

知識ベースを利用した自然言語処理システム

6Q-5

阿折義三
日本DEC 研究開発センター

1. はじめに

自然言語処理のボトルネックは解析対象文の曖昧性であり、従来は格文法や意味属性の工夫で曖昧性を解消しようとしてきたが、現実の自然言語を人間並みに正確に解析できるまでに至っていない。次のステップとして大規模例文知識ベースの利用が考えられているが、あらかじめ準備した例文知識ベースでは現実の多様な言語現象に対応しきれない。人間は所有する全ての知識を総合して言語の理解をしているだけでなく、困ったことに所有する知識を前提として自然言語を生成する。したがってコンピュータの場合も結局最後は常識、専門知識、そして文脈知識などあらゆる知識を総合的に利用しなければ満足な結果は得られないと考えられる。筆者は抽象化された一般知識を利用する方が例文ベースよりはるかに少ない知識で多様な言語現象に対応できるだけでなく、実は実現上もそれほど困難ではないことを主張する。

2. 多様な知識の総合的利用

自然言語処理への一般知識の利用という観点から考えると以下に示すようなAIや自然言語処理研究における伝統的知識表現が十分有効に利用できる。

言語知識

文法規則, 文法ヒューリスティック
慣用句, 一般例文
意味制約

一般常識/専門分野知識

概念階層, その他階層関係
三段論法的知識
分野固有単語およびその訳
文脈知識

重要な問題はこれらの多様な知識をどのように自然言語処理に統合的に利用するかである。筆者は1つの方法としてAIで一般的な方法である確信度に基づく曖昧性解消方法を提案する。1つ1つの知識に確信度を付与することにより多様な知識に対して利用可能性に関する統一的な規準ができ、前記すべての知識が統合的に利用可能となる。この際、確信度付与規準が重要な問題となるが、

○限定的表現, 一意性の高い表現には高い確信度,

○一般的表現, 多義的表現には低い確信度という規準によって多様な知識の利用可能性をクラス分けできる。たとえば文法レベルの知識は一般性があるが曖昧性が高いので確信度は最も低いレベルを付与する。一方, "Time flies like an arrow." を"時は金なり"と訳したい場合は、この文全体としてはかなりユニークな意味を持つと考えられるので、文全体に対してかなり高い確信度を付与すればよい。

知識が多様になり大規模になると解析対象の言語テキストに対して適用可能な知識は大幅に増大し、解析文全体の解釈可能な組み合わせは爆発的に増大する。したがって適切な競合解消機能なしには大規模知識ベースの利用はできない。筆者が開発したNLPシステムでは確信度に基づく競合解消機能を工夫することで、はじめて現実的性能になった。

3. 知識ベースを利用したNLP実現法

筆者が開発したシステムの実現法を許される紙面の範囲で紹介する。

3.1 知識表現

o パーマネント知識

どのような知識でも統合利用できることを示すために、概念階層、一般階層関係、一般ネットワーク関係、3段論法的知識、述語表現など考えられる知識表現をすべて利用可能とした。例として構成要素関係、所有関係を表わす知識表現を示す。

networkはclient, serverからなる(=構成要素関係: comp_of)。

```
($def_hi 'server 'comp_of 'network 0.9)
```

```
($def_hi 'client 'comp_of 'network 0.9)
```

MVSおよびVMはIBMのものである。(=所有関係:owned_by)

```
($def_hi "MVS" 'owned_by "IBM" 0.9)
```

```
($def_hi "VM" 'owned_by "IBM" 0.9)
```

o 一時知識(文脈知識)

解析対象テキストに現れる単語およびその一般化概念を文脈として記録/利用できるだけでなく、テキスト上の句を一般化した抽象化述語が記録/利用できる。またテキストを前記一般的知識表現に変換/記録でき、したがってテキストから獲得した一般知識もただちに文脈として利用できる。残念ながら紙面の都合で今回は詳細は紹介できない。

3.2 知識の利用法

重要な問題はテキスト現象と上記知識をどう結び付けるかであるが、筆者のシステムではテキストとのパターンマッチングに対する制約条件式によって実現した。具体例を以下に示す。

```
($if ($seq 'np1 "AND" 'np2 "ON" 'np3)
```

```
($memberp 'np1 'np2 'comp_of 'np3)) ...
```

'np1が's'client, 'np2が's'server, 'np3が's'networkの場合、前記networkに関する構成関係知識が

マッチし制約条件式である\$memberp関数は高い確信度を返す。

```
($if ($seq 'n1 'n2 "AND" 'n3)
```

```
($memberp 'n2 'n3 'owned_by 'n1))...
```

'n1がIBM, 'n2がMVS, 'n3がVMの場合、前記IBMに関する所有関係がマッチする。

機械翻訳への適用例を示す。下記例文のような複雑な係り受け関係を持つ文の構造を知識を利用することによって安定的に正しく決定できるようになった。ここで使用した知識は3.1および3.2で例示した知識だけである。一方、競合解消機能の改善により実行時間も5.1秒(1年前レポート)より1.37秒に短縮された。(30MIPSマシン使用)

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.

本システムによる翻訳文: (DECnet/SNA データ転送 ファシリテイ ソフトウェアは DECnet ネットワーク上の VMS/DTF サーバノードおよびそのクライアントをSNAネットワーク上のIBMの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: 1.37sec, Real Time: 1.90sec.

4. おわりに

人間並みのフレキシブルな自然言語理解ができる知識利用技術の確立は遠い先のことではなく、それは本アプローチでの細かい改善/工夫の積み重ねによって達成可能である。また必要なハードウェア処理能力は100MIPSのオーダーであり特別なハードはいらない。大規模知識ベースの構築技術がボトルネックとなり、そこに多数の研究開発者がチャレンジすることを期待する。

参 考 文 献

1) 阿折義三: 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, 情報処理46回全国大会, pp. 3-193(1993)