日本語処理基本システム（１）

2C-1

―全体構成―

長尾健司　菅野祐司　上田 謙一
松下電器産業株式会社　東京研究所

1．はじめに

最近、機械翻訳や質問応答、文書処理等の自然言語処理応用システムの開発者の間で、これらのシステムを支える基本部分の中に、共通に流用できる部分が少なくないという認識がでてきている。[1)]

言い換えると、そのような基本的なシステムを、あたかもツールのように予め用意しておけば、応用システム開発の労力は大幅に軽減されるということである。

本稿では、このような基本システムに要求される機能のうち、我々が構築中の主に文の解析機能を提供する部分について、その概要を報告する。

2．基本思想

システムを開発するに当たっての基本思想を以下に示す。

a）基本機能

べた書きの文を入力して、文の形態情報、構文情報、意味情報を出力する。

b）性能

（１）　**高精度**

日本語文の解析に特有である、解釈の曖昧性の問題に対処するため、可能性のある全ての解釈に対して、尤らしさの判定を、できるだけ文のグローバルな情報を用いて行うようにする。

（２）　**高速**

（１）を実現することは、とかく処理の高速性を損なうことにつながりがちであるが、解析の任意の段階で、複数ある解釈に含まれる共通の要素に対しては、処理を共通にすることにより、これに対処する。

（３）　**コンパクト**

共有する要素に対する処理の共通化は、おのずとデータ量の削減につながるが、これに加えて、各種のデータを表現する独自のシステム（処理系）を構築することにより、所要記憶量の低減をはかる。

3．全体構成

システムの全体構成を図１に示す。主要な構成要素の概略を以下に説明する。尚、形態素解析系についての詳細は２）を、自立語辞書検索系、構文解析系については３）を参照されたい。

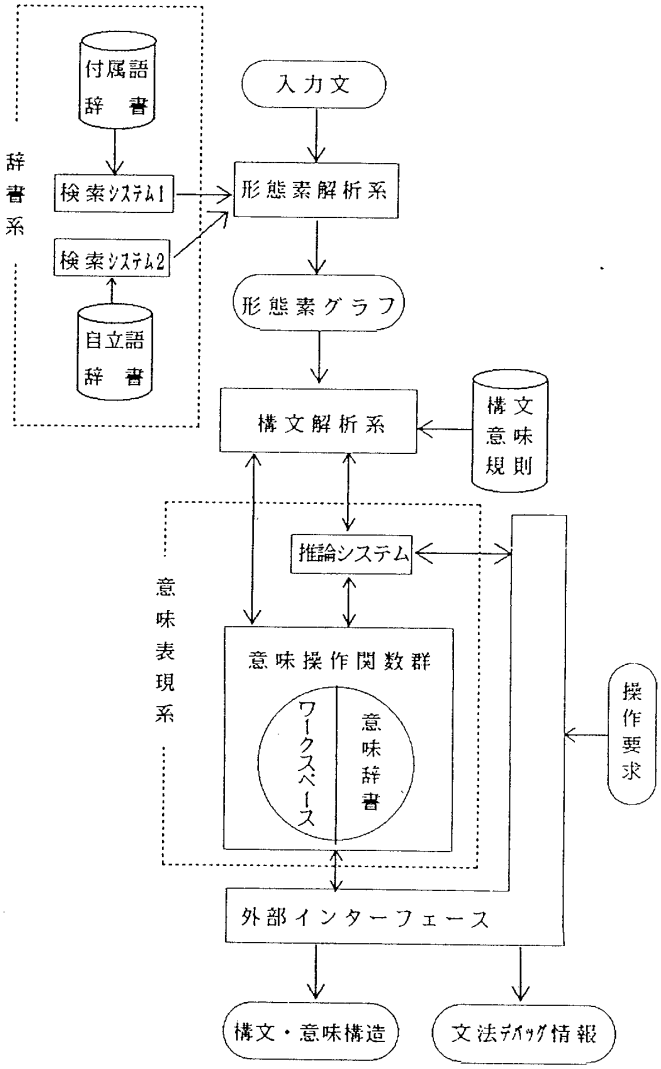

図１システム全体構成

Core System for Japanese Processing (1)
- Over all design -
Kenji Nagao, Yuji kanno, Kenichi Ueda
Tokyo Research Laboratory
Matsushita Electric Industrial Co., Ltd.

3－1．辞書系

辞書系は自立語辞書、付属語辞書、意味辞書及びそれぞれの検索部よりなる。自立語は自立語辞書と意味辞書にそれぞれ文法情報、意味情報を分担して格納する。これは文法情報は固定長データで表現することを想定しているのに対し、意味情報は3－4の意味表現系の提供する意味表現形式を用いて記述するため、可変長を想定していることによる。

又、付属語は語数が少ないため文法情報、意味情報ともに付属語辞書にまとめて格納するが、意味情報は自立語同様に意味表現形式を用いて記述している。

3－2．形態素解析系

形態素解析系は、日本語べた書き文を入力として受けつけ、可能な全ての解析を表現したグラフを出力する。

辞書検索による単語の切り出しの結果は、一般に複数あるため、形態素解析の問題領域はグラフ探索の問題として捉えることができる。入力文に曖昧性（文法的或いは意味的なものも含めて）が存在しない限り、正しい解析（正しい分かち書き）は一通りであり、即ちグラフ上ではひとつのパスを求めるだけでよい。しかし、実際には形態素解析のレベルでは、このような曖昧性を解消することは一般に不可能である。

そこで、本形態素解析系では、可能性のあるパス（解析）を全て探索して、後の構文解析や意味解析を経て解を絞りこんでいくという方針をとる。但し、ここで問題になるのは、これを単なる横型の探索により実現したのでは非常に大きな処理時間を要するという現実的な問題が表に出てくるということである。

本形態素解析系は、これらの問題に対処すべく、発見的なグラフ探索アルゴリズムを、指定した範囲内のコストを持つゴールパスを全て取り出すことができるように改良し、これを基本アルゴリズムとして用いる。

ここでグラフのノードに対応するものは、個々の単語ではなく、同一の入力文字列に対応し同一の品詞を持つ単語の集合であり、このことは処理効率の向上にかなり貢献している。これは、3－4に述べる意味表現形式のバックボーンによってサポートされたデータ構造である。

コストは、解析の確からしさを反映するものでなければならないが、現在のバージョンでは、文節数最小法と単語の頻度情報を考慮したものを採用している。プロトタイプ版の評価実験では精度、処理速度ともに実用的な性能を確認している。

3－3．構文解析系

構文解析系は、形態素解析系の出力結果である、可能なすべての分かち書きを表現したグラフ、を入力として受取り、文の全ての可能な解釈（文法的・意味的情報）を抽出し、3－4の意味表現系の提供する形式により表現し、外部インターフェースをとうして出力する。文法規則は、ＣＦＧルールと、これと対をなす意味規則により記述しており、意味規則は意味表現系の提供する意味操作関数を用いて記述する。処理方式の特徴としては、解析の任意の段階で全ての可能性を並列に処理し、共通の部分構造については一回の処理で済ませるようにしていることがいえる。

3－4．意味表現系

意味表現系は、意味操作関数群と推論システム及びこれらを支える独自の処理系により、文の意味構造を操作し管理する機能を提供する。意味構造は以下に示すような形式の意味フレームにより表現する。

＜意味フレーム＞：：＝（＜名前＞＜位置＞
＜値１＞＜値２＞・・・＜値ｎ＞

ここで、名前は文字列又は数値であるが、その中に”ＯＲ”という特殊な名前を用意しておりこれによって複数の意味フレームの実体を一つの意味フレームの構造で表すことができる。又、＜値ｊ＞は再び意味フレームよりなる。

意味構造をリストで表現した例を以下に示す。

```
（動詞　（０）（２　３）　．
　（（自／他　．（（自））
　　（ＯＲ　．（（ｎｉｌ．（（表記　．（（来る））
　　　　（活用型　．（（カ変））
　　　　（述語素　．（（Ａ））
　　　（ｎｉｌ．（（表記　．（（繰る））
　　　　（活用型　．（（五段））
　　　　（述語素　．（（Ｂ））））
```

4．おわりに

日本語処理基本システムの概要について述べた。
本システムの最大の課題は、意味規則を意味操作関数を用いて記述する際に、現在は、複数の解釈を考慮しなければならないという点を改善することである。

5．参考文献

1）鈴木：”ＭＥＬＣＯＭ　ＰＳＩ上の自然言語処理開発支援環境について”，
第３４回情報処理学会全国大会（１９８７）

2）長尾：
”文節数最小法・形態素解析の実現法について”，
第３５回情報処理学会全国大会（１９８７）

3）菅野：”日本語処理基本システム（２）”，本大会