自然言語処理 47－8
（1985. 1. 25）

自然言語による内容検索に向けて*
A STEP TOWARD CONTENT RETRIEVAL THROUGH NATURAL LANGUAGE DIALOGUE

杉山健司
（富士通研究所）

ABSTRACT: At SRI, we are developing a text access system that answers a request expressed in English by retrieving the relevant passages in a text. This paper first summarizes the merit and the developing method, and then focuses on the problem of the content retrieval, which is one of the key issues in the development of the system.

1．はじめに

近年，自然言語インタフェースの研究が盛んに行なわれ，リレーショナル・データベース等のように形式化されたデータに対する自然言語アクセスが可能になってきた。しかしながら，形式が整っていないようなデータ，例えば，テキストに対する自然言語インタフェースは未だない。

テキスト・アクセスに関する従来からの手法は，キーワードの論理的組合せによって質問式を作り，キーワードの出現場所から質問式に合致するテキストを検索するというものである。これを発展させた物として，ファジー情報を付加したキーワード間の階層関係をプロダクション・ルールの形で定義する手法も提案された［McCune et al. 83］。しかしながら，これらの手法は，いづれも，質問式を作ることが難しかったり，テキストの内容よりも，むしろ，テキスト文中に現れる単語の形に依存した検索になってしまう，という欠点がある。

そこで，SRI では，計算機言語学や人工知能の分野で開発された自然言語解析手法，並びに，知識表現，推論等の技術を使うことによって，内容に則した検索を自然言語で行なえる様研究を進めている［Walker et al. 81］。図1に我々が目標としているプロトタイプ・システムの応答例を示す。対象としているテキストは，肝炎に関する専門的医学書である。肝炎患者を扱っている医師が，患者の回復状態がどうも良くないので，システムに問い合せを行い，原因分析及び対処法の検討を行なっている例である。

このシステムを実現するための我々のアプローチは，

①テキスト内容の要約を表す論理式を自動的に作り出す手法の開発と，

②質問文を解析し，①の論理式とのマッチングを行ない，必要なテキストを取り出す手続の開発

との二つに分けられる［Hobbs et al. 82］。①では，パラグラフのcoherence 構造［Hobbs 78］を分析してテキストの要約表現を作り出す方法を提案し，②では，既存の自然言語解析システムDIALOGIC［Grosz et al. 82］を利用し，さらに推論と内容マッチングを行う手続を開発中である。プロトタイプ開発のため対象を肝炎に関する医学テキストとしているが，開発中の技術は，他の種類のテキストに対しても有効である［Walker 82］。

以下では，開発中のテキスト・アクセス・システムの概要を説明し，次に，システム開発上重要な位置を占める内容検索機能について議論する。

USER 「肝炎のとき，メレナを呈することがありますか」
SYSTEM 関連のあるテキストが2つあります。どちらにしますか？
1．その他の胃腸管系症状と所見(passage 967)
2．症状および／または所見(passage 1307)
USER 「1。」
SYSTEM 5 各種ウィルス性肝炎(passage 915)
5.2 典型的ウィルス性肝炎(passage 917)
5.2.3 臨床(passage 921)
5.2.3.5 その他の胃腸管系症状と所見
(passage 967)
(passage 974) 肝炎患者 432例中わずか1例（文献319），および黄熱ワクチン接種による肝炎患者838 例中1例（文献325）にメレナがみられた。
USER 「passage 1307」
SYSTEM 5.5 激症ウィルス性肝炎(passage 1291)
5.5.3 臨床(passage 1296)
5.5.3.3 症状および／または所見(passage 1307)
(passage 1320)フィブリノーゲン合成不良あるいはビ満性血管内凝固(DIC) の結果としての著しい血清フィブリノーゲン低下に伴い，胃腸管内出血が起るかもしれない。
USER 「プロトロンビン時間は重要ですか？」
SYSTEM (passage 1319)ビタミンK投与にもかかわらず，プロトロン時間は普通起ります。
USER 「腎不全は？」
SYSTEM (passage 1324)肝腎症候群が激症ウィルス性肝炎に併（続）発するかもしれません。その症状は，欠尿と窒素血症で，よく，浮腫，腹水，低ナトリウム血症を伴います。肝腎症候群が出現すると予後はあまりよくありません。
・・・
USER 「メレナの治療はどのようにすればいいですか。」
SYSTEM (passage 1352)出血の処理は，凝固因子欠乏を含め，適切な血液補充剤を使用すべきです。
（注 メレナ：胃腸管出血に関する症状）

図1 人間とテキスト・アクセス・システムの応答例

（但し，入出力文は本来英語であるが，図では和訳されている。）

＊ 本論文は，著者が SRI International滞在中に行った研究［Sugiyama 84］に基づいている。

2．テキスト・アクセス・システム

テキスト・アクセス・システムは，図2に示すような構成になっている。ユーザの質問文は，SRI のDIALOGIC [Grosz et al. 82] によって構文解析，並びに，意味表現(logical form 1)への変換が行われる。次に，このlogical form 1は，推論コンポーネント [Hobbs 80] に渡され，談話に係わる問題のうち，文単独で起る問題と以前の文脈を合せた時起る問題の2種類 [Hobbs 84b] が解決される。この中には，述語の暗黙引数の明示化や，転喩(metonymy)の問題がある [Hobbs et al. 82] 。これらの問題の解決は，意味表現(logical form 2)に反映される。最後に，このlogical form 2は，内容検索コンポーネントに渡され，テキストの要約表現とのマッチングが実施される。この結果，ユーザの質問に答えるようなpassage が肝炎テキストの中から探し出され，答としてユーザに返される。

内容検索および推論コンポーネントでは，それぞれの問題を解決するために一般的常識や医学に関する知識を共有して使う。この知識は，言語学的前提条件を拠にして集められている [Hobbs 84a] 。これらの知識，及び，意味表現(logical form 1,2)，要約表現は，'ontological promiscurity'と呼ばれるアプローチ [Hobbs 83] に基づいた述語論理式で表現されている。

3．内容検索

内容検索コンポーネントを開発するために我々は，まず，どんな質問文に対してどのpassage を検索すべきかを調査・検討し，次に，内容検索コンポーネントのインターフェイスである意味表現(logical form 2)，テキストの要約表現，知識ベースの3つをそれぞれ検討し，最後に，内容検索のアルゴリズムについて考察を加えた。

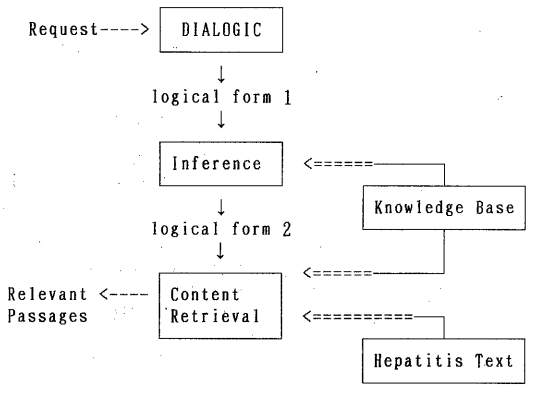


図2　自然言語によるテキスト・アクセス・システムの構成

3．1　質問文とpassage の関係

質問文とその答となるべきpassage の関係を調査した結果，質問文に対して答となりうるpassage は次の3つに分類されることがわかった。

① passage 全体として質問に答えている場合

② passage の一部分が質問の答になっている場合

③ passage の中に質問に関する答が間接的に含まれている場合

①，②を直接的答と呼ぶ。③は，答が陽にpassage 中にあるのではなく，passage の内容から推論によって答が得られる場合である。従って，①，②は，質問を発したユーザを満足させる可能性が高く，③は，直接的答がない時のみユーザを満足させる，と捉えることができる。

例えば，"Is passive immunization desirable for staff on hemodialysis units?"（人工透析従事者に受動免疫を行うのは望ましいことですか？）という質問を考える。Passage 1845(あるいは，6.7.4 とも呼ぶ) は，「受動免疫によく使われるガンマ・グロブリンという薬を使って，血液透析ユニットで働く人がB型肝炎になることを予防する」ことについて述べているので，上記①に分類される。Passage 6.7 は「血液透析ユニットばかりではなく，より一般的に肝炎ビールスの伝播が心配される場合の受動免疫」に関する記述であるので，②に分類される。Passage 6.3.7 は「道具を介して伝染する肝炎についての一般的予防法」となっており，免疫に関する情報は直接的には含まれていないので，③に分類される。

典型的な質問文に対して，上記の方法で質問文とpassage の関係を整理した。この資料は，内容検索コンポーネントの目標を表すと同時に，インプリメントするシステムの性能チェックとしても使われる。

この資料作成中に問題となった点が2つある。その1つは，passage の内容の非一様性である。例えば，質問文"What schedule of immunoprophylaxis is appropriate following exposure to HBV?"（B型肝炎ビールスに接触した後は，どのようなスケジュールで免疫学的予防を行うのが適当ですか？）に対して，passage 6.3.3 は上記の③，passage 6.3.4，6.3.5は上記の②というように分類される。Passage 6.3.4 が直接的答"immunoprophylaxis is good during the incubation period." （免疫学的予防法は，潜伏期間の間有効です）を含み，passage 6.3.5 も少し曖昧ではあるが直接的答"occasional use is good." （度々行うことが有効です）という文を含む。しかし，passage 6.3.3 は間接的答"preventing the primate contact and using ISG"（動物感染を予防し，免疫血清グロブリンを使うこと）という句しか含まない。しかしながら，これらのpassage のテキスト中の位置づけやpassage に付けられている標題から考えて，本来これらのpassage は，同じ種類の

情報が一様に含まれていることが予想されるが，実際のテキストはそうはなっていない。このような微妙な内容の違いは，現在のアプローチ（テキストの要約表現をベースに検索を行う）では扱えないので，内容の一様性が破られている場合は，テキスト作成者にフィードバックをかけ，一様になるようテキストを書き直してもらうか，それができない場合は，内容検索の際，ユーザに我慢してもらうことになる。

2番目の点は，質問が発せられる観点が，テキストが作られている観点と異なる場合があるということである。例えば，質問"What type of immunization is appropriate following exposure to HBV?"（B型肝炎ビールスに接触した後，どんな種類の免疫が適当ですか？）に答える情報は多くのpassage に分散してしまっている。これは，テキストが，どんな病原菌に対してどんな免疫法があるか，といった観点からは書かれてはおらず，各免疫法ごとにどんな病原菌に有効か，という観点から書かれているためである。質問が病原菌の種類を‘検索キー’にしているのに対し，テキストの方は免疫法の種類を‘主要キー’として構成されているため，と考えることができる。質問が自然言語なので検索キーは多種多様なものになることが予想され，この例のように必要な情報が分散していたり，また，全くある検索キーに関する情報がなかったりする場合がある。第1番目の点と合せて今後の課題である。

3．2　質問文のlogical form

2節で述べたようにユーザの質問文は，DIALOGICによってlogical formに変換され，次に，推論コンポーネントによって談話に係わる問題が解決される。例えば，次の例を考える。

質問文(3A)"To what contacts should immunoprophylaxis be administered following exposure to HBV?"
（B型肝炎ビールスにどのような接触をした場合，免疫学的予防を施すべきですか？）

は，DIALOGICにより次のようなlogical formに変換される。

```
should(Ea),administer'(Ea,Z,I,C),
immunoprophylaxis(I,X1,Y1),what(C),contact(C,W),
plural(C,Cs),follow(Ea,Ex),exposure(Ex,X2,Y2),
hbv(Y2)
```

このlogical formの意味は，「実行しなければならない事象Eaが存在し，その事象Eaとは，Z によるI のC への投薬であり，I はX1のY1に対する免疫学的予防法，C は病原菌との接触であり，しかも，この事象Eaは，X2がHBV という種類の病原菌にさらされたという事象Exの後に起る」というものである。次に，推論コンポーネントによって暗黙の引数の問題が解決される。質問文では，何に対する免疫学的予防法か（Y1は何か）が陽には表わされていない。しかし，内容検索を行うためには，このY1はHBV という病原菌であることを推論しなければならない。同様に，免疫学的予防法を適用しなければならない人X1は，その病原菌にさらされた人X2と同一であることを見つけ出さなければならない。

他の例として，上記の質問に続いて，次の質問が入力されたと仮定する。

質問文"What type of immunization is appropriate?"
（どのような種類の免疫が適当ですか？）

前例と同様，DIALOGICによってlogical formに変換され，推論コンポーネントによって談話に係わる問題が解決される。この例では，暗黙の引数以外に以前の文脈に係わる問題が解決される。すなわち，この質問文中の免疫とはHBV にさらされた後施すべきもの，という1文前の情報が追加される。暗黙引数の解決も含めて最終的に次のlogical formが得られる。

```
appropriate(T),what(T),type(T,I),
immunization(I,X,Y),follow(I,Ex),
exposure(Ex,X,Y);hbv(Y)
```

3．3　テキストの要約表現とテキスト構造

要約表現

［Hobbs et al. 82 ］で提案された方法論に基づいて各passage の要約表現を作り出す。但し，［Hobbs et al. 82 ］では，parallelというcoherent関係についてしか議論されていないので，ここでは，他のcoherent関係についてもこの方法論を拡張する。例えば，contrastという関係は，対照的な事柄を述べた2つの文（あるいは節）の間の関係であるが，この関係で結ばれた2文の要約としては，対照的事柄を包含する述語式(predication) を考える。

例として，次のpassage を考える。

(Passage 2.3.2)

Laboratory animals have been infected with hepatitis A virus but are not useful experimental animals because of absence of reliable and consistent biochemical or histological markers of infection. Hepatitis A virus has only recently been propagated in cell culture. Agar gel diffusion systems have not resulted in identification of serum antigens of type A virus; but convalescent serum from type A hepatitis patients does specifically aggregate virus particles in feces, and has been the basis for development of tests to identify hepatitis A antigen and antibodies to it. The effect on infectivity of the hepatitis A virus of exposure to ether, acid, temperature and chlorine is known.

第1文は，'but' で繋がれた2つの主節から出来ており，

前の主節では，「研究用動物がHAV に感染する」ことが，後の主節では，「研究用動物は役に立つ実験動物ではない」ことが述べられているが（contrast関係），全体としてこれらの事柄を包含する述語式「動物実験によって見つかるHAV の特徴」というものを考えることが出来る。第２文以下の文は，すべて，「研究室実験によって見い出されるHAV のいろいろな特徴」について述べているので，parallelという関係で結ばれ，「研究室実験によって見出されるHAV の特徴」がこれらの文の要約となる。

第１文，第２文以下とも，同じ種類の言明「～によって見つかるHAV の特徴」となり（parallel関係），全体としてのpassage の要約は，次のようになる。

characteristics(C,V)&hav(V)&find(-,C,M)&
(laboratory-test(M)|animal-experiment(M))
「研究室実験および動物実験によって見つかる
HAV の特徴」

テキストの各passage について同様の手順で要約を作り出している。現在のところ，作り出された要約表現は，ほとんどconjunction normal form である。

テキスト構造

一般に，ある内容でテキストを検索しようとする場合，テキストの全体構成を知っている方がより早く，的確に関連するpassage を検索することができる。そこで，先に得られた要約表現をもとに，テキストの内容構成を表すような構造を作り出すことにした（図３参照）。

このテキスト構造の重要な要素は，包含関係である。例えば，passage 5.2 はpassage 5.2.2 を包含する。5.2 は，典型症状性肝炎について，5.2.2 は典型症状性肝炎の中でも病原学について述べている。ここでは，包含関係は次のように定義される。即ち，「passage A がpassage B を包含する」とは，「任意の質問文Q に対して，もし，B の要約表現がQ と内容的に一致すれば，A の要約表現もQ とある程度内容的に一致する」ことである。包含関係の最も簡単な例は，A の要約表現がX&Y であり，B の要約表現がX&Y&Z である場合(passage 5.2と5.2.2 がその例）である。他の例として，passage 2 と2.3.4 がある。2 は肝炎の病原学についての記述であり，2.3.4 はnon-A non-B 型肝炎ビールス(HNANBV)の特徴を述べている。Non-A non-B 型肝炎ビールスは肝炎(hepatitis) の病原菌の一種であり，そのビールスに関する特徴は病原学(etiology)の一部と考えられるためである。

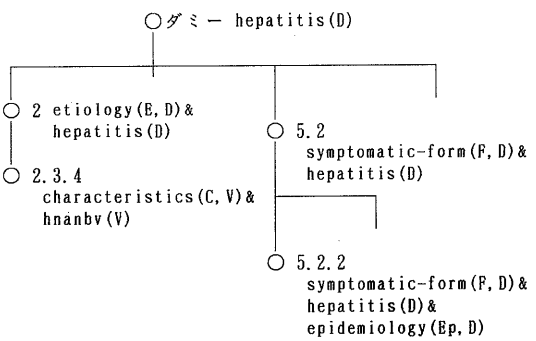


図３　テキスト構造

以上のようにして作られたテキスト構造は，一般に，lattice 構造となる。しかし，現在の対象テキストでは木構造になっている（図3)。図のノードの多くは，テキスト中のpassage に対応する。対応するpassage がないノードは，テキスト全体の構成を理解し易くするために追加されたダミーノードである。

このテキスト構造は，内容検索時に必要となる知識を集約したものと考えることもできる。例えば，ある質問内容がpassage 2.3.4 と一致する時に，passage 2 とも一致することを見け出すためには，non-A non-B 型肝炎ビールスとは何か，病原学とは何かといった諸々の知識が必要となるが，テキスト構造では，これらの知識がたった１つの関係として表されている。

以上のテキスト構造に重要な情報がさらに１つ加わる。それは，passage 間の排他関係である。例えば，passage 6.7.2 はB 型肝炎に対する受動免疫についてのpassage であるが，6.8.2 は同一の病気に対する能動免疫についてのpassage である。両passage は免疫法のタイプに関して排他的である。従って，質問文が能動免疫ではなく，受動免疫に関する情報を要求している場合は，6.8.2 以下のノード(passage) を無視し，6.7.2 以下のノードを調べるだけでよい。

３．４　知識ベース

本格的な知識ベースは［Hobbs 84a ］によって開発中である。ここでは，もっと小規模なもの，典型的例文を処理するに足るだけの知識ベースを検討する。

まず，この知識ベースに入れるべき事実を見つけ出す必要がある。その方法の１つとして，3.1，3.2節の結果をもとに，質問文と要約表現とを対比し，それらの間の関連性を見つけ出すために必要となる事実を抽出する方法がある。例えば，質問文"What is the incubation period of HBV?"（B型肝炎ビールスの潜伏期間はどれぐらいですか？）とpassage 2.4(症状による病原菌HAV,HBV,HNANBVの区別についてのpassage)が関連することを見つけるためには，「潜伏期間は症状の特徴の１つであり，病原体を区別する情報と成り得る」といった事実が必要となる。

事実を見つけ出す他の方法は，3.3 節で説明した包含関係を見つけ出す時に，前提として使われる事実を取り出す

方法である。例えば，3.3 節でpassage 2 と2.3.4 が包含関係にあることを見つける時，「Non-A non-B 型ビールスは，肝炎の病原菌の一種である」という事実が使われた。

以上のようにして得られた事実は，次に［Hobbs 84a ］と同様にして，自然な‘クラスタ’に分類される。これらのクラスタは，事物(matters) と動作／事象(actions/events)に大別され，それぞれさらに病気，ビールス，薬，動物，食物，道具と病原体の伝播，診断，テスト，治療，予防等に細分される。しかし，［Hobbs 84a ］とは異なり，一般的常識の範疇に入る知識は含まれていない。これは，例文の範囲では，常識レベルの知識は不要なためである。しかしながら，広範な質問文を受け付けるためには，常識レベルまで掘り下げた文の理解が必要となる。例えば，病気をある１つの実体と捉えるだけで十分な場合もあるが，より一般的に理解するためには，病気を「ある個体の状態であり，その状態は病原体という異物が入ったために起っている」というように捉え直さなければならない。

このようにして集められ，分類された事実を表現する方法として，シソーラス形式と述語論理形式との２つが考えられる。シソーラス形式では，各概念の上位下位，排他関係，密接な関連，希薄な関連の４つの関係によって大まかな事実を表現することができる。例えば，肝炎の下位概念としてＡ型肝炎，Ｂ型肝炎があり，Ａ型肝炎とＢ型肝炎は互いに排他的であったり，Ａ型肝炎(HA)はＡ型肝炎ビールス(HAV) と密接な関連を持つ，といった事実である。述語論理形式では，シソーラス形式では表現できない情報，例えば，Ａ型肝炎の病原菌はＡ型肝炎ビールスである(ha(D) <-> (Some V)hav(V)&etiologic-agent(V,D))，といった詳細な情報まで記述することができる。従って，シソーラス形式に基づく内容検索は，述語論理形式に基づく場合に比べ，粗いものになってしまう。しかしながら，述語論理形式が有効となる詳細な情報は，［Hobbs 84a ］や本節にあるように，人手による収集が必要となるが，シソーラス形式レベルの大まかな情報は，［Amsler 81 ］のようなアプローチによって既存の医学辞典から自動収集できる可能性がある。最終的にどちらの形式を採るかは将来の課題として，例文の範囲では述語論理形式を採用している。

３．５　内容検索アルゴリズム

アルゴリズム

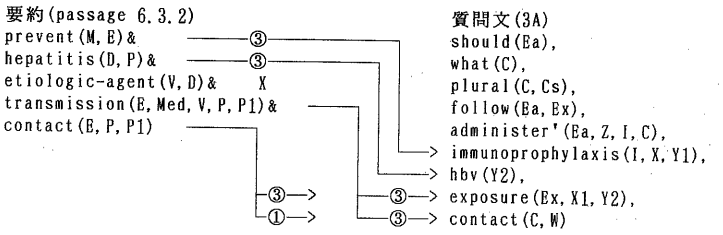
質問文と関連性の高いpassage をテキストの中から探し出す方法の一つは，テキストの先頭passage から順次，質問文と関連性があるかを判定していく方法である。しかし，関連性を見つけ出す手続は，知識ベース中のルールを使って内容マッチングを行うのでコストが高く，それをテキスト中の全passage に対して行うのは非能率である。他に考えられる方法として，質問文中の中心概念だけについてまず内容マッチングを行い，関連しそうなpassage だけを取り出し，その後，各passage について質問文中の他の概念との一致を調べる方法があり，この方が効率的である。現在，この方法に基づくアルゴリズムを検討している。

質問文の中心概念は，focus と一致すると考えられる。文のfocus が何かを判定する方法は，［Sidner 79 ］に従っている。例えば，「動詞の直後に続く名詞句がある場合は，それがfocus になる」という［Sidner 79 ］のルールに従って，質問文(3A)"To what contacts should immunoprophylaxis be administered following exposure to HBV?"のfocus は，immunoprophylaxis という名詞句となる。従って，検索の中心となるpredicate は，immunoprophylaxis(I,X1,Y1) (3.2 節のlogical form例を参照) となる。

この'focus predicate' と内容が一致するpassage を3.3 節のテキスト構造から探し出す。例えば，上記の質問に対しては，テキスト構造のpassage 6.4 以下のpassage グループ，6.5 以下のグループ等が見つかる。これらのpassage は，その要約としてimmunization(I,X,Y) やprophylaxis(I,X,Y)というpredicate を持っており，これらのpredicate は，知識ベース内のルール"immunoprophylaxis(I,X,Y) <-> immunization(I,X,Y)&prophylaxis(I,X,Y)"(免疫学的予防法は，免疫法であると同時に予防法である) によって，focus predicate と直接的にマッチングする。さらに，6.2 以下のpassage のように間接的なマッチングとなるpassage もある。これらのpassage は，prevent(-,D)というpredicate を持っており，immunoprophylaxis(I,X,Y) とは，上記ルール及びルール"immunization(I,P,D) -> (Some E)tcause(I,E)&prevent(E,D)&disease(D,P)" (免疫は，ある病気を予防するために行われる) によって間接的に関連づけられる。

以上のようにしてfocus predicate と内容的に一致したpassage のうち，3.3 節で説明した排他的関係によって質問文と合わないpassage グループが，検索対象から外される。例えば，上記の例では，passage 6.4 以下のグループがＡ型肝炎に対する免疫に関するテキストであるが，質問文がＢ型肝炎に対する免疫についてであるので，取り除かれ，passage 6.7.2, 6.8.2以下のグループが残る。

以上の手続によって複数のpassage が残った場合は，評価関数（後述）によって質問文との関連深さを評価し，一番関連深いと評価されたpassage を内容検索の結果とする。もし，passage が残らない場合は，質問文のfocus を移動して，再度同様の手続で関連するpassage を探す。Focus 移動は，［Grosz 77］のモデルに従う。例えば，質問文"What tests establish immunity to HBV?" （Ｂ型肝炎ビールスに対する免疫力を調べる検査はどのようなものですか？）のfocus はimmunityであるが，それに一致するpassage が見当らない。そこでfocus をimplicit focusに移動す



[①: 一番強い一致, ②:2番目に強い一致, ③: 一番弱い一致, X:不一致]
図4　passage の要約と質問内容との一致

る。[Grosz 77] に従うと,「implicit focusは, focus space に含まれる動詞 (例ではestablish)と格関係を持つ名詞句 (例ではtests)となる」ので, tests が新たなfocus となる。このfocus に対して上記と同様の手続を踏むことによって関連性のあるpassage 5 等が見つかる。

評価関数

3.1 節で得られたpassage と質問文との関連性を反映するような評価関数を作り出すことを試みる。そのため, 3.1 節の例文に対して上述のアルゴリズムを適用し, 関連性がありそうだとして残されたpassage の要約表現と質問文のlogical formとを対比し, それぞれのpredicate の一致関係を分析した。Predicate の一致は, 3.4 節で説明したシソーラスのレベルで捉え, 一致の強さを3つに分類した。即ち, 一番強い一致 (強度①) が同義関係にあるpredicate 同士, 2番目に強い一致 (強度②) が上位下位関係及び密接な関連, 一番弱い一致 (強度③) が希薄な関連である。これらの一致の強さや不一致等の特徴を各要約表現を中心として整理すると, 例えば, 図4のようになる。

次に, このpredicate 間の一致の特徴がどのようにpassage 対質問文の関連性と関係するか (相関関係) を分析した結果, 次のような評価関数を得た。

(評価関数)= 2*(カバー度)+ (一致の強度)- (不一致度)

カバー度は, 上記の3つの要素のうち最も相関性が高いもので, 質問logical form中のpredicate が幾つマッチされたかを示す数字である。図4の例では, immunoprophylaxis 以下のpredicate がそれに当り, カバー度は4となる。

一致の強度は, 次の式で表わされる。

(一致の強度)=

(Focus predicate について

強度①なら,3; 強度②なら,2; 強度③なら,1) +

(Focus predicate 以外の質問predicate について

合計する。一致が複数あるときは一番強い一致が

強度①なら,2; 強度②なら,1; 強度③なら,0)

図4の例では, focus predicate: immunoprophylaxisが強度③でマッチされているので, 上式の前半は 1である。上式の後半については, 質問の方のhbv が強度③でマッチされているので 0, exposureが二つのマッチングとも強度③なので 0, contact が二つのマッチングのうち強い方が強度①なので 2となり, 合計して 3となる。

不一致度は, 要約表現中のpredicate で一致が図れなかったものの数で, 図4では1 となる。

この評価関数を作り出すために使った例文以外の例について, この評価関数を適用した結果, 分析した例文と同程度の精度で, passage 対質問文の関連性を予測することができた。

4. おわりに

本稿では, テキスト・アクセス・システムというフレームワークの中で, 自然言語による内容検索について議論した。本稿で提案された方法は, 内容検索システムへの1つのアプローチを示している。

内容検索で現在残されている問題は, どのように知識ベースから必要なルールを取り出し, 要約表現と質問logical form間のマッチングを行うかである。基本的には, 推論コンポーネントと同じような方式 [Hobbs 80] を採ることになろう。[Hobbs 80] では, 談話の問題を解決するルールが優先的に選択されるが, 内容検索では, 内容のギャップを埋めるルールが優先されることになろう。この内容のギャップを定義する1つの可能性は, ギャップの逆である内容の一致性・関連性をmeaning-connectionという概念 [TINLUNCH 82] で定義する方法である。

本稿で報告した研究成果, 及び, SRI 既存の研究成果を基にして, 自然言語テキスト・アクセス・システムのインプリメントが1985年から開始される予定である。

謝辞

本研究を進めるに当り, 御指導御協力頂いたBell Communications ResearchのRobert Amsler, Donald Walker両博士, SRI のJerry Hobbs 博士, カーネギメロン大のArmar Archbold氏, 並びに, 山口大学医学部の小西久典助教授に感謝致します。また, 本論文の原稿にコメントを頂いた富士通研究所の秋山幸司, 亀田雅之両氏, 並びに, SRI 留学

に際してお世話になったSRI Norm Nielsen博士，富士通久保宏志部長，富士通研究所林達也部長を始めとするSRI，富士通，及び，富士通研究所の方々に感謝致します。

参考文献

Amsler, R. A. 1981 "A Taxonomy for English Nouns and Verbs" Proc. 19th Annual Meeting of the Association for Computational Linguistic pp. 133-138.

Grosz, B. J. 1977 "The Representation and Use of Focus in a System for Understanding Dialogues" Technical Note 150, Artificial Intelligence Center, SRI International, Menlo Park, California.

Grosz, B. ; Haas, N. ; Hobbs, J. ; Martin, P. ; Moore, R. ; Robinson, J. ; Rosenshein, S. 1982 "DIALOGIC: A Core Natural Language Processing System" Proc. 9th Int. Conf. on Computational Linguistics (COLING-82).

Hobbs, J. R. 1978 "Why is Discourse Coherent?", Technical Note 176, Artificial Intelligence Center SRI International, Menlo Park, California.

Hobbs, J. R. 1980 "Selective Inferencing", Proc. 3rd National Conf. of Canadian Society for Computational Studies of Intelligence, pp. 101-114.

Hobbs, J. R. ; Walker, D. E. ; Amsler, R. A. 1982 "Natural Language Access to Structured Text", Proc. 9th Int. Conf. on Computational Linguistics (COLING-82), pp. 127-134.

Hobbs, J. R. 1983 "An Improper Treatment of Quantification in Ordinary English" Proc. of 21st Annual Meeting of the Association for Computational Linguistics, pp. 57-63.

Hobbs, J. R. 1984a "Building a Large Knowledge Base for a Natural Language System" Proc. 10th Int. Conf. on Computational Linguistics (COLING-84), pp. 283-286.

Hobbs, J. R. 1984b "Discourse and Inference" Unpublished Manuscript, Artificial Intelligence Center, Menlo Park, California.

McCune, B. A. ; Tong, R. M. ; Dean, J. S. ; Shapiro, D. G. 1983 "RUBRIC: A System for Rule-Based Information Retrieval" Proc. IEEE 7th Int. Computer Software & Applications Conf. (COMPSAC-83), pp. 166-172.

Sidner, C. L. 1979 "Disambiguating References and Interpreting Sentence Purpose in Discourse" in Artificial Intelligence: An MIT Perspective, Vol. 1 P. H. Winston and R. H. Brown eds., The MIT Press.

Sugiyama, K. 1984 "Progress Report on Content Matching I～IV" Unpublished Memos, Advanced Computer Systems Department, SRI International, Menlo Park, California.

TINLUNCH 1982 "Notes on Relevance Readings" TINLUNCH Discussion Paper, Stanford Univ., August.

Walker, D. E. ; Hobbs, J. R. 1981 "Natural Language Access to Medical Text" Proc. the 5th Annual Symposium on Computer Applications in Medical Care pp. 269-273, IEEE, New York.

Walker, D. E. 1982 "Natural-Language-Access System and the Organization and Use of Information" Proc. 9th Int. Conf. on Computational Linguistics (COLING-82), pp. 407-417.