類似検索を用いた情報検索システム

美馬 秀樹　　隅田 英一郎　　飯田 仁
ＡＴＲ音声翻訳通信研究所

1．まえがき

近年のインターネットの普及や，今後期待されている電子図書館の実現等により，膨大な量の全文データベースが利用可能となるため，必要な情報を効率的に探し出す情報検索技術が望まれている．データベースに対する検索では，大きく分けて，限定的かつ詳細な情報を要求する特定検索と，「情報散策」[4]といった明確な目的の薄い大まかな情報集合への検索を要求する広域検索[3]が考えられるが，特定検索においては基本的に全データを検索する必要があるため，検索対象が膨大になった場合，検索コストが極めて大きくなる可能性がある．

このような特定検索においては，従来より，検索キーやカテゴリーによる検索指定と AND/OR 等の関係式よりなる検索式を用いるもの，自然言語による検索要求から検索キー等を抽出し検索式を作成するもの[5]，さらに，それら検索キーに対し，シソーラスを用いた拡張検索[9]等が用いられてきたが，検索式においては必要な情報に対する検索キーを明示する必要があるうえ，その作成は初心者には困難である．また，自然言語による要求は非常に柔軟なアプローチと言えるが，入力文よりユーザの検索要求を正確に抽出するのが難しい．さらに，シソーラスを用いた拡張検索では，検索のもれや検索キーのゆれをある程度吸収できるが，キーとシソーラスのみの組み合わせでは逆に意図しない文脈等におけるノイズを多く拾ってしまう可能性がある．

このような問題に対し，用例に基づく枠組み[6][7]における類似検索技術の有用性に対する認識が高まっている．類似検索では，入力文に意味的に類似した用例をテキストデータベースからすばやく検索する．よって，類似検索を文書データ等に対する情報検索に応用することにより，入力文に意味的に類似したテキストに関係する文書が能率的に検索できる．さらに，類似検索には従来技術と比べ次の特徴がある．１）ユーザは検索キーや検索式を考える必要がなく，任意のフレーズや文そのままを入力とすることができる．２）意味距離計算により，入力文と検索対象との表層のばらつきが自動的に吸収されるため，ユーザは基本的に入力語彙等のバリエーションを工夫する必要がない．３）意味距離によりマッチングの度合いの指標が得られる．

本稿では，このような類似検索技術を用いた情報検索システムの実現方法について述べる．また，入力文と意味的に近い検索候補において，ユーザの意図する情報へ効率的に絞り込むため，意味距離計算により得られた意味距離を視覚化し，任意の意味距離にある情報を効果的に提示する枠組みを提案する．本手法では，検索入力文と当該用例を含むテキストとの意味距離を段階的に色分けし，ユーザに視覚的に提示する．ユーザはこの視覚化されたスペースの任意の位置を指し示すことにより，対応する意味的近さにあるテキストに動的にアクセス可能となる．本手法に基づく新聞記事の検索に対するプロトタイプシステムを作成し，実験によりその有効性を確かめた．

以下，２．で，類似検索を用いた情報検索の概要，３．で，類似検索を用いた情報検索システムの構成，４．で，プロトタイプシステムによる検索実験，５．で実験に対する考察，６．でまとめと今後の課題について述べる．

2．類似検索を用いた情報検索

類似検索に基づく枠組みで実現される情報検索では，入力語句と用例との間の意味距離計算という一貫した方法により，入力文に意味的に類似した用例を含むテキスト（以下，検索対象文とする）を検索する．これらの検索対象文に関連される情報（雑誌や新聞記事等の文書データ，写真や音のマルチメディアデータ等；以下，目的情報とする）を抽出することで，入力文に意味的に関連した情報をユーザに提示できる．また，意味距離計算は単純であり，高速な検索処理を実現できる．本章では，このような類似検索を用いた情報検索の概要を述べる．

2.1　用例を使ったインデックスの記述

類似検索に対するインデックスは，検索対象文の用例†を意味的にまとまった単位で抽出したものである．用例に対する原言語表現は，検索の鍵となる部分を持ち，この部分を具体化する語句を基に検索される検索対象文と，検索対象文に関連される目的情報が決定される．そこで，検索対象文の用例と，目的情報を示す検索レコード情報の組を，あらかじめ検索対象文となる実際のテキストデータから抽出し，検索の際のインデックス情報として次のように記述する．

原言語表現　＝＞
　(E_1, [検索レコード1]),
　　:
　(E_n, [検索レコードn])

この記述では，用例が E_i である時，検索レコード i に記述された情報が検索対象文や，関連する目的情報であることを示す．尚，E_i は語句の組である．検索処理において，入力表現の語句とN-ベスト（又は，意味距離のある閾値以下）でマッチする用例を意味距離計算により求め，用例が含まれる検索対象文と，関連される目的情報を検索結果として選択する．つまり，入力表現の語句

† 本稿では，用例という言葉を「原言語表現の検索の鍵となる部分についての具体例」という意味で使用する．また，説明を簡潔にするために，用例の記述を省略することがある．

に最も意味的に近い用例が E_i であれば，検索レコード i の記述を原言語表現に最も関連の大きい目的情報とみなすわけである．

2．2 意味距離計算

入力文の語句と用例との間の意味的な近さを求める方法として，本検索システムでは，隅田らの意味距離計算[7]を採用している．この方法では，まず，シソーラス（類語辞典[2]に準拠）の概念階層における意味概念間の位置関係によって，入力文の単語 i と用例の単語 e の間の意味距離 $d(i, e)$を計算する．意味距離は，0 から 1 までを値域とし，0 に近いほど i と e は意味的に類似していることを示す．意味距離はシソーラスの与え方によって値が変わるが，プロトタイプシステムで現在使用しているシソーラスに基づいた意味距離の例を示す．

d (関西新空港，空港) = 0.00.
d (提唱，諮問) = 0.33333.
d (経済，美馬) = 1.00.

シソーラス中の単語と exact-match しない複合語については，複合語の主とする意味が終端方向に表れる頻度が多いことを考慮して，単語間の左方向最長一致による partial-match を採用している．また，動詞等の活用語については，シソーラスとの整合を取るため終止形に変換してマッチングを行っている．

原言語表現に対する検索の鍵となる部分についての入力文と用例の表現を，それぞれ，I と E とする．I と E の間の意味距離は，I と E を構成する単語間の意味距離を基にして計算する．I および E が，次のように t 個の語句の組として構成されているとする．

$$I = (i_1, .., i_t) \ .$$
$$E = (e_1, .., e_t) \ .$$

I と E の間の意味距離を次のように計算する．

$$d(I, E) = d((i_1, .., i_t), (e_1, .., e_t))$$
$$= \sum_{k=1}^{t} d(i_k, e_k) \cdot w_k$$

w_k は，検索における k 番目の要素の重みを示し，0から1までを値域とする†．

2．3 用例に対する検索対象文の決定処理

本システムでは検索対象文に現れる意味的役割を担う表現のパタンにより用例を分類してインデックス化し，検索入力のパタンに応じて，対応する用例との意味距離計算により検索対象文を絞り込む．

検索対象文に対し，文法属性を表現しない X のような記号（以下，可変部と呼ぶ）と表層語句とにより原言語表現を表わす．可変部は検索の鍵となる部分であり，可変部を具体化する語句により用例を記述する．例えば，次の用例は，検索対象文“運輸省が関西新空港の位置を諮問した”より得られた「X が Y する」，「XのY」のパタンを含む用例記述の例である．

X がYする ＝>
((運輸省，諮問) [文番号 128]),
:
XのY ＝>
((関西新空港，位置) [文番号 128]),
:

(運輸省，諮問) では，「XのY」に対し，X=「運輸省」，Y=「諮問」による「運輸省が諮問する」という用例 を表わす．可変部 X，Y を具体化する入力文の語句の組を I とする．例えば，入力文が「政府が提唱する」の場合，I は(政府$_{[ガ]}$，提唱$_{[スル]}$) となる．

w_k の値を一律に 0.5 とすると，I と用例(運輸省，諮問)の間の意味距離は次のようになる．

d ((政府$_{[ガ]}$，提唱$_{[スル]}$), (運輸省$_{[ガ]}$，諮問$_{[スル]}$))
$= d$ (政府，運輸省) $\cdot w_1 + d$ (提唱，諮問) $\cdot w_2$
$= 0.00002 \times 0.5 + 0.33333 \times 0.5$
$= 0.16667$

また，I と他の用例との距離として以下のようなものも考えられる．

d ((政府$_{[ガ]}$，提唱$_{[スル]}$), (政府$_{[ガ]}$，提案$_{[スル]}$)) =0.16 .
d ((政府$_{[ガ]}$，提唱$_{[スル]}$), (県議会$_{[ガ]}$，聞く)) = 0. 29.
:
d ((政府$_{[ガ]}$，提唱$_{[スル]}$), (町長$_{[ガ]}$，話す)) = 0.67 .
d ((政府$_{[ガ]}$，提唱$_{[スル]}$), (企業$_{[ガ]}$，融資$_{[スル]}$)) = 1.00.

以上のような計算により（運輸省$_{[ガ]}$，諮問$_{[スル]}$）を用例として持つ ”運輸省が関西新空港の位置を諮問した ” 等が意味的に近い検索対象文として検索され，入力「政府が提唱する」の検索結果として，関連づけられている新聞記事や画像データを得ることができる．

2．4 意味距離の総和による検索対象文の決定

入力文の原言語表現と検索対象文に対する用例とのマッチングに複数の組み合わせ方が存在する場合，各検索対象文に対するトータルな意味距離に何らかの基準が必要である．例えば，「政府が新空港の場所を提唱する」は原言語表現「X の Y」と「XがYする」の二つに対する用例とマッチするが，

(1) “運輸省が関西新空港の位置を諮問した。”
(2) “政府が提案している売上税の・・・・・。”
(3) “現場はリーム空港の近くで・・・・・。”

のような検索対象文に対しての適切な意味的近さを定義しなければならない．このような場合，a) それぞれのパタンに対する意味距離の最小のものを検索対象文とのトータルな意味距離とする．b) 各検索対象文において，入力文との照合する用例が多いほどトータルの意味距離を小さくする．c) 文中での用例の現れる位置や，用例の出現する頻度などにより用例との意味距離に重みを設定する．等の処理が考えられる．プロトタイプシステムでは，

† W_k の値は単語の頻度情報に基づいて計算する等が考えられる．ただし，プロトタイプシステムでは，古瀬らの TDMT[7] 同様，W_k の値を一律に1／t としている．

a)と b)に対して選択可能とし，検索実験により評価を行っている．

2．5 意味距離の視覚化による目的情報の選択

実際の新聞等のデータには様々な類義語や言い回しが存在するうえ，ユーザによる解釈の違い等により必ずしも意味距離の最小のものがユーザの意図する情報とは限らない．したがって，ユーザには目的情報がリアルタイムに提示され，かつ対話的に選択できることが望ましい．また，検索候補が大量になった場合にも，ユーザとの対話等により選択操作を適切に支援する必要がある．従来の検索システムでは，得られた候補の見出しやサマリーを列挙するものが中心であり，大量の候補に対してユーザへの適切な支援環境とは言い難い．このような問題に対し，情報の視覚化(information visualization)が重要視されている[3][4]．本システムでは，検索手法として意味距離計算を用いるが，意味距離計算の特徴として，意味距離により入力文と検索対象文との意味的近さの度合いが得られる．よって，意味距離を視覚化することにより，意味的な近さを基準としたユーザへの対話的な選択支援が行える．本報告では，このような意味距離を効果的に視覚化し，任意の意味距離にある情報をリアルタイムに提示する枠組みを提案する．本手法では，入力文と検索対象文との意味距離を用例空間として段階的に色分けし(図2)，ユーザに視覚的に提示する．ユーザはこの色分けしたスペースの任意の位置を指し示すことで，対応する意味的近さにある目的情報に動的にアクセス可能となる．

3．検索システムの構成

図1に，本検索システムの構成を示す．

1）言語処理部では，検索要求である入力文に対し，形態素解析，及び用例における原言語表現とのマッチングにより意味距離計算の対象となるインデックスデータ(用例データの集合）を選択する．

2）意味距離計算では，先に説明した計算法により，入力と検索対象文との意味距離を計算する．

3）候補選択においては，本報告で提案する意味距離の視覚化による情報への動的アクセス法に基づき，ユーザとの対話的なインタフェースを用いて，検索候補よりユーザの意図する情報に絞り込む．

4）検索・表示部では3）により指示された検索対象文に関連する情報をデータベースより検索し，ユーザにリアルタイムに提示する．

4．プロトタイプシステムによる検索実験

上記システム構成に基づき，新聞記事の検索を対象としたプロトタイプシステムを Sparc Station 10，メモリ128MB のワークステーション上に実装し，検索実験を行った．実験には「XのY」，「X {が，は，に，を，で} Yする」の原言語表現に対し，約4 万文の検索対象文(EDR 日本語コーパス[1]を使用)より抽出した約14 万の用例を用意し，記事数約2 万のデータに対して任意のフレーズを入力

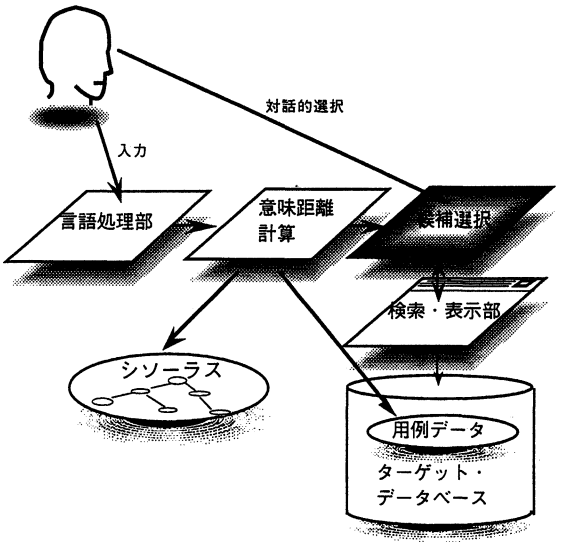

図1．類似検索を用いた検索システムの構成

し，逐次アルゴリズムを用いた意味距離計算により検索を行った．

尚，現在のプロトタイプシステムでは，意味距離計算以降の処理の実験評価を目的とし，実験には言語解析済みのフレーズを入力とした．以下に，検索実験の入力例と検索された用例（波下線），その用例を含む文を示す．

入力：「日本の経済」
検索結果：“ニューヨークの株式暴落をきっかけに、世界同時に大暴落となっています。”
入力：「マスコミの問題」
検索結果：“ ～ エイズをめぐる報道とプライバシーの問題などが次々起こり、～ 。”
入力：「問題が生じる」
検索結果：“韓国労働省の調べによると、 ～ 新たに労使紛争が発生、８４件が解決した。”

5．実験結果に対する考察

5．1 検索結果

類似検索を用いた検索では，漠然とした検索要求に対しても思いつくフレーズを並べるだけで検索が行えるため，比較的安易に検索要求を満たすことができる場合が多い．

しかし，「政府が提唱する」と「政府の提唱」の例のように，原言語表現が異なっていても同義と考えられるものや，“で”格と“に”格のような場合によっては交換可能な格については，現状の固定化された用例に対する意味距離計算のみではカバーできない．したがって，事前にパラフレーズ等のフィルタを通す必要がある．

また，類似検索については，意味距離の定義がシソーラスの構造によって変化し，さらにユーザの意図や状況等によっても動的に変化すると考えられるが，このようなゆれの問題に対して複雑なシソーラスの探索や，文脈状況の解析を行なうことは，計算コストの面を考慮する

と必ずしも有効な手法とは言えない．しかし，シソーラスにおける概念の抽象度を比較的大きくとる（階層を浅く一様に分布させる）こと及び検索候補に対する視覚化の効果により，ある程度このようなゆれの問題を回避できると考えられる．類語辞典[2]を用いた実験においても良好な結果を得た．ただし，検索候補が多くなる場合についてはシソーラスの選択等にさらなる考察の余地があろう．例えば，大井ら[10]は，大規模シソーラスを用い，単語の類似度，関連度，頻度に基づいた検索手法を提案し，実験を行っている．また，文脈を考慮することで，文書中の単語の意味的曖昧性を解消する実験を行い，検索精度が向上することを確認している．

意味距離の視覚化による効果については，意味距離の同じ候補が多くなった場合の距離空間の細分化等に考察の余地がある．しかし，１）意味的近さという比較的直感的な基準による視覚化，２）距離空間上の任意の位置へのマウス・クリックによるダイレクトなインタフェース，３）目的情報へのリアルタイム・アクセスによる対話的操作環境，などを実現することで，検索要求に対し概ね快適で良好な選択が可能となった．検索候補を列挙するのみの選択操作と比較しても，候補の絞り込みは直接的かつ迅速に行えるため，本手法は，候補選択の困難さに対する一解決策になりえると考えられる．

5．2　検索速度

本検索システムでは，検索手法として入力と目的言語表現との意味距離計算を用いているが，検索対象文（目的情報）が大規模になった場合には処理時間が膨大になってしまうという懸念がある．しかし，インデクシングやクラスタリングなどの高速化手法が適用でき，さらに，超並列計算による高速化の効果が報告されている[8]．よって，検索対象が大量になった場合でも常に高速な検索処理を実現することが可能である．

6．まとめと今後の課題

本報告では，類似検索を用いた情報検索システムの実現手法について述べた．類似検索では，意味距離計算を用い，入力に意味的に類似した言語表現を検索する．したがって，入力に意味的に関連された目的情報を検索することが可能になる．

また，検索結果に対する効率的な候補の選択方法として，意味距離を視覚化し，任意の意味距離にある情報を提示する枠組みを提案した．本手法により，ユーザは任意の意味的近さにある情報への動的アクセスが可能となる．

本システムで採用した意味距離計算は単純であり，大規模なデータベースに対する特定検索においても高速な検索が可能である．また，近年の用例に基づく言語翻訳においても入力文に類似した用例の検索手法として有効であるとの報告もなされている[7]．

本手法に基づくプロトタイプシステムを作成し，実験によりその有効性を確かめた．今後の課題としては，ユーザモデルの利用[5]によるユーザの個別要求への対応

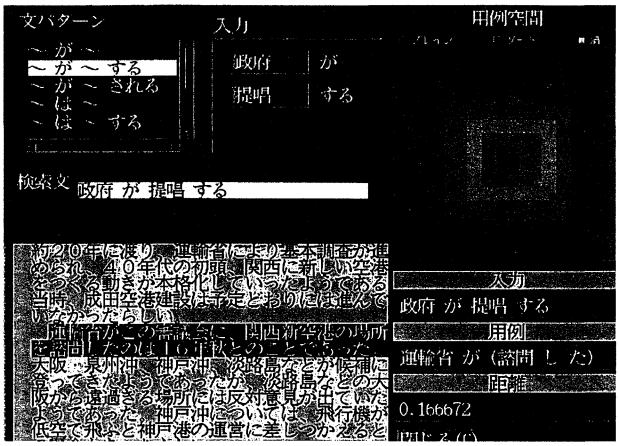

図2．プロトタイプシステムにおける検索例

等が考えられる．また，アプリケーションとして，音声入力との結合や，インターネット上で特に必要とされる多言語入力に対する検索システムの実現等も興味深い課題である．

謝辞

類語新辞典の利用を許可していただいた角川書店に深謝致します．

参考文献

(1)日本電子化辞書研究所: “EDR 電子化辞書仕様説明書”, (1993).
(2)大野 晋, 浜西正人: “類語新辞典”, 角川書店 (1984).
(3)武田浩一, 建石由佳: “情報の可視化と自然言語処理”, 自然言語処理の応用に関するシンポジウム, pp.57-64 (1995).
(4)有田英一, 安井照昌, 津高新一郎: “情報空間の可視化による「情報散策」方式”, Int. Sym. Info-Tech ‘95, Osaka, Japan, pp.20-26 (1995).
(5)加納康男, 岸野文朗: “ユーザモデルを用いた知的文献検索インタフェース”, 信学論(D-I), J74-D-I, 8, pp.567-576 (1991).
(6)隅田英一郎, 堤 豊: “翻訳支援のための類似用例の実用的検索法”, 信学論(D-II), J74-D-II, 10, pp.1437-1447 (1991).
(7)古瀬 蔵, 隅田英一郎, 飯田 仁: “経験的知識を活用する変換主導型機械翻訳”, 情処学論, Vol.35, No.3, pp.414-425 (1994).
(8) Eiichiro Sumita, Kozo Oi, Osamu Furuse, Hitoshi Iida and Tetsuya Higuchi: “Example-Based Machine Translation using Associatve Processors”, Journal of Natural Language Processing, Vol.2, No.3, pp.27-48 (1995).
(9) Chris D. Paice: “A Thesaural Model of Information Retrieval”, Information Processing & Management, Vol.27, No.5, pp.433-447 (1991).
(10)大井耕三, 隅田英一郎, 飯田 仁: “単語間の意味的類似度に基づく文書検索手法”, 言語処理学会第２回年次大会, A5-2 (1996).