ICOT Technical Memorandum: TM-0474

TM-0474

テキスト情報の知的検索における諸問題

秋山幸司

March, 1988

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

テキスト情報の知的検索における諸問題

秋山　幸司

（富士通株式会社）

近年，いわゆるデータベースに対して，従来の手続き的な検索コマンド言語を使わずに，エンドユーザにより良い検索環境を提供する試みとして，自然言語質問文による検索を実現する研究・開発が行われてきた．これらのデータベース自然言語インタフェースの研究で確立してきた技術をテキストベースの検索に活かす目的で，簡単な日本語で入力された検索要求について，テキストベースからその回答となる意味内容を持つテキスト群を検索する実験的システムＩＲＩＳプロトタイプ版を構築し評価した．この結果，分野移行性および検索処理能力の点でデータベース自然言語インタフェースに比べてなんら遜色のないシステムが構築できたものの，分野知識の抽出に膨大な工数が必要であり，何らかの支援環境を用意しない限り実用化は難しいことが判明した．

THE ISSUE OF INTELLIGENT INFORMATION RETRIEVAL TO TEXTBASE

Kohji AKIYAMA

FUJITSU LIMITED

1015, Kami-kodanaka, Nakahara-ku, Kawasaki, 211 Japan

This paper describes the evaluation of an intelligent information retrieval system : IRIS, which retrieves texts related to user's natural language query. The evaluation was done by comparing IRIS to a natural language interface system to databases, because IRIS uses the basic concept and system design of the system. Although IRIS is comparable with the natural language system in terms of transportability and retrieval power, the cost of domain knowledge acquisition in the text-base environment is at least ten times larger than the cost in the database environment. We conclude that it requires much more effort to put IRIS into practical use.

はじめに

知的情報検索システムＩＲＩＳは，簡単な日本語で入力された検索要求について，テキストベースからその回答となる意味内容を持つテキスト群を検索することを目的とする実験的システムである〔杉山　86〕．

近年，何らかの基準に従って加工・抽出された特定のデータの集合であるいわゆるデータベースに対して，従来のＳＱＬのような検索コマンド言語を使わずに，エンドユーザにより良い検索環境を提供する試みとして，自然言語質問文による検索インタフェースを構築する研究・開発が行われてきた〔放送局85〕〔Ishikawa　86〕．最近では実用レベルに達したシステムも幾つか出現するなど，ＡＩ応用製品としてある程度の成功を収めつつある〔Yoshino 87〕．

本稿は，これらのデータベース自然言語インタフェース（以下，ＤＢ自然言語Ｉ／Ｆと略す）の研究で確立してきた技術を活かして，ＩＲＩＳプロトタイプ版を構築し評価した結果を，テキストベースの性質やその利点・欠点などを交えながら，ＤＢ自然言語Ｉ／Ｆと対比させて述べるものである．

1. テキストベースとは

普通に言うところのＤＢ自然言語Ｉ／ＦとＩＲＩＳとの相違点を述べる前に，ＩＲＩＳの検索対象であるテキストベースが通常のＤＢとどのように異なるのかを述べておく必要があろう．

1.1　ＤＢとテキストベース

普通に言うところの狭義の「データベース」（以下，ＤＢと略す）は，対象とする分野の実体の属性や実体間の関係を表している数値や文字列を格納する．たとえ情報源（１次情報）がテキストであっても，そこからデータベースの用途に合わせて情報（２次情報）を抽出・加工して格納するか，もしくは，各テキストについてそれ自体を一つの値と考えてその中身には言及しない．言い換えれば，ＤＢ中のデータはすべてアトム的である．

一方，テキストベースとは，何らかの目的で収集されたデータ（１次情報）がテキスト情報（文の集合）であって，それらが特別に加工されずに１次情報のまま格納されており，テキスト情報の最小単位（通常は文）を高速に検索するために，何らか（通常は各テキストに対して設定されたキーワード）の索引を持つような特殊なデータベースのことである．勿論，ＤＢのようなデータを合わせ持っていてもよい．

そして，肝心なことは，ＤＢＭＳと違ってテキストベース管理システムでは，程度の差はあるが，格納された各テキスト文の中身について関心を持つことである．すなわち，格納された各テキストをアトム的には扱わない．現在に至るまで，テキストベース管理システムでは，計算機能力や言語解析手法などの点で，各テキストの内容の取り扱いを計算機が自動的に行うことができず，システム維持・管理者が手作業で行わなければならない処理も多い．しかし，検索ユーザからシステムを眺めた場合，システム管理者をも含めたテキストベース管理システムでは，格納されたテキストの内容に対して，単なるデータの訂正などにとどまらない何らかの知的処理が介在している．

このような特徴は，情報を有機的に結合し高速に検索するという目的から見た場合，テキストベースの実現を通常のＤＢとは違ったものにする．例えば，ＤＢでは情報を高速に検索するための索引は，いわゆるレコードの特定のフィールドの値をそのまま使った転置ファイルが使われる．一方，テキストベースにおいては，各テキストに対してその内容を表現する値を何らかの方法で生成し，この値をキーとする索引ファイルを構成する．

1.2　テキストベースの実現手法

現時点では，テキストの内容を言い表す値として，キーワードを使うことが多い．これは，人間がテキストを検索する際に，検索したい内容を表す単語を用いることが最も直接的であり簡単であるからであろう．各テキストに対するキーワードの生成方法は，大きく分けて２つある〔亀田　87〕．

一つは，テキストベースが対象とする分野の分類表を用意しておき，各テキストの内容に近い分類表項目の言葉をキーワードとする方法（分類統制方式）である．この方法では，通常，計算機による自動化は無理で，テキストの人手による分類作業が必要となる．また，検索時には分類表を見ながら検索したいテキスト内容を指定するのであるが，検索者のニーズに対して分類が適切であれば良いが，そうでない場合には的確な検索ができない．

もう一つは，辞書などを用いた形態素解析によって各テキスト文を単語に分割し，この中からキーワードとなりうる単語（通常は名詞）を抽出する方法（フリーターム方式）である．英文などでは辞書が無くても簡単に単語を抽出でき，また，日本語文でも最近は比較的容易に形態素解析が行えるため，この方法を採用するテキストベースが多いが，当然，的確なキーワードが付く保証はなく，検索漏れや正解率の低下が問題になる．特に日本語テキストでは，形態素解析に使用する辞書の整備の程度が及ぼす影響が大きい．

このように，ＤＢでは検索の直接の目的がデータそのものにある（勿論，最終目的は，例えばそのデータを使って意志決定をすることかもしれない）が，テキストベースでは，検索の直接の目的はテキスト文の内容にあり，しかも検索に用いる索引がテキストの内容を正確に反映するものとは限らない．このことから，テキストベースに対する自然言語インタフェースは，通常のＤＢ自然言語Ｉ／Ｆには無い機能を必要とする．

2. アプローチ

2.1　ＤＢ自然言語Ｉ／Ｆの機能構成

テキストベースに対する自然言語インタフェースとしてのＩＲＩＳのアプローチの仕方は，ＤＢ自然言語Ｉ／Ｆのそれを基本としている．

例えば，ＫＩＤにおいては，基本的に図２－１に示すような機能構成を採っている〔Ishikawa　86〕．すなわち，日本語質問文によって表されたユーザの検索要求を，形態素解析および構文意味解析

によって理解し，この意味構造からユーザが求めるデータを検索するためのＤＢコマンドを生成する．この解釈の過程で各種の知識（単語辞書，統語的知識，分野知識，ＤＢ写像知識）が利用される．このようにして生成されたＤＢコマンドを実行し，その結果として得られたデータをユーザに提示する．

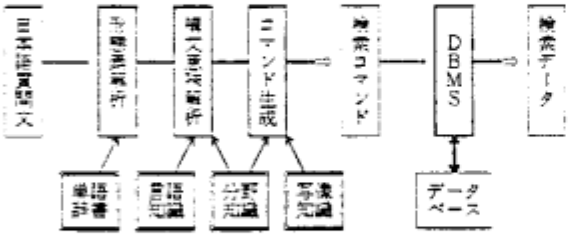

図２－１　ＫＩＤの基本的機能構成

2.2　テキストベースへの応用

ＩＲＩＳでも，図２－１の基本的機能構成に類似の構成を採った．実際，検索要求に対する構文意味解析までは同じ構成である．しかし，その先の処理がＤＢ自然言語Ｉ／Ｆとは異なってくる．

(1)　検索コマンドの生成

まず第一に，ＤＢ自然言語Ｉ／ＦではＤＢに対する検索コマンドを生成するが，ＩＲＩＳではテキストベースに対する検索コマンドを生成しなければならない．ところが，ＤＢの検索コマンド生成がほぼ自明な処理であるのに対し，テキストベースの検索コマンド生成は自明な処理ではない．

ＤＢにおいては，質問文の意味内容を正しく理解してユーザの求めているデータが判れば，あとはＤＢ中の各種ＤＢファイルをどのように関連付けて検索するかということを考えればよい．この種の知識は図２－１にも示されているように，意味構造からＤＢへの写像という形で表現でき，この写像はシステム生成時にＤＢのスキーマなどから容易に作成できる．実際，ＫＩＤにおいては，検索コマンドの生成という機能は，システム構築全体から見ればやや自明な処理として位置づけられている．また，ユーザが質問文中で使う言葉とＤＢ中に格納されているデータとの対応付けも比較的容易であり，システム構築時に定めておいて多少チューニングする程度で，ユーザに不満を抱かせないシステムとすることが可能である．

一方，現在のテキストベース検索システムでは，各テキストを高速に検索するための手段として，1.2項で述べたようなキーワード索引を提供するものがほとんどである．システムがまず行わなければならない処理は，ユーザが質問文中で使った検索内容を指定する言葉を，実際の検索において有効に働くキーワード群に展開することである．例えば，分類統制方式テキストベースでは，ユーザが述べたテキスト内容指定が分類表のどのキーワードに対応するものかを決定しなければならないし，この対応は一意ではなく曖昧さを含むものである．フリータ－ム方式テキストベースでは，ユーザが述べた単語群のそれぞれについて，それに類似な単語を多数集めてこなければ検索漏れが多くなるし，あまり集め過ぎると検索結果にゴミ情報が多くなる．しかも，両方式とも，検索結果をユーザの検索要求に沿う内容・件数とするために，複数のキーワード間の論理的結合関係（AND，OR など）を調整する必要がある．

(2)　検索結果の吟味

第二に，テキストベースにおいては，生成された検索コマンドを実行した結果得られる情報は，ユーザの検索目的を完全に反映したものとは限らない．

ＤＢにおいては，生成された検索コマンドによって表現される内容を忠実に反映した検索結果（データ）が得られるというある種の論理性が期待できる．もし結果がユーザの意図に反していたならば，それは検索コマンドの間違い，従って，通常は質問文の解釈の間違いが原因である．（写像知識の誤りは，質問文とは独立してチェックしたり修正したりできるはずである．）

一方，テキストベースにおいては，ユーザの検索目的はテキスト文によって表される内容である．前述のように，現在のテキストベースの索引の基であるキーワードあるいはその論理的組み合わせは，テキスト内容を正確に反映しうるとは限らない．そこで，もし仮に，質問文の解釈が妥当であり，生成された検索コマンドが最善のものであったとしても，結果として得られるテキスト文によって表される内容がユーザの目的を満たす保証はない．

(3)　解決手法

このような問題点を解決するため，ＩＲＩＳでは，図２－２のような基本的機能構成とする．第一に，ＤＢ自然言語Ｉ／Ｆにおける自明なコマンド生成部のかわりに，人間の検索専門家が検索要求からキーワードを抽出・展開して検索式を生成する過程を模倣するエキスパートシステムを構築し，これによってテキストベースに対するキーワード検索式を生成する．第二に，検索された結果得られるテキスト群をただ表示するのではなく，これらについても構文意味解析を行ってその内容を求め，質問文で表される検索要求の内容と比較照合することによって，ユーザが求めている内容に近いテキストから順に提示する．

このようなアプローチによって，ＩＲＩＳは，通常のＤＢ自然言語Ｉ／Ｆを使う場合と同様の感覚でテキストベースに対する検索を行うことのできる環境を提供しようと試みている．

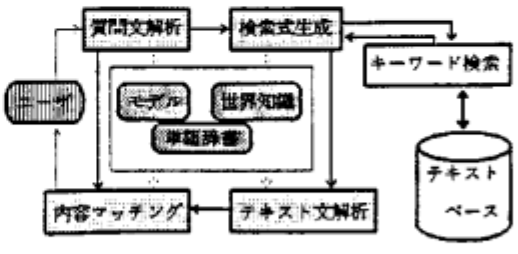

図２－２　ＩＲＩＳの構成概要

3. システム概要

図2－2に示したように，ＩＲＩＳは，大きく分けて，質問文解析部，キーワード検索式生成部，テキスト文解析部，および内容照合部の4つのモジュールから構成される．知識としては，対象分野の実体・概念やそれらの間の関係を記述した分野モデル（統語的知識も含む），分野モデルについて成り立っている事実関係を記述した世界知識，単語辞書，検索専門家の知識，テキスト内容の一致の程度に関する知識，などを用いている．以下の各項では，ＩＲＩＳプロトタイプ版の対象分野，各種モジュール，各種知識などの概要を示す．詳しくは文献を参照されたい．

3.1 対象分野

ＩＲＩＳプロトタイプ版で扱うテキストベースの対象分野は「情報産業界の新聞記事見出し」である．選定に当たっては，テキスト内容の独立性，テキストの用途，およびテキストの対象分野，の3点を考慮した．

テキストの独立性とは，格納の単位となるテキスト群の各内容がそれぞれ互いにどの程度の関連があるかを示すものである．例えば，一冊の書籍に含まれるテキスト群はすべて何らかの意味でお互いの内容が関連性を持っている．逆に新聞記事見出しでは，一部の記事群の間で関連性や因果関係が認められる程度であり，個々の内容の独立性が高い．プロトタイプ版としては，テキスト間の関係のような文脈情報を扱いたくなかったので，なるべく独立性の高いテキスト群を選ぶことになる．

テキストの用途とは，その名のとおりどのような目的でテキストが使われるかを示すものである．例えば文献の抄録と新聞記事とでは用途は違う．いずれにしても，検索対象としてのニーズがあるテキストでなければ面白味がない．

テキストの対象分野とは，テキストの内容が示す話題がどの分野に関わるものであるかを示すものである．ＩＲＩＳは，テキストが対象とする分野の知識を使うことによって処理を進める知識駆動型のシステムであるから，プロトタイプ版として適切な分野とは，ＩＲＩＳの開発者が詳しい知識を有し，かつ，分野知識を容易に収集できる分野のことである．

以上の項目を検討した結果，情報産業界の新聞記事見出し文を選定した．これに基づいて，テキスト文を約八百件，質問文を約百件収集し，分析および知識抽出を行った．テキストベース化はフリーターム方式で行い，各テキストに対するキーワードは，テキストを形態素解析した結果から機能語（助詞・助動詞・区切り記号など）を除いたすべての単語とした．なお，用言はすべてその終止形に変換してキーワードとした．

3.2 知識

(1) 分野モデル

分野モデルとは，テキストベースが対象とする分野における実体や概念，およびそれらの間に存在する関係を記述したものである〔杉山 86〕．収集した質問文およびテキスト文から分野の実体や述語概念を抽出し，筆者らの過去の経験やＤＢ自然言語Ｉ／Ｆでの採用実績から，意味ネットワークとフレーム理論を合わせたようなオブジェクト指向による知識表現法を用いて記述した．また，各概念に対応する単語群の統語上の知識も，この分野モデルの一部として記述してある．

(2) 世界知識

世界知識とは，分野モデルで規定された世界で成り立っている事実，例えば，実体についての属性値データ，命題や事象の間の関連性，ある概念に属する単語群相互の関係，などを記述したものであり，検索式を生成する際のキーワードの選択・選定，および内容照合における一致度の評価にとって必須の知識である〔杉山 86〕．ごく小規模な机上シミュレーションから，必要とされる知識の種類を洗い出し，前述の約八百件のテキスト文から抽出されるだけの知識を整備した．現状では表3－1に示すような知識が存在する．

表3－1　世界知識の種類

知識の種別	例
命題間関連知識	製造⇒販売，参入⇔製造　など
命題間相反知識	参入¬撤退，良い¬悪い　など
地理的知識	国家⊃都市，経済的集団⊃国家　など
製品名称知識	FM16βはパソコンで富士通製　など
組織体名称知識	電算機6社∋〈富士通，日電，…〉など
客体シソーラス	パソコン⊏コンピュータ⇔16ビット　など

(3) 単語辞書

単語辞書の各エントリには，見出し，品詞，形態素情報，構文解析情報，分野モデル上で対応する概念，などが記述されている．前述の約八百件のテキスト文で使われている単語をすべて抽出して整備したところ，約3300個の単語が収集された．その内訳を図3－1に示す．

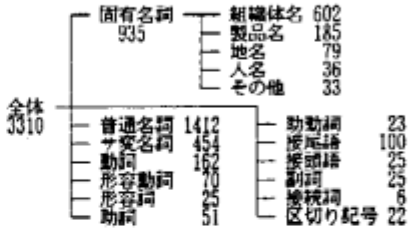

図3－1　ＩＲＩＳの単語辞書の内訳

(4) 検索専門家の知識

検索専門家の知識は，検索処理を一歩進めるためになされる基本動作をモデル化した検索戦術〔Bates 79〕，それをどのように適用するかを示す検索戦略，キーワードとなりうる単語を認定するためのキーワード抽出規則などから構成される〔秋山 87a〕．いずれも，検索専門家の技能を抽象化したものであり，基本的には各戦術がルールセットに対応し，その中の細かな知識はホーン節を用いた if-

then型あるいは手続きとして記述されている．

(5) 照合規則

照合規則は，実体の記述の間の一致度を計算する規則，世界知識を用いた述語間の関連性の推論を行う規則，などから構成される．プロトタイプ版では，あまり複雑な知識は記述されておらず，基本的な一致度算出を行うアルゴリズムと，それが人間の感覚から外れる場合に修正を行うアドホックな規則からなる．

3.3 質問文解析部

質問文解析部は，文節数最小法に基づく形態素解析を行った後で，各自立語に対応する分野モデル上の概念をインスタンス化し，分野モデルに記述された意味的制約や統語的知識を用いてこれらの間の関係付けを行うことにより，質問文に対応する意味構造を生成する．解析部の概念構成は基本的にはシフトリデュースパーザであるが，幾つかの簡略化を行っている．また，暗黙に成り立つ関係を顕在化する機構を持っている〔伊吹 87a〕．

3.4 テキスト文解析部

基本機構は質問文解析部と全く同じであるが，対象テキストである見出し文の特徴（助詞の省略，体言止め，連用中止，サ変動詞の語尾省略）などに対処するためのヒューリスティクスを付加してある．また，一つの記事に対応する見出しは，通常，主見出しや副見出しなど，複数の構成要素を持つので，これらの構成要素間にまたがる関係付けを行う機構を持つ〔伊吹 87a〕．

3.5 検索式生成部

検索式生成部は，検索専門家の知識と世界知識を用いて，質問文の内容を示す意味構造から，その内容に合致すると思われるテキスト文が持っているはずのキーワード（あるいはその論理的結合）を推論して検索式を生成するエキスパートシステムである．入力された意味構造を大域的作業領域として使う所謂黒板モデルを採用しており，各戦術やキーワード抽出規則をタスクとするアジェンダ制御となっている〔秋山 87a〕．

3.6 内容照合部

内容照合部は，照合を行う質問文およびテキスト文の各意味構造を，述語概念の部分を述語論理形式で表現し，その引数が実体部分となるように変換する．その上で，照合規則を用いて，まず質問文とテキスト文の対応する述語間の照合を行い，次に，その述語の対応する引数値となっている実体間の照合を行って，最終的な一致度を計算する〔杉山 87〕．

4. 評価

ＤＢ自然言語Ｉ／Ｆの評価は難しいが，一般には，処理能力，分野移行性（システムの枠組みの独立性），システム構築の容易度，といった観点から行われているようである．そこで，ＩＲＩＳプロトタイプ版を上記のような観点から評価してみた．

4.1 処理能力

処理能力の尺度としては，システムの処理速度，入力文の解析能力，検索能力などが考えられる．

(1) 処理速度

ＩＲＩＳプロトタイプ版は，ＤＢ自然言語Ｉ／Ｆを原点としてそれをテキストベース向きに改良した枠組みが，ＤＢ自然言語Ｉ／Ｆと同じような成功を収めるかどうかを評価するために試作されたので，率直に言えば，処理速度は十分ではない．

最も遅いのは検索式生成部であるが，エキスパートシステムとしてまだ未熟であることが原因である．例えば，ユーザが入力した言葉を基にキーワードを求める処理は一種の探索であるが，これが最良探索ではなくほとんど横型全探索に近い．また，形態素解析部についても，辞書検索が最適化されていないことや実現技術が未熟であることなどから，一般的な塩の５倍程度遅い．

一方，構文意味解析については，パーザの動作を決定的にしたことなどから十分な処理速度が得られている．また，内容照合処理については，照合アルゴリズムが幼稚であることや手続き的に実現したことなどから妥当な速度となっている．

(2) 文解析能力

質問文例を調査した結果，質問文解析部では，受け付ける文の構造を"高々１つの埋め込み文を持つ単文"としている．一般に新聞記事見出しという対象分野では，単一の事象，すなわち，「～が…を…した」という文で言い表されるような内容が幾つか集まって，一つの記事が構成されている．従って，検索要求はこのような事象を指定するための「～は…を…したか？」とか「…を…した～は？」という質問文になりやすいのである．

プロトタイプ版では，質問文 100文に対して，79文の解析が成功し，９文の解析は不完全（しかし誤りではない）で，残りは失敗した（解析結果の誤り，受理不能など）．不完全あるいは誤りの原因として，解析規則の適用の不都合，分野モデルの情報不足による等位接続解析の失敗，代名詞の照応参照の未解決などがある〔伊吹 87b〕．

一方，テキスト文解析部では，一つの記事に対応する幾つかの見出しが，一つの事象の一部分だけを述べていることが多く，しかも，助詞の省略が多いため，見出し文の特徴を利用したヒューリスティクスを用いて解析能力を維持している．見出し間の関係については，構文的には何ら情報が無いことが普通であるので，一つの見出しの解析の後で，意味的関係およびその制約や優先順序だけをたよりに関係付けを行っている．

この結果，プロトタイプ版では，テキスト文約八百件に対して，約60%の解析に成功し，約15%の解析は不完全で，残りは失敗した．不完全あるいは誤りの原因として，質問文解析部における原因の他に，見出し間にまたがる関係付けの誤りを挙げることができる〔伊吹 87b〕．

(3) 検索能力

前述のように，ＤＢ自然言語Ｉ／Ｆの場合には，検索能力は，ＤＢ本来の検索能力が十分であって写像知識にも誤りがなければ，質問文の解釈の精度で決まってくる．従って，普通に受理できた質問文については，人間がＤＢ検索コマンドを生成した場合と同等の検

索結果を保証することが可能である．

一方，ＩＲＩＳにおいては，質問文を正しく解釈するだけではなく，その解釈結果から適切なキーワード検索式を生成する能力，および，検索結果の内容を吟味して質問文が要求する内容に合致したテキストから表示する能力が必要となる．キーワード検索式の生成はＤＢコマンドの生成よりも高度な技能であり，実際，世の中にはキーワード検索の専門家が存在する〔Guide 83〕．

さて，キーワード検索システムの検索能力を評価する場合には，通常，検索結果に含まれるべき正解テキストをどれだけ漏らさずに検索できるかを示す再現率，および，検索結果に含まれる正解テキストが検索結果全体に占める割合を示す適合率，の２つの指標が使われる〔伊藤　85〕．

検索式生成部の評価を行うに当たっては，まず，各質問文に対してその回答となるべき正解テキストを約八百件のテキストベースから人手で抽出した．これをＩＲＩＳの検索式生成部によって検索されたテキスト群と比較することにより，検索式生成部の再現率，適合率を計算することができる．この結果，解析に成功した79個の質問文に対する平均再現率が約６割，平均適合率が約５割であった．人間の検索専門家の場合，フリータ－ム方式のテキストベースに対するこれらの値は，再現率，適合率共に約６割というところが普通であるから，検索結果にゴミがやや多いものの，専門家と同等の検索能力を有すると言える．

内容照合部は，検索結果を順序付けするものであるから，単なる再現率および適合率ではその評価にはならないので，順序集合に関する正規化再現率および正規化適合率〔伊藤　86〕を求めた．内容照合方式はやや稚拙であるので，うまくツボにはまるようなテキスト群については非常に良い順序付けが行われるが，それを外れると人間の感覚に合わない．このため大規模な評価に耐えるだけの実力は無いと判断し，小規模な単体評価のみを行った状態である．百件弱のテキストベースに対する18の質問文による評価結果では，正規化再現率および正規化適合率ともに約８割の値を得ている．

4.2　分野移行性

分野移行性については，新聞記事見出しというテキスト文特徴は変えず，テキスト文の内容の対象分野を情報産業界から国家間外交や軍事問題を中心とする国際政治に変更して評価を行った．これに関しては〔秋山 87b〕にて報告済みであるので，ここではその概要のみを述べる．

テキスト文 210件およびその内容に対する質問文 130件を収集し，これを基に知識を抽出した．3.2項で述べた主な知識のうち，分野モデルと世界知識は当然のことながら全面的に書き直した．単語辞書については，不足している単語を約六百ほど追加し，不要な単語を削除した結果，約千五百語を持つ辞書となった．検索専門家の知識については，キーワード抽出規則の書き換え，および，検索戦術の適用条件の分野依存部分の書き換えを行った．照合規則については特に変更した部分はない．

一方，システムの枠組み部分については，内容照合部を除いて全く変更する必要はなかった．内容照合部では意味構造を内部形式に変換しているため，分野モデルの書き換えに伴ってこの内部形式を変更しなければならず，分野移行性が損なわれている．

このように，ＩＲＩＳにおける分野移行性（あるいは枠組みの分野独立性）はある程度達成されているということができる．

4.3　構築の容易度

ここで述べる“構築”とは，システムの枠組み自体の構築ではなく，その枠組みに対して，対象分野の知識を組み込み，デバッグやチューニングを行って，所定の機能を果たす応用システムを完成することである．ＤＢ自然言語Ｉ／Ｆにおける標準的な構築の工程と，ＩＲＩＳにおける工程とを定性的に比較すると，おおよそ次のようになる．

(1)　ＤＢ自然言語Ｉ／Ｆの場合

ここではＫＩＤを例にして考える．というのも，筆者がＫＩＤの開発者と直接会話をできる立場にいたことと，ＫＩＤの実用化試験版（開発者以外の人間が応用システムを構築することができるように，支援環境やマニュアルを整備したもの．ただし製品ではない）が，先進ユーザのもとに試験的に導入されて評価を受けたことがあるからである．

ＫＩＤにおける標準的な応用システム開発工程は次のようなものである．

①　質問文例（ 200文程度）の収集と分析・整理
②　対象とするＤＢのスキーマ情報の整理
③　質問文で使われている単語の抽出
④　スキーマを基にした分野モデルの作成・整理
⑤　分野モデルからＤＢスキーマへの写像の作成
⑥　③で抽出された単語や，分野モデル作成およびスキーマ情報の整理で使われた単語を，辞書へ登録する．
⑦　システムの動作試験および調整

質問文例はなくべく広範囲のユーザから集めた偏りの少ないデータであることを期待している．単語抽出においては，助詞や助動詞などの機能語については既に辞書登録済みであり，⑥のフェーズが完了した時点での標準的な単語数は 500語程度である．スキーマを基にした分野モデルの作成・整理には，ある程度標準的な手順が確立されており，生成される概念クラスの数はＤＢの大きさにもよるが，実用ＤＢで 100～ 150個程度である．なお，文例 200文程度という数字は経験的な基準であり，少なすぎると分野モデルが不完全になって後の調整が困難になるが，多すぎると分析・整理が困難になる．筆者の個人的な意見としては，10の２乗のオーダを超えたものは純粋な手作業では解決が難しいという感触を抱いている．

応用システムの利用者との対話がスムーズに行われる場合，このような処理にかかる標準的工数〔石川　85〕は，表４－１に示すように全体で約20人日，すなわち（土，日は休みとして）約１人月で応用システムが動き出す．この１人月は長いと感じられるかもしれないが，企業の採算ベースには合うと思われる．

(2)　ＩＲＩＳの場合

ＩＲＩＳにおける開発工程はシステムの枠組みの開発およびモジュールに固有の分野独立性の高い知識（検索専門家の知識，内容照合規則，単語辞書における助詞・助動詞などの機能語）の記述や整備を除くと，概ね次のようなステップであった．

① 質問文例（約百文）およびテキスト文例（数百件）の収集と分析・整理

② 質問文とテキスト文で使われている単語の抽出

③ 主にテキスト文を基にした分野モデル作成・整理

④ テキスト文と専門知識を基にした世界知識の収集

⑤ ②で抽出された単語を，それに対応する分野モデルでの概念と共に辞書へ登録する．

⑥ 典型的質問文的文程度およびテキスト文百件程度の小規模パイロットモデルでの動作試験および調整

⑦ 本格的なシステムの動作試験および調整

ＩＲＩＳでは，検索対象がテキストベースであるので，ＤＢのようなスキーマ情報は無い．従って，対象とするテキストを収集・分析して分野モデルを作成し世界知識を収集することになるが〔杉山86〕，この作成手順は明確ではない．質問文は比較的単純であるが，テキスト文はありとあらゆる文例があり，数百件集めても飽和した感じはしない．また，単語も豊富であり，ＤＢと違って固有名詞がＤＢ中にデータとして格納されているようなことは期待できないため，単語辞書はＤＢ自然言語Ｉ／Ｆに比べて一桁大きくなる．また，システムのテスト・調整は，文解析だけでなく検索式生成部や内容照合部に対しても行う必要がある．（この評価では，検索式生成部や内容照合部の分野独立性の高い知識の調整の手間は考慮していない．）

ＩＲＩＳを応用システムに仕立てる工数は，情報産業界と国家間外交問題を対象とした場合のそれぞれを平均すると表４－２のようになる．この４～６人月という数字が実用ベースであるかどうかは疑問である．

表４－１　ＫＩＤにおける応用システム典型的構築工数

項番	内　　容	工　数
①	質問文例の収集と分析・整理	3人日
②	対象とするＤＢのスキーマ情報の整理	3
③	質問文で使われている単語の抽出	2
④	スキーマを基にした分野モデルの作成・整理	5
⑤	分野モデルからＤＢスキーマへの写像の作成	3
⑥	単語辞書への登録	2
⑦	システムの動作試験および調整	5

表４－２　ＩＲＩＳにおける応用システム典型的構築工数

項番	内　　容	工　数
①	質問文例とテキスト文例の収集と分析・整理	0.5人月
②	質問／テキスト文で使われている単語の抽出	0.5
③	テキストを基にした分野モデルの作成・整理	1～2
④	テキストなどからの世界知識の収集	0.5
⑤	単語辞書への単語と概念の対の登録	0.5
⑥	小規模パイロットシステムでの調整	0.5
⑦	本格的システムでの動作試験および調整	1～2

(3) 比較

以上，(1)と(2)を比較して明らかに判る事実は，分野知識（分野モデル・世界知識・単語辞書）の作成および本格的なシステムの動作試験や調整にかかる工数が十倍以上違うということである．しかも，これには，検索式生成部や内容照合部用の知識の調整工数は含まれていないのである．このことは，ＤＢ自然言語Ｉ／Ｆとテキストベース自然言語Ｉ／Ｆとでは，応用システム構築における困難さが大きく異なることを示している．

次に，仕立てられた応用システムの質を比較してみよう．ＤＢ自然言語Ｉ／Ｆでは，ユーザの検索要求を満足させる質問文を受け付けるシステムが一度構築されてしまえば，ＤＢのデータの量がいくら増加しても本質的に検索品質は低下しない．ところが，テキストベースでは，テキスト件数を増加させた場合に検索品質（例えば再現率や適合率）が低下し得るのである．例えば，年代の移り変わりに伴って，使用される用語の変化やテキスト文の文体の変化は容易に起こり得るが，これをシステムの分野モデル，世界知識，および単語辞書（極端な場合は検索専門家の知識や照合規則）に反映しない限り，検索できないテキストが発生することは自明である．このことは，テキストベース自然言語Ｉ／Ｆの維持管理がＤＢのそれに比べて困難であることを示している．

さらに，当初の質問文例とは全く異なる検索要求が発生した場合の対処を考えてみよう．単なる統語的知識の欠如ならば，どちらも解析規則（文法）や分野モデルの統語的情報（あるいは言語的知識）を修正すれば良い．しかし，検索要求が異なるという場合，分野モデルの修正が必要である．ＤＢにおいては，ＤＢに格納されているデータによって表される概念は，ＤＢのスキーマ以上のものではありえないことが多い．特に関係モデルを基礎とするＤＢではそうである．従って，ＤＢの概念スキーマを基にして分野モデルを作成しておけば，修正は既存の概念間の関係を追加・変更する程度で済むと思われる．然るに，テキストベースでは概念そのものが欠けている場合が少なくない．従って，概念の追加を伴う大規模な変更になる可能性が高い．分野モデルの作成工数の大きさを考えると，テキストベース自然言語Ｉ／Ｆの拡張性はＤＢのそれに比べて困難になろう．

5. 結論およびＩＲＩＳの今後

以上，ＤＢ自然言語Ｉ／Ｆの基本機能構成をテキストベース自然言語Ｉ／Ｆに適用した結果を述べた．結論を端的にまとめると次のようになる．

① 分野移行性は，現行のＤＢ自然言語Ｉ／Ｆに対して遜色のない程度を達成できる．

② 処理能力についても，既存のＡＩ手法を的確に応用することにより，実用レベルを達成できそうである．

③ 応用システムを構築する際の知識の作成・維持に必要な工数を劇的に縮める支援環境を研究・開発しない限り，ＤＢ自然言語Ｉ／Ｆのような実用化は困難である．

DBは，元来，用途が明確で定型的なデータ（数値・文字列）を論理的なモデル（スキーマ）の基で維持管理する目的で発展してきた．すなわち，DBに格納されているデータはセマンティクスが明確である．一方，テキストベースは，用途が不明確で非定型的なためにテキストからの抽出が困難な情報を，1次情報をそのまま格納することによって何とかユーザに提供するために生まれたといってもよい．従って，テキストベースのセマンティクスは，自然言語のセマンティクスをそのまま踏襲しており，DBのような狭い分野に閉じるということがない．本稿で述べた結論は，DBとテキストベースの間にあるこのような本質的相違を端的に反映していると言える．

現在，IRISは，数十～数百件のテキストベースを対象としたプロトタイプ版の評価をほぼ終了し，数百～数千件のテキストベースを対象とした拡張版を開発中である．IRISプロトタイプ版，および，既存のDB自然言語I／Fでは，質問文あるいはテキスト文の一つ一つに対応した調整が（何とか）可能であった．しかし，数千件のテキストに対しては，このような個別対応による調整によって検索品質を維持することは困難である．

拡張版では，プロトタイプ版で提供できた検索品質を，テキスト文の個別対応に頼らずに維持するために必要な分野モデルの詳細度の検証，および，検索品質のチェックにかかる工数の削減に必要なデバッグ支援環境の構築を目的としている．この目的が達成された場合，個別対応による調整を最小限にするための分野モデルの詳細度が明確になると期待できる．これを基に，分野モデル作成基準が（DB自然言語I／Fのように）定式化できれば，分野モデル作成工数を減らすことが可能になり，大規模テキストベースに対して，ある程度の品質を保証した検索を提供できるテキストベース自然言語インタフェースを実用化することができるであろう．

謝辞　本研究は第5世代コンピュータプロジェクトの一環として行われ，ICOT第2研究室の内田，吉川の両氏を始めとする方々に御支援頂きました．また，KID開発者の一員である吉野，牧之内の両氏からは有益な助言を頂きました．ここに印して感謝致します．

【参考文献】

〔Bates 79〕 Bates, M. J. "Information Search Tactics", Journal of the American Society for Information Science, pp. 205-214. 1979.

〔Guida 83〕 Guida, G. and Tasso, C. "IR-NLI: An Expert Natural Language Interface to Online Databases", Proc. of ACL'83, pp. 31-38. 1983.

〔牧之内85〕 牧之内，泉田 「知識に基づいた自然言語インタフェースKIDの開発」，情報処理学会第30回全国大会論文集，pp. 1421-2. 1985.

〔石川 85〕 石川，鳴海，甲田，神田 「自然言語インタフェースKIDの評価」，情報処理学会第30回全国大会論文集，pp. 1429-30. 1985.

〔Ishikawa 86〕 Ishikawa, H. et al. "A Knowledge-based Approach To Design A Portable Natural Language Interface To Database Systems", Proc. IEEE COMPDEC Conf., pp. 134-143. 1986

〔伊藤 86〕 伊藤 「情報検索」，昭晃堂，ソフトウェア講座19. 1986.

〔杉山 86〕 杉山，秋山，伊吹，川崎，内田 「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研究会資料58-8. pp. 1-8. 1986.

〔伊吹 87a〕 伊吹，杉山，鈴木，玉田，川崎 「自然言語インタフェースとしてのIRIS」，情報処理学会第34回全国大会論文集，pp. 1325-6. 1987.

〔伊吹 87b〕 伊吹，杉山，鈴木，玉田，川崎 「内容検索のための自然言語パーザ」，日本ソフトウェア科学会第4回全国大会論文集，pp. 347-350. 1987.

〔杉山 87〕 杉山，秋山，川崎 「内容検索システムとしてのIRIS」，情報処理学会第34回全国大会論文集，pp. 1329-1330. 1987.

〔秋山 87a〕 秋山，杉山 「従来型情報検索システムへの知的インタフェースとしてのIRIS」，情報処理学会第34回全国大会論文集，pp. 1327-8. 1987.

〔秋山 87b〕 秋山，杉山，伊藤，小野寺 「知的情報検索システムIRISの分野移行性の評価」，情報処理学会第35回全国大会論文集，pp. 1429-1430. 1987.

〔亀田 87〕 亀田，藤崎 「テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム」，情報処理学会論文誌，Vol. 28. No. 11. pp. 1103-11. Nov. 1987

〔Yoshino 87〕 Yoshino, T. et al "A Practical Natural Language Interface To Databases", Proc. International Conf. on Artifitial Intelligence (in 2nd World Basque Congress), pp. 173-183. Sep. 1987.