7T-5

知的情報検索システムＩＲＩＳの分野移行性の評価

秋山　幸司*，　杉山　健司*，　伊藤　博樹**，　小野寺　浩**

（*富士通株式会社，　**富士通 エフ・アイ・ピー 株式会社）

1. 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，その回答となる意味内容を持つテキスト群をテキストベースから検索することをめざす実験的システムである．

本稿では，対象分野を最初のプロトタイプとは異なる分野に変更して小規模なパイロット版を作成し，ＩＲＩＳの分野移行性を評価した結果を報告する．ＩＲＩＳの構成概要を図１に示す．ＩＲＩＳは，限定された対象分野の記述（分野モデル）およびそこで成り立つ事実（世界知識）を使うが，これらは，すべてのサブシステムから独立した形で実現されている．ＩＲＩＳの詳細については文献〔杉山(1986, 1987)，伊吹(1987)，秋山(1987)〕を参照願いたい．

2. 移行先の分野

ＩＲＩＳの最初のプロトタイプにおける対象分野は「情報産業界の新聞記事見出し」であった〔杉山(1986)〕．これは，テキストの収集が容易で検索ニーズがあり，かつ，開発者にとって分野モデルの構築および世界知識の収集が比較的容易であったからである．汎用性評価のための移行先分野については，このような条件を満たしていることに加えて，分野モデルや世界知識が大きく異なることが望ましい．検討の結果，国家間の外交や軍事問題を中心とする国際政治面の新聞記事見出し文を取り上げることにした．国際政治面には，外国の一国内の動向を示す雑多な記事も混在しているが，ほとんどすべての分野を包含すると言えるこの種の記事群は除外し，純粋に多国間の外交に関連する記事のみに対象を限った．

3. ＩＲＩＳの分野依存部分とその変更の程度

ＩＲＩＳにおける分野依存知識の大半は，分野モデルおよび世界知識として各サブシステムから独立な形で実現されているが，各サブシステムに局所的な分野依存の知識や手続きも存在している．これらの多くは，曖昧さを解決したり条件判定を行うために，分野モデルの特定の概念を参照したり世界知識を独自に解釈するものである．なお，以下に述べる状況は，収集された例文に対し，典型的テキスト文49件からなるテキストベースについて，典型的質問文10文を受け付ける小規模な評価用システム（以下，パイロット版と呼ぶ）の構築過程で認められたものである．

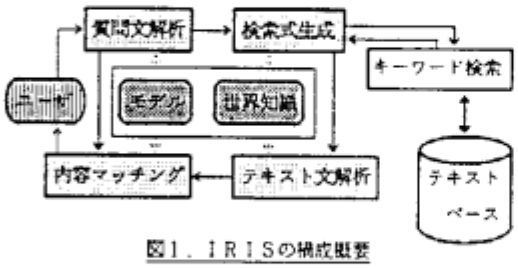

図１．ＩＲＩＳの構成概要

(1) 分野モデルおよび世界知識

分野移行に際しては，分野モデルおよび世界知識の変更は必然的である．まず，検索要求を表す質問文をアンケート調査により 130文，検索対象である新聞記事見出し文を朝日新聞約１ヶ月分の１面と７面から 210件，それぞれ収集した．次に，最初のプロトタイプと同様の手法〔杉山(1986)〕を用いて，これらの文から分野における実体およびそれらについて成り立つ述語（例．図２，図３）を抽出し分野モデルを生成した．

一方，必要な世界知識の検討およびその収集は，一般に完全を期することは不可能である．パイロット版では，図４のような世界知識について，前述の新聞記事 210件からできる限りの知識を抽出し，地理情報などの一般知識については別途文献からも入手した．

このように，分野モデルおよび世界知識は，分野移行に際してほぼ全面的に変更を要した．しかし，ＩＲＩＳの最初のプロトタイプの作成の経験から，分野モデルの抽象度や推論に必要とされる世界知識の種類がある程度把握できたため，最初のプロトタイプでは，分野モデルの生成に２ヶ月，世界知識の収集に２週間を要したものが，本プロトタイプではそれぞれ１ヶ月および１週間であった．

中心的実体　：　国家，軍，主権者，資源，兵器，問題，会議　など

中心的述語　：　賛意，協力，攻撃，非難，交渉，対立　など

図２．分野の中心的な実体・述語の例

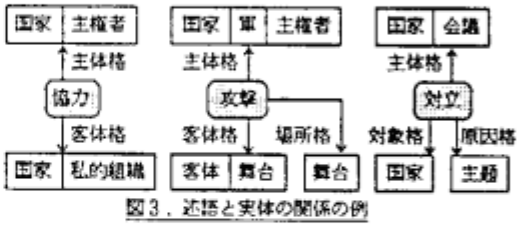

図３．述語と実体の関係の例

(1)命題間関連知識　…　例．協力≒賛意≒合意，対立≒攻撃　など

(2)命題間相反知識　…　例．賛意⇔非難，攻撃⇔和解　など

(3)地理的知識　…　例．日本⊃東京，ＥＣ⊃ドイツ　など

(4)要人・役職の知識…　例．日本∋中曽根，ソ連∋書記長　など

(5)客体シソーラス　…　例．兵器⊃核兵器，兵器⊃ICBM　など

(6)主題シソーラス　…　例．軍縮⊃＊禁止(A,O)，国家(A)，兵器(O)＊

図４．世界知識の種類

Initial study on transportability of IRIS (an intelligent information retrieval system)

Kohji AKIYAMA*, Kenji SUGIYAMA*, Hiroki ITOH** and Hiroshi Onodera**　(*FUJITSU Limited, **FUJITSU FACOM Information Processing Corp.)

(2) 質問文／テキスト文解析部

質問文およびテキスト文の構文意味解析部における分野依存部分は，①単語とそれに対応した分野モデルでの概念，および，②文中の句の表層格から述語概念の引数（いわゆる深層格）への写像，である．前者は単語辞書中に記述され，後者は分野モデルの述語概念に対応するクラスノードに記述される．

未登録語を辞書に登録し，各単語の意味情報（モデル上で対応する概念）を設定する単語辞書変更作業は，検証を含めて約2週間であった．次に，表層格と述語の引数との写像を，モデル中のクラスノードに記述する．この作業は約2日であった．なお，パーサのアルゴリズム部分の変更は一切行わずに済んだ．

(3) 検索式生成エキスパート

検索式生成エキスパートにおける分野依存部分は，分野モデルの各概念に対応した初期キーワード抽出規則，分野に依存した一部の検索戦術，および，検索戦術適用規則である．

最も単純なキーワード抽出規則は，意味構造中の各概念ノードに対応するレキシコン群をすべてキーワードとすることであるが，分野の話題の中心になる実体については，キーワードとなりうるレキシコンの概念クラスに制約があったり，あらかじめ特定の展開をすべきであることが多く，分野に応じたこれらの知識を組み込むことが必要となる．パイロット版には，現時点までに抽出された知識を組み込んだがまだ十分ではない．一方，検索戦術および適用規則については，とりあえず，ＩＲＩＳの最初のプロトタイプから分野に依存した戦術および適用規則を除去した部分をパイロット版に組み込んでいる．勿論，本分野に依存した検索戦術も入れるべきであるが，これは，パイロット版の結果を評価した後で行うこととした．

以上のように，検索式生成エキスパートの調整度はまだ十分ではないが，これに要した期間は約3日である．

(4) 内容照合部

内容照合部における分野依存部分は，意味構造を標準形式に変換する部分，および，標準形式の内部構造である．

現在の内容照合部では，標準形式の内部構造が分野モデルに依存する形となっており，分野移行に際しては全面的な書き換えが必要である．これに伴って，標準形式への変換プログラム部の一部も書き換える必要がある．このため，変更には，テスト期間も含めて約2週間を要した．

4. 新たな機能の追加

ＩＲＩＳが最初に扱った「情報産業界の新聞記事」では，＂会社が製品を発売する＂，＂会社群が提携する＂などのように，主体の行動が記事の主題になっていることが多く，主体の行動を十分に分類した上で述語の世界知識を整備することにより，検索漏れを極小にしたり内容照合を述語中心に行うことが可能であった．

ところが，「外交問題の新聞記事」では，＂兵器の禁止や撤去＂，＂軍事的条約の締結＂などが実は＂軍縮＂という主題である，というように，記事の主題が，単なる述語記号に対応せず，複雑な意味構造で表される場合が多い．従って，述語中心の比較を行う内容照合部で＂軍縮＂と＂兵器の削減＂を意味的に関連が深いと判断するためには，意味構造と主題の関係を世界知識として記述する枠組みが必要である．検索式生成では，シソーラスに＂撤去＂の関連語として＂軍縮＂を記述する方式でも，ある程度の再現率を達成できるが，主題を基にした用語展開を組み込めば，より知的な検索式生成を実現できる．また，「主題」を陽に意識すれば，単なる「関連語」に比べて世界知識抽出の際の漏れが少なくなるであろう．

そこで，「外交問題」におけるＩＲＩＳパイロット版では，意味構造に対応した一階述語論理表現をデータとする「主題シソーラス」を設け，このような世界知識の枠組みを試作してみた．現時点では，検索式生成において有効に動作することを確認し，内容照合部での推論に利用する機構を構築中である．

5. 評価および今後の課題

分野移行の結果，内容照合部の分野依存性（あるいは分野モデルへの依存性）が強く，移行性にやや難がある他は，おおむね良好な分野移行性を持っていることが判った．このことから，ＩＲＩＳの枠組みはある程度汎用的であると言える．なお，内容照合部は，今後，照合のための標準形式を用いることの是非を含めて検討し，再設計により汎用性の向上を図る予定である．

一方，汎用性によりシステムの枠組みを変えずに分野移行が可能ということは，必ずしも分野移行が容易であることを意味しない．実際，ＩＲＩＳにおける分野モデル作成，世界知識収集，および単語辞書の整備の工数は，他の移行作業を無視しうる程大きく，データベースの自然言語インタフェース（例えば〔石川(1985)〕）における工数の十倍以上である．データベースのデータが持つセマンティクスは人間が論理的に定義した単純明快なものであるのに比べ，テキストベースのそれは，曖昧で巨大で複雑であり，テキストベース中の各テキストを分析して初めてセマンティクスの一端が判るということが原因の一つである．従って，ＩＲＩＳを実用化するためには，対象分野の調査および知識獲得の支援ツール（例えば，未登録語の自動抽出，モデルエディタ，世界知識の自動獲得など）を考案・実現することが課題となる．

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．御支援頂いた ICOT の方々に深謝致します．

【参考文献】

〔石川(1985)〕　石川ほか「自然言語インタフェースＫＩＤの評価」情報処理学会第30回全国大会予稿集 pp.1429-30. 1987.

〔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.

〔伊吹(1987)〕　伊吹ほか「自然言語インタフェースとしてのＩＲＩＳ」情報処理学会第34回全国大会予稿集 pp.1325-6. 1987.

〔秋山(1987)〕　秋山ほか「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」　同上　pp.1327-8. 1987.

〔杉山(1987)〕　杉山ほか「内容検索システムとしてのＩＲＩＳ」　同上　pp.1329-30. 1987.