \sigma_{im})$$

ただし、

$$A_{ij} = \alpha * W_{ij}/Q_i$$

$$B_i = \beta * E_i$$

$$\delta_{ijm} = \begin{cases} 1: \text{検索指示文の単語 } IW_i \text{ に意味属性 } IS_{ij} \\ \text{が付与され、} IS_{ij} = OS_{im} \text{ なるペアが存在} \\ \text{するとき} \\ 0: \text{その他} \end{cases}$$

$$\sigma_{im} = \begin{cases} 1: IW_i \text{ が自立語で、かつ、} IW_i = OW_{im} \text{ と} \\ \text{なるペアが存在するとき} \\ 0: \text{その他} \end{cases}$$

第1項目は、意味属性に基づく項であり、第2項目は、表記に基づく項である。 A_{ij} は、検索対象文には依存せず、検索指示文の意味属性から計算できる。したがって、 S_m の初期値を0にしておき、 $\delta_{ijm}=1$ となる m についてのみ S_m に A_{ij} を累積することにより第1項が求まる。同様に、 B_i も検索指示文中の単語から計算でき、 $\sigma_{im}=1$ となる m についてのみ S_m に B_i を累積することにより第2項が求まる。任意の意味属性 OS に対して、 $OS = OS_{im}$ となる文番号 m の集合を取り出すことができるインデックステーブルと、任意の単語 OW に対して、 $OW = OW_{im}$ となる文番号 m の集合を取り出すことができるインデックステーブルを作成しておくことにより、累積すべき m を高速に決定できる。このように、類似度計算は、累積値の計算、インデックステーブルの検索、累積値の加算で行え、検索対象文の個数 L に対して $\log_2(L)$ のオーダーである。

検索指示文 頭痛がして、嘔吐もある。
 検索対象文1 頭が痛くて、吐いた。(60点)
 検索対象文2 頭痛がする。(35点)
 ($\alpha = 30, \beta = 5$ とした場合)

検索指示文			文1		文2	
単語	意味属性	得点	意味	表記	意味	表記
頭痛	【頭】	$\frac{1}{2}\alpha$	○	×	○	○
	【痛み】	$\frac{1}{2}\alpha$	○	×	○	○
嘔吐	【吐く】	α	○	×	×	×
合計得点			2α		$\alpha + \beta$	

図4 類似度算出の例

Fig. 4 Example of similarity computation.

6. テキストベース検索システム (TBRIS)

本方式の有効性を確認するため、テキストベース検索システムをパーソナルコンピュータ上で試作した。各テキストは見出し文と本文とで構成され、見出し文を検索対象文、本文を検索後にユーザが参照する部分（詳細内容と呼ぶ）とする。プログラムはC言語で記述され、メモリ 640 kB 内で動作するように、形態素解析処理部と類似度計算部はオーバーレイ構造（処理が必要な時にプログラムをロードする）としている。

図5に本システムの実行イメージを示す（図5では、意味属性に基づく処理のみを示しており、表記に関する処理は、省略している）。

テキスト登録時に、システムは、検索対象文と詳細内容に分割し、検索対象文には文番号を付与し、文番号とテキストとの対応関係を表すテーブルを作成する。検索対象文に対しては、意味属性および表記をインデックスとするインデックステーブルを作成する。

ユーザが、検索指示文を入力すれば、検索対象文の類似度が算出される。検索対象文は、類似度順に並べられて、候補文として表示される。ユーザは候補文の中から、自分の意図に合った文を選択することにより、それに対応する詳細内容を見ることができる。検索対象文を類似度順に表示できるため、ユーザは自分の意図にあったものを容易に見つけることができる。

また、本システムは、検索された結果に対して新たな文または単語列を入力して検索結果を絞り込む機能も有している。この機能は、以下の処理で実現しており、図6のような検索が行える。

- (1) 絞り込みの指示があった時点の検索結果の文番号と類似度を記憶する。
- (2) 絞り込みのための検索指示文により、独立に文番号と類似度を得る。
- (3) (1)と(2)で得られた各々の文番号のうち共通に含まれる文番号についてのみ、両方の類似度を合計する。

- (4) (3)で得られた類似度でソートを行い、類似度順に表示する。

7. 評価実験結果と考察

6章で述べたシステムを用いて評価実験を行った。評価の対象として文献10)および11)を基に作成した「けがや病気の応急手当」に関するテキストを用いた。

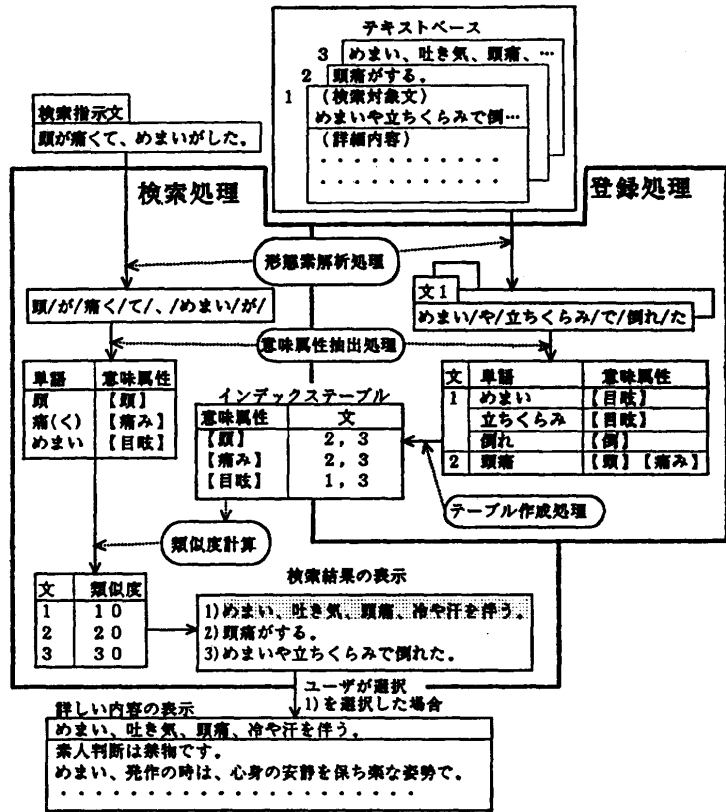


図5 本システムの実行イメージ
Fig. 5 Execution image of the text-base retrieval system.

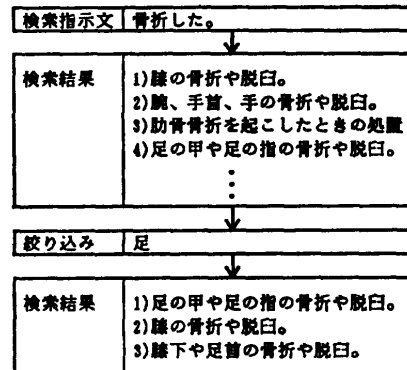


図6 絞り込みの例
Fig. 6 Example of two-step retrieval.

検索対象文数は、約 150 文で、各文は平均 15 文字である。図 7 に検索対象文の例を示す。また、単語辞書として、システム辞書（約 15 万語、約 2,800 種の意味属性）と分野辞書を用いた。分野辞書には、約 350 の単語を登録し、約 150 の意味属性の組合せでその意味を記述している。このうち、システム辞書で用意されている意味属性は約 80 で、残りの 70 が新たに用意した意味属性である。図 8 に評価実験で用いた意味属性の例を示す。実験は、以下の方法により行った。検索対象文を被験者に見せ、検索対象文とは異なる表現で検索指示文を作成してもらい、その文で検索を行い、検索結果の 4 位以内（1 画面で見れる範囲）に、元の検索対象文が入る場合を検索が成功したとみなして、検索成功率を求めた。その結果、検索成功率は約 90% であった（検索指示文数 100 文）。意味属性を用いず、同義語の定義のみで類似度計算を行った場合には、検索成功率は約 80% であり、約 10% の向上分は、意味属性を用いた本方法の効果である。

1. 咯血、吐血した。
2. 頭を強く打った。
3. 目に異物が入った。
4. 乗り物酔いをした。
5. しゃっくりが止まらない。
6. 胸を強く打った。
7. 出血多量の場合の対処法（止血）。
8. 病気や、怪我が重くて寒気を覚える時（保温法）。
9. かすり傷ができた（切り傷、すり傷）。
10. 犬や猫などの動物に咬まれた。
11. 歯が痛い。
12. 植物にかぶれた。
13. 魚の骨がのどに刺さった。

図 7 検索対象文の例

Fig. 7 Example of target sentences.

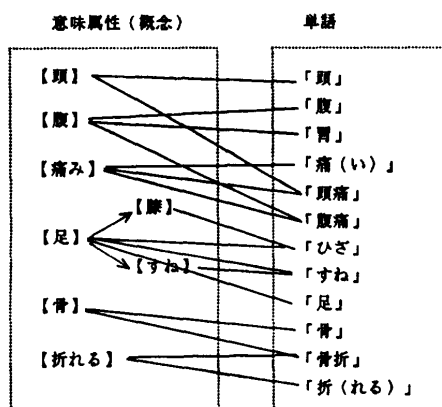


図 8 評価実験で用いた意味属性の例

Fig. 8 Semantic categories for validity examination.

この方法を用いても検索できない場合は、例えば、下痢に関する文を検索することを目的に、「腹の調子がおかしい。」という検索指示文を入力した場合である。このような検索に対しては、本方法のような単語の意味属性を使った方法だけでは対処することができない。これを解決するためには、文全体の意味を考慮することと具体的な事象を表す概念とより抽象的な概念とを結び付ける知識が必要となる。

この実験では、登録されている文を見せて行ったため、その文に影響された表現の検索指示文が多かった。実際にユーザが検索する場合には、どんな文が登録されているか知らないため、もっと多様な表現の文が入力されるであろう。本方式では 1 つの単語に複数の意味属性を付与するため、意味が完全に一致しなくても関連ある単語を含む文を検索でき、ユーザにとっての有効性が高い。

検索時間は、同じテキストベースを用いて、10 単語以内の検索指示文に対して、パーソナルコンピュータ上で 5 秒以内で検索できた。検索時間のうち、形態素解析時間とオーバレイによる時間がほとんどを占めており、類似度計算時間は 1 秒以内であった。

8. おわりに

テキスト情報を検索する方式として、意味属性を基に検索指示文と見出し文との間の類似度を算出して検索する方式を提案した。各見出し文の意味を意味属性の集合として近似的に扱うことにより、「頭痛がする」と「頭が痛い」との関係のような、多様な表現の類義関係を考慮した検索が可能となり、同義語のみを考慮した場合に比べ、検索成功率が 10% 向上でき、約 90% の検索成功率を得ることができた。また、インデックステーブルを用いた高速な類似度計算方法により、パーソナルコンピュータ上で実用時間内の検索が可能とした。

本方式は、文の意味を意味属性の集合として、大胆に近似した方法であり、文の意味を正しく理解しているわけではない。しかし、人間が検索を行う場合、できるだけ簡潔な表現を用いることが多い。このため、検索指示文に含まれる単語の意味が検索のためのキーとなっている場合が多く、実用的には有用な方法だと言える。

文の意味を正しく理解して意味構造を生成し検索することも、適用対象によっては必要となるが、文型が異なる文間の類似性をいかに判定するかが問題となる

ばかりでなく、すべての文ごとに意味構造の照合を行っていたのでは、実用的な時間内での検索は困難となる。これを解決するには、本論文で提案した方法により、厳密な解釈の対象となる文を絞り込んでおき、絞り込んだ文に対してのみ詳細な処理を行う方法が有効であると考えている。

現在、議事録のタイトル文等を本方式に適用することにより、実際の業務で使用する場合の有効性を検証中である。今後は、辞書への登録支援処理を充実させるとともに、文の構文情報も利用する方法を検討していきたい。

謝辞 日頃ご指導いただく NTT 情報通信処理研究所の石野所長、ご討論いただいた、坂井陽一メッセージシステム研究部長、坂間保雄グループリーダー、池原悟主幹研究員ならびに同研究部の皆様方に深謝します。

参 考 文 献

- 1) 木本晴夫：キーワード自動抽出と重要度評価，情報処理学会自然言語処理研究会報告，64-1 (1987)。
- 2) Salton, G. and McGill, M. J.: *The SMART and SIRE Experimental Retrieval Systems, Introduction to Modern Information Retrieval*, pp. 118-156, McGraw-Hill Book Company (1983)。
- 3) 沓沢尚明：日本語自然文による新聞記事検索，情報処理学会情報システム研究会報告，18-1 (1988)。
- 4) 杉山健司，秋山幸司，伊吹 潤，川崎正博，内田祐士：自然言語理解に基づく情報検索システム (IRIS)，情報処理学会自然言語処理研究会報告，58-5 (1986)。
- 5) 秋山幸司：テキスト情報の知的検索における諸問題，情報処理学会データベース・システム研究会報告，64-3, pp. 1-8 (1988)。
- 6) 高松 忍，藤田米春，西田富士夫：係り受け関係に基づく文献の検索，情報処理，Vol. 19, No. 12, pp. 1150-1157 (1978)。
- 7) 長尾 真：言語工学，p. 246，昭晃堂 (1983)。
- 8) 宮崎正弘：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol. 25, No. 6, pp. 970-979 (1984)。
- 9) 池原 悟，宮崎正弘，白井 諭，林 良彦：言語における話者の認識の多段翻訳方式，情報処理学会論文誌，Vol. 28, No. 12, pp. 1269-1279 (1987)。
- 10) 最新現代医学百科，主婦の友社 (1985)。
- 11) 救急と看護百科，緒方出版 (1981)。

(平成 2 年 12 月 3 日受付)

(平成 3 年 6 月 13 日採録)



松尾比呂志 (正会員)

1955 年生。1978 年九州大学電気工学科卒業。1980 年同大学院修士課程修了。同年日本電信電話公社 (現，NTT) 入社。以来，大規模集積回路設計技術，文字認識処理技術，自然言語処理技術の研究に従事。現在，NTT 情報通信処理研究所主任研究員。電子情報通信学会会員。



内野 一 (正会員)

1964 年生。1987 年茨城大学工学部情報工学科卒業。1989 年同大学院工学研究科情報工学専攻修士課程修了。同年日本電信電話 (株) 入社。現在 NTT 情報通信処理研究所において，日英機械翻訳の研究に従事。