5M－2

日本語対話システムAnyoneにおける形態素・構文解析

大樫 仁司　高梨 郁子　辻 秀一
三菱電機(株) 情報システム研究所

１．はじめに

日本語対話システムAnyoneはオフコン(MELCOM80)において、自然言語によるデータベース検索を実現するシステムである[1)2)]。その日本語解析部は形態素解析・構文解析・意味解析[3)]および時間処理の４モジュールから成っている。本論文では、形態素解析および構文解析について述べる。今回、形態素解析では処理不能な表現の入力が問題点となったが、使用頻度の高い表現を収集し個別の応答を用意することで、ユーザへの適切なガイダンスを可能にした。また、基本的な表現を名詞句連続としたため、構文解析では係り受けの曖昧さの爆発が問題点となったが、初期処理において各形態素を表・項目名および項目値といったデータベース検索分野に特化したクラス分けを行ない、この情報を使った制約規則により曖昧さを抑えた。

２．形態素解析

形態素解析は、速度性能、メモリ性能、および正解率をバランスよく実用的な値とすることに主眼を置いた。設計方針として、用語辞書と接続行列を用いて横型全解探索を行なった後、文節数最小法で第１解を選びそれのみを出力することとした。用語辞書は、自立語、活用語尾および付属語がすべて１つの辞書に入っている。辞書検索は、先頭文字による２次記憶検索の後、メモリ上で候補選択を行なう。接続行列はビット行列であり、文節の開始、文節の終了、活用語と語尾、自立語と付属語など２見出し語間のすべての接続可否を決定し、また文節候補の作成も同時に行なう。その後、文節数最小法で第１解を選択するが、曖昧さが残った場合には自立語長を使った評価によって第１解を決定している。出力は文節の列であり、文節内の見出し語分割の曖昧さや同じ見出し語に対する品詞の曖昧さは１つの解の中に抱え込んだ形とした。

このような方針の結果、特に性能に関する問題点は発生しなかった。また正解率についても、分割ミスに関する問題点は発生しなかった。これは、対象データベースを限定した辞書構築[4)]をしたために、多義語は多いものの語彙数がかなり限られたためと考えている。

問題点は、未知語であった。タイプミス、存在しない表・項目名や値、処理不能な表現が原因である。未知語処理や同義語のサポートを行なうだけでは不十分であり、システムが扱えない表現に対しては、《未知語》という応答ではなく、《扱えない》という応答が必要であると判断した。処理不能表現を《未知語》と応答してしまうと、ユーザが同義表現を入力してみるという努力をしてしまい、《扱えない》ということになかなか気がつかないからである。

そこで、開発期間中に試使用者がしばしば入力した処理不能な表現を処理不能語として辞書に登録し、これが使用されたかどうかを形態素解析段階でチェックし、個別の応答を返すようにした。応答は、《扱えない》ということのみを表現するのではなく、図１のように推奨表現を必ず示すようにした。その結果、処理不能であることが即座にユーザに伝わるので、スムーズに推奨表現へ移行することができた。

「一番」は使わないでください。
『～が多い（少ない）～』のように入力すれば、順番に表示されます。

図１．処理不能語への応答例

Morphological & syntactic analysis in "Anyone" a Japanese language interactive system
Hitoshi OHGASHI, Ikuko TAKANASHI, Hidekazu TSUJI
Computer & Information Systems Lab., MITSUBISHI Electric Corporation.

３．構文解析

Anyoneでの日本語は、ユーザにとっての文の組み立て易さに主眼を置き、基本的な表現として「～の～の～の～は」という名詞句連続を推奨している。そのため構文解析においては、単純に名詞句間の係り受けを全部認めてしまうと、句の数が増えるに従って係り受けの曖昧さが急速に爆発するという問題点がある。

そこで、データベース検索という分野の特徴を生かした制約規則により曖昧さを減少させることを考えた。

まず、形態素を品詞等と意味素を使って、形態素解析時点とは異なる構文解析時点固有のクラス分けを行なうようにした。具体的には、表・項目名、項目値、時間表現、およびその他の大きく４種類に分ける。普通名詞の多くが表・項目名に分類されるが、一部は時間表現やその他に分類される。固有名詞はすべて項目値に分類されるが、一部の数量表現なども項目値に分類される。

次に制約規則について述べる。

まず項目値の名詞句連続は、図２a)に示す通りお互いには係り受けさせず、その後に出現する表・項目名に係り受けさせる。データベース検索においては項目値は表・項目名に対するそれぞれ独立した検索条件にできるため、項目値同士の関係を解析しても無意味だからである。たとえば、「横浜支店の大樫の」は、独立した別条件とする検索式としてもなんら問題はない。この制約規則によって、図２a)のケースでは単純には５個の曖昧な構文木が有り得るところを、１個に制限することができる。

また表・項目名の名詞句連続は、図２b)に示す通り順番に次へ次へと係り受けさせる。データベース検索においては、表・項目名は検索内容であり、順番に後ろの表・項目名に対する条件となるからである。この制約規則によって、図２b)のケースでは２個の曖昧な構文木が有り得るところを、１個に制限できる。

さらに、時間表現は「～から～までの」を推奨し、係り先は表・項目名に限定した。

以上の３つの制約規則によって、名詞句数が増えても係り受けの曖昧さが爆発しないように抑えている。

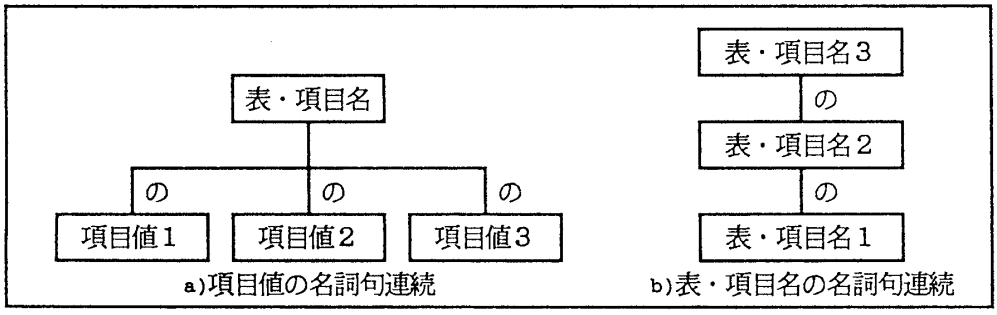

図２．係り受けの制約

４．おわりに

以上、日本語対話システムAnyoneにおける形態素・構文解析について述べた。形態素解析においては処理不能表現に対する適切なガイダンスを可能にした。また、データベース検索特有の形態素のクラス分けを行ない、それを用いた制約規則によって構文解析での係り受けの爆発を抑えた。

参考文献

1)板橋美子ほか：日本語対話システム「Anyone」自然言語によるエンドユーザコンピューティング，情報処理学会情報システム研究会資料，45-2(1993).

2)永松靖朗ほか：日本語対話システムAnyoneにおけるユーザーインタフェース，情報処理学会第47回全国大会，5M-1(1993).

3)高梨郁子ほか：日本語対話システムAnyoneにおける意味解析，情報処理学会第47回全国大会，5M-3(1993).

4)清水英弘ほか：日本語対話システムAnyoneにおける辞書構築，情報処理学会第47回全国大会，5M-4(1993).