

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

秋 山 幸 司, 杉 山 健 司

(富士通株式会社)

1. はじめに

日本語質問文を理解し、回答となる内容を持つテキスト群をテキストベースから検索する知的情報検索実験システム IRIS [杉山(1986)] では、テキストベースの検索方式として、従来技術であるキーワード検索を採用している。これは、(1)巨大なテキストベースに対して単なるキーワードよりも知的な構造を持つ索引を構築するコストは膨大であること、(2)既存のテキストベースの多くがキーワード検索であり、これら既存の資源を活かしたいこと、および、(3)検索の専門家は、キーワード検索レベルでも十分に知的な検索を実行していること、などによるものである。

対象分野のモデルを用いて日本語質問文を解析して得られた意味構造 [杉山(1986)] と従来型キーワードとの間には大きな溝が存在する。そこで IRIS では、キーワード検索の専門家が持つ技能を模倣するエキスパートシステム AREX (Advanced Retrieval Expert) をサブシステムとして持つことにより、質問の意味構造からキーワード検索式を生成して回答の候補となるテキスト群を検索する。本稿では、キーワード検索テキストベースへの知的アクセスの実現という立場から、AREX の概要および評価を述べる。

2. システム構成

AREX の構成を図1に示す。動作の基本は、大域的作業領域を中心とし、種々の規則がそのデータを更新していくことにより検索式が生成されるいわゆる黒板モデルである。しかし、後述の理由から

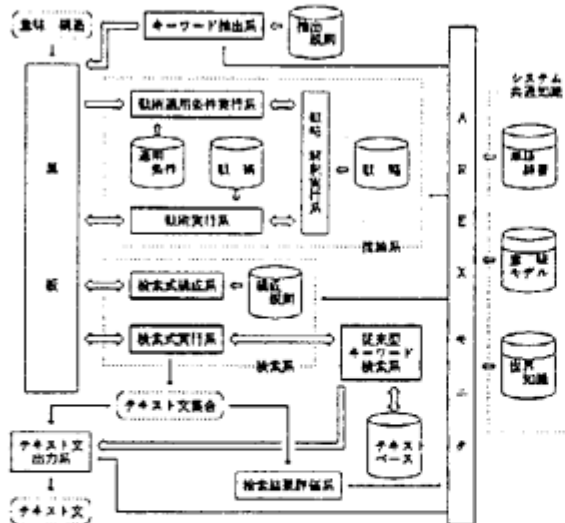


図1 AREXの構成

IRIS as an Intelligent Interface to
Traditional Information Retrieval Systems

Kohji AKIYAMA, Kenji SUGIYAMA

Fujitsu Limited

AREX は一様な形式の規則群からなるプロダクションシステムではなく、各種知識とその解釈実行系が黒板を共有しながらある種の同期をとりつつ並行動作する複合的システムであり、内部的にはタスクを実行単位とするアジェンダ制御となっている。

AREX の入力には IRIS の質問文解析部によって生成された質問文意味構造である。IRIS における対象分野のモデルはオブジェクト指向知識表現を基礎としており、実現に使用したマシン P S I の言語 E S P [近山(1984)] のクラスオブジェクトとして記述される。意味構造はこのモデルのインスタンスオブジェクトのネットワークで表現される [杉山(1986)]。黒板を構成する要素として、モデルのインスタンスがそのまま使われる。黒板上の意味構造を変形したりその部分構造に対応するキーワードを抽出・展開する規則を適用することがいわゆるエキスパートシステムにおける推論に相当する。

3. 知識

AREX が用いる知識としては、IRIS の各部に共通の (大域的な) 対象分野のモデル・世界知識・単語辞書 [杉山(1986)] の他に、(AREX 専用の) 検索技能を表現する初期キーワード抽出規則・検索戦略・検索戦略といった知識がある。

初期キーワード抽出規則は、実行に先立って、意味構造中の各インスタンスに対応する単語群の中からキーワードとなりうる単語群を抽出するための分野依存知識である。次に述べる検索戦略の一種とも考えられるが、分野依存性が高いので独立に記述している。

検索戦略とは、検索処理を一步進めるためになされる基本動作のことであり、人間の情報検索活動の観察から、シソーラスを引く用語戦略、キーワード間の論理結合関係を操作する検索式戦略、検索目的のチェックやコストを評価する監視戦略など種々の戦略が整理されている [Bates (1979)]。AREX においては図2に示すような戦略を定式化した。並列戦略は、関連語などを元の単語と or 結合して厳密さを減少させる戦略の総称である。縮約戦略は、and 結合されているキーワード群のどれかをなんらかの規準に従って削って厳密さを減少させる戦略の総称である。一般に戦略は、それを黒板の部分構造に適用する際の制約条件を持つが、分野独立な条件 (例: “2つ以上のキーワードが and 結合されていること”) のみを戦略自体に記述し、分野に依存する条件 (例: “「規格」クラスに属するキーワードには適用不可”) は戦略適用条件として独立に記述して、分野移行性を高めている。

検索戦略とは、検索目的を達成するために検索戦略を何時どのような条件の下でどのように使用するかを定めた計画である。従って、検索目的や検索されるテキストの性質などから最適な戦略を動的に生成することが望ましい。しかし、検索専門家が行うような意志決定を含む高度な計画立案処理は現段階では実現困難であるため、



図2 AREXが用いる戦術

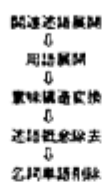


図3 AREXの戦略

AREX においては、図3に示すように検索式の厳密さを徐々に緩めるように戦術を配置した固定的戦略を用い、所望の件数に達した所で停止する単純なものとなっている。

このように AREX における規則はプロダクションシステムのような平坦で一律な形式ではないため、個々の戦術・適用条件・抽出規則などはタスクとして記述され、戦略解釈実行系がこれらのタスクを適当に活性化することで黒板が更新されて検索式が変化する。

4. 適用例

例えば、「富士通と提携した会社？」という質問文に対する意味構造は図4のようになる。対象分野のモデルは述語概念と名詞概念に大別され、個々の述語概念はそれぞれ名詞概念との意味関係を規定する固有の「格フレーム」を持つ[杉山(1986)]。图中的「協力」は、組織体が他の組織体と協力する行動を表す述語概念であり、「組織体」は企業等を表す名詞概念である。

まず、初期キーワード抽出規則により、「協力」という述語概念を表すすべての単語が辞書より検索されて or 結合され、また「組織体」に対応する単語「富士通」が固有名詞なのでキーワードとされる。等位接続以外は、述語概念とその引数である名詞概念の持つキーワードの論理積が全体としてのキーワード検索式になり、図4の下段に示されるような検索式が生成される。

この検索式で十分な記事件数が得られない場合、検索戦略(図3)に基づいて関連述語展開戦術が適用される。この戦術は、世界知識の一種である命題間関係知識[杉山(1986)]を用いて、「協力」を「(技術等を)与える」や「共同」という概念を持つなんらかの述語に変換する。この結果、図5(a)(b)のような2つの意味構造が黒板に追加され、図4と同様にしてそれぞれ検索式が生成される。図4と図5から得た3つの検索式の論理和が全体の検索式となる。

この例では、戦略中のさらなる戦術はすべて適用不可能となるので、まだ十分な記事件数が得られなくても検索はここで終了するが、「16ビットパソコン」などの「製品」を含む意味構造では、さらに、用語展開戦術によるシソーラスを用いた関連語の追加や、名詞単語削除戦術による「16ビットパソコン⇔16ビット」といった不要語の削除などが行われ、記事件数が次第に増加していく。

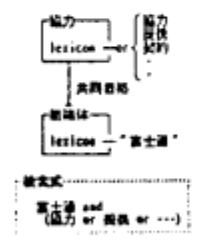


図4 「富士通と提携した会社？」に対する意味構造

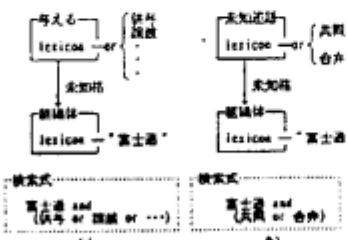


図5 図4を関連述語展開して得られる意味構造

5. 評価

知識駆動型システムでは、評価結果は知識の整備の程度に依存するので、次のような評価を行った。まず、現在の適用分野である情報産業界の新聞記事見出しの中から典型的な見出し72件を抽出して作成した小規模テキストベースに対し、IRIS に対する典型的検索要求を表す質問文70文についてはほぼ完全な検索結果が得られるように AREX の使う知識を整備した。つぎに、日本情報産業新聞(週刊)、週刊電波コンピュータ、および日経産業新聞の情報産業界関連記事、を約1.5ヶ月分収集して838件の記事見出しテキストベースを作成し、質問文70文に対して前述の72件のテキストベースのかわりにこの838件のテキストベースを検索した結果を取得した。一方、情報産業界で働く3人の被験者が、各質問文に対する回答となると思われる見出し群を838件のテキストベースから手作業で抽出した。以上の結果から、各質問文に対して AREX が検索した結果の再現率および適合率が算出される。

結果は、再現率の平均値が59%、適合率の平均値が48%であった。27の質問文(2/5)が再現率80%以上であるが、14の質問文(1/5)が再現率20%以下である。また、15の質問文(1/5)が結果にゴミを20件以上含むが、これらの質問文を除くと適合率平均値は59%になる。

5. 今後の課題

検索結果を悪化させている要因の詳しい解析はこれからであるが、概して、(1)組織体や製品の名称が未登録であることによる単語分割の失敗、(2)製品名とそれが属する製品カテゴリ(例、パソコン・FAXなど)の知識の欠如によるキーワードの展開不足、の2つに集約される。従って再現率の改善には、単語分割部に知識を持たせて未登録語処理を高度化することや、組織体や製品の名称を自動抽出する技術の開発が必要である。また、適合率の改善には、(1)売上・用途・形状・性能といった製品属性を明らかにすること、(2)供給関係・設備投資・人事移動・景気などの複雑な概念の表現方法の検討、などといったモデルの根本的な詳細化が必要である。

謝辞：本研究は第五世代コンピュータプロジェクトの一環として行われた。御支援いただいたICOTの横井俊夫、岩下安男両室長に深謝致します。

【参考文献】

- [杉山(1986)] 杉山ほか「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資料58-8, 1986.
- [近山(1984)] Chikayama, T. "ESP Reference Manual", ICOT Technical Report TR-044, 1984.
- [Bates (1979)] Bates, M. J. "Information Search Tactics", Journal of the American Society for Information Science, pp. 205-214, 1979.