8B−2

エキスパートシステム作成支援ツール技術の自然言語処理への適用

阿折義三
日本DEC研究開発センター

1. はじめに

自然言語処理システムの研究開発の1つのボトルネックは生産性の高い開発支援ツールがないということであった。従来,自然言語処理システムはLISPやPROLOGなど,人工知能向けとはいいながら,かなり基礎的なプログラミング言語で構築しなければならなかった。そのため,重要なアイデアが生まれたとしても,小規模な実験システムでしか検証することができなかった。エキスパートシステム(以降ESと略す)がその作成支援ツールの出現によって飛躍的に発展したように,自然言語処理システムの研究開発に対しても,有効な支援ツールの提供が望まれる。私はES作成支援ツール技術,すなわちフレーム型やルール型などの知識表現,推論,推論結果のトレース,確信度による競合解消などの技術が自然言語処理システム作成支援ツールとして非常に有効であることを実証できたので報告する。

2. 自然言語処理システム支援ツールへの要件

自然言語処理システム支援ツールに要求される機能／性能条件を一般のエキスパートシステムとの違いの点から考えると以下のような要件を考慮する必要がある。

- o 大規模辞書へのアクセスが高性能であること
- o 自然言語という特殊なデータの構造を推論の対象とできること,すなわち順序付けられた文字,単語,および概念の列が高性能で解析できること
- o 最近の自然言語処理技術を利用できること
  - 意味素性体系の自由な定義とそれに基づく選択制限,
  - 選好(preference),すなわちより望ましい解釈の選択に関する各種アイデアが試行できる,
  - 実例ベースの利用に関する各種アイデアが試行できる,
  - 解析の意味的深さの選択の自由度がある,(見出しレベル,品詞レベル,核文法レベル,概念レベル,etc)
- o よく構造化された大規模辞書ソースとそのカスタマイズ機能の提供
- o 常識ベースを利用した自然言語処理システムへのスムーズな発展性など

3. ES技術の自然言語処理への適応

多くの自然言語処理のための要素技術はESの要素技術に容易にマッピングできた。(図1参照)

しかし, ESの技術を前記要件に適応させるためには以下のような工夫が必要であった。

3.1 順序性のあるデータに特殊化した推論エンジン

文字列,単語列,および概念列を効率よく推論するために順序性のあるデータに特殊化した推論エンジンを開発した。本推論エンジンは順序条件を満たすデータの列のみを推論対象とする。多くのアプリケーションではすべてのデータの直積を評価する必要はなく,順序づけられたデータの組のみに限定してもよい場合が多い。このような場合, 本推論エンジンは効率的に推論を実行する。

3.2 自然言語処理向け確信度制御

句構造などの確からしさを表わすのに確信度を利用し,以下のように制御した。

確信度初期値付与の原則:
より限定的な句構造,よりユニークな句構造にはより高い確信度を,
より一般的句構造,より多義的句構造には低い確信度を与える。

現状は人間が確信度の初期値を設定しているが,技術的には辞書／事例ベースの簡単な統計解析により自動的に決定することも本ツールを利用した研究対象となりうる。

推論エンジンでの確信度計算の原則:
列の要素の確信度を加味しながらデータ列の長さに対して単調増加関数となるよう工夫した。したがって単純な最長一致ではない。

競合解消:
デフォルトは一番高い確信度の句構造を選択し,指定により確信度の高いものを指定個数選択できるようにした。

確信度の1つのメリットはルールベースによる翻訳,事例ベースによる翻訳,最終的には常識ベースによる翻訳を確からしさという統一尺度で融合できることである。

Applying Expert System Development Tool Technology to Natural Language Processing
Yoshizo Aori
DEC Japan

3.3 知識のオンデマンドローデイング

大規模辞書を少ないメモリで効率よくアクセスするために，必要な知識をオンデマンドでロードするようにした。ロードは文字列中に新しい単語を検出した時点で実行され，その単語に関する知識群をメモリ上の知識ベースに組入れる。その知識群には単語の属性や訳のようなフレーム型知識だけでなく，成句や文型などのルール型知識も含めることができる。組み入れられた知識群は直ちに推論に使われる。(図2参照)

3.4 知識獲得ユーテイリテイ

市販の電子辞書を自然言語処理へ利用するために，テキストファイル構造の市販電子辞書を本ツールの知識構造に変換するユーテイリテイを開発した。変換は2フェーズで行う。第1フェーズは厳密に定義されたリスト構造への一般的変換，第2フェーズは本ツールの知識表現への部分的変換を行う。研究開発者は第2フェーズを変更することにより，電子辞書の利用の仕方，利用の範囲を容易に変更できる。この変換ツールの利用により，極めて少ない人的リソースで大規模な自然言語処理システムの構築ができるようになった。

4. 本ツールの評価

以下に本ツールの環境と評価結果を示す。

開発言語: VAX LISP(=Common Lisp)
開発規模: 10 K ステップ
本ツールの利用環境:
　VAX システム，
　日本語VMS V5.5 以降，
　日本語VAX LISP V3.1，
　VMSのISAM アクセス機能，
機械翻訳に使用した知識ベースの規模:
　形態素解析ルール数: 約100ルール
　文法ルール数: 約200ルール
　単語辞書の大きさ: 約100単語
上記知識ベース構築工数: 3 人月
翻訳例: 文法ルールに対応する約200個の例文，およびソフトウエアのユーザズガイドからの抜粋1ページについて実行したが，翻訳品質は非常に良好であった。(図3参照)
性能測定環境: 30VUPS の VAXシステム
1文当たりの翻訳CPU時間: 0.3 秒 - 7秒
所要メモリ: 処理系: 8MB，知識ベース: 1MB，作業域: 4MB以上
市販電子辞書からの知識ベース構築工数:
(5000 語の重要単語について品詞別訳と文型ルールを自動生成，一部手修正あり。) 1 人月

5. おわりに

3章で述べた幾つかの工夫により大規模な自然言語処理システムを効率よく構築できる見通しが立った。なお本ツールの適用領域は機械翻訳，談話理解などの自然言語処理だけてなく，コンピュータプログラムのソースコードの解析やテスト結果のログの解析など，文字列／単語列／概念列を解析対象とした極めて一般的なものである。今後の予定としては以下のことを考えている。

- o 性能，使い易さ，ユーザエラーへの耐久性の改善などツールとしての完成度向上
- o 常識ベースを利用した自然言語処理研究開発のための フレームワークの提供
  常識ベースの知識表現，常識ベースを利用するための後向き推論の追加など
- o クライアント，サーバ インタフェースによる一般アプリケーションへの自然言語処理インタフェースの提供 (=ミドルウエア化)
- o 最新のRISC プラットフォームへの移植

```
一般オブジェクト定義関数: ($o <親クラス名> <オブジェクト名> <属性リスト>)
        ;本ツールは文字,単語,概念,句,節などのすべての言語要素をオブジェクトとして扱う
単語オブジェクト定義関数: ($d <親クラス名> <見出し名> <概念名> <属性リスト> <訳リスト>)
属性オブジェクト定義関数: ($atr <親属性クラス名> <属性名>)
ルール定義関数: ($rule ($if ($seq <オブジェクト リスト1>))      ;<-- 検出すべき変数/定数(=オブジェクト)の順序列
                    ($then ($phrase <列名>)                    ;<-- 検出した列全体に付ける名前(=オブジェクト)
                        ($trans <列名> <オブジェクトリスト2>))) ;<-- 機械翻訳のための列順序変換関数
```

図1. 自然言語処理向け知識定義機能

```
"CONNECT"                                                    ;<-- 見出し
(block a
 ($d '$vt nil "CONNECT" '("&sahen" $now) '("接続"))          ;&sahen は訳の活用形属性, $now は時制属性
 ($d '$vi "CONNECT" "&vi_connect" '("&ra_5" $now) '("つなが"))
 .....
 ($cf $very_high)                                            ;ルールの確信度が非常に高いことを示す。
 ($rule ($if ($seq "CONNECT" '$obj "TO" '$np)) ($then ($phrase '$vp) ($trans '$vp '$obj "を" '$np "へ接続")))
 ..... )
```

図2. オンデマンドローデイングの単位の例

原文: The DECnet/SNA Data Transfer Facility software is a DECnet/SNA access routine that connects a VMS/DTF server node and its clients on a DECnet network to IBM MVS and VM client systems on an SNA network.

翻訳例: (DECnet/SNA データ転送 ファシリテイ ソフトウエア は DECnet ネットワーク の上の VMS/DTF サーバ ノード および そのクライアント を SNA ネットワーク の上の IBM MVS および VM クライアント システム に 接続する DECnet/SNA アクセス ルーチン です.)

CPU Time: 5.11 sec., Real Time: 6.02 sec. GC CPU Time: 0.47 sec., GC Real Time: 0.50 sec.

図3. 英日 翻訳に適用した例