自 然 言 語 処 理 79－2
（1990．9．21）

統合パーサによるノイズを含んだ文の理解

劉 学敏　西田 豊明　堂下 修司
京都大学工学部情報工学教室

本稿では，我々が作成中のノイズを含んだ文に対する統合的自然言語理解システムについて報告する．このシステムは統合パーサという推論管理システムの上で構築される．統合パーサは，推論過程の論理的な整合性を維持する機能と，処理を確からしい方向に誘導する機能の融合によって構成される．システムは統合パーサの機能を利用して，自然言語処理の過程における仮定の生成とそれに基づく推論，仮定の切換えによる言明集合の変更，そして，仮定の確からしさによる推論方向の選択などを効率的に行なうことができる．また，本システムは，入力誤りを効率良く訂正するために，断片的な解析という解析方法を用いる．この方法では，与えられた入力文に対して，システムはまずそれをいくつかの解析断片に分けて，その中の誤りのない断片を優先して処理する．次いで，システムは誤りのない断片に関する解析結果を利用して，誤りを含んだ断片の構文構造と意味概念を予測し，その予測結果に基づいて，入力誤りの訂正を行なう．このような解析方法によって，入力誤りによって生じた不確定性による無駄な処理を軽減することが期待できる．

Understanding Erroneous Sentence by an Integrated Parsing Engine

Xuemin LIU　Toyoaki NISHIDA　Shuji DOSHITA
Department of Information Science, Kyoto University

This paper describes a natural language understanding system which is designed for processing a sentence that contains some input errors. The system is based on an integrated parsing engine IPE. The IPE is constructed for maintaining consistency of the belief set and guiding the processing to the most plausible direction. Using IPE as a subsystem, the natural language understanding system can do a guided search in an easy way. In this system, we use a new analyzing method called fragmental analysis for handling input errors involved in the given sentence. When a sentence is given, the system divides this sentence into some fragments, each of them is marked as a fragment which contains some input errors, or a fragment which contains no input errors. The fragments which contain no input errors will be analyzed first. Using the results got from the analyzing about these fragments, the system can guess the syntax structures and the meanings about the erroneous fragments and generate various kinds of information which can be used to correct the input errors in an efficient way.

1 はじめに

自然言語を計算機によって処理する場合, その自然言語処理システムへの入力には, ある程度の入力誤りを含んでいることが多い. 例えば, キーボードによる入力の場合, 作業者のミスタイプによる入力誤りがある. また, 音声による入力あるいは手書き文書の画像による入力の場合, まず, 入力に対して, 音声認識処理あるいは画像処理を行なって, その出力を自然言語処理システムに渡すことが普通である. この場合, 現在の音声認識の認識率あるいは画像処理の精度によって, その出力には多数の認識誤りが含まれている. この認識誤りは自然言語処理システムにとって, 一種の入力誤りとして考えられる. このような入力誤りは一般的に文字レベルで発生するので, 本稿では, このような入力誤りをノイズと呼ぶ.

入力にノイズが含まれている場合, 自然言語処理システムは単に辞書引きによって入力文の中から単語を検出できないことが多い. この場合, 自然言語処理システムは入力に対して何かの修正を行なわなければならない. この修正によって, 入力誤りを訂正してはじめて処理が構文解析と意味解析に進められる. しかし, 入力誤りを訂正することによって, 一つの入力位置から数多くの単語候補が生成される. これらの単語候補を正しい単語に絞り込むのは非常に困難である. また, これらの単語候補の存在はシステムの構文解析と意味解析にとって大きな負担となる.

入力に含んだ誤りを訂正する方法に関して, いくつかの研究があげられる [7, 8, 9]. しかし, これらの方法では, 入力を訂正する際に, 単語レベルの知識だけを利用することが多かった. 一方, 自然言語自身の曖昧性によって, 単語レベルの知識だけでは諸多の単語候補から一意に単語を確定できない場合が多い. この問題を解決するため, 自然言語処理システムは構文的な知識と意味概念に関する知識, そして, 今までの処理によって得られた部分的な解析結果と文脈情報などを統合して利用しなければならない.

本研究ではいま, 我々が提案した統合パーサ [3] という方法をベースにして, ノイズを含んだ文に対する自然言語処理システムの作成を行なっている. このシステムにおいて, 我々は自然言語処理過程を仮説に基づく推論過程として取り扱い, このような推論過程を効率よく実現するため, 統合パーサの整合性維持する機能と仮定の確からしさによって推論方向を誘導する機能を利用する. システムは処理の過程で, 各レベルの解析を並行して行ない, その解析結果の統合によって, 自然言語の曖昧性や入力誤りによる不確定性を早期に解消することができる. これによって, 無駄の少なく, より人間に近い処理方法を追求する.

また, ノイズによる入力誤りを訂正するために, 本システムでは断片的な解析方法を利用する. この方法では, システムは入力文をいくつかの解析断片を分けて, まず, 入力誤りが含まれていないと想定する断片の解析を優先して行ない, この解析結果を利用して, 入力誤りを含んだ断片に対して訂正処理を行なう. 最後に, 各断片の解析結果を結合し, 文に対する最適な解釈を生成する.

本稿では, このシステムの構成, 基本的な解析方法とノイズによる入力誤りの対処方法などについて報告する.

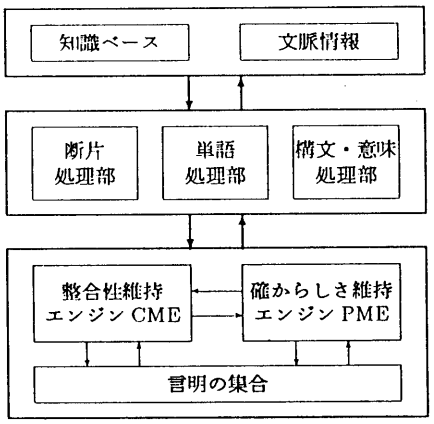

図 1: 統合パーサによる自然言語理解システムの構成.

2 システムの構成

本研究では, 図 1に示すように統合的自然言語理解システムを構築する.

ここで, システムの各部分と入出力について説明する.

2.1 入力と出力

本システムは, 我々の研究室で開発中の音声認識システムの処理結果を入力にする. この音声認識システムは不特定話者の連続音声を入力として音声認識処理を行ない, その処理結果を音素 Lattice の形で出力する (各音素ごとに第三候補まで出力する). 本研究では, この音素 Lattice の第一候補音素の列を仮名文字列に変換して, システムに与える. また, 音声認識の結果を有効に利用するために, 各文字に対して, 可能な候補も生成し, その文字に付加する. 従って, 入力は実際に文字の Lattice とも考えられる. ここで, 音声認識システムに対して, その出力は仮名文字列に変換可能であることを要求する. また, 促音も特別な音素として出力することを要求する.

現在の音声認識の精度がまだ不十分であるため, 音声認識システムの出力には相当の認識誤りが含まれている. その種類として, 音素の誤認識, 音素の挿入と音素の脱落の 3 種類がある. これらによって, 自然言語処理システムへの入力には次のような入力誤りが含まれている.

- 置換誤り : 母音の誤認識, 或は母音の前における子音の誤認識, 脱落, 挿入によって生じる.
- 挿入誤り : 母音の挿入または子音と母音の同時挿入によって生じる.
- 脱落誤り : 母音の脱落によって生じる.

ただし, 本システムでは, もともとの音声入力には誤りが含まれていないと仮定する. つまり, 上の入力誤りの全ては音声認識の精度によって生じるとする.

システムからの出力は入力文に対する意味解釈で,意味表現リストの形をとる.この意味表現リストは文の種類(例えば,説明文,命令文,質問文など)と文の構文構造に従う意味構造からなる.また,文の中の各単語に対して,その意味概念を明示した形で出力する.

本システムは文を処理単位にする.一方,自然言語には,文の意味は文脈情報に依存することが多いから,システムは処理の過程で先行文脈に関する情報も参照する.このため,システムは一つの文の処理を行なった後,それ処理結果を文脈情報として記憶する.

2.2 知識ベース

システムの知識ベースには次のようなものからなる.

- 文法規則とその処理規則
  日本語の構文規則と各規則に関する解析処理の指定を格納する.
- 辞書
  辞書として,単語辞書と意味辞書がある.単語辞書として,単語とその品詞また各品詞に対応する意味概念などを登録する.一方,意味辞書は各意味概念の間の関連関係などを登録する.
- 対象世界に関する記述
  名詞で表す概念と対象世界に存在するものの対応関係,また,動作で表す動作は対象世界において実行可能であるかどうかなどの知識を格納する.
- 音声認識に関する知識
  音声認識処理において,各子音と母音の認識率,また,音素の挿入や脱落はどのような場合に発生しやすいかに関する知識を格納する.

2.3 統合パーサ

本研究では,自然言語理解を仮説に基づく推論として取り扱い,システムは統合パーサと呼ぶ推論管理機構の上に構築する.統合パーサは図1の下部に示すように整合性維持エンジンCME(Consistency Maintenance Engine)と確からしさ維持エンジンPME(Plausibility Maintenance Engine)の融合によって構成されている.

(1) 整合性維持エンジンCME

CMEは言明と仮定または言明の間の依存関係を利用して,言明の集合の論理的な整合性を維持する.推論の過程において,CMEは推論システムの要求に応じて,仮定を生成したり,矛盾した仮定の撤去とそれに伴う作業記憶の切換えを行なったりする.これらの点について,CMEは従来の推論管理システム例えばTMS[1]やATMS[2]と同じような機能を持っている.一方,TMSやATMSと異なって,CMEは言明の集合の整合性を維持するだけではなく,システムの可能な推論方向も管理する.

仮説推論において,推論方向は仮定の組合せに対応する.ATMSに従って,我々は仮定の組合せを環境と呼ぶ.自然言語の曖昧性が多いため,特に,入力誤りによる不確定性のため,処理の過程で数多くの仮定が生成されることが多い.一方,生成された仮定の間に排他性があり,その全ての組合せを考える必要がない.そこで,我々のCMEは推論可能な環境だけを取り出して,一つの木の形

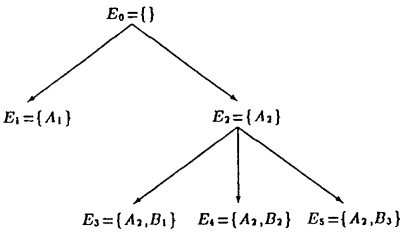

図 2: 環境の木の例.

で管理する.この木を環境の木と呼び,仮定の生成に伴って動的に構成する.環境の木の任意の葉節点は一つの可能な処理方向を指示する.これによって,必要性のない環境における無駄な推論を避けることができる.

(2) 確からしさ維持エンジンPME

PMEはシステムが処理の過程で生成した仮定の確からしさに関する情報を維持し,処理を最も確からしい方向に誘導する.PMEは仮定の確からしさを利用して,環境の確からしさを決定する.また,環境の確からしさと環境の木を利用して,一つの信念ネットワーク(Belief Network)[4]を構成する.これを利用して,PMEは環境の木から最も確からしい環境を選出できる.また,矛盾が生じた場合,PMEは信念ネットワークを更新し,仮定の確からしさを維持する.PMEによって,システムは確からしい推論環境の処理を優先して行なう.これによって,無駄な推論を避けることができ,また,唯一な解釈を導出できない場合,最も確からしい解釈を出力できる.

統合パーサを利用する自然言語処理システムにおいて,その推論処理は常にPMEによって選択された最も確からしい推論環境において行なう.また,処理の過程で得られた処理結果は全てCMEによって管理される.処理の過程で曖昧性や不確定性が生じると,システムはCMEによって過程を生成し,環境の木を拡張する.同時に,システムは生成された仮定に対して,その確からしさを決定し,PMEに与える.PMEはこれを利用して,環境の木から確からしい推論環境を選択する.処理によって現在の推論環境が矛盾したことがわかったら,システムはCMEによって矛盾した環境を削除し,また,PMEによって仮定の確からしさを修正する.その後,PMEは修正された確からしさに基づいて,再び新しい推論環境を選択する.システムの処理はこ環境に移って続ける.これを示すために,次の推論過程を考えよう.

例 1. (a) $X_1 \Rightarrow Y_1 \oplus Y_2$,
(b) $Y_1 \Rightarrow Z_1 \oplus Z_2 \oplus Z_3$.

ここで,ある曖昧性(a)によって,既存の言明X_1から排他的に新しい言明Y_1とY_2を導出することができ,さらに,曖昧性(b)によって,言明Y_1から,排他的に言明Z_1, Z_2, Z_3を導出できると意味する.いま,曖昧性(a)に対処するため,仮定A_1とA_2が生成され,また,曖昧性(b)に対処するため,仮定B_1とB_2とB_3が生成されたとする.これらの仮定を利用して,生成された環境の木は図2に示す.

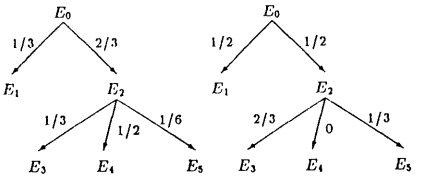

図 3: 信念ネットワークの例.

上の五つの仮定から生成可能な推論環境の数は 2^5 である.ATMS において, これらを全て考えなければならない. 一方,CME において, 環境の木の葉節点だけを考えればよいから, 考慮すべきものはわずか四つである.

さらに, システムが知識と現在までの処理結果によって, 仮定 A_1 と A_2 に確からしさ 1/3 と 2/3 を与え, また, 仮定 B_1,B_2,B_3 に確からしさ 1/3,1/2,1/6 を与えたとしよう.PME はこれらの確からしさを利用して, 図 2に示す環境の木から, 図 3(a) のような信念ネットワークを構成できる. これによって,PME は環境 E_4 を確からしい環境として選出できる. また, その後の処理において, 環境 E_4 に矛盾が起きたら,PME は仮定の確からしさを図 3(b) のように修正する. 修正した後, 環境 E_3 は確からしい環境になる. その後の処理は E_3 において続ける.

2.4 推論処理部

システムの推論処理部は図 1の真中の部文に示すように, 断片処理部と単語処理部そして構文・意味処理部と三つの部分からなる. 本システムは, 入力誤りを効率的に訂正するため, 断片的解析方法を用いる. 断片処理部は断片の生成, 管理などの処理を行なう. 一方, 単語処理部は単語の検出とそれに伴う単語の品詞や意味概念の決定などの処理を行なう. また, 構文・意味処理部は入力文に対して, 構文解析と意味解析を行なう. これらの各部分の詳細について, 次の章から述べる.

3 断片的解析

自然言語処理において, その解析は left-to-right の形で進めることが多い. しかし, 入力に誤りがある場合, この方法では, 誤りが発見されるとき, その後ろの部分はまだ解析していないから, 誤り訂正のために利用できる情報は少ない. 特に, 入力誤りが文の先頭で発生する場合, 誤りの訂正は一層難しくなる.

一方, 与えられた入力文にはノイズが含まれているといっても, その中の全ての単語にノイズを含んだことは少ない. もしシステムが入力文に対して, 誤りのない単語を優先して処理し, その処理結果を利用して, 誤りのある部分の構文構造と意味概念を推定できれば, 入力誤りの訂正を効率的に行なうことが期待できる.

そこで, 本研究では, 従来の left-to-right の解析方法のかわりに, 断片的解析という方法を採用する. この方法では, まず, 与えられた文をいくつかの断片に分けて, 各断片において, 部分的な解析を行う. 次に, 各断片の部分的な解析結果を結合して, 全体的な解釈を構成する. 解析の過程で, システムは誤りのない断片を優先して解析を行なう. こうすると, 入力文の全体的な構造を早く把握することができ, 入力誤りの訂正に対して, 有用な情報を提供できる.

3.1 断片分け

断片は基本的に単語と推定する文字の部分列である. 但し, 日本語には, 同音異義の単語が多いから, 本研究では, 断片のレベルではそれを区別しない. また, 助詞は通常独立的な意味を持っていないから, 本研究では助詞に対して, 断片を生成せず, それを前の断片に帰着する.

断片は単語検出操作によって分けられる. 辞書引きによって単語を検出した場合, システムは検出された単語に対して, その語境界情報を利用して, 一つの断片を生成する. 一方, 辞書引きによる単語検出が失敗した場合, システムはまず現在の単語検出位置を記憶し, 次に, 新しい検出位置を見つける. その新しい検出位置と記憶された検出位置によって, 一つの新しい断片を生成する. 新しい検出位置を次のように見つける.

(1) その位置から辞書引きによって単語が検出できる.

(2) その位置の前の部分文字列は指定された特徴パターンである.

ここで, 特徴パターンは日本語において頻繁に出現する文字列である. 本研究では, 下に列挙するものを特徴パターンとして利用する.

- 助詞
- 動詞〃する〃とその活用形
- 動詞の語尾の活用形
- 形容詞の語尾の活用形

一方, ノイズのため, 上の特徴パターンにも入力誤りが含まれている可能性がある. これに対処するため, 本研究では, 特徴パターンのの検出は単語マッチングという方法 (4.1に参照) によって行なう.

上の方法によって分けられた断片は, 次の二種類に区分できる.

(1) ノイズが含まれていない断片

この種類の断片は辞書引き操作によって単語検出できる断片である. システムはこの種の断片に関する解析を優先して行なう. このような断片から得られた解析結果を利用して次のノイズのある断片の構文構造と意味概念を予測する.

(2) ノイズが含まれている断片

この種類の断片は辞書引きによって単語の検出ができない断片である. このような断片において, 単語を検出するために, システムはまず入力に対して何かの方法によって修正しなければならない.

ここで, 考えなければならないのは, ノイズによって, ある単語は別の単語になることがある. これに対処するため, 本研究では, ノイズのない断片に対して, もしこの断片から得られた単語は処理によって拒否されるならば, システムはこの断片をノイズのある断片に変更する. こうすると, 例えノイズによって, 誤った単語を構成して

も，システムはこのようなノイズに対処することもできる．

例2. あしたかぎをひらきます．
(明日会議を開きます.)

ここで，単語［かぎ］は実に［会議］から脱落誤りによって構成されたものである．この文の処理において，単語［開きます］を検出するまで，システムは［かぎ］を含む断片をノイズのない断片として取り扱う．一方，動詞［開く］を検出した後，意味解析を行なうと，［鍵］は［開く］の動作対象にならないから，矛盾が生じる．この場合，システムは単語［かぎ］あるいは［開く］どちらかを拒否しなければならない．例えば，短い単語をまず拒否対象として選択すれば，システムは［かぎ］に対応する断片をノイズのある断片に変更し，再びこの断片から単語の検出を行なう．

3.2 断片の管理

断片的な解析方法において，互いに重なっている断片同士の間の排他性のため，断片の選択する必要がある．また，構文解析と意味解析を行なう場合，システムは断片が他の断片と隣接するかどうかを判断する必要がある．

一方，ノイズのため，実際に隣接した単語は挿入誤りによって隣接しなくなることがあるし，また，脱落誤りによって，重なっている断片が実際に意味上に隣接する可能性がある．従って，まず，断片の間の排他的な関係と隣接関係を判断するための規準を決めなければならない．本研究では，

(1) 一つの断片が文字列として他の断片の部分文字列になる場合，あるいは，入力には連続して発生する脱落誤りの文字数をnとする場合，二つの断片が互いに重なった部分の長さはnより大きいならば，これらの断片を排他的であるとみなす．また，

(2) 入力には連続して発生可能な挿入誤りの文字数をnとする場合，離れた二つの断片の間の文字数がn以下であるならば，これらの断片を隣接していると見なす．

ここで，挿入誤りと脱落誤りの連続発生可能な文字数nは予め知識としてシステムに与える.(本研究では当面n=2にする)

断片の管理の目標として，下の二つの機能を達成することである．

(1) 断片の間の論理的な整合性の管理

断片の間の排他性の保持，すなわち，互いに排他的な断片を同時に採用しないことを保証する．また，既に処理が失敗した断片を再び処理しないことを保証する．

(2) 断片の集合 (系列) の確からしさの管理

排他的な断片が存在するため，与えられた入力文からいくつかの断片の系列が生成できる．これらのうち，最も確からしいものを取り出す．

これらの管理を統合パーサによって行なう．互いに排他的な断片を生成するとき，システムはCMEによって，各断片に対して，仮定を生成する．これらの仮定を利用して，環境の木を拡張する．また，各断片に対して，確からしさを設定し,PMEに与える．よって,PMEは断片の最も確からしい系列を選択することができる．

4 単語解析

単語解析は断片の内で行なわれる．単語解析において，システムはまず単語検出を行なう．次いで，検出された各単語の品詞と意味概念を調べ，不整合な単語を排除する．残る単語候補に対して，仮定を生成し，確からしさを設定する．

4.1 単語検出

誤りを含んだ入力から単語を検出するための方法は二つある．一つは単語マッチングという方法である，この方法では，入力文字列を辞書内の各単語と比較して，その類似度の高いものを単語候補にする．もう一つは修正補正と言う方法である．この方法の要旨として，入力を単語 (列) になるように適当に修正する．単語マッチング方法は入力誤りの種類に関係なく，与えられた文字列に類似する全ての単語を検出できる．しかし，辞書の中の単語の数が多い場合，かなりの処理時間がかかる．それに対して，修正補正方法は誤りの種類によって異なる対処方法を利用することが可能であるから，効率的な処理が期待できる．しかし，有力な情報がない場合，入力をどのように修正するかを決定するのは難しい．

一方，入力には，各種類の誤りが同じ確率で生じるのではない．その発生する原因を調べると，一番多いのは置換誤りと想定できる．また，母音の持続時間が長いから，それが脱落された可能性は挿入された可能性より低い．このため，本研究では，誤りを含んだ部分に関する単語検出を三つの場合に分けて，各々に異なる方法で対処する．

(1) 置換誤りだけの場合

この場合，単語の長さに関する情報が残っているから，比較的に容易に対処できる．これに対して，システムは置換補正，即ち，入力文字列の文字を適当に置換することによって誤りを訂正する．入力文字をどのように置換するかを決めるために，システムは各文字に付加された候補文字を利用する．

(2) 置換と挿入誤りが同時に発生する場合

この場合，システムはまず削除補正 (適当に文字を削除すること) によって，入力文字列を修正し，次に，修正した文字列を置換誤りだけを含んだ文字列とみなして,(1)のように単語検出を行なう．ここで，削除対象となる文字を決めるために，音声認識に関する情報 (例えば，母音の挿入誤りの発生特性など) を利用する．

(3) 置換，挿入，脱落誤りが同時に発生する場合 この場合，脱落誤りによって，情報の紛失は大きいから，修正補正の方法によって対処は困難である．本研究では，この場合の単語検出は単語マッチング方法によって行なう．一方，単語マッチングによる処理速度の低下を避けるため，本研究では，まず，現在の断片と隣接する断片の解析結果を利用して，出現可能な単語集合を予測する．次に，その予測された単語集合に対して，単語マッチング操作を行なう．これについて，第6章で述べる．

4.2 単語候補の絞り込み

前節で述べた単語検出のいずれの場合に，複数の単語が検出されることが予想できる．単語候補集合を絞り込むため，システムはまずそれぞれの単語に対して，その品詞を調べる．構文的な知識と隣接断片の処理結果を利

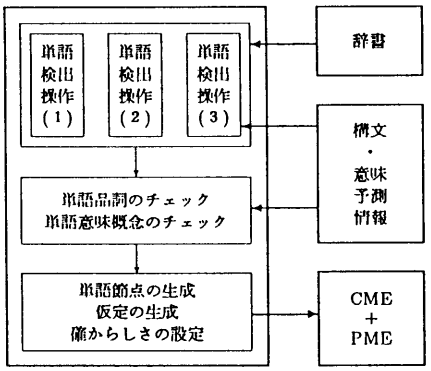

図 4: 単語処理部の処理過程の概要.

用して, その品詞が期待されていない単語を排除する. 次いで, 残る単語に対して, その意味概念を生成する. もしこの時点でまた複数の単語候補が存在すれば, システムはこれらの単語の各々に対して, 前述の断片管理の同じように, 仮定の生成そして確からしさの設定を行なう. また, 生成された仮定を利用して, 局所的な環境の木を生成し, これによって, 断片内の処理過程を管理する.

単語候補の確からしさは次のようなことの統合によって決定する.

- 単語の中の誤った文字数,
- 単語の品詞と隣接断片の関連の強さ,
- 単語の意味概念と先行文脈または隣接断片の関連の強さ.

単語解析によって得られた処理結果を単語節点というデータ構造で表す. 単語節点には, 単語の品詞, その意味概念, また, 単語を構成する文字リストなどの情報が記録される.

上に述べたノイズのある断片における単語解析の過程の概要を図 4に示す. ここで, 単語検出操作 (1) ～ (3) は各々前節で述べた単語検出の三つの場合に対応する.

5 構文・意味解析

自然言語の曖昧性の単にそれが発生した解析レベルの知識だけでは解決できないことが多い. 特に構文的な曖昧性は意味解析を行なうまで解消するのは不可能であり, また, 意味解析を行なう場合構文的な知識と解析結果を利用する必要がある. このため, 本研究では, 構文解析と意味解析を併せて一つの処理レベルにして, これらの解析処理を並行して行なう. この処理を構文・意味解析と呼ぶ.

構文・意味解析は統合パーサによって選択する断片の集合 (系列) に対して行なう. この解析において, システムはまず現在の推論環境で成り立つ構文要素を利用して, 与えられた文法に従って, 新しい構文要素を生成する. 次いで, その構文構造に従って, 対応する意味表現を生成する. ここで, もっとも基本的な構文要素 (品詞) とその意味表現は単語節点によって生成する.

構文解析と意味解析の過程でいろいろな曖昧性が生じることがある. しかし, 構文解析と意味解析は基本的に検出された単語の列に対して行なうので, これらの曖昧性に対して, 入力誤りのない場合と同じように対処できる. 一方, 本研究では既にノイズのない入力に関する統合的自然言語処理システムを実現した [10]. 従って, この場合, 参考文献 [10] で述べた構文解析と意味解析の場合の曖昧性に関する基本的な対処方法を利用すればよい.

構文・意味解析によって得られた結果を一つのペアで表現する.

＜構文要素 ・ 意味解釈＞

ここで, [構文要素] は構文解析によって得られた部分的パーザ木に対応し, [意味解釈] はその部分的なパーザ木のの意味解釈に対応する.

解析の過程において, このペアの二つの部分のどちらか一方だけが空であることは許される. この場合, これを未確定なペアと呼ぶ. その未確定する部分は解析の進行によって, より多くの情報が得られるときに決める. 例えば, 意味解析をある理由によって遅延しなければならない場合, このような未確定なペアを利用すれば良い.

例 2. らえしゅのすいようびにけんきゅうかいがある
(来週の水曜日に研究会がある)

ここで, 単語 [水曜日] で指す日を決めるために, その前の修飾部分の解析結果が必要である. しかし, 前の単語に入力誤りがあるので, 単語 [水曜日] の解析は先に行なうことになる. このとき, その指す日を唯一に確定できないため, システムはこの解析結果をしばらく次のように記録する.

＜ NOUN ・ ϕ ＞

その後, 例えば誤りの修正によって, 単語 [来週] が検出されるとき, あるいは, 前の単語が [水曜日] の修飾語ではないことを判明する時, システムは単語 [水曜日] に対して意味解析を再開する.

一方, 与えられた断片の系列の中で, 単語処理をまだ行なわれていない断片 (実は誤りを含んだ断片) が含まれている場合, システムはこのような断片に対して, 既に処理した断片の解析結果を利用して, その構文構造と意味構造を予測することが必要である. この予測結果も未確定なペアで表現する. これについて, 次の章で述べる.

6 予測情報の生成と利用

入力誤りを訂正するために, 誤りのある断片に対して, 予測情報を生成することは重要である. 本システムにおいて, 予測情報は単語解析と構文・意味解析の二つの処理レベルにおいて生成する.

単語解析レベルの予測は入力誤りの発生原因に対する予測である. その予測結果に基づいて, 入力を修正する. 第4章で述べたように, 本研究では置換, 挿入, 脱落誤りが同時に発生する場合, 単語マッチング方法によって単語の検出を行なう. また, 置換, 挿入誤りを同時に発生す

NOUN = ADJE + NOUN	(1)
NOUN = NPNO + NOUN	(2)
NOUN = SENT + NOUN	(3)
NPNO = NOUN + PPNO	(4)
CASE = NOUN + POST	(5)
SENT = VERB	(6)
SENT = ADJE	(7)
SENT = NOUN + ENDV	(8)
SENT = CASE + SENT	(9)
SENT = SENT + ENDV	(10)
SENT = ADVE + SENT	(11)

記号の説明: NOUN= 名詞(句), ADJE= 形容詞,
NPNO= 連体修飾句, SENT= 文, PPNO= 助詞"の",
CASE= 格, POST= 助詞, VERB= 動詞, ENDV= 助動詞,
ENDP= 終助詞, ADVE= 副詞.

表 1: 構文規則表

る場合,削除補正によって,それを置換誤りに変換する.従って,入力の修正は主として置換誤りに対する訂正である.一方,置換誤りを訂正する時,単に音素 Lattice だけでは正しい候補文字を生成できないことがある.この場合,本研究では,音声認識処理の認識率を利用する.すなわち,その認識率を一種の条件付き確率とみなして,それを利用して,事後確率を計算する.その事後確率によってもとの音素を予測する.

構文・意味解析レベルにおける予測は誤りを含んだ断片の構文構造(主として単語の品詞)と意味概念に関する予測である.その目標として,誤りのある断片に出現可能な単語集合を取り出すことである.本システムではこの予測を次のように行なう.

まず,いまの構文解析結果を利用して,誤りのある断片の可能な構文構造を予測する.その予測の結果を未確定なペア

＜構文構造 ・ ϕ＞

によって表す.次は,未定の意味概念 ϕ を決めるために,意味解析の結果によって,この断片の可能な意味概念を予測する.この予測過程を次の例によって示す.

例 4. かいぎしつきってくたさ.
(会議室 に 来て下さい)

この文において,単語[に]が脱落され,そして,[来て下さい]と言う断片にノイズが含まれている.いま,システムはこの断片を処理しようとする.システムはまず[会議室]の後続単語の品詞特性を予測する.本システムは与えられた文法に従って,各種類の構文要素の前後に出現可能な構文要素を予測する機能を持っている.　例えば,表 1に示す文法を利用する場合,名詞の後ろに出現可能なものは(PPNO POST ENDV)であり,助詞または格の後ろに出現可能なものは(VERB ADJE NOUN RENT SENT CASE ADVE)である.この予測によって,[会議室]の後続単語は助詞あるいは助動詞であることがわかった.一方,この断片の長さは 6 文字であるから,それは一つの単語である可能性が高い.　従って,もし助詞が脱落されたら,この断片の可能な品詞として,名詞,形容詞,動詞,そして副詞のいずれかである.一方,もしこれは名詞あるいは副詞ならば,全体的に一つの文にならない.従って,可能性は動詞と形容詞だけである.　そこで,システムはまず構文的な予測情報として,次の仮定と構文・意味ペアを生成する.

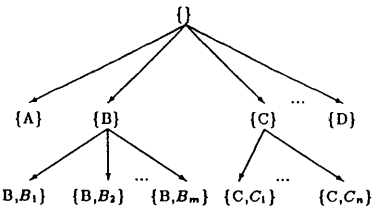

図 5: 例 4 の処理における局所的探索を制御するための環境の木.

A: 次の単語は助動詞: ＜ ENDV ・ ϕ＞
B: 次の単語は動詞: ＜ VERB ・ ϕ＞
C: 次の単語は形容詞: ＜ ADJE ・ ϕ＞
D: その他

次にシステムは上の仮定に基づいて,さらに単語の意味概念を予測し,可能な単語候補を取り出す.例えば,動詞ならば,次のようなものを生成し,

B_1: 会議室に対する動作: 締める,開ける,行く,来る,予約する...
B_2: 会議室の状態: 締めている,空いている,...
.....
B_m: その他

また,形容詞ならば,次のようなものを生成する.

C_1: 会議室の属性: 広い,暗い,明るい,...
.....
C_n: その他

最後,システムはこの予測によって得られた単語集合の中で,単語マッチング方法によって単語の検出を行なう.単語の意味概念から関連する単語を取り出すために,本研究ではいま,Waltz らのマイクロ素性[5]と似ているいるような意味概念の関連ネットワーク[10]を利用している.なお,本システムの関連研究として,本研究室では,Schank らのスクリプト方法[6]を用いて,単語の意味概念から,関連した単語集合を予測する研究を行なっている.これを将来このシステムに組み込む予定である.

上の処理過程は図 5に示す環境の木によって制御する.

7　解析過程の制御

図 1に示すように,システムの推論部は断片処理部と単語処理部そして構文・意味処理部の三つの部分からなる.システムは処理の過程で,まず,断片処理部によって生成された断片系列の集合から一つを選んで,次に,この

断片の系列に対して，単語解析と構文・意味解析を行なう．

単語処理部と構文・意味処理部は各々統合パーサによって管理されている言明の集合の上で処理を行ない，その処理によって得られた結果はまた統合パーサによって管理するので．各処理部は他の処理部に対して制御上の制約は与えない．従って，単語解析処理と構文・意味解析の処理順序によって，システムはボトムアップ的な処理方式とトップダウン的な処理方式と両方とも可能である．一方，どのような処理方式がよいかは静的に決められない場合があるから，本研究では，処理を知識ベースによって制御するという動的な制御方式を採用する．

具体的には，システムの各推論環境は，次のような操作リストというデータ構造を持っている．

{ <操作，期待値>，・・・}

ここで，操作は断片分け操作と単語処理の各段階の単語検出操作，そして，構文解析操作と意味解析操作などである．期待値はこの操作の実行に得られるべき結果である．期待値はいままでの処理結果と知識によって予測したものである．例えば，単語検出操作に対して，次のようなものを期待値に含むべきである．

1) 単語の品詞
2) 単語で表す意味概念のリスト

システムは操作リストから一つの操作要素を取り出して，指示される操作を実行し，その実行結果を期待値と比較する．処理結果と期待値に一致するならば，現在の推論環境での処理を続ける．一方，処理結果は期待値と一致しない場合，システムは現在の推論環境の処理を中断し，PME の確からしさ修正機能を起動する．仮定の確からしさを修正した後，推論を他の環境に移って続ける．

また，一つの操作を実行した後，システムはこの操作の結果によって，次の操作を決定し，操作リストに加える．この操作の決定は知識ベースによって決める．この知識ベースの中の知識をどのように表現するかは現在検討中である．

8 おわりに

本稿では，我々が提案した統合パーサという方法を，ノイズを含んだ文の理解への適用について検討し，それに関する自然言語理解システムの構成方法について報告した．

このシステムは統合パーサをベースにして構成される．処理の過程で，システムは統合パーサによって，言明の集合整合性を維持し，処理を確からしい方向に誘導する．また，本システムでは，ノイズによる入力誤りを効率的に訂正するために，断片的な解析方法を用いる．与えられた入力文に対して，システムは単語検出の結果によって，入力文をいくつかの解析断片を分ける．処理はまず誤りのない部分に対して行なう．次いで，誤りのない部分に関する解析結果と構文的知識と意味的知識の統合によって，誤りのある断片に対して，その可能な構文構造と意味構造を予測する．それに基づいて，入力誤りの訂正を行なう．これによって，入力誤りによる不確定性を効率的に解消することと，また，その不確定性による無駄な処理を避けることが期待できる．

なお，このシステムはいま本研究室の Sun Work Station の上で Common Lisp を用いて作成中である．

参考文献

[1] Doyle, J.: A Truth Maintenance System. *Artificial Intelligence*, 12:231–272, 1979.

[2] de Kleer, J.: An assumption-based TMS. *Artificial Intelligence*, 28:127–162, 1986.

[3] Liu, X,. Nishida, T., and Doshita, S.: Maintaining Consistency and Plausibility in Integrated Natural Language Understanding, *Proceedings of ICSC'88*, 360–367 (1988).

[4] Pearl, J.: Distributed Revision of Composite Beliefs, *Artificial Intelligence*, 33:173–215, 1987.

[5] Waltz, D., and Pollack. J. B.: Massively parallel parsing: a strongly interactive model of natural language interpretation. *Cognitive Science*, 9:51–74, 1985.

[6] Schank, R. C. and Abelson, R.: Scripts, Plans, Goals, and Understanding, *Lawrence Erlbaum Associates* (1977).

[7] 荒木, 宮永, 栃内: 多段階分割復元法による誤りの多い文字列からの原文の復元, 情報処理学会論文誌,Vol. 30 No. 2, pp. 169–178, (1989).

[8] 島崎, 安田, 高木, 池原: 日本語訂正支援システムにおける評価法の検討, 第 36 回情報処理学会全国大会論文集, 5U-3, pp. 1283–1284 (1988).

[9] 鈴木, 武田: 日本語文書校正支援システムの設計と評価, 情報処理学会論文誌, Vol. 30 No. 11, pp. 1402–1412, (1989).

[10] 劉, 西田, 堂下: 統合パーサによる統合的自然言語解析, 情報処理学会論文誌,Vol. 31 No. 9, (1990). (掲載予定)